IBM® DB2® Universal Database

Administration Guide: Planning

Version 7

SC09-2946-00

IBM® DB2® Universal Database

Administration Guide: Planning

Version 7

SC09-2946-00

Before using this information and the product it supports, be sure to read the general information under "Appendix F. Notices" on page 421.

Order publications through your IBM representative or the IBM branch office serving your locality or by calling `1-800-879-2755` in the United States or `1-800-IBM-4YOU` in Canada.

Contents

Part 5. Appendixes 329

About This Book

The Administration Guide in its three volumes provides information necessary to use and administer the year 2000 ready, DB2* relational database management system (RDBMS) products, and includes:

- Information about database design (found in *Administration Guide: Planning*)
- Information about implementing and managing databases (found in *Administration Guide: Implementation*)
- Information about configuring and tuning your database environment to improve performance (found in *Administration Guide: Performance*).

Many of the tasks described in this book can be performed using different interfaces:

- The **Command Processor**, which allows you to access and manipulate databases from a graphical interface. From this interface, you can also execute SQL statements and DB2 utility functions. Most examples in this book illustrate the use of this interface. For more information about using the command processor, see the *Command Reference*.
- The **application programming interface**, which allows you to execute DB2 utility functions within an application program. For more information about using the application programming interface, see the *Administrative API Reference*.
- The **Control Center**, which allows you to graphically perform administrative tasks such as configuring the system, managing directories, backing up and recovering the system, scheduling jobs, and managing media. The Control Center also contains `Replication Administration` to graphically set up the replication of data between systems. Further, the Control Center allows you to execute DB2 utility functions through a graphical user interface. There are different methods to invoke the Control Center depending on your platform. For example, use the `db2cc` command on a command line, (on OS/2) select the Control Center icon from the DB2 folder, or use start panels on Windows platforms. For introductory help, select **Getting started** from the **Help** pull-down of the Control Center window. The **Visual Explain** and **Performance Monitor** tools are invoked from the Control Center.

There are other tools that you can use to perform administration tasks. They include:

- The `Script Center` to store small applications called scripts. These scripts may contain SQL statements, DB2 commands, as well as operating system commands.

- The `Alert Center` to monitor the messages that result from other DB2 operations.
- The `Tool Settings` to change the settings for the Control Center, Alert Center, and Replication.
- The `Journal` to schedule jobs that are to run unattended.
- The `Data Warehouse Center` to manage warehouse objects.

Who Should Use This book

This book is intended primarily for database administrators, system administrators, security administrators and system operators who need to design, implement and maintain a database to be accessed by local or remote clients. It can also be used by programmers and other users who require an understanding of the administration and operation of the DB2 relational database management system.

How This Book is Structured

This book contains information about the following major topics:

The World of DB2 Universal Database

- Chapter 1. Administering DB2 Universal Database, presents an introduction to, and overview of, DB2 Universal Database.

Database Concepts

- Chapter 2. Basic Relational Database Concepts, presents an overview of database objects, including recovery objects, storage objects, and system objects.
- Chapter 3. Federated Systems, discusses federated systems, which are database management systems (DBMSs) that support applications and users submitting SQL statements referencing two or more DBMSs or databases in a single statement.
- Chapter 4. Parallel Database Systems, provides an introduction to the types of parallelism available with DB2.
- Chapter 5. About Data Warehousing, provides an overview of data warehousing and data warehousing tasks.
- Chapter 6. About Spatial Extender, introduces Spatial Extender by explaining its purpose and discussing the data that it processes.

Database Design

- Chapter 7. Logical Database Design, discusses the concepts and guidelines for logical database design.

- Chapter 8. Physical Database Design, discusses the guidelines for physical database design, including considerations related to data storage.
- Chapter 9. Designing Distributed Databases, discusses how you can access multiple databases in a single transaction.
- Chapter 10. Designing for Transaction Managers, discusses how you can use your databases in a distributed transaction processing environment, such as CICS.
- Chapter 11. Designing for High Availability, presents an overview of the high availability failover support that is provided by DB2.

High Availability Systems

- Chapter 12. High Availability Cluster Multi-processing, Enhanced Scalability (HACMP ES) for AIX, discusses DB2 support for high availability failover recovery on AIX.
- Chapter 13. High Availability in the Windows NT Environment, discusses DB2 support for high availability failover recovery on Windows NT.
- Chapter 14. DB2 and High Availability on Sun Cluster 2.2, discusses DB2 support for high availability failover recovery on the Sun Solaris operating system.

Appendixes

- Appendix A. Using the DB2 Library, provides information about the structure of the DB2 library, including wizards, online help, messages, and books.
- Appendix B. Naming Rules, presents the rules to follow when naming databases and objects.
- Appendix C. Planning Database Migration, provides information about migrating databases to Version 7.
- Appendix D. Incompatibilities Between Releases, presents the incompatibilities introduced from release to release up to, and including, Version 7.
- Appendix E. National Language Support (NLS), introduces DB2 National Language Support, including information about countries, languages, and code pages.

A Brief Overview of the Other Volumes of the Administration Guide

Administration Guide: Implementation

The *Administration Guide: Implementation* is concerned with the implementation of your database design. The specific chapters and appendixes in that volume are briefly described here:

Administering Using the Control Center

- ″Administering DB2 Using GUI Tools″ introduces the Graphical User Interface (GUI) tools used to administer the database.

Implementing Your Design

- ″Before Creating a Database″ discusses the prerequisites before you create a database.
- ″Creating a Database″ presents those tasks associated with the creation of a database and related database objects.
- ″Altering a Database″ discusses what must be done before altering a database and those tasks associated with the modifying or dropping of a database or related database objects.

Database Security

- ″Controlling Database Access″ describes how you can control access to your database's resources.
- ″Auditing DB2 Activities″ describes how you can detect and monitor unwanted or unanticipated access to data.

Moving Data

- ″Utilities for Moving Data″ is a one-page introduction to the different ways to move data and to direct you to the *Data Movement Utilities Guide and Reference* book.

Recovery

- ″Recovering a Database″ discusses factors to consider when choosing database and table space recovery methods, including backing up and restoring a database or table space, and using the roll-forward recovery method.

Appendixes

- ″Using Distributed Computing Environment (DCE) Directory Services″ provides information about how you can use DCE Directory Services.
- ″User Exit for Database Recovery″ discusses how user exit programs can be used with database log files, and describes some sample user exit programs.
- ″Issuing Commands to Multiple Database Partition Servers″ discusses the use of the *db2_all* and *rah* shell scripts to send commands to all partitions in a partitioned database environment.
- ″How DB2 for Windows NT Works with Windows NT Security″ describes how DB2 works with Windows NT security.
- ″Using the Windows NT Performance Monitor″ provides information about registering DB2 with the Windows NT Performance Monitor, and using the performance information.

- ″Working with Windows NT or Windows 2000 Database Partition Servers″ provides information about the utilities available to work with database partition servers on Windows NT or Windows 2000.
- ″Configuring Multiple Logical Nodes″ describes how to configure multiple logical nodes in a partitioned database environment.
- ″High Speed Inter-node Communications″ describes how to enable Virtual Interface Architecture for use with DB2 Universal Database.
- ″Lightweight Directory Access Protocol (LDAP) Directory Services″ provides information about how you can use LDAP Directory Services.
- ″Extending the Control Center″ provides information about how you can extend the Control Center by adding new tool bar buttons including new actions, adding new object definitions, and adding new action definitions.

Administration Guide: Performance

The *Administration Guide: Performance* is concerned with performance issues; that is, those topics and issues concerned with establishing, testing, and improving the performance of your application, and that of the DB2 Universal Database product itself. The specific chapters and appendixes in that volume are briefly described here:

Introduction to Performance

- ″Elements of Performance″ introduces concepts and considerations for managing and improving DB2 UDB performance.
- ″Architecture and Processing Overview″ introduces underlying DB2 Universal Database architecture and processes.

Tuning Application Performance

- ″Application Considerations″ describes some techniques for improving database performance when designing your applications.
- ″Environmental Considerations″ describes some techniques for improving database performance when setting up your database environment.
- ″System Catalog Statistics″ describes how statistics about your data can be collected and used to ensure optimal performance.
- ″Understanding the SQL Compiler″ describes what happens to an SQL statement when it is compiled using the SQL compiler.
- ″SQL Explain Facility″ describes the Explain facility, which allows you to examine the choices the SQL compiler has made to access your data.

Tuning and Configuring Your System

- ″Operational Performance″ provides an overview of how the database manager uses memory and other considerations that affect run-time performance.

- ″Using the Governor″ provides an introduction to the use of a governor to control some aspects of database management.
- ″Scaling Your Configuration″ introduces some considerations and tasks associated with increasing the size of your database systems.
- ″Redistributing Data Across Database Partitions″ discusses the tasks required in a partitioned database environment to redistribute data across partitions.
- ″Benchmark Testing″ provides an overview of benchmark testing and how to perform benchmark testing.
- ″Configuring DB2″ discusses the database manager and database configuration files and the values for the configuration parameters.

Appendixes

- ″DB2 Registry and Environment Variables″ presents profile registry values and environment variables.
- ″Explain Tables and Definitions″ provides information about the tables used by the DB2 Explain facility and how to create those tables.
- ″SQL Explain Tools″ provides information on using the DB2 explain tools: db2expln and dynexpln.
- ″db2exfmt — Explain Table Format Tool″ provides information on using the DB2 explain tool to format the explain table data.

Part 1. The World of DB2 Universal Database

Chapter 1. Administering DB2 Universal Database

DB2 provides the flexibility for you to run a wide range of hardware configurations. It allows you to choose how to best match your hardware and application requirements with a specific DB2 product configuration.

DB2 also supports many different levels of complexity in database environments, and there are considerations and tasks specific to each environment. These are discussed in detail in both the *Administration Guide* and other books in the DB2 library (see “Appendix A. Using the DB2 Library” on page 331). In some cases, entire sections of these books are only appropriate for a specific environment. After reading the preface to this book (″About This Book″), you will understand which chapters in this and the other volumes of the *Administration Guide* (the *Administration Guide: Implementation*, and the *Administration Guide: Performance*) are appropriate for your business needs.

If you are new to relational database management systems (RDBMSs), or to DB2, you will find the section entitled ″Basic Relational Database Concepts″ helpful. If you are familiar with these concepts, or do not need to review them, you can skip this section and move directly to the sections detailing more advanced topics, such as:

- Federated systems. This sections discusses database management systems (DBMSs) that support applications and users submitting SQL statements referencing two or more DBMSs or databases in a single statement.
- Parallel database systems. This section provides an introduction to the types of parallelism available with DB2. Components of a task, such as a database query, can be run in parallel to dramatically enhance performance.
- Distributed transaction processing. This section discusses how you can access multiple databases in a single transaction, and how you can use your databases in a distributed transaction processing environment.
- High availability systems. This section presents an overview of the high availability failover support that is provided by DB2. Failover capability allows for the automatic transfer of workload from one processor to another when there is hardware failure.

DB2 can address your most specialized data management needs, such as:

- *Replication*, which allows you to copy data on a regular basis to multiple remote databases. If you need updates from a master database to be copied automatically to other databases, you can use the replication features of DB2 to specify what data should be copied, which database tables the data should be copied to, and how often the updates should be copied. If you want to use the replication features of DB2, refer to the *Replication Guide and Reference*. It introduces the concepts of DB2 data replication, and it describes how to plan, configure, and administer a replication environment.

- *Data warehousing*, in which you can create stores of "informational data", or data that is extracted from operational data and then transformed for end-user decision making. For example, a data warehousing tool might copy all the sales data from the operational database, perform calculations to summarize the data, and write the summarized data to a target in a separate database. You can query the separate database (the *warehouse*) without impacting the operational databases. For detailed information about data warehousing, refer to the *Data Warehouse Center Administration Guide*.
- A *geographic information system* (GIS), which can be created through Spatial Extender. A GIS is a complex of objects, data, and applications that allows you to generate and analyze spatial information about geographic features. In Spatial Extender, a geographic feature can be represented by a row in a table or view, or by a portion of such a row. For detailed information about using Spatial Extender, refer to the *Spatial Extender User's Guide and Reference*.

The *Administration Guide: Planning* also covers database design, including logical database design and physical database design considerations for DB2. Other planning issues, such as planning database migration, identifying incompatibilities that might impact your applications (an *incompatibility* is a part of DB2 Universal Database that works differently than it did in a previous release of DB2; if used in an existing application, it will produce an unexpected result, necessitate a change to the application, or reduce performance), and exploiting national language support (NLS), are also discussed.

The *Administration Guide: Implementation* covers the details of implementing your database design. Topics include creating and altering a database, database security, database recovery, and administering DB2 using the Control Center, a DB2 graphical user interface.

The *Administration Guide: Performance* covers topics and issues concerned with establishing, testing, and improving the performance of your application and of DB2 itself.

Part 2. Database Concepts

Chapter 2. Basic Relational Database Concepts

This section covers the following topics:

- “Overview of Database Objects”
- “Overview of Recovery Objects” on page 12
- “Overview of Storage Objects” on page 13
- “Overview of System Objects” on page 18
- “Business Rules for Data” on page 20
- “Recovering a Database” on page 24
- “Reorganizing Tables in a Database” on page 49
- “Overview of DB2 Security” on page 49

Overview of Database Objects

This section provides an overview of the following key database objects:

- Instances
- Databases
- Nodegroups
- Tables
- Views
- Indexes
- Schemas
- System catalog tables

Figure 1 on page 8 illustrates the relationship among some of these objects. It also shows that tables, indexes, and long data are stored in table spaces.

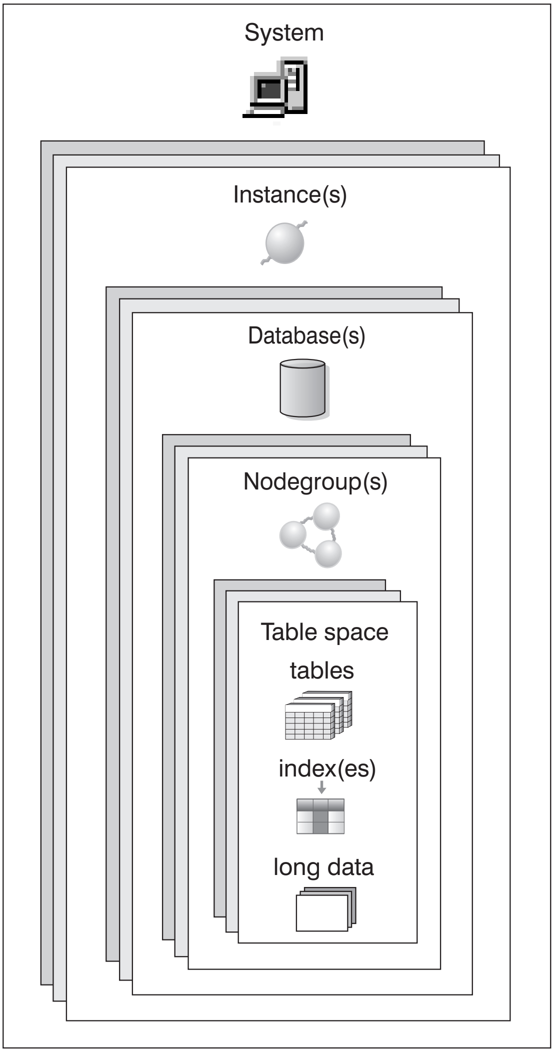


Figure 1. Relationships Among Some Database Objects

Instances

An *instance* (sometimes called a *database manager*) is DB2 code that manages data. It controls what can be done to the data, and manages system resources assigned to it. Each instance is a complete environment. It contains all the database partitions defined for a given parallel database system (see "Chapter 4. Parallel Database Systems" on page 57). An instance has its own databases (which other instances cannot access), and all its database partitions share the same system directories. It also has separate security from other instances on the same machine (system).

Databases

A *relational database* presents data as a collection of tables. A table consists of a defined number of columns and any number of rows. Each database includes a set of system catalog tables that describe the logical and physical structure of the data, a configuration file containing the parameter values allocated for the database, and a recovery log with ongoing transactions and archivable transactions.

Nodegroups

A *nodegroup* is a set of one or more database partitions. When you want to create tables for the database, you first create the nodegroup where the table spaces will be stored, then you create the table space where the tables will be stored. See "Nodegroups and Data Partitioning" on page 57 for more information about nodegroups. See "Chapter 4. Parallel Database Systems" on page 57 for the definition of a database partition. See "Table Spaces" on page 13 for more information about table spaces.

Tables

A relational database presents data as a collection of tables. A *table* consists of data logically arranged in columns and rows. All database and table data is assigned to table spaces. See "Table Spaces" on page 13 for more information about table spaces. The data in the table is logically related, and relationships can be defined between tables. Data can be viewed and manipulated based on mathematical principles and operations called *relations*.

Table data is accessed through Structured Query Language (SQL, see the *SQL Reference*), a standardized language for defining and manipulating data in a relational database. A *query* is used in applications or by users to retrieve data from a database. The query uses SQL to create a statement in the form of

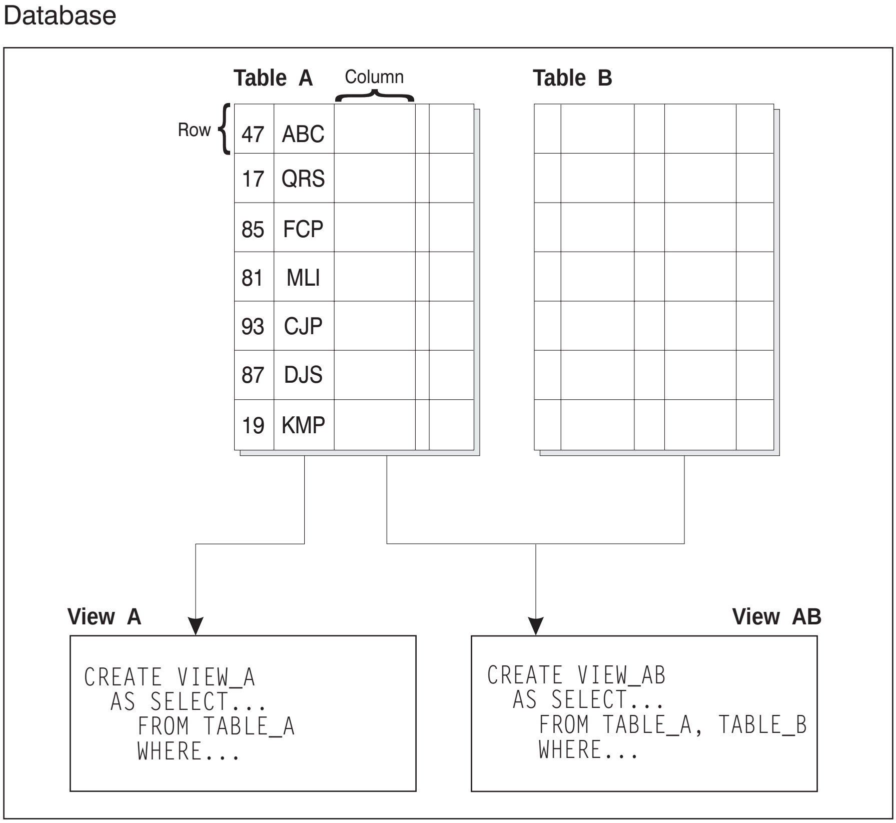
```
SELECT <data_name> FROM <table_name>
```

Views

A *view* is an efficient way of representing data without needing to maintain it. A view is not an actual table and requires no permanent storage. A ″virtual table″ is created and used.

A view can include all or some of the columns or rows contained in the tables on which it is based. For example, you can join a department table and an employee table in a view, so that you can list all employees in a particular department.

Figure 2 on page 10 shows the relationship between tables and views.



47	ABC			
17	QRS			
85	FCP			
81	MLI			
93	CJP			
87	DJS			
19	KMP			

```
CREATE VIEW_A
  AS SELECT...
    FROM TABLE_A
    WHERE...
```

```
CREATE VIEW_AB
  AS SELECT...
    FROM TABLE_A, TABLE_B
    WHERE...
```

Figure 2. Relationship Between Tables and Views

Indexes

An *index* is a set of keys, each pointing to rows in a table. For example, table A in Figure 3 on page 11 has an index based on the employee numbers in the table. This key value provides a pointer to the rows in the table: employee number 19 points to employee KMP. An index allows more efficient access to rows in a table by creating a direct path to the data through pointers.

The SQL *optimizer* automatically chooses the most efficient way to access data in tables. The optimizer takes indexes into consideration when determining the fastest access path to data.

Unique indexes can be created to ensure uniqueness of the index key. An *index key* is a column or an ordered collection of columns on which an index is defined. Using a unique index will ensure that the value of each index key in the indexed column or columns is unique. "Business Rules for Data" on page 20 describes keys and indexes in more detail.

Figure 3 shows the relationship between an index and a table.

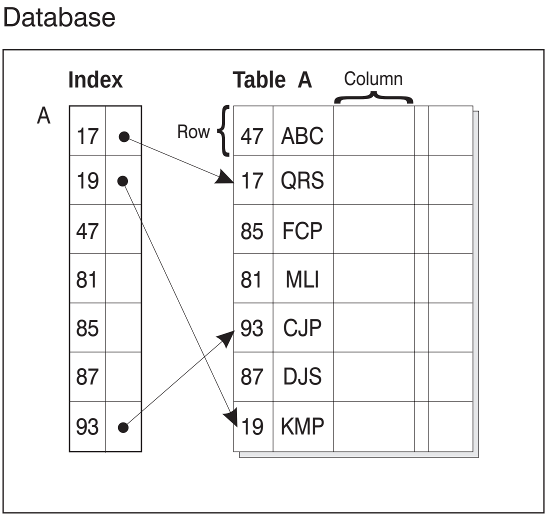


Figure 3. Relationship Between an Index and a Table

Schemas

A *schema* is an identifier, such as a user ID, that helps group tables and other database objects. A schema can be owned by an individual, and the owner can control access to the data and the objects within it.

A schema is also an object in the database. It may be created automatically when the first object in a schema is created. Such an object can be anything that can be qualified by a schema name, such as a table, index, view, package, distinct type, function, or trigger. You must have IMPLICIT_SCHEMA authority if the schema is to be created automatically, or you can create the schema explicitly.

A schema name is used as the first part of a two-part object name. When an object is created, you can assign it to a specific schema. If you do not specify a schema, it is assigned to the default schema, which is usually the user ID of the person who created the object. The second part of the name is the name of the object. For example, a user named Smith might have a table named SMITH.PAYROLL.

System Catalog Tables

Each database includes a set of *system catalog tables*, which describe the logical and physical structure of the data. DB2 creates and maintains an extensive set of system catalog tables for each database. These tables contain information about the definitions of database objects such as user tables, views, and indexes, as well as security information about the authority that users have on these objects. They are created when the database is created, and are updated during the course of normal operation. You cannot explicitly create or drop them, but you can query and view their contents using the catalog views.

Overview of Recovery Objects

Log files and the recovery history file are created automatically when a database is created (Figure 4). You cannot directly modify a log file or the recovery history file; however, they are important should you need to use your database backup image to recover data that is lost or damaged.

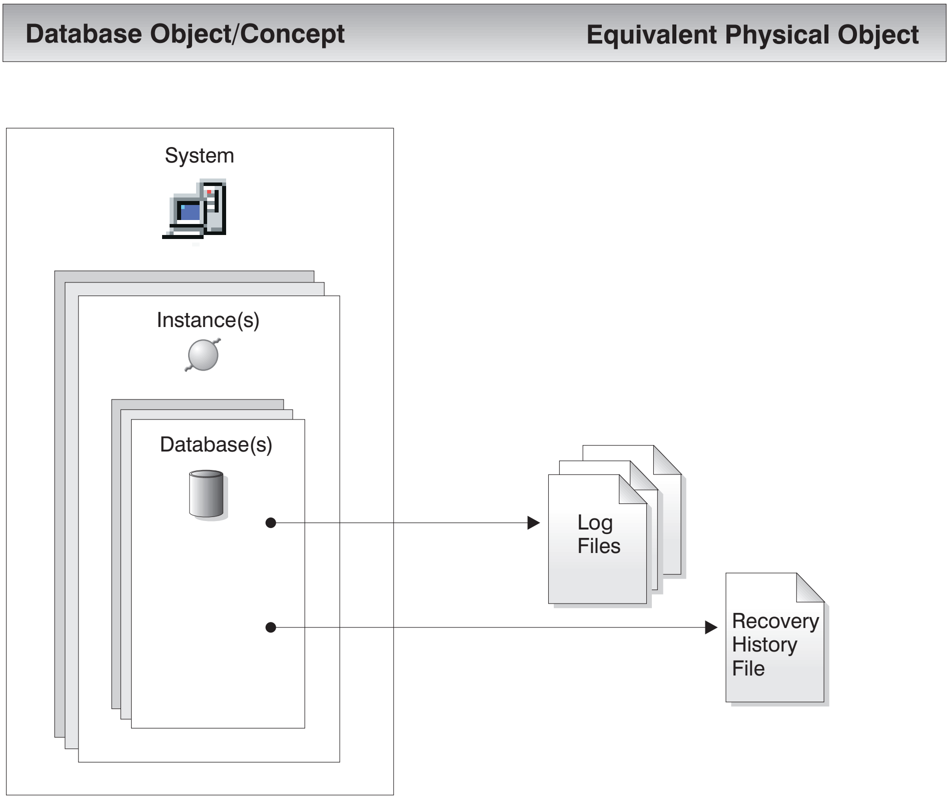


Figure 4. Log Files and the Recovery History File

Recovery Log Files

Each database includes *recovery logs*, which are used to recover from application or system errors. In combination with the database backups, they are used to recover the consistency of the database right up to the point in time when the error occurred. Database recovery is discussed in more detail in "Recovering a Database" on page 24.

Recovery History File

The *recovery history file* contains a summary of the backup information that can be used in case all or part of the database must be recovered to a given point in time. It is used to track recovery-related events such as backup, restore, and load operations. The procedure for backing up and restoring a database is described in "Recovering a Database" on page 24. The load utility is described in the *Data Movement Utilities Guide and Reference*.

Overview of Storage Objects

The following database objects let you define how data will be stored on your system, and how performance (related to accessing the data) can be improved:

- Table space
- Container
- Buffer pool

Table Spaces

A database is organized into parts called *table spaces*. A table space is a place to store tables. When creating a table, you can decide to have certain objects such as indexes and large object (LOB) data kept separately from the rest of the table data. A table space can also be spread over one or more physical storage devices. The following diagram shows some of the flexibility you have in spreading data over table spaces:

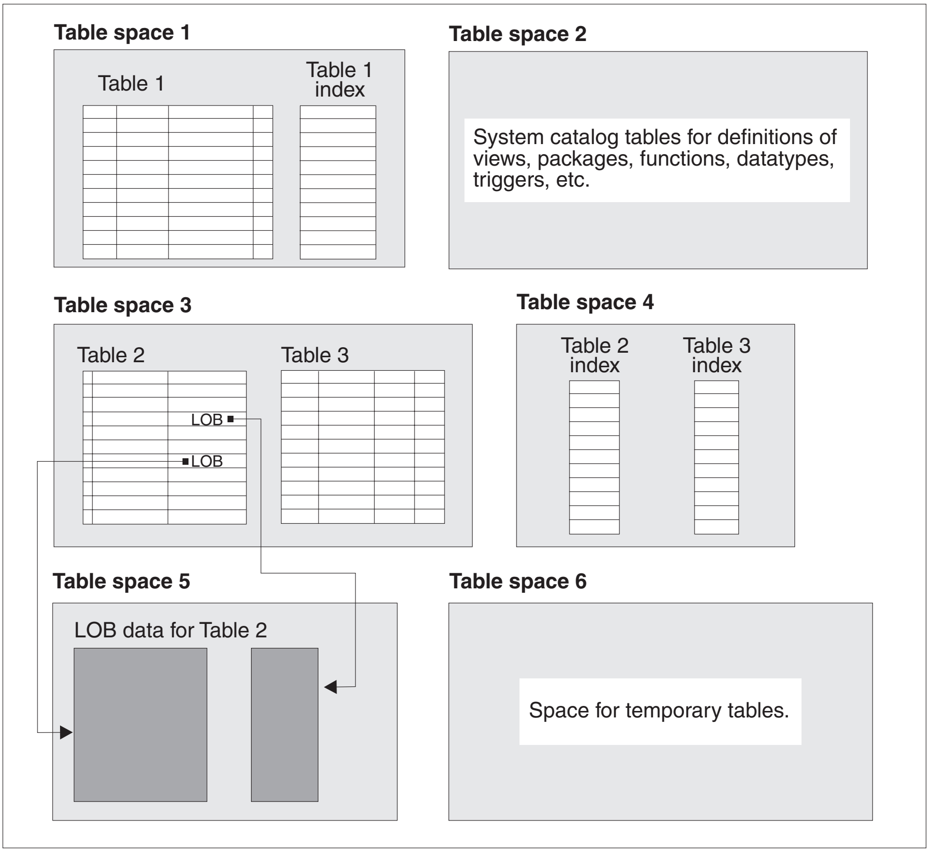


Figure 5. Table Spaces

Table spaces reside in nodegroups (see “Nodegroups” on page 9). Table space definitions and attributes are recorded in the database system catalog (see “System Catalog Tables” on page 11).

Containers are assigned to table spaces. A *container* is an allocation of physical storage (such as a file or a device).

A table space can be either system managed space (SMS), or database managed space (DMS). For an SMS table space, each container is a directory in the file space of the operating system, and the operating system’s file manager controls the storage space. For a DMS table space, each container is either a fixed size pre-allocated file, or a physical device such as a disk, and the database manager controls the storage space.

Figure 6 illustrates the relationship between tables, table spaces, and the two types of space. It also shows that tables, indexes, and long data are stored in table spaces.

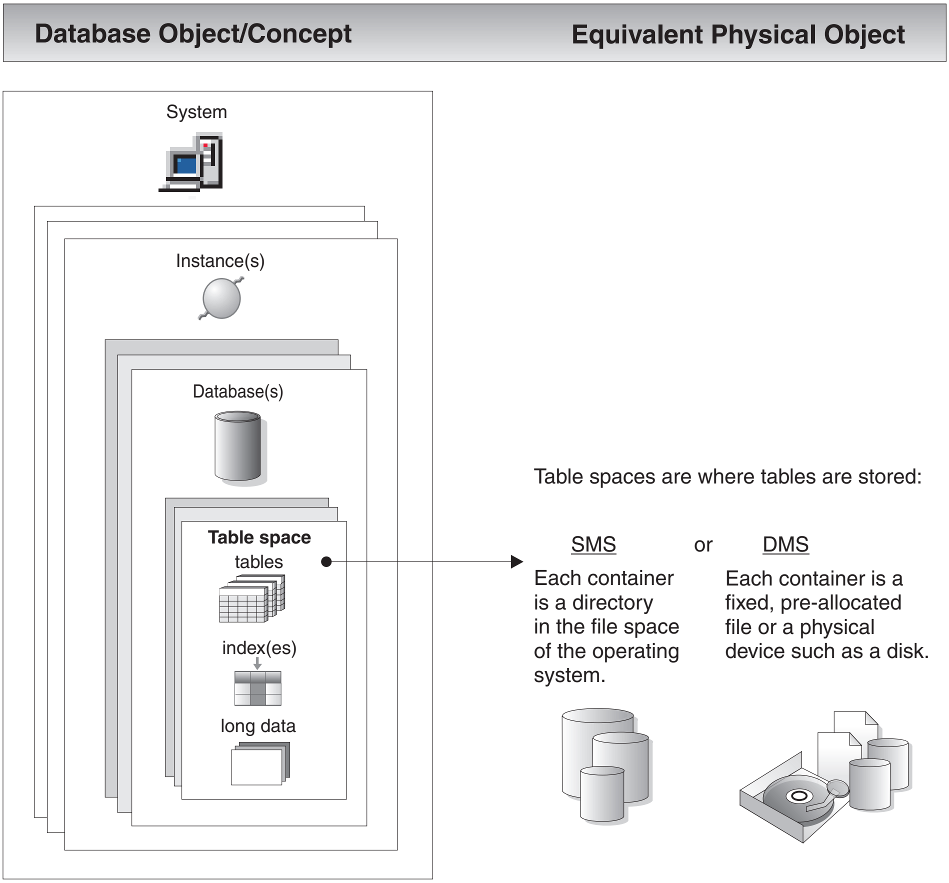

Database Object/Concept
Equivalent Physical Object
System
Instance(s)
Database(s)
Table space
tables
index(es)
long data
Table spaces are where tables are stored:
SMS
or
DMS
Each container is a directory in the file space of the operating system.
Each container is a fixed, pre-allocated file or a physical device such as a disk.

Figure 6. Table Spaces and Tables

Figure 7 on page 16 shows the three table space types: *regular*, *temporary*, and *long*.

Tables containing user data exist in regular table spaces. The default user table space is called USERSPACE1. Indexes are also stored in regular table spaces. The system catalog tables exist in a regular table space. The default system catalog table space is called SYSCATSPACE.

Tables containing long field data or long object data, such as multi-media objects, exist in long table spaces.

Temporary table spaces are classified as either system or user. *System temporary table spaces* are used to store internal temporary data required during SQL operations such as sorting, reorganizing tables, creating indexes, and joining tables. Although you can create any number of system temporary table spaces, it is recommended that you create only one, using the page size that the majority of your tables use. The default system temporary table space is called TEMPSPACE1. *User temporary table spaces* are used to store declared global temporary tables that store application temporary data. User temporary table spaces are *not* created by default at database creation time.

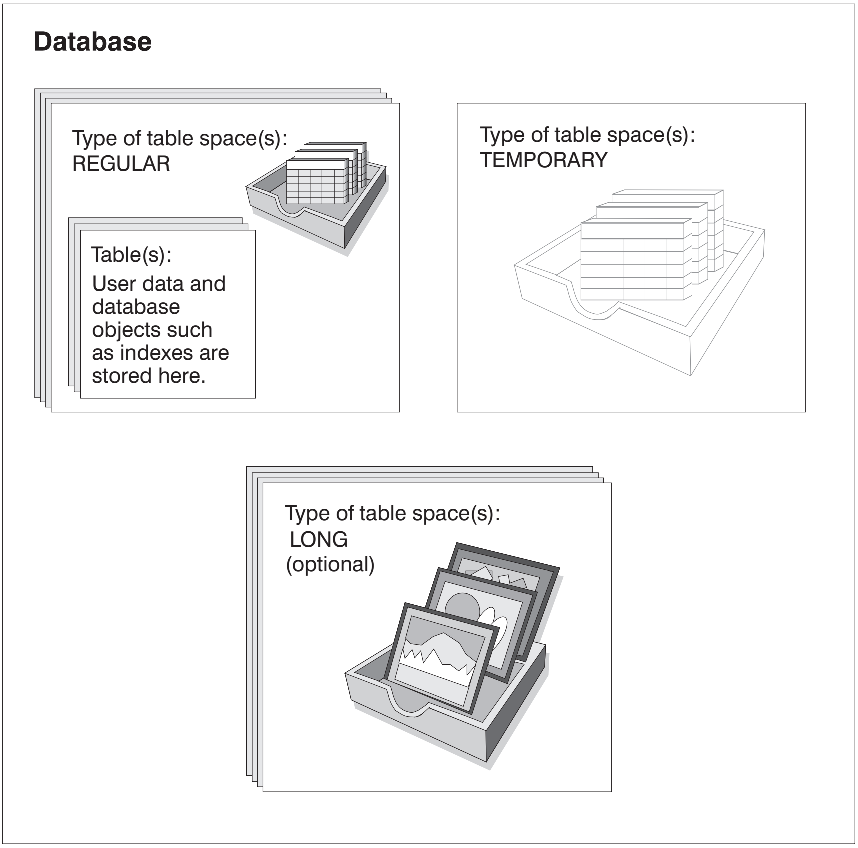


Figure 7. Three Table Space Types

Containers

A *container* is a physical storage device. It can be identified by a directory name, a device name, or a file name.

A container is assigned to a table space. A single table space can span many containers, but each container can belong to only one table space.

Figure 8 illustrates the relationship between tables and a table space within a database, and the associated containers and disks.

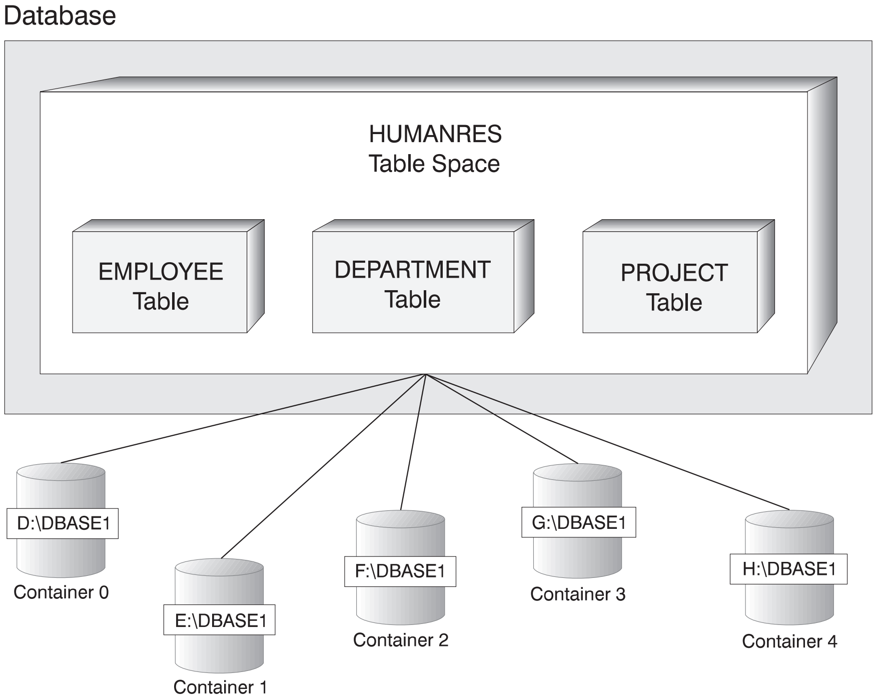


Figure 8. Table Spaces and Tables Within a Database

The EMPLOYEE, DEPARTMENT, and PROJECT tables are in the HUMANRES table space which spans containers 0, 1, 2, 3, and 4. This example shows each container existing on a separate disk.

Data for any table will be stored on all containers in a table space in a round-robin fashion. This balances the data across the containers that belong to a given table space. The number of pages that the database manager writes to one container before using a different one is called the *extent size*.

Buffer Pool

A *buffer pool* is the amount of main memory allocated to cache table and index data pages as they are being read from disk, or being modified. The purpose of the buffer pool is to improve system performance. Data can be accessed much faster from memory than from disk; therefore, the fewer times the database manager needs to read from or write to a disk (I/O), the better the performance. (You can create more than one buffer pool, although for most situations only one is required.)

The configuration of the buffer pool is the single most important tuning area, because you can reduce the delay caused by slow I/O.

Figure 9 illustrates the relationship between a buffer pool and containers.

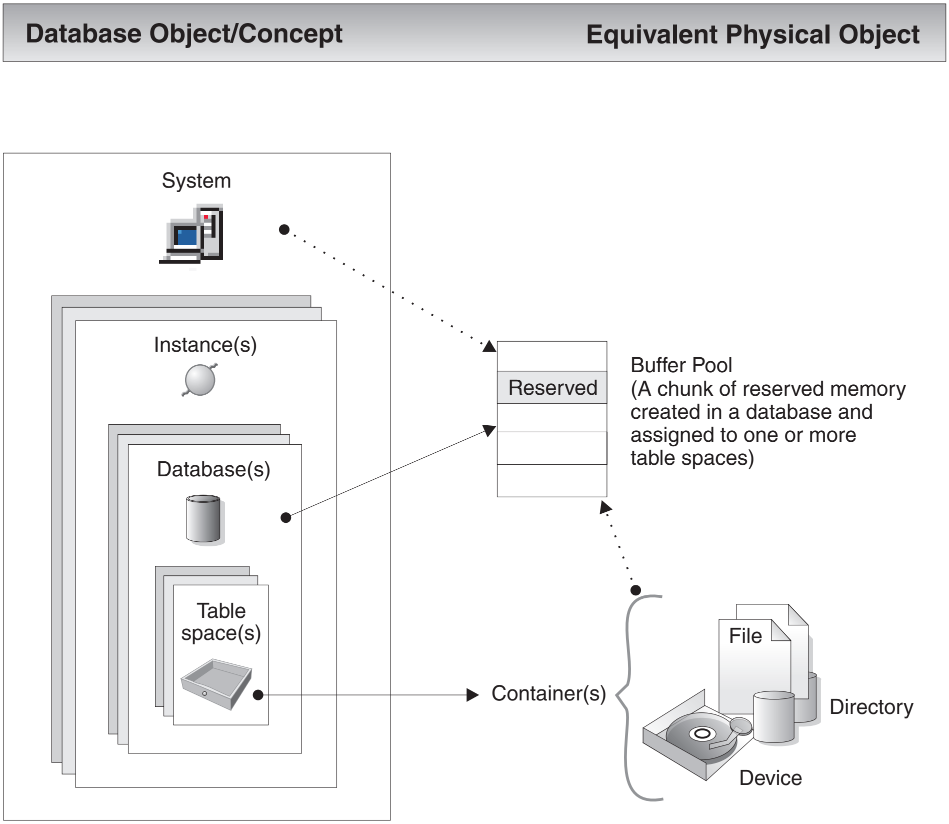

Database Object/Concept
Equivalent Physical Object
System
Instance(s)
Database(s)
Table space(s)
Reserved
Buffer Pool
(A chunk of reserved memory created in a database and assigned to one or more table spaces)
Container(s)
File
Directory
Device

Figure 9. Buffer Pool and Containers

Overview of System Objects

When a DB2 instance or a database is created, a corresponding configuration file is created with default parameter values. You can modify these parameter values to improve performance.

Configuration Parameters

Configuration files contain parameters that define values such as the resources allocated to the DB2 products and to individual databases, and the diagnostic level. There are two types of configuration files: the database manager configuration file for each DB2 instance, and the database configuration file for each individual database (see Figure 10 on page 20).

The *database manager configuration file* is created when a DB2 instance is created. The parameters it contains affect system resources at the instance level, independent of any one database that is part of that instance. Values for many of these parameters can be changed from the system default values to improve performance or increase capacity, depending on your system's configuration.

There is one database manager configuration file for each client installation as well. This file contains information about the client enabler for a specific workstation. A subset of the parameters available for a server are applicable to the client.

A *database configuration file* is created when a database is created, and resides where that database resides. There is one configuration file per database. Its parameters specify, among other things, the amount of resource to be allocated to that database. Values for many of the parameters can be changed to improve performance or increase capacity. Different changes may be required, depending on the type of activity in a specific database.

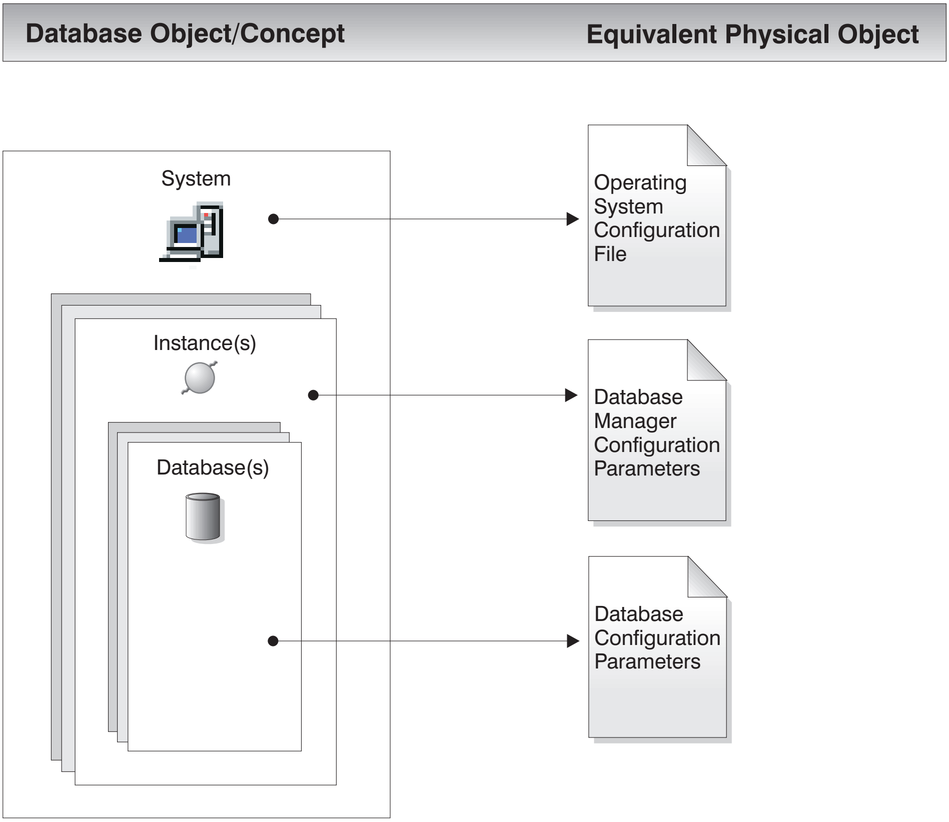

Database Object/Concept
Equivalent Physical Object
System
Instance(s)
Database(s)
Operating System Configuration File
Database Manager Configuration Parameters
Database Configuration Parameters

Figure 10. Configuration Parameter Files

Business Rules for Data

Within any business, data must often adhere to certain restrictions or rules. For example, an employee number must be unique. DB2 provides *constraints* as a way to enforce such rules.

DB2 provides the following types of constraints:

- NOT NULL constraint
- Unique constraint
- Primary key constraint
- Foreign key constraint
- Check constraint

NOT NULL constraint

NOT NULL constraints prevent null values from being entered into a column.

unique constraint

Unique constraints ensure that the values in a set of columns are unique and not null for all rows in the table. For example, a typical unique constraint in a DEPARTMENT table might be that the department number is unique and not null.

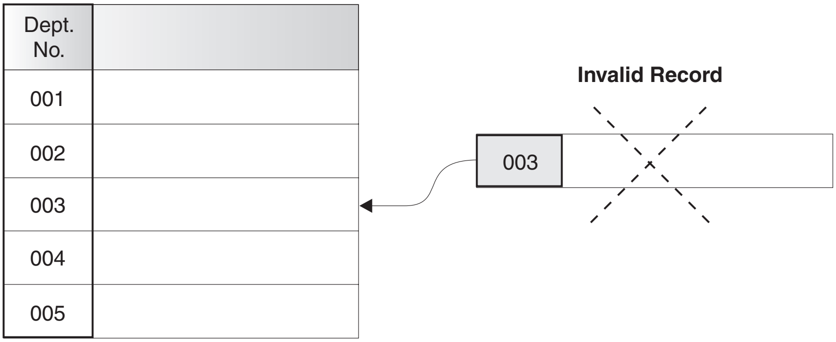


Figure 11. Unique Constraints Prevent Duplicate Data

The database manager enforces the constraint during insert and update operations, ensuring data integrity.

primary key constraint

Each table can have one primary key. A primary key is a column or combination of columns that has the same properties as a unique constraint. You can use a primary key and foreign key constraints to define relationships between tables.

Because the primary key is used to identify a row in a table, it should be unique and have very few additions or deletions. A table cannot have more than one primary key, but it can have multiple unique keys. Primary keys are optional, and can be defined when a table is created or altered. They are also beneficial, because they order the data when data is exported or reorganized.

In the following tables, DEPTNO and EMPNO are the primary keys for the DEPARTMENT and EMPLOYEE tables.

Table 1. DEPARTMENT Table

DEPTNO (Primary Key)	DEPTNAME	MGRNO
A00	Spiffy Computer Service Division	000010

Table 1. DEPARTMENT Table (continued)

DEPTNO (Primary Key)	DEPTNAME	MGRNO
B01	Planning	000020
C01	Information Center	000030
D11	Manufacturing Systems	000060

Table 2. EMPLOYEE Table

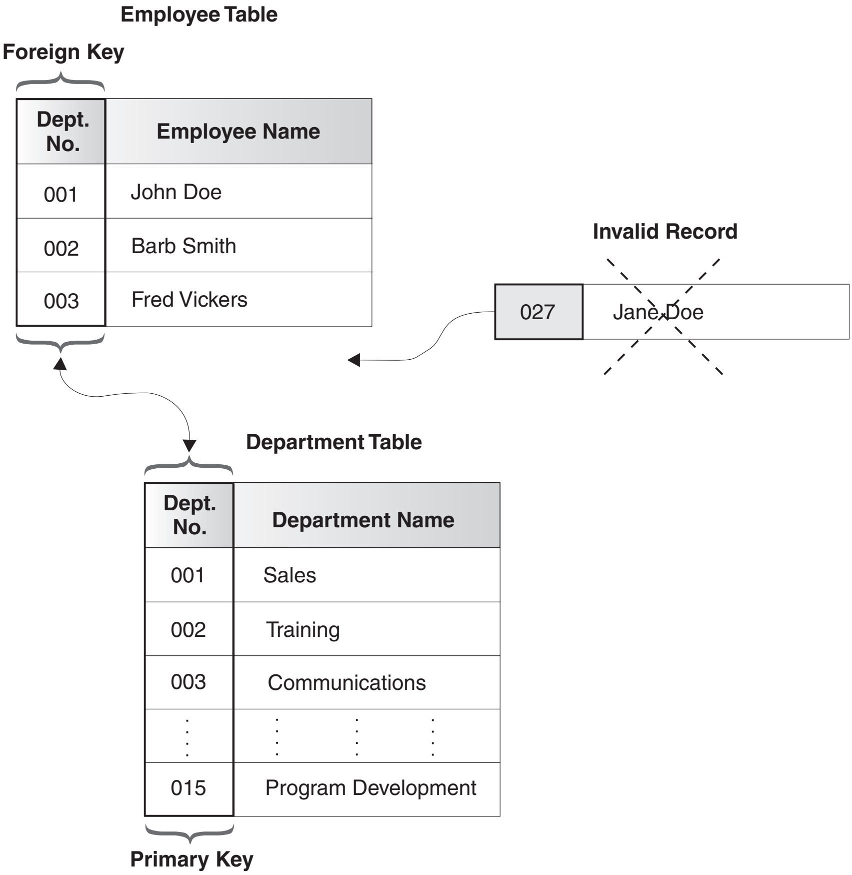
EMPNO (Primary Key)	FIRSTNAME	LASTNAME	WORKDEPT (Foreign Key)	PHONENO
000010	Christine	Haas	A00	3978
000030	Sally	Kwan	C01	4738
000060	Irving	Stern	D11	6423
000120	Sean	O'Connell	A00	2167
000140	Heather	Nicholls	C01	1793
000170	Masatoshi	Yoshimura	D11	2890

foreign key constraint

Foreign key constraints (also known as referential integrity constraints) enable you to define required relationships between and within tables.

For example, a typical foreign key constraint might state that every employee in the EMPLOYEE table must be a member of an existing department, as defined in the DEPARTMENT table.

To establish this relationship, you would define the department number in the EMPLOYEE table as the foreign key, and the department number in the DEPARTMENT table as the primary key.



Dept. No.	Employee Name
001	John Doe
002	Barb Smith
003	Fred Vickers

Dept. No.	Department Name
001	Sales
002	Training
003	Communications
⋮	⋮ ⋮ ⋮
015	Program Development

Figure 12. Foreign and Primary Key Constraints Define Relationships and Protect Data

check constraint

A check constraint is a database rule that specifies the values allowed in one or more columns of every row of a table.

For example, in an EMPLOYEE table, you can define the Type of Job column to be ″Sales″, ″Manager″, or ″Clerk″. With this constraint, any record with a different value in the Type of Job column is not valid, and would be rejected, enforcing rules about the type of data allowed in the table.

You can also use *triggers* in your database. Triggers are more complex and potentially more powerful than constraints. They define a set of actions that are executed in conjunction with, or triggered by, an INSERT, UPDATE, or DELETE clause on a specified base table. You can use triggers to support general forms of integrity or business rules. For example, a trigger can check a customer's credit limit before an order is accepted, or be used in a banking application to raise an alert if a withdrawal from an account did not fit a customer's standard withdrawal patterns. For more information about triggers, refer to the *Application Development Guide*.

Recovering a Database

A database can become unusable because of hardware or software failure (or both), and different failure scenarios may require different recovery actions. You should have a rehearsed strategy in place to protect your database against the possibility of failure.

This section discusses different recovery methods, and shows you how to determine which recovery method is best suited to your business environment. The following topics are covered:

- "Overview of Recovery"
- "Factors Affecting Recovery" on page 30
- "Disaster Recovery Considerations" on page 44
- "Reducing the Impact of Media Failure" on page 45
- "Reducing the Impact of Transaction Failure" on page 47
- "System Clock Synchronization in a Partitioned Database System" on page 47

Overview of Recovery

You need to know the strategies available to you when there are problems with the database. These include problems with media and storage, power interruptions, and application failures. You can back up your database, or individual table spaces, and then rebuild them should they be damaged or corrupted in some way. The concept of a database *backup* is the same as any other data backup: taking a copy of the data and storing it on a different medium in case of failure or damage to the original. The simplest case of a backup involves shutting down the database to ensure that no further transactions occur, and then simply backing it up.

The rebuilding of the database is called *recovery*. *Crash recovery* automatically attempts to recover the database after a failure. There are two ways to recover a damaged database: *version recovery* and *roll-forward recovery*.

Non-recoverable databases have both the *logretain* and the *userexit* database configuration parameter disabled. This means that the only logs that are kept are those required for crash recovery. These logs are known as *active logs*, and they contain current transaction data. Version recovery using *offline* backups is the primary means of recovery for a non-recoverable database. (An offline backup means that no other application can use the database when the backup operation is in progress.) Such a database can only be restored offline. It is restored to the state it was in when the backup image was taken.

Recoverable databases have either the *logretain* database configuration parameter set to "RECOVERY", the *userexit* database configuration parameter enabled, or both. Active logs are still available for crash recovery, but you also have the *archived logs*, which contain committed transaction data. Such a database can only be restored offline. It is restored to the state it was in when the backup image was taken. However, with roll-forward recovery, you can roll the database *forward* (that is, past the time when the backup image was taken) by using the active and archived logs to either a specific point in time, or to the end of the active logs.

Recoverable database backup operations can be performed either offline or *online* (online meaning that other applications can connect to the database during the backup operation). The database restore and roll-forward operations must always be performed offline. During an online backup operation, roll-forward recovery ensures that *all* table changes are captured and reapplied if that backup is restored.

If you have a recoverable database, you can back up, restore, and roll forward individual table spaces, rather than the entire database. When you back up a table space online, it is still available for use, and simultaneous updates are recorded in the logs. When you perform an online restore or roll-forward operation on a table space, the table space itself is not available for use until the operation completes, but users are not prevented from accessing tables in other table spaces.

Crash recovery protects a database from being left in an inconsistent, or unusable, state. Transactions (or units of work) against the database can be interrupted unexpectedly. If a failure occurs before all of the changes that are part of the unit of work are completed and committed, the database is left in an inconsistent and unusable state.

The database then needs to be moved to a consistent and usable state. This is done by rolling back incomplete transactions and completing committed transactions that were still in memory when the crash occurred (Figure 13 on page 26).

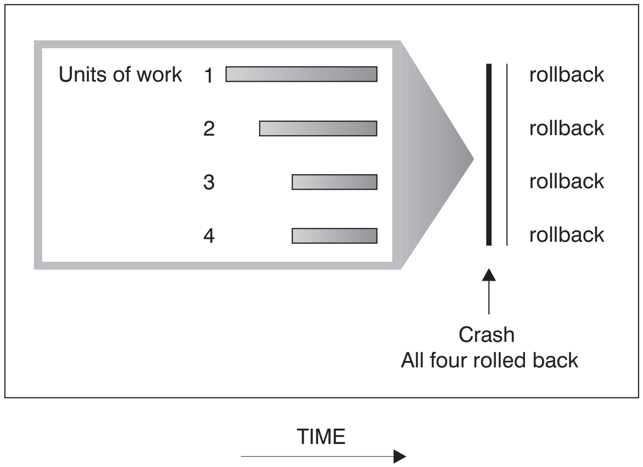


Figure 13. Rolling Back Units of Work

When a database is in a consistent and usable state, it has attained what is known as a ″point of consistency″. An offline database backup represents a point of consistency. When a point of consistency is reached, all transactions have been resolved and the data is available to other users or applications.

You can move to a point of consistency following a crash by invoking the RESTART DATABASE command (refer to the *Command Reference*). If you want this done in every case of a failure, you should consider the use of the **automatic restart enable** (*autorestart*) configuration parameter. The default behavior for this database configuration parameter is to invoke the RESTART DATABASE command whenever it is needed. When *autorestart* is enabled, the next connect request to the database after a failure causes the RESTART DATABASE command to be invoked.

Crash recovery moves the database to a consistent and usable state. If, however, crash recovery is applied to a database that is enabled for forward recovery (that is, the *logretain* configuration parameter is set to ″RECOVERY″, or the *userexit* configuration parameter is enabled), and an error occurs during crash recovery that is attributable to an individual table space, that table space must be taken offline, and cannot be accessed until it is repaired. Crash recovery continues. At the completion of crash recovery, the other table spaces in the database are still usable, and connections to the database can be established. (There are exceptions involving the table spaces that have temporary tables or the system catalog tables. These are discussed under roll-forward recovery.)

As mentioned earlier, DB2 provides two methods to recover a damaged database:

- *Version recovery* is the restoration of a previous version of the database, using an image that was created during a backup operation.

  A database restore operation will rebuild the entire database using a backup of the database made earlier. A backup of the database allows you to restore a database to a state identical to the one at the time that the backup was made. Every unit of work from the time of the backup to the time of the failure is lost (see Figure 14).

  Using the version recovery method, you must schedule and perform full backups of the database on a regular basis.

  In a partitioned database environment, the database is located across many database partition servers (or nodes). You must restore all partitions, and the backup images that you use for the restore database operation must all have been taken at the same time. (Each database partition is backed up and restored separately.) A backup of each database partition taken at the same time is known as a *version backup*.

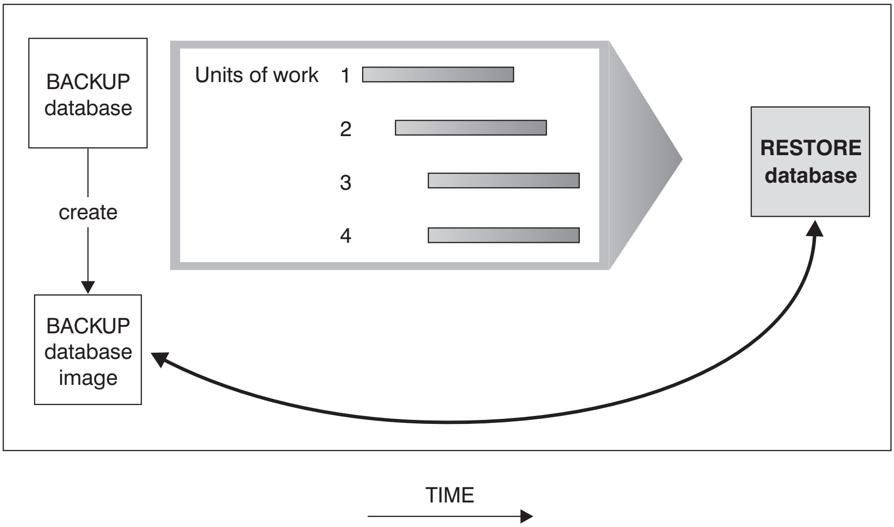


Figure 14. Restoring a Database

- To use the *roll-forward recovery* method, you must have taken a backup of the database, and archived the logs (by enabling either the *logretain* or the *userexit* database configuration parameters, or both. For information on the decisions that you must make regarding the logging procedure that you use, see “Database Logs” on page 31.) Restoring the database and specifying the WITHOUT ROLLING FORWARD option is equivalent to using the version recovery method. The database is restored to a state identical to the one at the time that the offline backup image was made. If you restore the database and do *not* specify the WITHOUT ROLLING FORWARD option for the restore database operation, the database will be in roll-forward pending state at the end of the restore operation. This allows roll-forward recovery to take place.

The two types of roll-forward recovery to consider are:

- *Database roll-forward recovery*. In this type of roll-forward recovery, transactions recorded in database logs are applied following the database restore operation (see Figure 15). The database logs record all changes made to the database. This method completes the recovery of the database to its state at a particular point in time, or to its state immediately before the failure (that is, to the end of the active logs.)

  In a partitioned database environment, the database is located across many database partitions. If you are performing point-in-time roll-forward recovery, all database partitions must be rolled forward to ensure that all partitions are at the same level. If you need to restore a single database partition, you can perform roll-forward recovery to the end of the logs to bring it up to the same level as the other partitions in the database.

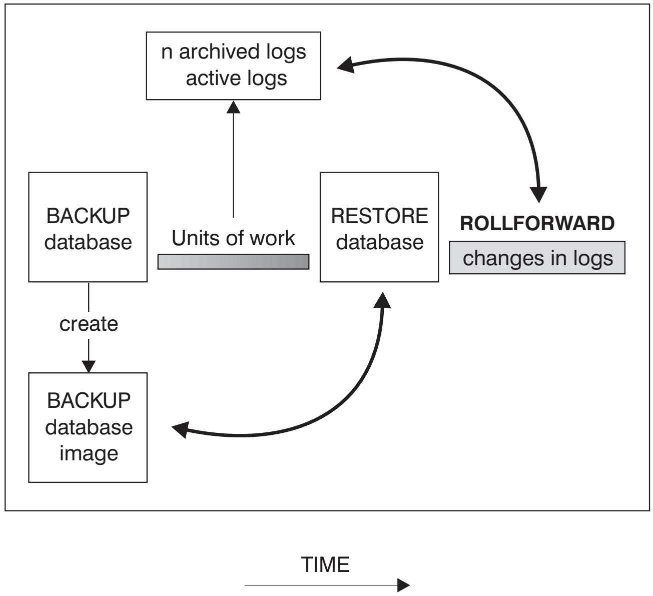


Figure 15. Database Roll-forward Recovery

- *Table space restore and roll forward*. If the database is enabled for forward recovery, it is also possible to back up, restore, and roll forward table spaces. To perform a table space restore and roll-forward operation, you need a backup image of either the entire database (that is, all of the table spaces), or one or more individual table spaces. You also need the log records that affect the table spaces that are to be recovered. You can roll forward through the logs to one of two points:

- The end of the logs; or,
- A particular point in time (called *point-in-time* recovery).

Notes:

1. Table spaces that are not selected at the time of the backup operation will not be in the same state as those that were restored.
2. When using the roll-forward recovery method with table spaces, you must identify ″key″ table spaces in the database to be recovered, as well as schedule and perform a backup of the database (or the ″key″ table spaces) on a regular basis.

Table space roll-forward recovery can be used in the following two situations:

- After a table space restore operation, the table space is always in roll-forward pending state, and it must be rolled forward. Invoke the ROLLFORWARD DATABASE command (refer to the *Command Reference*) to apply the logs against the table spaces to either a point in time, or to the end of the logs.
- If one or more table spaces are in *roll-forward pending* state after crash recovery, first correct the problem with the table space. In some cases, correcting the problem with the table space does not involve performing a restore database operation. For example, a power loss could leave the table space in roll-forward pending state. If the problem is corrected before crash recovery, crash recovery may be sufficient to take the database to a consistent, usable state. A restore database operation is not required in this case. Once the problem with the table space is corrected, you can use the ROLLFORWARD DATABASE command to apply the logs against the table spaces to either a point in time, or to the end of the logs.

  Note: If the table space in error contains the system catalog tables, you will not be able to start the database. You must restore the SYSCATSPACE table space, then perform roll-forward recovery to the end of the logs.

In a partitioned database environment, if you are rolling forward a table space *to a point in time*, you do not have to supply the list of nodes (database partitions) on which the table space resides. DB2 submits the roll-forward request to all partitions. This means the table space must be restored on all database partitions on which the table space resides.

In a partitioned database environment, if you are rolling forward a table space *to the end of the logs*, you must supply the list of database partitions if you do *not* want to roll the table space forward on all partitions. If you want to roll forward all table spaces on all partitions that are in roll-forward pending state to the end of the logs, you do not have to supply the list of database partitions. By default, the database roll-forward request is sent to all partitions.

Factors Affecting Recovery

To decide which database recovery method to use, you must consider the following key factors:

- Will the database be recoverable or non-recoverable?
- How near to the time of failure will you need to recover the database (the point of recovery)?
- How much time can be spent recovering the database? This would include:
  - Time between backups (will affect roll-forward recovery)
  - Time the database is usable or accessible (backing up online or offline based on data availability needs)
- How much storage space can be allocated for backup copies and archived logs?
- Will you be using table space level or full database level backups?

In general, a database maintenance and recovery strategy should ensure that all information is available when it is required for database recovery. The strategy should include a regular schedule for taking database backups, as well as scheduled backups when a database is created, or in the case of a partitioned database system, when the system is scaled by adding or dropping database partition servers (nodes). In addition to these basic requirements, a good strategy will include elements that reduce the likelihood and impact of database failure.

The following topics provide additional information:

- “Recoverable and Non-Recoverable Databases” on page 31
- “Database Logs” on page 31
- “Reducing Logging on Work Tables” on page 36
- “Point of Recovery” on page 37
- “Frequency of Backups and Time Required” on page 37
- “Recovery Time Required” on page 39
- “Storage Considerations” on page 39
- “Keeping Related Data Together” on page 40
- “Restrictions on Using Different Operating Systems” on page 41
- “Damaged Table Space Recovery” on page 41
- “Recovery Performance Considerations” on page 43.

While the general focus of this section is on the database, your overall recovery planning should also include recovering:

- The operating system and DB2 executables
- Applications, UDFs, and stored procedure code in operating system libraries
- Commands for creating DB2 instances and non-DB2 resources
- Operating system security
- Load copies from a load operation (if you specify COPY YES on the LOAD command)

Recoverable and Non-Recoverable Databases

If you can recreate data easily, the database holding that data can be a non-recoverable database. For example:

- Tables that hold data from an outside source that is used for read-only applications (and the data is not mixed with existing data) should be considered for placement within a non-recoverable database.
- Tables with small amounts of data. Here recovery is not a problem. Rather, there is just not enough logging done for the data to justify the added complexity of managing log files and rolling forward after a restore.
- Large tables where small numbers of rows are periodically added. Again, there is not enough volatility to justify managing log files and rolling forward after a restore operation.

If you cannot recreate data easily, the database holding that data should be a recoverable database. The following are examples of data that should be part of a recoverable database:

- Data that you cannot recreate. This includes data whose source is destroyed after the data is loaded, and data that is manually entered into tables.
- Data that is modified by application programs or workstation users after it is loaded into the database.

Database Logs

All databases have logs associated with them. These logs keep records of database changes. If a database needs to be restored to a point beyond the last full, offline backup, then logs are required to roll the data forward to the point of failure.

There are two types of DB2 logging: *circular* and *archive*, each providing a different level of recovery capability.

Circular logging is the default behavior when a new database is created. With this type of logging, only full, offline backups of the database are valid. As the name suggests, circular logging uses a "ring" of online logs to provide recovery from transaction failures and system crashes. The logs are used and retained only to the point of ensuring the integrity of current transactions. Circular logging does not allow you to roll forward a database through prior transactions from the last full backup. Recovery from media failures and disasters is done by restoring from a full, offline backup. All changes since the last backup are lost. The database must be offline (inaccessible to users) when a full backup is taken. Since this type of restore recovers your data to the specific point in time of the full backup, it is called *version recovery*.

Figure 16 shows that the active log uses a ring of log files when circular logging is active.
Active logs are used during crash recovery to prevent a failure (system power or application error) from leaving a database in an inconsistent state. The RESTART DATABASE command uses the active logs, if needed, to move the database to a consistent and usable state. During crash recovery, changes recorded in these logs that were not committed because of the failure are rolled back. Changes that were committed but were not physically written from memory (buffer pool) to disk (database containers) are redone. These actions ensure the integrity of the database. The ROLLFORWARD DATABASE command may also use the active logs, if needed, during a point-in-time recovery, or a recovery to the end of the logs. Active logs are located in the database log path directory.

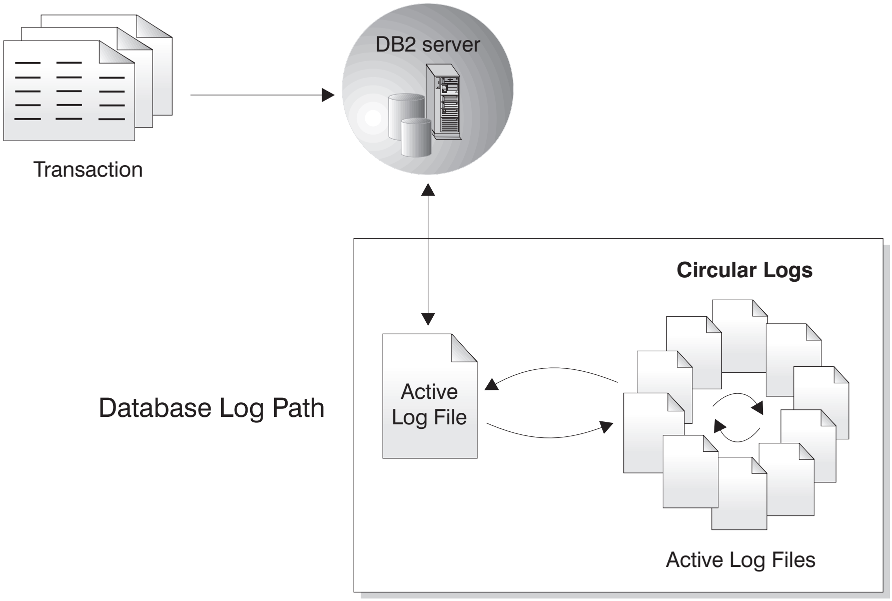


Figure 16. Circular Logging

Archived logs are used specifically for roll-forward recovery. They can be:

online archived logs

When changes in the active log are no longer needed for normal processing, the log is closed, and becomes an archived log. An archived log is said to be *online* when it is stored in the database log path directory (see Figure 17).

offline archived logs

An archived log is said to be *offline* when it is no longer found in the database log path directory (see Figure 18 on page 34). You can also store archived logs in a location other than the database log path directory by using a user exit program. (For additional information, see "User Exit for Database Recovery" in the *Administration Guide: Implementation*.)

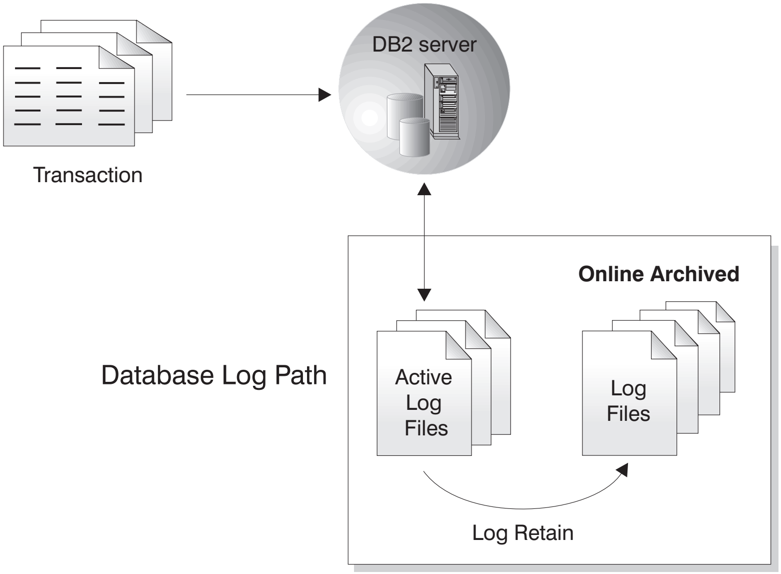


Figure 17. Archive Logging

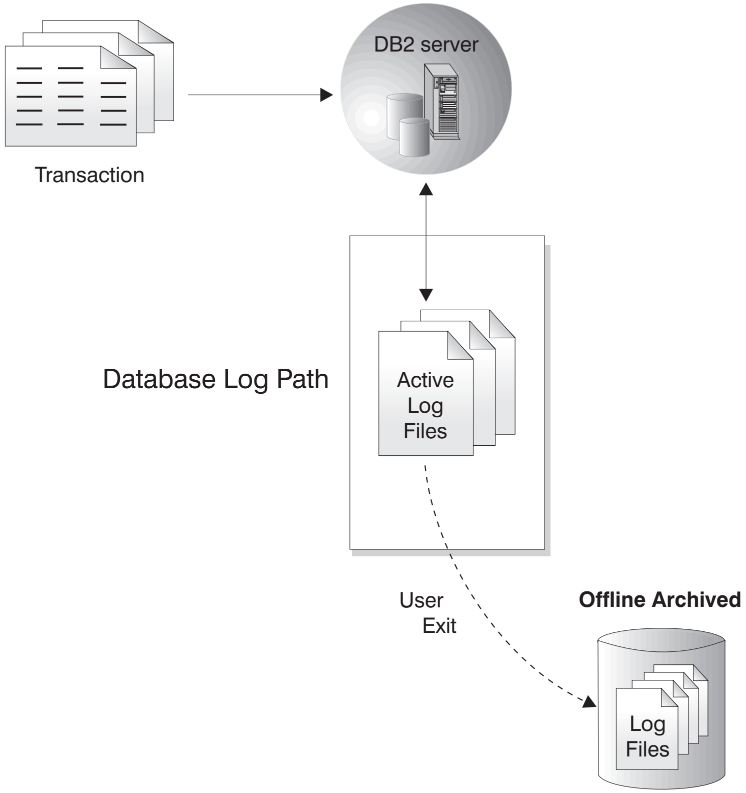
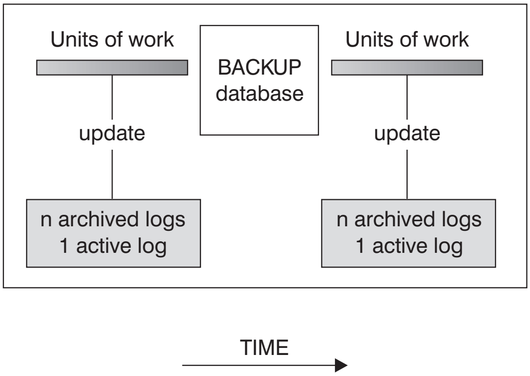


Figure 18. Offline Archived Logs

Roll-forward recovery can use both archived logs and active logs to rebuild a database either to the end of the logs, or to a specific point in time. The roll-forward function achieves this by reapplying committed changes found in the archived and active logs to the restored database.

Roll-forward recovery can also use logs to rebuild a table space by re-applying committed updates in both archived and active logs. You can recover a table space to the end of the logs, or to a specific point in time.

During an online backup, all activities against the database are logged. When an online backup is restored, the logs must be rolled forward at least to the point in time at which the backup was completed. For this to happen, you must archive the logs and make them available when the database is to be restored. The log file used at backup time may continue to be open long after the backup operation completes. The FLUSH LOG option for online backup on the BACKUP DATABASE command will close the active log when an online backup completes. This will allow the active log to be archived, so that you will have a complete backup, as well as all of the logs required for the restoration of that backup.



Logs are used between backups to track the changes to the databases.

Figure 19. Active and Archived Database Logs in Roll-forward Recovery

Two database configuration parameters allow you to change where archived logs are stored: The *newlogpath* parameter, and the *userexit* parameter. Changing the *newlogpath* parameter also affects where active logs are stored. Refer to *Administration Guide: Performance* for more information about these configuration parameters.

To determine which log *extents* (see "Containers" on page 16) in the database log path directory are archived logs, check the value of the *loghead* database configuration parameter. This parameter indicates the lowest numbered log that is active. Those logs with sequence numbers less than *loghead* are archived logs and can be moved. You can check the value of this parameter by using the Control Center; or, by using the command line processor and the GET DATABASE CONFIGURATION command to view the "First active log file". Refer to *Administration Guide: Performance* for more information about this configuration parameter.

Notes:

1. If you erase an active log, the database becomes unusable and must be restored before it can be used again. You will be able to roll forward only up to the first log that was erased.
2. If you are concerned that your active logs may be damaged (as a result of a disk crash), you should consider mirroring the volumes on which the logs are stored.

Reducing Logging on Work Tables

If your application creates and populates work tables from master tables, and you are not concerned about the recoverability of these work tables because they can be easily recreated from the master tables, you may want to create the work tables specifying the NOT LOGGED INITIALLY parameter on the CREATE TABLE statement. The advantage of using the NOT LOGGED INITIALLY parameter is that any changes made on the table (including insert, delete, update, or create index operations) in the same unit of work that creates the table will not be logged. This not only reduces the logging that is done, but may also increase the performance of your application. You can achieve the same result for existing tables by using the ALTER TABLE statement with the NOT LOGGED INITIALLY parameter.

Notes:

1. You can create more than one table with the NOT LOGGED INITIALLY parameter in the same unit of work.
2. Changes to the catalog tables and other user tables are still logged.

Because changes to the table are not logged, you should consider the following when deciding to use the NOT LOGGED INITIALLY parameter:

- *All* changes to the table must be flushed out to disk at commit time. This means that the commit may take longer.
- An error returned for any operation in a unit of work in which the table is created will result in the rollback of the entire unit of work (SQLCODE -1476, SQLSTATE 40506).
- You cannot recover these tables when rolling forward. If the roll-forward operation encounters a table that was created with the NOT LOGGED INITIALLY option, the table is marked as unavailable. After the database is recovered, any attempt to access the table returns SQL1477N.

  Note: When a table is created, row locks are held on the catalog tables until a COMMIT is done. To take advantage of the no logging behavior, you must populate the table in the same unit of work in which it is created. This has implications for concurrency. For more information, refer to ″Concurrency″ in the *Administration Guide: Performance*.

For more information about creating tables, refer to the *SQL Reference*.

If you plan to use declared temporary tables as work tables, note the following:

- Declared temporary tables are not created in the catalogs; therefore locks are not held.
- Logging is not performed against declared temporary tables, even after the first COMMIT.

- Use the ON COMMIT PRESERVE option to keep the rows in the table after a COMMIT; otherwise, all rows will be deleted.
- Only the application that creates the declared temporary table can access that instance of the table.
- The table is implicitly dropped when the application connection to the database is dropped.
- Errors in operation during a unit of work using a declared temporary table do not cause the unit of work to be completely rolled back. However, an error in operation in a statement changing the contents of a declared temporary table will delete all the rows in that table. A rollback of the unit of work (or a savepoint) will delete all rows in declared temporary tables that were modified in that unit of work (or savepoint).

For more information about declared temporary tables and their limitations, refer to the DECLARE GLOBAL TEMPORARY TABLE statement in the *SQL Reference*.

Point of Recovery

The version and roll-forward recovery methods provide different points of recovery. The version method involves making an offline, full database backup copy of the database at scheduled times. With this method, the recovered database is only as current as the backup copy that was restored. For instance, if you make a backup copy at the end of each day, and you lose the database midway through the next day, you will lose a half-day of changes.

In the roll-forward recovery method, changes made to the database are retained in logs. With this method, you first restore the database or table spaces using a backup copy; then you use the logs to reapply changes that were made to the database since the backup copy was created.

With roll-forward recovery enabled, you can take advantage of online backup and table space level backup. For full database and table space roll-forward recovery, you can choose to recover to the end of the logs, or to a specified point in time. For instance, if an application corrupted the database, you could start with a restored copy of the database, and roll forward changes up to just before that application started. No units of work written to the logs after the time specified are reapplied.

You can also roll forward table spaces to the end of the logs, or to a specific point in time.

Frequency of Backups and Time Required

Your recovery plan should allow for regularly scheduled backups, since backing up a database requires time and system resources.

You should take full database backups regularly, even if you archive the logs (which allows for roll-forward recovery). If your recovery strategy includes roll-forward recovery, a recent full database backup will mean that there are fewer archived logs to apply to the database, which reduces the amount of time required by the ROLLFORWARD utility to recover the database.

You should also consider not overwriting backups and logs, saving more than one full database backup and its associated logs as an extra precaution.

If the amount of time needed to apply archived logs when recovering and rolling forward a very active database is a major concern, consider the cost of backing up the database more frequently. This reduces the number of archived logs you need to apply when rolling forward.

You can perform a backup while the database is either *online* or *offline*. If it is online, other applications or processes can continue to connect to the database, as well as read and modify data while the backup operation is running. If the backup is performed offline, only the backup operation can be connected to the database; the rest of your organization cannot connect to the database while the backup task is running.

To reduce the amount of time that the database is not available, consider using online backups. Online backups are supported only if roll-forward recovery is enabled. If roll-forward recovery is enabled and you have a complete set of logs, you can rebuild the database, should the need arise.

Notes:

1. You can only use an online backup if you have the database log (or logs) that span the time taken for the backup operation.
2. Offline backups are faster than online backups.

If a database contains large amounts of long field and LOB data, backing up the database could be very time-consuming. The BACKUP command provides the capability of backing up selected table spaces. If you use DMS table spaces, you can store different types of data in their own table spaces to reduce the time required for backup operations. You can keep table data in one table space, long field and LOB data in another table space, and indexes in another table space. By storing long field and LOB data in separate table spaces, the time required to complete the backup can be reduced by choosing not to back up the table spaces containing the long field and LOB data. If the long field and LOB data is critical to your business, backing up these table spaces should be considered against the time required to complete the restore operation for these table spaces. If the LOB data can be reproduced from a separate source, choose the NOT LOGGED option when creating or altering a table to include LOB columns.

If you reorganize a table, you should back up the affected table spaces after the operation completes. If you have to restore the table spaces, you will not have to roll forward through the data reorganization.

Note: If you back up a table space that does not contain all of the table data, you cannot perform point-in-time roll-forward recovery on that table space. All the table spaces that contain any type of data for a table must be rolled forward simultaneously to the same point in time.

Recovery Time Required

The time required to recover a database is made up of two parts: the time required to complete the restoration of the backup; and, if the database is enabled for forward recovery, the time required to apply the logs during the roll-forward operation. When formulating a recovery plan, you should take these recovery costs and their impact on your business operations into account. Testing your overall recovery plan will assist you in determining whether the time required to recover the database is reasonable given your business requirements. Following each test, you may want to increase the frequency with which you take a backup. If roll-forward recovery is part of your strategy, this will reduce the number of logs that are archived between backups and, as a result, reduce the time required to roll forward the database after a restore operation.

Note: The setting of the ″enable intra-partition parallelism″ (*intra_parallel*) database manager configuration parameter does not affect the performance of either backup or restore operations. Multiple processes will be used for each of these operations, regardless of the setting of the *intra_parallel* parameter.

Storage Considerations

When deciding which recovery method to use, consider the storage space required.

The version recovery method requires space to hold the backup copy of the database and the restored database. The roll-forward recovery method requires space to hold the backup copy of the database or table spaces, the restored database, and the archived database logs.

If a table contains long field or large object (LOB) columns, you should consider placing this data into a separate table space. This will affect your storage space considerations, as well as affect your plan for recovery. With a separate table space for long field and LOB data, and knowing the time required to back up long field and LOB data, you may decide to use a recovery plan that only occasionally saves a backup of this table space. You may also choose, when creating or altering a table to include LOB columns, not to log changes to those columns. This will reduce the size of the required log space and the corresponding log archive space.

The backup of an SMS table space that contains LOBs can be larger than the size of the original table space. The backup can be as much as 40 per cent larger, depending on the LOB data size in the table space. For example, if you take a backup of a 1 GB SMS table space (with LOBs), you will need more than 1 GB of disk space when you restore it. This only occurs on file systems that support sparse allocation (for example, on UNIX based operating systems).

To prevent media failure from destroying a database and your ability to rebuild it, keep the database backup, the database logs, and the database itself on different devices. For this reason, it is highly recommended that you use the *newlogpath* configuration parameter to put database logs on a separate device once the database is created. (This and other configuration parameters related to logging are discussed in ″Rolling Forward Changes in a Database″ in the *Administration Guide: Implementation*.)

The database logs can use up a large amount of storage. If you plan to use the roll-forward recovery method, you must decide how to manage the archived logs. Your choices are the following:

- Dedicate enough space in the database log path directory to retain the logs.
- Manually copy the logs to a storage device or directory other than the database log path directory after they are no longer in the active set of logs.
- Use a user exit program to copy these logs to another storage device in your environment. (For more information, see ″User Exit for Database Recovery″ in the *Administration Guide: Implementation*.)

Note: On OS/2, DB2 supports a user exit program to handle the storage of both database backup images and database logs on standard and non-standard devices. For more information, see ″User Exit for Database Recovery″ in the *Administration Guide: Implementation*.

Keeping Related Data Together

As part of your database design, you will know the relationships that exist between tables. These relationships can be expressed at the application level, when transactions update more than one table, or at the database level, where referential integrity exists between tables, or where triggers on one table affect another table. You should consider these relationships when developing a recovery plan. You will want to back up related sets of data together. Such sets can be established at either the table space or the database level. By keeping related sets of data together, you can recover to a point where all of the data is consistent. This is especially important if you want to be able to perform point-in-time roll-forward recovery on table spaces.

Restrictions on Using Different Operating Systems

When working in an environment that has more than one operating system, you must consider that the backup and recovery plans cannot be integrated. That is, you may not use the BACKUP DATABASE command on one operating system, and the RESTORE DATABASE command on another operating system. You should keep the recovery plans for each operating system separate and independent.

If you must move tables from one operating system to another, use the **db2move** command, or use the EXPORT with the IMPORT or LOAD commands. For more information, refer to the *Data Movement Utilities Guide and Reference*.

Damaged Table Space Recovery

A damaged table space has one or more containers that cannot be accessed. This is often caused by media problems that are either permanent (for example, a bad disk), or temporary (for example, an offline disk, or an unmounted file system).

If the damaged table space is the system catalog table space, the database cannot be restarted. If the container problems cannot be fixed leaving the original data intact, the only available options are:

- To restore the database; or,
- To restore the catalog table space.

Note: Table space restore is only valid for recoverable databases, since the database must be rolled forward.

If the damaged table space is not the system catalog table space, DB2 attempts to make as much of the database available as possible; success in this case depends on the logging strategy.

If the damaged table space is a sole temporary table space, you should create a new temporary table space as soon as a connection to the database is made. Once created, the new temporary table space can be used, and normal database operations requiring temporary table space can resume. You can, if you wish, drop the offline temporary table space. There are special considerations for table reorganization using a system temporary table space:

- If the database or database manager configuration parameter *indexrec* is set to ″RESTART″, all invalid indexes must be rebuilt during database activation; this includes indexes from reorganization that crashed during the build phase.
- If there are incomplete reorganization requests in a damaged temporary table space, you may have to set the *indexrec* configuration parameter to ″ACCESS″ to avoid restart failures.

Table Space Recovery for Recoverable Databases: The damaged table space is put in offline and not accessible state, and in roll-forward pending state, because crash recovery is necessary. The restart operation will succeed if there is no additional problem. The damaged table space can be used again once you:

- Fix the damaged containers without losing the original data, and then complete a table space roll-forward operation. (The roll-forward operation will first attempt to bring it from offline to normal state.)
- Perform a table space restore operation after fixing the damaged containers (with or without losing the original data), and then a roll-forward operation.

Table Space Recovery for Non-recoverable Databases: Since crash recovery is necessary, and logs are not kept indefinitely, the restart operation can only succeed if the user is willing to drop the damaged table spaces. (Successful completion of recovery means that the log records necessary to recover the damaged table spaces to a consistent state will be gone; therefore, the only valid action against such table spaces is to drop them.)

You can do this by invoking an unqualified restart database operation. It will succeed if there are no damaged table spaces. If it fails (SQL0290N), you can look in the `db2diag.log` file for a complete list of table spaces that are currently damaged.

- If you are willing to drop all of these table spaces once the restart database operation is complete, you can initiate another restart database operation, listing all of the damaged table spaces under the DROP PENDING TABLESPACES option. If a damaged table space is included in the DROP PENDING TABLESPACES list, the table space is put into drop pending state, and your only option after recovery is to drop the table space. The restart operation continues without recovering this table space. If a damaged table space is *not* included in the DROP PENDING TABLESPACES list, the restart database operation fails with SQL0290N.
- If you are unwilling to drop (and thus lose the data in) these table spaces, your options are to:
  - Wait and fix the damaged containers (without losing the original data), and then try the restart database operation again
  - Perform a database restore operation.

Note: Putting a table space name into the DROP PENDING TABLESPACES list does not mean that the table space will be in drop pending state. This will occur only if the table space is found to be damaged during the restart operation. Once the restart operation is successful, you should issue DROP TABLESPACE statements to drop each of the table spaces that are in drop pending state (invoke the LIST TABLESPACES command to find out which table spaces are in this state). This way the space can be reclaimed, or the table spaces can be recreated.

Recovery Performance Considerations

The following should be considered when thinking about recovery performance:

- You can improve performance for databases that are frequently updated by placing the logs on a separate device. In the case of an online transaction processing (OLTP) environment, often more I/O is needed to write data to the logs than to store a row of data. Placing the logs on a separate device will minimize the disk arm movement that is required to move between a log and the database files.

  You should also consider what other files are on the disk. For example, moving the logs to the disk used for system paging in a system that has insufficient real memory will defeat your tuning efforts.
- To reduce the amount of time required to complete a restore operation:
  - Adjust the restore buffer size. The buffer size must be a multiple of the buffer size that was used during the backup operation.
  - Increase the number of buffers.

    If you use multiple buffers and I/O channels, you should use at least twice as many buffers as channels to ensure that the channels do not have to wait for data. The size of the buffers used will also contribute to the performance of the restore operation. The ideal restore buffer size should be a multiple of the extent size for the table spaces.

    If you have multiple table spaces with different extent sizes, specify a value that is a multiple of the largest extent size.

    The *minimum* recommended number of buffers is the number of media devices or containers plus the number specified for the PARALLELISM option.
  - Use multiple source devices.
  - Set the PARALLELISM option for the restore operation to be at least one (1) greater than the number of source devices.
- If a table contains large amounts of long field and LOB data, restoring it could be very time consuming. If the database is enabled for roll-forward recovery, the RESTORE command provides the capability to restore selected table spaces. If the long field and LOB data is critical to your business, restoring these table spaces should be considered against the time required to complete the backup task for these table spaces. By storing long field and LOB data in separate table spaces, the time required to complete the restore operation can be reduced by choosing not to restore the table spaces containing the long field and LOB data. If the LOB data can be reproduced from a separate source, choose the NOT LOGGED option when creating or altering a table to include LOB columns. If you choose not to restore the

table spaces that contain long field and LOB data, but you need to restore the table spaces that contain the table, you must roll forward to the end of the logs so that all table spaces that contain table data are consistent.

Note: If you back up a table space that contains table data without the associated long or LOB fields, you cannot perform point-in-time roll-forward recovery on that table space. All the table spaces for a table must be rolled forward simultaneously to the same point in time.

- The following apply for both backup and restore operations:
  - Multiple I/O buffers and devices should be used.
  - Allocate at least twice as many buffers as devices being used.
  - Do not overload the I/O device controller bandwidth.
  - Use more buffers of smaller size rather than a few large buffers.
  - Tune the number and the size of the buffers according to the system resources.

Disaster Recovery Considerations

The term *disaster recovery* is used to describe the activities that need to be done to restore the database in the event of a fire, earthquake, vandalism, or other catastrophic events. A plan for disaster recovery can include one or more of the following:

- A site to be used in the event of an emergency
- A different machine on which to recover the database
- Off-site storage of database backups and archived logs.

If your plan for disaster recovery is to recover the entire database on another machine, you require at least one full database backup and all the archived logs for the database. You may choose to keep a standby database up to date by applying the logs to it as they are archived. Or, you may choose to keep the database backup and log archives in the standby site, and perform restore and roll-forward operations only after a disaster has occurred. (In this case, a recent database backup is clearly desirable.) With a disaster, however, it is generally not possible to recover all of the transactions up to the time of the disaster.

The usefulness of a table space backup for disaster recovery depends on the scope of the failure. Typically, disaster recovery requires that you restore the entire database; therefore, a full database backup should be kept at a standby site. Even if you have a separate backup image of every table space, you cannot use them to recover the database. If the disaster is a damaged disk, a table space backup of each table space on that disk can be used to recover. If you have lost access to a container because of a disk failure (or for any other reason), you can restore the container to a different location. For additional information, see "Redefining Table Space Containers During RESTORE" in the *Administration Guide: Implementation*.

Both table space backups and full database backups can have a role to play in any disaster recovery plan. The DB2 facilities available for backing up, restoring, and rolling forward data provide a foundation for a disaster recovery plan. You should ensure that you have tested recovery procedures in place to protect your business.

Reducing the Impact of Media Failure

To reduce the probability of media failure, and to simplify recovery from this type of failure:

- Mirror or duplicate the disks that hold the data and logs for important databases.
- In a partitioned database environment, set up a rigorous procedure for handling the data and the logs on the catalog node. Because this node is critical for maintaining the database:
  - Ensure that it resides on a reliable disk
  - Duplicate it
  - Make frequent backups
  - Do not put user data on it.

Protecting Against Disk Failure

If you are concerned about the possibility of damaged data or logs due to a disk crash, consider the use of some form of disk fault tolerance. Generally, this is accomplished through the use of a *disk array*. A disk array consists of a collection of disk drives that appear as a single large disk drive to an application.

Disk arrays involve *disk striping*, which is the distribution of a file across multiple disks, the mirroring of disks, and data parity checks.

Disk arrays are sometimes referred to simply as RAID (Redundant Array of Independent Disks). The specific term RAID generally applies only to hardware disk arrays. Disk arrays can also be provided through software at the operating system or application level. The point of distinction between hardware and software disk arrays is how CPU processing of I/O requests is handled. For hardware disk arrays, I/O activity is managed by disk controllers; for software disk arrays, this is done by the operating system or an application.

Hardware Disk Arrays (RAID): In a RAID disk array, multiple disks are used and managed by a disk controller, complete with its own CPU. All of the logic required to manage the disks forming this array is contained on the disk controller; therefore, this implementation is operating system independent.

There are five types of RAID architecture, RAID-1 through RAID-5, and each provides disk fault tolerance. Each varies in function and performance. In general, RAID refers to a redundant array. RAID-0, which provides only data striping (and not fault-tolerant redundancy), is excluded from this discussion. Although the RAID specification defines five architectures, only RAID-1 and RAID-5 are typically used today.

RAID-1 is also known as disk mirroring or duplexing. Disk mirroring duplicates data (a complete file) from one disk onto a second disk, using a single disk controller. Disk duplexing is the same as disk mirroring, except that disks are attached to a second disk controller (like two SCSI adapters). Data protection is good. Either disk can fail, and data is still accessible from the other disk. With duplexing, a disk controller can also fail without compromising data protection. Performance with RAID-1 is also good, but the trade-off in this implementation is that the required disk capacity is twice that of the actual amount of data, because data is duplicated on pairs of drives.

RAID-5 involves data and parity striping by sectors, across all disks. Parity is interleaved with data information, rather than stored on a dedicated drive. Data protection is good. If any disk fails, the data can still be accessed by using the information from the other disks, along with the striped parity information. Read performance is good, though write performance is considerably worse than that of RAID-1 or normal disk. A RAID-5 configuration requires a minimum of three identical disks. The amount of extra disk space required for overhead varies with the number of disks in the array. In the case of a RAID-5 configuration of 5 disks, the space overhead is 20 percent.

When using a RAID (but not RAID-0) disk array, a failed disk will not prevent you from accessing data on the array. When hot-pluggable or hot-swappable disks are used in the array, a replacement disk can be swapped with the failed disk while the array is in use. With RAID-5, if two disks fail at the same time, all data is lost (but the probability of simultaneous disk failures is very small).

You might consider using RAID-1 or software-mirrored disks (see “Software Disk Arrays” on page 47) for your logs, because this provides for recoverability to the point of failure, and offers good write performance, which is important for logs. In cases where reliability is critical (time cannot be lost recovering data following a disk failure), and write performance is not so critical, consider using RAID-5 disks. Alternatively, if write performance is critical, and you are willing to achieve this despite the cost of additional disk space, consider RAID-1 for your data, as well as for the logs.

Software Disk Arrays: A software disk array accomplishes much the same as does a hardware disk array (see "Hardware Disk Arrays (RAID)" on page 45), but the management of disk traffic is done by either an operating system task, or an application program running on the server. Like other programs, the software array must contend for CPU and system resources. This is not a good option for a CPU-constrained system, and it should be remembered that overall disk array performance is dependent on the server's CPU load and capacity.

A typical software disk array provides disk mirroring (see "Hardware Disk Arrays (RAID)" on page 45). Although redundant disks are required, a software disk array is comparatively inexpensive to implement, because costly RAID disk controllers are not required.

Note: Having the operating system boot drive in the disk array prevents your system from starting if that drive fails. If the drive fails before the disk array is running, the disk array cannot allow access to the drive. A boot drive should be separate from the disk array.

Reducing the Impact of Transaction Failure

To reduce the impact of a transaction failure, try to ensure:

- An uninterrupted power supply
- Adequate disk space for database logs
- Reliable communication links among the database partition servers in a partitioned database environment
- Synchronization of the system clocks in a partitioned database environment (see "System Clock Synchronization in a Partitioned Database System").

System Clock Synchronization in a Partitioned Database System

You should maintain relatively synchronized system clocks across the database partition servers to ensure smooth database operations and unlimited forward recoverability. Time differences among the database partition servers, plus any potential operational and communications delays for a transaction should be less than the value specified for the *max_time_diff* (maximum time difference among nodes) database manager configuration parameter.

To ensure that the log record time stamps reflect the sequence of transactions in a partitioned database system, DB2 uses the system clock on each machine as the basis for the time stamps in the log records. If, however, the system clock is set ahead, the log clock is automatically set ahead with it. Although the system clock can be set back, the clock for the logs cannot, and remains at the *same* advanced time until the system clock matches this time. The clocks are then in synchrony. The implication of this is that a short term system clock error on a database node can have a long lasting effect on the time stamps of database logs.

For example, assume that the system clock on database partition server A is mistakenly set to November 7, 1999 when the year is 1997, and assume that the mistake is corrected *after* an update transaction is committed in the partition at that database partition server. If the database is in continual use, and is regularly updated over time, any point between November 7, 1997 and November 7, 1999 is virtually unreachable through roll-forward recovery. When the COMMIT on database partition server A completes, the time stamp in the database log is set to 1999, and the log clock remains at November 7, 1999 until the system clock matches this time. If you attempt to roll forward to a point in time within this time frame, the operation will stop at the first time stamp that is beyond the specified stop point, which is November 7, 1997.

Although DB2 cannot control updates to the system clock, the *max_time_diff* database manager configuration parameter reduces the chances of this type of problem occurring:

- The configurable values for this parameter range from 1 minute to 24 hours. Refer to *Administration Guide: Performance* for more information about setting *max_time_diff*.
- When the first connection request is made to a non-catalog node, the database partition server sends its time to the catalog node for the database. The catalog node then checks that the time on the node requesting the connection, and its own time are within the range specified by the *max_time_diff* parameter. If this range is exceeded, the connection is refused.
- An update transaction that involves more than two database partition servers in the database must verify that the clocks on the participating database partition servers are in synchrony before the update can be committed. If two or more database partition servers have a time difference that exceeds the limit allowed by *max_time_diff*, the transaction is rolled back to prevent the incorrect time from being propagated to other database partition servers.

To correct and prevent an incorrect time stamp in a database log from being propagated further:

1. Adjust the system clock to the correct time.
2. Restore the database partition on the appropriate database partition server with a backup that was taken before the time was incorrectly set.
3. Roll forward the changes to the end of the log for the database partition.
4. Take a backup copy of the database partition immediately after the changes are rolled forward.

After you complete these actions, the log time will be adjusted, the incorrect time stamp will not be propagated, and you will be able to do point-in-time recovery from the last backup taken on the database partition.

Reorganizing Tables in a Database

A table can become fragmented after many updates, causing performance to deteriorate. If you collected statistics and did not notice a visible performance improvement, reorganizing table data may help. When you reorganize table data, you are rearranging the data into a physical sequence according to a specified index, and removing the free space that is inherent in fragmented data. This can provide faster access to the data, thereby improving performance.

Before you reorganize tables, it is recommended that you invoke the REORGCHK command, and collect statistics on the table. Running this command will help you determine whether a reorganization of the table data is appropriate. Refer to the *Command Reference* for information about the REORGCHK command.

Overview of DB2 Security

To protect data and resources associated with a database server, DB2 uses a combination of external security services and internal access control information. To access a database server, you must pass some security checks before you are given access to database data or resources. The first step in database security is called *authentication*, where you must prove that you are who you say you are. The second step is called *authorization*, where the database manager decides if the validated user is allowed to perform the requested action, or access the requested data.

Authentication

Authentication of a user is completed using a security facility outside of DB2. The security facility can be part of the operating system, a separate product or, in certain cases, may not exist at all. On UNIX based systems, the security facility is in the operating system itself. DCE Security Services is a separate product that provides the security facility for a distributed environment. There are no security facilities on the Windows 95 or the Windows 3.1 operating system.

The security facility requires two items to authenticate a user: a user ID and a password. The user ID identifies the user to the security facility. By supplying the correct password (information known only to the user and the security facility) the user’s identity (corresponding to the user ID) is verified.

Once authenticated:

- The user must be identified to DB2 using an SQL authorization name or *authid*. This name can be the same as the user ID, or a mapped value. For example, on UNIX based systems, a DB2 *authid* is derived by transforming to uppercase letters a UNIX user ID that follows DB2 naming conventions. Within the DCE Security Services product, the DB2 *authid* is contained in the DCE registry, and is extracted from there once authentication has successfully completed.
- A list of groups to which the user belongs is obtained. Group membership may be used when authorizing the user. Groups are security facility entities that must also map to DB2 authorization names. This mapping is done in a method similar to that used for user IDs.

  DB2 will obtain a list of groups up to a maximum of 64 groups. If a user is a member of more than 64 groups, only the first 64 that map to valid DB2 authorization names are added to the DB2 group list. No error is returned, and any groups after the first 64 are ignored by DB2.

DB2 uses the security facility to authenticate users in one of two ways:

- DB2 uses a successful security system login as evidence of identity, and allows:
  - Use of local commands to access local data
  - Use of remote connections where the server trusts the client authentication.
- DB2 accepts a user ID and password combination. It uses successful validation of this pair by the security facility as evidence of identity and allows:
  - Use of remote connections where the server requires proof of authentication
  - Use of operations where the user wants to run a command under an identity other than the identity used for login.

DB2 administrators can allow others to change passwords on AIX and Windows NT EEE systems through the profile registry variable DB2CHGPWD_EEE. The default value for this variable is `NOT SET` (disabled). DB2CHGPWD_EEE accepts the standard boolean values used by other DB2 profile variables.

The DB2 administrator is responsible for ensuring that the passwords for all nodes are maintained centrally, using either a Windows NT Domain Controller on Windows NT, or NIS on AIX.

Note: If the passwords are not maintained centrally, enabling the DB2CHGPWD_EEE variable may result in passwords not being consistent across all nodes. That is, if you use the ″change password″ feature, your password will only be changed at the node to which you are connected.

DB2 UDB on AIX can log failed password attempts with the operating system, and detect when a client has exceeded the number of allowable login tries, as specified by the LOGINRETRIES parameter.

For additional information about the system entry validation checking that is particularly relevant if you have remote clients accessing the database, see ″Selecting an Authentication Method for Your Server″ in the *Administration Guide: Implementation*.

Authorization

Authorization is the process whereby DB2 obtains information about an authenticated DB2 user, indicating the database operations that user may perform, and what data objects may be accessed. With each user request, there may be more than one authorization check, depending on the objects and operations involved.

Authorization is performed using DB2 facilities. DB2 tables and configuration files are used to record the permissions associated with each authorization name. The authorization name of an authenticated user, and those of groups to which the user belongs, are compared with the recorded permissions. Based on this comparison, DB2 decides whether to allow the requested access.

There are two types of permissions recorded by DB2: privileges and authority levels. A *privilege* defines a single permission for an authorization name, enabling a user to create or access database resources. Privileges are stored in the database catalogs. *Authority levels* provide a method of grouping privileges and control over higher-level database manager maintenance and utility operations. Database-specific authorities are stored in the database catalogs; system authorities are associated with group membership, and are stored in the database manager configuration file for a given instance.

Groups provide a convenient means of performing authorization for a collection of users without having to grant or revoke privileges for each user individually. Unless otherwise specified, group authorization names can be used anywhere that authorization names are used for authorization purposes. In general, group membership is considered for dynamic SQL and non-database object authorizations (such as instance level commands and utilities), but is not considered for static SQL. The exception to this general case occurs when privileges are granted to PUBLIC: these are considered when static SQL is processed. Specific cases where group membership does not apply are noted throughout the DB2 documentation, where applicable.

For more information, see ″Privileges, Authorities, and Authorization″ in the *Administration Guide: Implementation*.

Federated Database Authentication and Authorization Overview

Because a DB2 federated database system can access information in multiple database management systems, additional steps may be required to secure your data.

When planning your approach to authentication, consider the fact that users may need to pass authentication checks at data sources as well as at DB2. In a federated system, authentication can take place at DB2 client workstations, DB2 servers, data sources (DB2, DB2 for OS/390, other DRDA servers, Oracle), or a combination of DB2 (client or DB2 server) and data sources. Even in DCE environments, specific steps may be necessary if data sources require a user ID and password. For more information, see ″Federated Database Authentication Processing″ in the *Administration Guide: Implementation*.

Similarly, users must pass authorization checking at data sources and at DB2. Each data source (DB2, Oracle, DB2 for OS/390, and so on) maintains the security of the objects under its control. When a user performs an operation against a nickname, that user must pass authorization checking for the table or view referenced by the nickname.

Chapter 3. Federated Systems

A *federated database system* or *federated system* is a database management system (DBMS) that supports applications and users submitting SQL statements referencing two or more DBMSs or databases in a single statement. An example is a join between tables in two different DB2 databases. This type of statement is called a *distributed request*.

A DB2 Universal Database federated system provides support for distributed requests across databases and DBMSs. You can, for example, perform a UNION operation between a DB2 table and an Oracle view. Supported DBMSs include DB2, members of the DB2 family (such as DB2 for OS/390 and DB2 for AS/400), and Oracle.

A DB2 federated system provides *location transparency* for database objects. If information (in tables and views) is moved, references to that information (called *nicknames*) can be updated without any changes to applications that request the information. A DB2 federated system also provides *compensation* for DBMSs that do not support all of the DB2 SQL dialect, or certain optimization capabilities. Operations that cannot be performed under such a DBMS (such as recursive SQL) are run under DB2.

A DB2 federated system functions in a *semi-autonomous* manner: DB2 queries containing references to Oracle objects can be submitted while Oracle applications are accessing the same server. A DB2 federated system does not monopolize or restrict access (beyond integrity and locking constraints) to Oracle or other DBMS objects.

A DB2 federated system consists of a DB2 UDB instance, a database that will serve as the *federated database*, and one or more *data sources*. The federated database contains catalog entries identifying data sources and their characteristics. A data source consists of a DBMS and data. Applications connect to the federated database just like any other DB2 database. See Figure 20 on page 54 for a visual representation of a federated database environment.

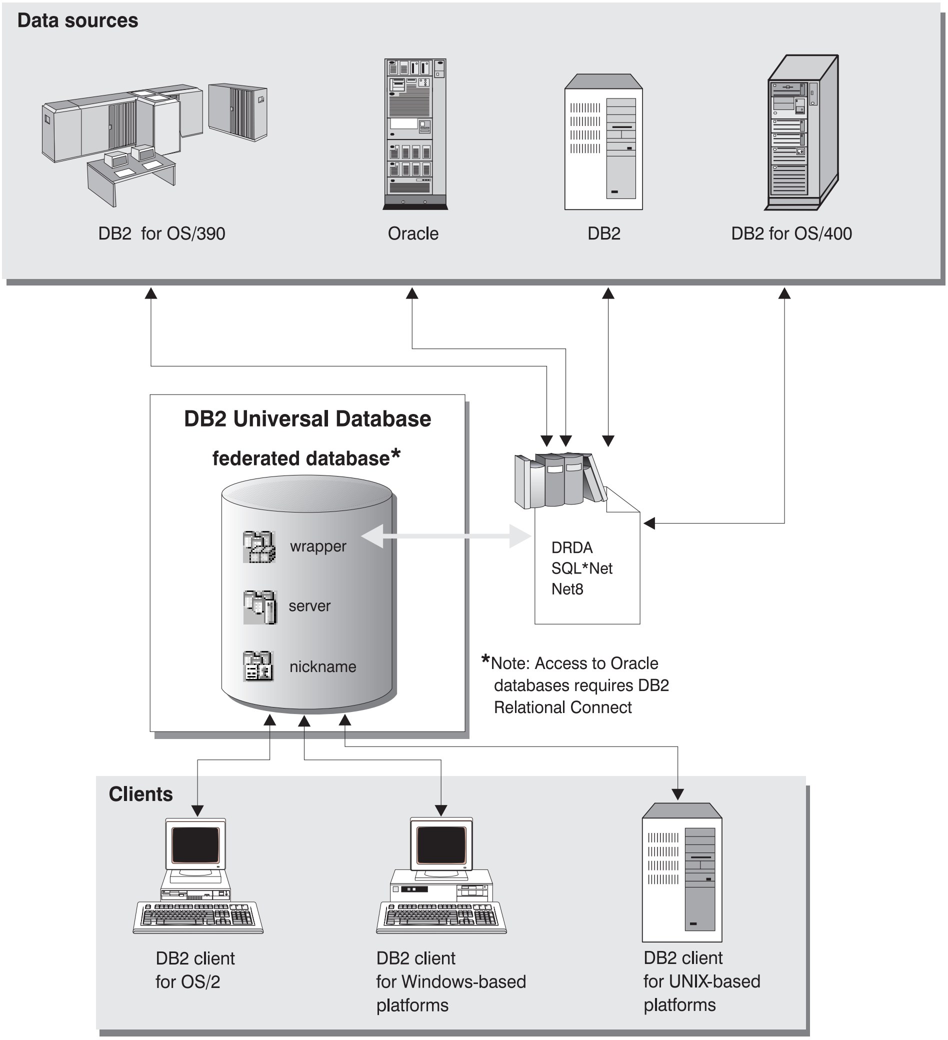


Figure 20. A Federated Database System

DB2 federated database catalog entries contain information about data source objects: what they are called, what information they contain, and conditions under which they can be used. Because this DB2 catalog stores information about objects in many DBMSs, it is called a *global catalog*. Object attributes are stored in the catalog. The actual DBMSs being referenced, modules used to communicate with the data source, and DBMS data objects (such as tables) that will be accessed are outside of the database. (One exception: a federated database can be a data source for the federated system.) You can create federated objects using the Control Center or SQL DDL statements. Required federated database objects are:

Wrappers
Identify the modules (DLL, library, and so on) used to access a particular class or category of data source.

Servers
Define data sources. Server data includes the wrapper name, server name, server type, server version, authorization information, and server options.

Nicknames
Identifiers stored in the federated database that reference specific data source objects (tables, aliases, views). Applications reference nicknames in queries just like they reference tables and views.

Depending on your specific needs, you can create additional objects:

- User mappings, to address authentication issues
- Data type mappings, to customize the relationship between a data source type and an DB2 type
- Function mappings, to map a local function to a data source function
- Index specifications, to improve performance.

After a federated system is set up, the information in data sources can be accessed as though it were in one large database. Users and applications send queries to one federated database, which then retrieves data from DB2 family and Oracle systems as needed. User and applications specify nicknames in queries; these nicknames provide references to tables and views located in data sources. From an end-user perspective, nicknames are similar to aliases.

There are many factors affecting federated system performance. The most critical factor is to ensure that accurate and up-to-date information about data sources and their objects is stored in the federated database global catalog. This information is used by the DB2 optimizer, and can affect decisions to push down operations for evaluation at data sources. Refer to the *Administration Guide: Performance* for additional information about federated system performance.

A DB2 federated system operates under some restrictions. Distributed requests are limited to read-only operations. In addition, you cannot execute utility operations (LOAD, REORG, REORGCHK, IMPORT, RUNSTATS, and so on) against nicknames.

You can, however, use a pass-through facility to submit DDL and DML statements directly to database managers using the SQL dialect associated with that data source.

Federated systems tolerate parallel environments. Performance gains are limited by the extent to which a federated database query can be semantically broken down into local object (table, view) references and nickname references. Requests for nickname data are processed sequentially; local objects can be processed in parallel. For example, given the query `SELECT * FROM A, B, C, D`, where A and B are local tables, and C and D are nicknames referencing tables at Oracle data sources, one possible plan would join tables A and B with a parallel join. The results are then joined sequentially with nicknames C and D.

Enabling a Federated System

DB2 Enterprise Edition (EE) and DB2 Enterprise - Extended Edition (EEE) can support federated databases. To enable a federated system:

1. Select the *Distributed Join for DB2 Databases* installation option of DB2 EE or EEE during installation.
2. If including Oracle databases in your federated system, install DB2 Relational Connect. For more information, refer to the *Installation and Configuration Supplement*.
3. Set the database manager configuration parameter *federated* to ″YES″.
4. Create wrappers, servers, and nicknames (see ″Creating a Database″ in the *Administration Guide: Implementation* for more information).
5. Create additional objects, or set options as required (see ″Implementing Your Design″ in the *Administration Guide: Implementation* for more information).

Chapter 4. Parallel Database Systems

DB2 extends the database manager to the parallel, multi-node environment. A *database partition* is a part of a database that consists of its own data, indexes, configuration files, and transaction logs. A database partition is sometimes called a node or a database node. (Node was the term used in the DB2 Parallel Edition for AIX Version 1 product.)

A *single-partition database* is a database having only one database partition. All data in the database is stored in that partition. In this case nodegroups (see “Nodegroups” on page 9), while present, provide no additional capability.

A *partitioned database* is a database with two or more database partitions. Tables can be located in one or more database partitions. When a table is in a nodegroup consisting of multiple partitions, some of its rows are stored in one partition, and other rows are stored in other partitions.

Usually, a single database partition exists on each physical node, and the processors on each system are used by the database manager at each database partition to manage its part of the total data in the database.

Because data is divided across database partitions, you can use the power of multiple processors on multiple physical nodes to satisfy requests for information. Data retrieval and update requests are decomposed automatically into sub-requests, and executed in parallel among the applicable database partitions. The fact that databases are split across database partitions is transparent to users issuing SQL statements.

User interaction occurs through one database partition, known as the *coordinator node* for that user. The coordinator runs on the same database partition as the application, or in the case of a remote application, the database partition to which that application is connected. Any database partition can be used as a coordinator node.

Nodegroups and Data Partitioning

You can define named subsets of one or more database partitions in a database. Each subset you define is known as a *nodegroup*. Each subset that contains more than one database partition is known as a *multi-partition nodegroup*. Multi-partition nodegroups can only be defined with database partitions that belong to the same instance.

Figure 21 shows an example of a database with five partitions in which:

- A nodegroup spans all but one of the database partitions (Nodegroup 1).
- A nodegroup contains one database partition (Nodegroup 2).
- A nodegroup contains two database partitions.
- The database partition within Nodegroup 2 is shared (and overlaps) with Nodegroup 1.
- There is a single database partition within Nodegroup 3 that is shared (and overlaps) with Nodegroup 1.

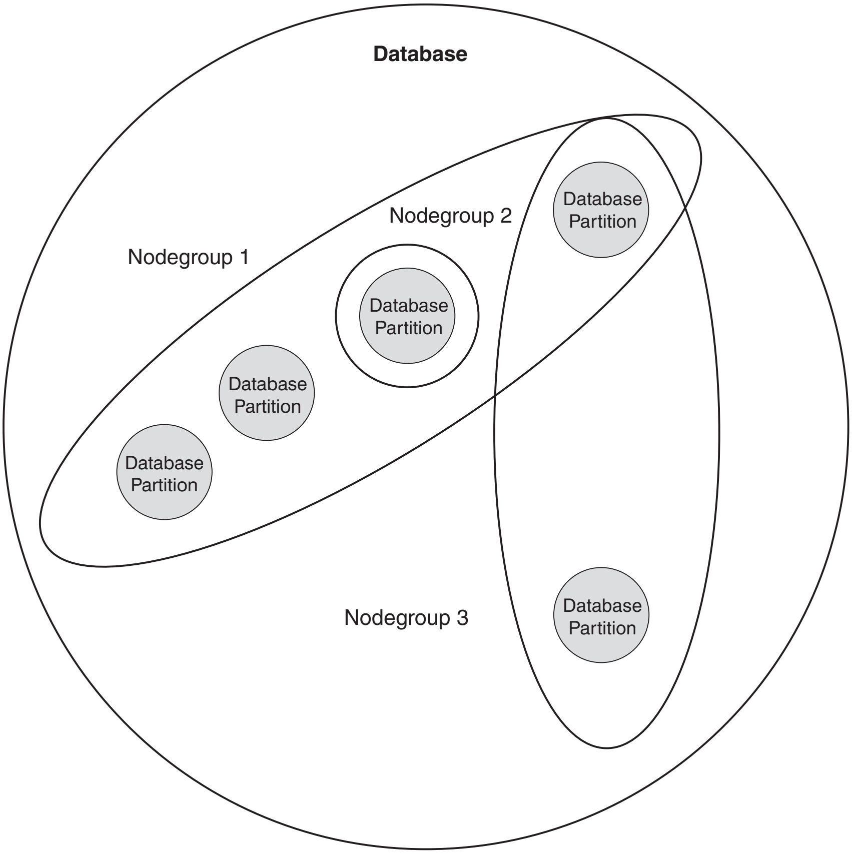


Figure 21. Nodegroups in a Database

You create a new nodegroup using the CREATE NODEGROUP statement. Refer to the *SQL Reference* for more information. Data is divided across all the partitions in a nodegroup. If you are using a multi-partition nodegroup, you must look at several nodegroup design considerations. For more information, see “Designing Nodegroups” on page 124.

Types of Parallelism

Components of a task, such as a database query, can be run in parallel to dramatically enhance performance. The nature of the task, the database configuration, and the hardware environment, all determine how DB2 will perform a task in parallel. These considerations are interrelated, and should be considered together when you work on the physical and logical design of a database. This section describes the following types of parallelism that are supported by DB2:

- I/O
- Query
- Utility

I/O Parallelism

When there are multiple containers for a table space, the database manager can exploit *parallel I/O*. Parallel I/O refers to the process of writing to, or reading from, two or more I/O devices simultaneously; it can result in significant improvements in throughput.

I/O parallelism is a component of each hardware environment described in “Hardware Environments” on page 62. Table 3 on page 70 lists the hardware environments best suited to I/O parallelism.

Query Parallelism

There are two types of query parallelism: inter-query parallelism and intra-query parallelism.

Inter-query parallelism refers to the ability of multiple applications to query a database at the same time. Each query executes independently of the others, but DB2 executes all of them at the same time. DB2 has always supported this type of parallelism.

Intra-query parallelism refers to the simultaneous processing of parts of a single query, using either *intra-partition parallelism*, *inter-partition parallelism*, or both.

The term *query parallelism* is used throughout this book.

Intra-partition Parallelism

Intra-partition parallelism refers to the ability to break up a query into multiple parts. (Some of the utilities also perform this type of parallelism. See “Utility Parallelism” on page 62.)

Intra-partition parallelism subdivides what is usually considered a single database operation such as index creation, database loading, or SQL queries into multiple parts, many or all of which can be run in parallel *within a single database partition*.

Figure 22 shows a query that is broken into four pieces that can be run in parallel, with the results returned more quickly than if the query were run in serial fashion. The pieces are copies of each other. To utilize intra-partition parallelism, you must configure the database appropriately. You can choose the degree of parallelism or let the system do it for you. The degree of parallelism represents the number of pieces of a query running in parallel.

Table 3 on page 70 lists the hardware environments best suited for intra-partition parallelism.

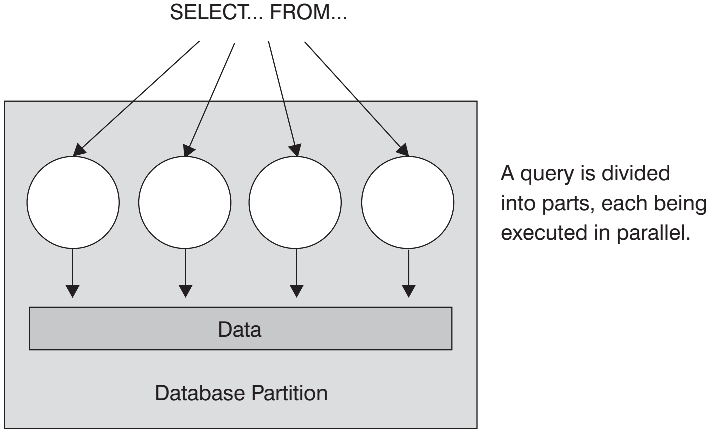

SELECT... FROM...
A query is divided into parts, each being executed in parallel.
Data
Database Partition

Figure 22. Intra-partition Parallelism

Inter-partition Parallelism

Inter-partition parallelism refers to the ability to break up a query into multiple parts across multiple partitions of a partitioned database, on one machine or multiple machines. The query is run in parallel. (Some of the utilities also perform this type of parallelism. See “Utility Parallelism” on page 62.)

Inter-partition parallelism subdivides what is usually considered a single database operation such as index creation, database loading, or SQL queries into multiple parts, many or all of which can be run in parallel *across multiple partitions of a partitioned database on one machine or on multiple machines*.

Figure 23 on page 61 shows a query that is broken into four pieces that can be run in parallel, with the results returned more quickly than if the query were run in serial fashion on a single partition.

The degree of parallelism is largely determined by the number of partitions you create and how you define your nodegroups.

Table 3 on page 70 lists the hardware environments best suited for inter-partition parallelism.

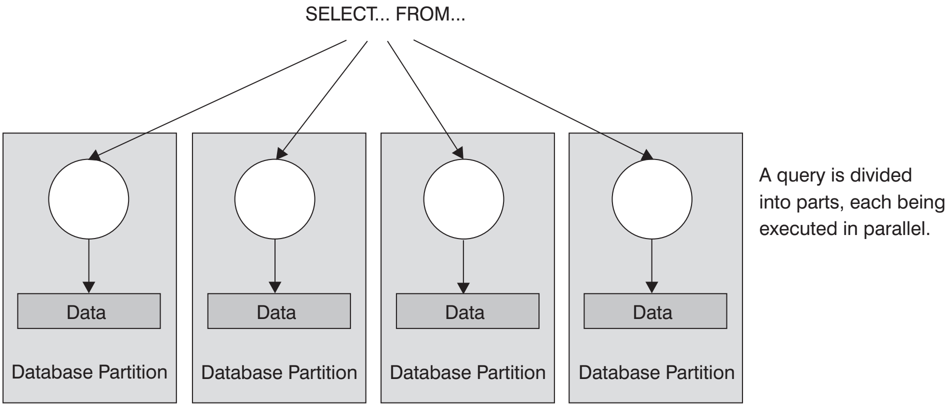


Figure 23. Inter-partition Parallelism

Simultaneous Intra-partition and Inter-partition Parallelism

You can use intra-partition parallelism and inter-partition parallelism at the same time. This combination provides two dimensions of parallelism, resulting in an even more dramatic increase in the speed at which queries are processed:

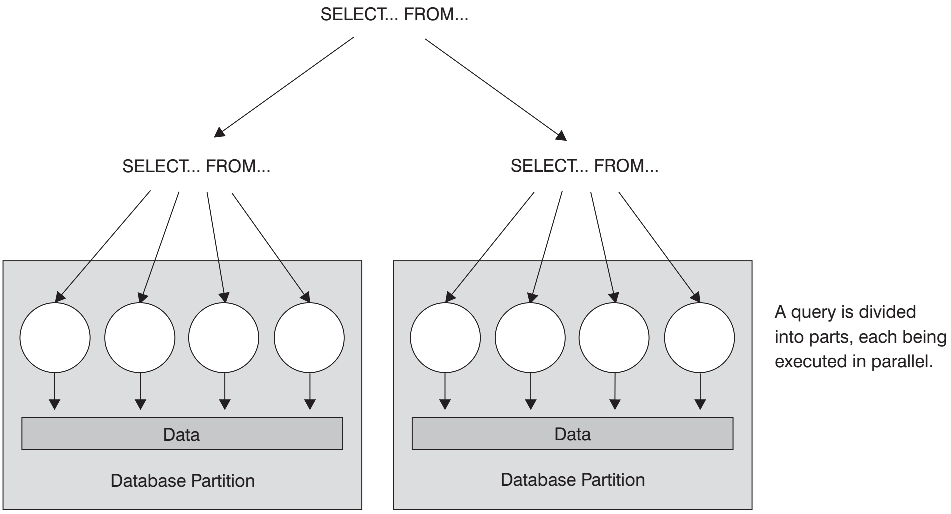


Figure 24. Simultaneous Inter-partition and Intra-partition Parallelism

Utility Parallelism

DB2 utilities can take advantage of intra-partition parallelism. They can also take advantage of inter-partition parallelism; where multiple database partitions exist, the utilities execute in each of the partitions in parallel.

The load utility can take advantage of intra-partition parallelism and I/O parallelism. Loading data is a CPU-intensive task. The load utility takes advantage of multiple processors for tasks such as parsing and formatting data. It can also use parallel I/O servers to write the data to containers in parallel. Refer to the *Data Movement Utilities Guide and Reference* for information on how to enable parallelism for the load utility.

In a partitioned database environment, the AutoLoader utility takes advantage of intra-partition, inter-partition, and I/O parallelism by parallel invocations of the LOAD command at each database partition where the table resides. Refer to the *Data Movement Utilities Guide and Reference* for more information about the AutoLoader utility.

During index creation, the scanning and subsequent sorting of the data occurs in parallel. DB2 exploits both I/O parallelism and intra-partition parallelism when creating an index. This helps to speed up index creation when a CREATE INDEX statement is issued, during restart (if an index is marked invalid), and during the reorganization of data.

Backing up and restoring data are heavily I/O-bound tasks. DB2 exploits both I/O parallelism and intra-partition parallelism when performing backup and restore operations. Backup exploits I/O parallelism by reading from multiple table space containers in parallel, and asynchronously writing to multiple backup media in parallel. Refer to the BACKUP DATABASE command and the RESTORE DATABASE command in the *Command Reference* for information on how to enable parallelism for these utilities.

Hardware Environments

This section provides an overview of the following hardware environments:

- Single partition on a single processor (uniprocessor)
- Single partition with multiple processors (SMP)
- Multiple partition configurations
  - Partitions with one processor (MPP)
  - Partitions with multiple processors (cluster of SMPs)
  - Logical database partitions (also known as Multiple Logical Nodes, or MLN, in DB2 Parallel Edition for AIX Version 1)

Capacity and scalability are discussed for each environment. *Capacity* refers to the number of users and applications able to access the database. This is in large part determined by memory, agents, locks, I/O, and storage management. *Scalability* refers to the ability of a database to grow and continue to exhibit the same operating characteristics and response times.

Single Partition on a Single Processor

This environment is made up of memory and disk, but contains only a single CPU (see Figure 25). It is referred to by many different names, including stand-alone database, client/server database, serial database, uniprocessor system, and single node or non-parallel environment.

The database in this environment serves the needs of a department or small office, where the data and system resources (including a single processor or CPU) are managed by a single database manager.

Table 3 on page 70 lists the types of parallelism best suited to take advantage of this hardware configuration.

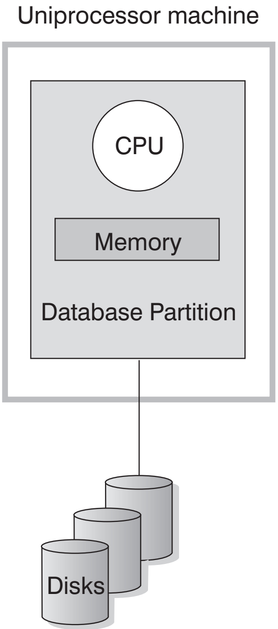


Figure 25. Single Partition On a Single Processor

Capacity and Scalability

In this environment you can add more disks. Having one or more I/O servers for each disk allows for more than one I/O operation to take place at the same time. You can also add more hard disk space to this environment.

A single-processor system is restricted by the amount of disk space the processor can handle. However, as workload increases, a single CPU may not be able to process user requests any faster, regardless of other components, such as memory or disk, that you may add. If you have reached maximum capacity or scalability, you can consider moving to a single partition system with multiple processors.

Single Partition with Multiple Processors

This environment is typically made up of several equally powerful processors within the same machine (see Figure 26 on page 65), and is called a *symmetric multi-processor (SMP)* system. Resources, such as disk space and memory, are *shared*.

With multiple processors available, different database operations can be completed more quickly. DB2 can also divide the work of a single query among available processors to improve processing speed. Other database operations, such as loading data, backing up and restoring table spaces, and creating indexes on existing data, can take advantage of multiple processors.

Table 3 on page 70 lists the types of parallelism best suited to take advantage of this hardware configuration.

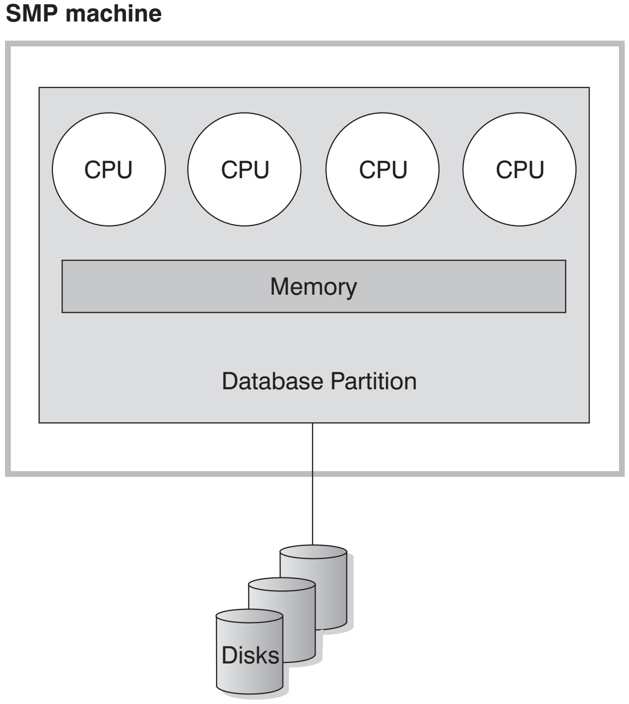


Figure 26. Single Partition Database Symmetric Multiprocessor System

Capacity and Scalability

In this environment you can add more processors. However, since the different processors may attempt to access the same data, limitations with this environment can appear as your business operations grow. With shared memory and shared disks, you are effectively sharing all of the database data.

You can increase the I/O capacity of the database partition associated with your processor by increasing the number of disks. You can establish I/O servers to specifically deal with I/O requests. Having one or more I/O servers for each disk allows for more than one I/O operation to take place at the same time.

If you have reached maximum capacity or scalability, you can consider moving to a system with multiple partitions.

Multiple Partition Configurations

You can divide a database into multiple partitions, each on its own machine. Multiple machines with multiple database partitions can be grouped together. This section describes the following partition configurations:

- Partitions on systems with one processor
- Partitions on systems with multiple processors
- Logical database partitions

Partitions with One Processor

In this environment, there are many database partitions. Each partition resides on its own machine, and has its own processor, memory, and disks (Figure 27 on page 67). All the machines are connected by a communications facility. This environment is referred to by many different names, including cluster, cluster of uniprocessors, massively parallel processing (MPP) environment, and shared-nothing configuration. The latter name accurately reflects the arrangement of resources in this environment. Unlike an SMP environment, an MPP environment has no shared memory or disks. The MPP environment removes the limitations introduced through the sharing of memory and disks.

A partitioned database environment allows a database to remain a logical whole, despite being physically divided across more than one partition. The fact that data is partitioned remains transparent to most users. Work can be divided among the database managers; each database manager in each partition works against its own part of the database.

Table 3 on page 70 lists the types of parallelism best suited to take advantage of this hardware configuration.

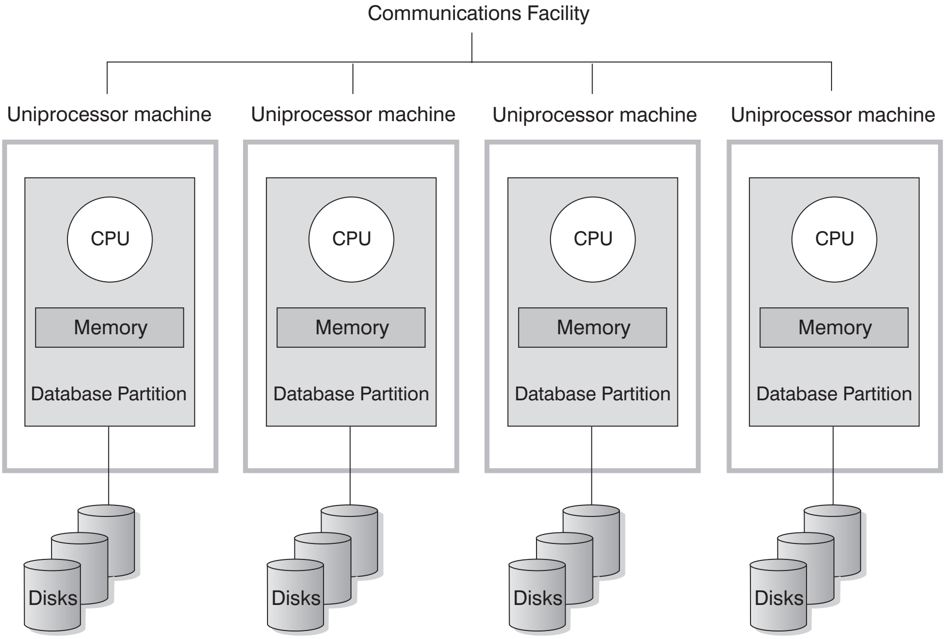


Figure 27. Massively Parallel Processing System

Capacity and Scalability: In this environment you can add more database partitions (nodes) to your configuration. On some platforms, for example the RS/6000 SP, the maximum number is 512 nodes. However, there may be practical limits on managing a high number of machines and instances.

If you have reached maximum capacity or scalability, you can consider moving to a system where each partition has multiple processors.

Partitions with Multiple Processors

An alternative to a configuration in which each partition has a single processor, is a configuration in which a partition has multiple processors. This is known as an *SMP cluster* (Figure 28 on page 68).

This configuration combines the advantages of SMP and MPP parallelism. This means that a query can be performed in a single partition across multiple processors. It also means that a query can be performed in parallel across multiple partitions.

Table 3 on page 70 lists the types of parallelism best suited to take advantage of this hardware configuration.

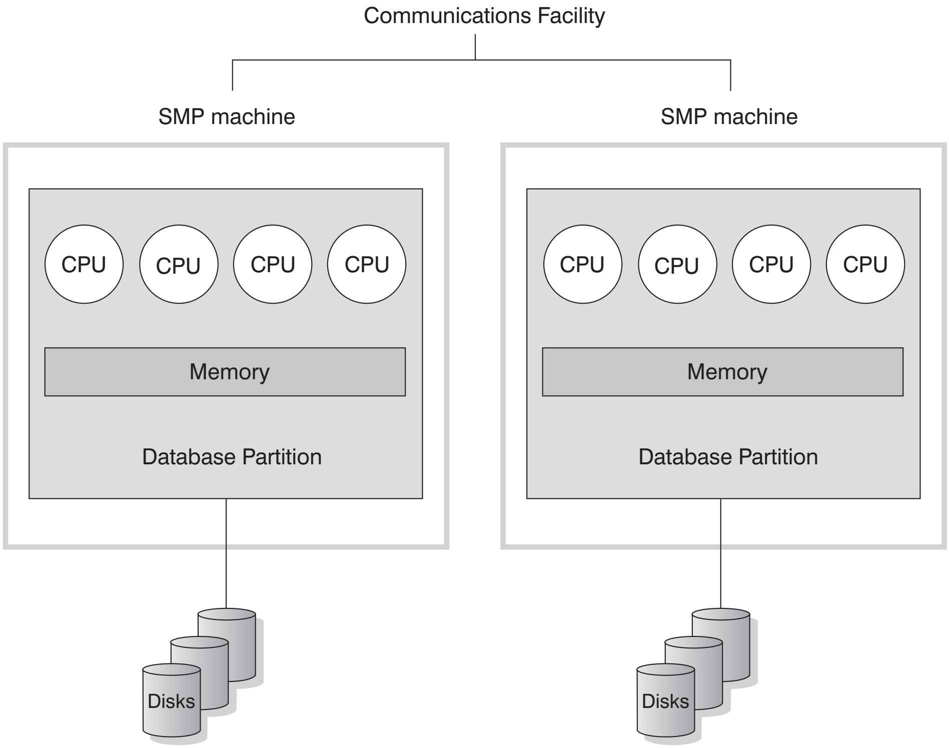


Figure 28. Cluster of SMPs

Capacity and Scalability: In this environment you can add more database partitions, and you can add more processors to existing database partitions.

Logical Database Partitions

A logical database partition differs from a physical partition in that it is not given control of an entire machine. Although the machine has shared resources, database partitions do not share the resources. Processors are shared but disks and memory are not.

Logical database partitions provide scalability. Multiple database managers running on multiple logical partitions may make fuller use of available resources than a single database manager could. Figure 29 on page 69 illustrates the fact that you may gain more scalability on an SMP machine by adding more partitions; this is particularly true for machines with many processors. By partitioning the database, you can administer and recover each partition separately.

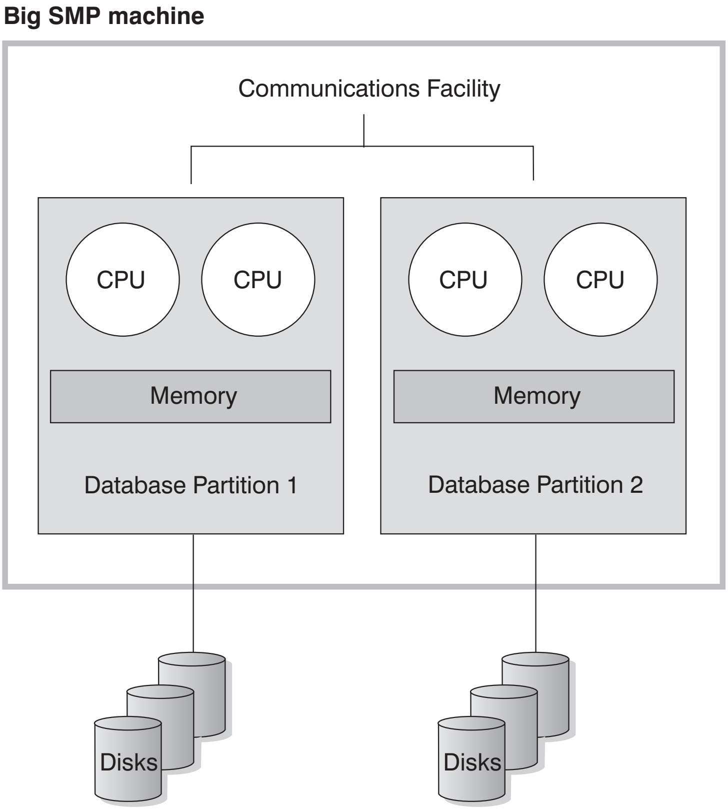


Figure 29. Partitioned Database, Symmetric Multiprocessor System

Figure 30 on page 70 illustrates the fact that you can multiply the configuration shown in Figure 29 to increase processing power.

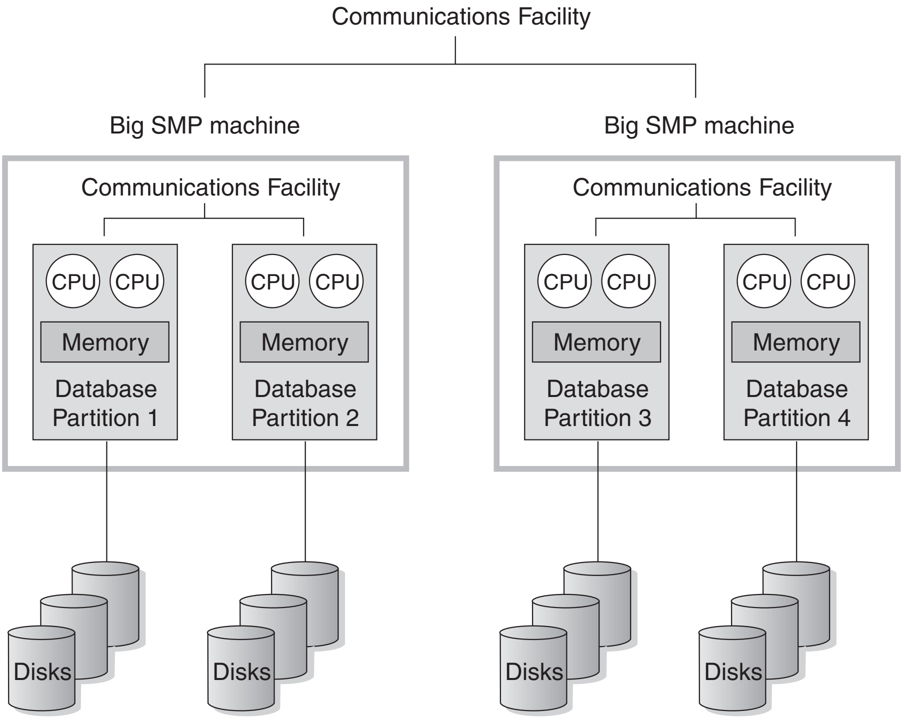


Figure 30. Partitioned Database, Symmetric Multiprocessor Systems Clustered Together

Table 3 lists the types of parallelism best suited to take advantage of this hardware environment.

Note: The ability to have two or more partitions coexist on the same machine (regardless of the number of processors) allows greater flexibility in designing high availability configurations and failover strategies. Upon machine failure, a database partition can be automatically moved and restarted on a second machine that already contains another partition of the same database. For more information, see “Chapter 11. Designing for High Availability” on page 203.

Summary of Parallelism Best Suited to Each Hardware Environment

The following table summarizes the types of parallelism best suited to take advantage of the various hardware environments.

Table 3. Types of Parallelism Possible in Each Hardware Environment

Hardware Environment	**I/O Parallelism**	**Intra-Query Parallelism**	
		Intra- Partition Parallelism	**Inter- Partition Parallelism**
Single Partition, Single Processor	Yes	No(1)	No

Table 3. Types of Parallelism Possible in Each Hardware Environment (continued)

Hardware Environment	**I/O Parallelism**	**Intra-Query Parallelism**	
		Intra- Partition Parallelism	**Inter- Partition Parallelism**
Single Partition, Multiple Processors (SMP)	Yes	Yes	No
Multiple Partitions, One Processor (MPP)	Yes	No(1)	Yes
Multiple Partitions, Multiple Processors (cluster of SMPs)	Yes	Yes	Yes
Logical Database Partitions	Yes	Yes	Yes
Note: (1) There may be an advantage to setting the degree of parallelism (using one of the configuration parameters) to some value greater than one, even on a single processor system, especially if the queries you execute are not fully utilizing the CPU (for example if they are I/O bound).			

Chapter 5. About Data Warehousing

DB2 Universal Database offers the Data Warehouse Center, a component that automates data warehouse processing. You can use the Data Warehouse Center to define the data to include in the warehouse. Then, you can use the Data Warehouse Center to automatically schedule refreshes of the data in the warehouse.

This section provides an overview of data warehousing and data warehousing tasks. For more detailed information about warehousing, and for information on using the Data Warehouse Center, refer to the *Data Warehouse Center Administration Guide* and the Data Warehouse Center online help.

What is Data Warehousing?

The systems that contain *operational data*—the data that runs the daily transactions of your business—contain information that is useful to business analysts. For example, analysts can use information about which products were sold in which regions at which time of year to look for anomalies or to project future sales.

However, there are several problems with analysts accessing the operational data directly:

- They might not have the expertise to query the operational database. For example, querying IMS databases requires an application program that uses a specialized type of data manipulation language. In general, those programmers who have the expertise to query the operational database have a full-time job in maintaining the database and its applications.
- Performance is critical for many operational databases, such as databases for a bank. The system cannot handle users making ad hoc queries.
- The operational data generally is not in the best format for use by business analysts. For example, sales data that is summarized by product, region, and season is much more useful to analysts than the raw data.

Data warehousing solves these problems. In *data warehousing*, you create stores of *informational data*—data that is extracted from the operational data, and then transformed for end-user decision making. For example, a data warehousing tool might copy all the sales data from the operational database, perform calculations to summarize the data, and write the summarized data to a target in a separate database from the operational data. End users can query the separate database (the *warehouse*) without impacting the operational databases.

The following sections describe the objects (subject areas, warehouse sources, warehouse targets, agents, agent sites, steps, and processes) that you will use to create and maintain your data warehouse.

Subject Areas

A *subject area* identifies and groups the processes that relate to a logical area of the business. For example, if you are building a warehouse of marketing and sales data, you can define a Sales subject area and a Marketing subject area. You can then add the processes that relate to sales underneath the Sales subject area. Similarly, you can add the definitions that relate to the marketing data underneath the Marketing subject area.

Warehouse Sources

Warehouse sources identify the tables and files that will provide data to your warehouse. The Data Warehouse Center uses the specifications in the warehouse sources to access and select the data. The sources can be nearly any relational or nonrelational source (table, view, or file) that has connectivity to your warehouse.

Warehouse Targets

Warehouse targets are database tables or files that contain data that has been transformed so that end users can use it. Like a warehouse source, warehouse targets can also provide data to Data Warehouse Center steps.

Warehouse Agents and Agent Sites

Data Warehouse Center *agents* manage the flow of data between the data sources and the target warehouses. Agents are available on the Windows NT, AIX, OS/2, OS/390, OS/400, and SUN Solaris operating systems. The agents use Open Database Connectivity (ODBC) drivers or DB2 CLI to communicate with different databases.

Several agents can handle the transfer of data between sources and target warehouses. The number of agents that you use depends on your existing connectivity configuration and the volume of data that you plan to move through your warehouse. Additional instances of an agent can be generated if multiple processes that require the same agent are running simultaneously.

Agents can be local or remote. A *local warehouse agent* is an agent that is installed on the same machine as the warehouse server. A *remote warehouse agent* is an agent that is installed on another machine that has connectivity to the warehouse server.

An *agent site* is a logical name for a workstation where agent software is installed. The agent site name is not the same as the TCP/IP host name. A single physical machine can have only one TCP/IP host name. However, you can define multiple agent sites on a single machine. A logical name identifies each agent site.

The *default agent site*, named the Default VW AgentSite, is a local agent on Windows NT that Data Warehouse Center defines during initialization of the warehouse control database.

Steps and Processes

A *step* is a logical entity in the Data Warehouse Center that defines:

- The structure of the output table or file.
- The mechanism (either SQL or a program) for populating the output table or file.
- The schedule by which the output table or file is populated.

Steps move data and transform data by using SQL statements or by calling programs. When you run a step, the transfer of data between the warehouse source and the warehouse target, and any transformation of that data, takes place.

A *process* contains a series of steps that perform transformation and movement tasks. In general, a process populates a warehouse target in a warehouse database by extracting data from one or more warehouse sources, which can be database tables or files. However, you can also define a process for launching programs that does not specify any warehouse sources or targets.

You can run a step on demand, or you can schedule a step to run at a set time. You can schedule a step to run one time only, or you can schedule it to run repeatedly, such as every Friday. You can also schedule steps to run in sequence, so that when one step finishes running, the next step begins running. You can schedule steps to run upon (successful or unsuccessful) completion of another step. If you schedule a process, the first step in the process runs at the scheduled time.

When a step or a process runs, it can save data by:

- Replacing all the data in the warehouse target with new data.
- Appending the new data to the existing data.
- Appending a separate edition of data.

Suppose that you want Data Warehouse Center to perform the following tasks:

1. Extract data from different databases.
2. Convert the data to a single format.
3. Write the data to a table in a data warehouse.

You would create a process that contained individual steps. Each step would perform a separate task, such as extracting the data from the databases, or converting it to the correct format. You would then use another step to populate the target table, which contains the transformed data.

The following sections describe the various types of steps that you will find in the Data Warehouse Center. For more information about steps, refer to the *Data Warehouse Center Administration Guide*.

SQL Steps

An SQL step uses an SQL SELECT statement to extract data from a warehouse source, and generates an INSERT statement to insert the data into the warehouse target table.

Program Steps

There are several types of program steps: DB2 for AS/400 Programs, DB2 for OS/390 Programs, DB2 for UDB Programs, Visual Warehouse 5.2 DB2 Programs, OLAP Server Programs, File Programs, and Replication. These steps run predefined programs and utilities.

Transformer Steps

Transformer steps are stored procedures and user-defined functions that specify statistical or warehouse transformers that you can use to transform data. You can use transformers to clean, invert, and pivot data; generate primary keys and period tables; and calculate various statistics.

In a transformer step, you specify one of the statistical or warehouse transformers. When you run the process, the transformer step writes data to one or more warehouse targets.

User-defined Program Steps

A *user-defined program step* is a logical entity within the Data Warehouse Center that represents an application that you want the Data Warehouse Center to start. A warehouse agent can start a user-defined program step:

- During the population of a warehouse target.
- After the population of a warehouse target.
- By itself.

For example, you can write a user-defined program that will perform the following process:

1. Export data from a table.
2. Manipulate that data.
3. Write the data to an interim output resource or a warehouse target.

Warehousing Tasks

Creating a data warehouse involves the following tasks:

- Defining a subject area that identifies and groups the processes that you will use in your warehouse.
- Exploring the source data (or operational data), and defining warehouse sources.
- Creating a database to use as the warehouse, and defining warehouse targets.
- Specifying how to move and transform the source data into its format for the warehouse database by defining a process.
- Testing the steps that you have defined, and scheduling them to run automatically.
- Administering the warehouse by defining security, and monitoring database usage.
- Creating an information catalog of the data in the warehouse, if you have the DB2 Warehouse Manager package. An information catalog is a database that contains business metadata. The metadata helps users identify and locate data and information available to them in the organization. End users of the warehouse can search the catalog to determine which tables to query.
- Defining a star schema model for the data in the warehouse. A star schema is a specialized design that consists of multiple dimension tables (which describe aspects of a business) and one fact table (which contains the facts about the business). For example, if you manufacture soft-drinks, some dimension tables are products, markets, and time. The fact table contains transaction information about the products that are ordered in each region by season.
- Joining the fact table and dimension tables to combine details from the dimension tables with the order information. For example, you could join the product dimension with the fact table to add information about how each product was packaged to the orders.

You can learn more about these tasks and others by using the *Business Intelligence Tutorial*, viewing the *DB2 Universal Database Quick Tour*, or reading the *Data Warehouse Center Administration Guide*.

Chapter 6. About Spatial Extender

This section introduces Spatial Extender by explaining its purpose and discussing the data that it processes. For detailed information about using Spatial Extender, refer to the *Spatial Extender User's Guide and Reference*.

The Purpose of Spatial Extender

Use Spatial Extender to create a *geographic information system* (GIS): a complex of objects, data, and applications that allows you to generate and analyze spatial information about geographic features. *Geographic features* include the objects that form the surface of the earth and the objects that occupy it. They make up both the natural environment (examples are rivers, forests, hills, and deserts) and the cultural environment (cities, residences, office buildings, landmarks, and so on).

Spatial information includes facts such as:

- The location of geographic features with respect to their surroundings (for example, points within a city where hospitals and clinics are located, or the proximity of the city's residences to local earthquake zones)
- Ways in which geographic features are related to each other (for example, information that a certain river system is enclosed within a specific region, or that certain bridges in that region cross over the river system's tributaries)
- Measurements that apply to one or more geographic features (for example, the distance between an office building and its lot line, or the length of a bird preserve's perimeter)

Spatial information, either by itself or in combination with traditional database output, can help you to design projects and make business and policy decisions. For example, suppose that a welfare district manager needs to verify which welfare applicants and recipients actually live within the area that the district services. Spatial Extender can derive this information from the serviced area's location and from the addresses of the applicants and recipients.

Or suppose that the owner of a restaurant chain wants to do business in nearby cities. To determine where to open new restaurants, the owner needs answers to such questions as: Where in these cities are there concentrations of the type of people who would frequent my restaurants? Where are the major highways? Where is the crime rate lowest? Where are my competitor's restaurants located? Spatial Extender can produce spatial information in visual displays to answer such questions, and the underlying database management system can generate labels and text to explain the displays.

Data that Represents Geographic Features

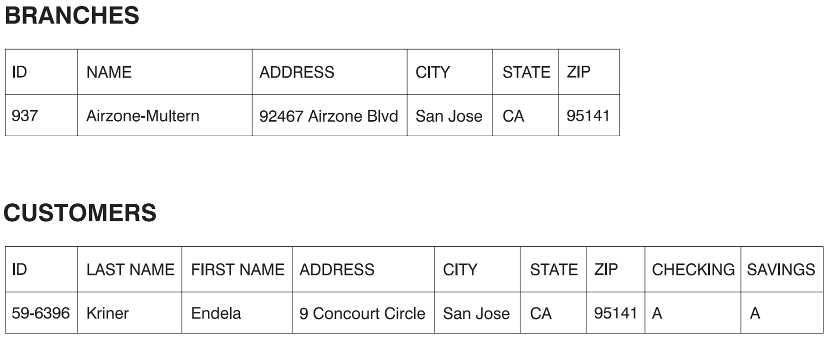
This section provides an overview of the data that you generate, store, and operate upon to obtain spatial information. The topics covered are:

- How data represents geographic features
- The nature of spatial data
- Ways to produce spatial data

How Data Represents Geographic Features

In Spatial Extender, a geographic feature can be represented by a row in a table or view, or by a portion of such a row. For example, consider the following two geographic features: office buildings and residences. In Figure 31, each row of the BRANCHES table represents a branch office of a bank. As a variation, each row of the CUSTOMERS table, taken as a whole, represents a customer of the bank; but part of each row—specifically, the cells that contain a customer's address—can be viewed as representing the customer's residence.

These tables contain data that identifies and describes the bank's branches and customers. Such data is called *attribute data*.



BRANCHES

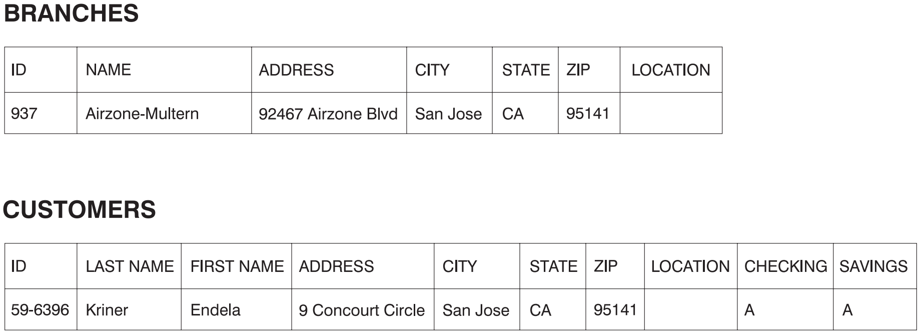
ID	NAME	ADDRESS	CITY	STATE	ZIP
937	Airzone-Multern	92467 Airzone Blvd	San Jose	CA	95141

CUSTOMERS

ID	LAST NAME	FIRST NAME	ADDRESS	CITY	STATE	ZIP	CHECKING	SAVINGS
59-6396	Kriner	Endela	9 Concourt Circle	San Jose	CA	95141	A	A

Figure 31. Table row that represents a geographic feature; table row whose address data represents a geographic feature. The row of data in the BRANCHES table represents a branch office of a bank. The cells for address data in the CUSTOMERS table represent the residence of a customer.

A subset of the attribute data (the values that represent branch and customer addresses) can be translated into values that yield spatial information. For example, as shown in the figure, one branch office address is 92467 Airzone Blvd., San Jose CA 95141. A customer address is 9 Concourt Circle, San Jose CA 95141. Spatial Extender can translate these addresses into values that indicate where the branch and the customer's home are situated with respect to their respective surroundings. Figure 32 shows the BRANCHES and CUSTOMERS tables with new columns that contain such values.



BRANCHES

ID	NAME	ADDRESS	CITY	STATE	ZIP	LOCATION
937	Airzone-Multern	92467 Airzone Blvd	San Jose	CA	95141	

CUSTOMERS

ID	LAST NAME	FIRST NAME	ADDRESS	CITY	STATE	ZIP	LOCATION	CHECKING	SAVINGS
59-6396	Kriner	Endela	9 Concourt Circle	San Jose	CA	95141		A	A

Figure 32. Tables with spatial columns added. In each table, the LOCATION column will contain coordinates that correspond to the addresses.

When addresses and similar identifiers are used as the starting point for spatial information, they are called *source data*. Because the values derived from them yield spatial information, these derived values are called *spatial data*. The next section describes spatial data and introduces its associated data types.

The Nature of Spatial Data

Much spatial data is made up of coordinates. A *coordinate* is a number that denotes a position that is relative to a point of reference. For example, latitudes are coordinates that denote positions relative to the equator. Longitudes are coordinates that denote positions relative to the Greenwich meridian. Thus, the position of Yellowstone National Park is defined by its latitude (44.45 degrees north of the equator) and its longitude (110.40 degrees west of the Greenwich meridian).

Latitudes, longitudes, their points of reference, and other associated parameters are referred to collectively as a *coordinate system*. Coordinate systems based on values other than latitude and longitude also exist. These coordinate systems have their own measures of position, points of reference, and additional distinguishing parameters.

The simplest spatial data item consists of two coordinates that define the position of a single geographic feature. The term *data item* refers to the value or values that occupy a cell in a relational table. A more extensive spatial data item consists of several coordinates that define a linear path, such as a road or a river. A third kind consists of coordinates that define the perimeter of an area, such as the rim of a land parcel or flood plain.

Each spatial data item is an instance of a spatial data type. The data type for two coordinates that mark a location is ST_Point; the data type for coordinates that define linear paths is ST_LineString; and the data type for coordinates that define perimeters is ST_Polygon. These types, together with the other data types for spatial data, are structured types that belong to a single hierarchy.

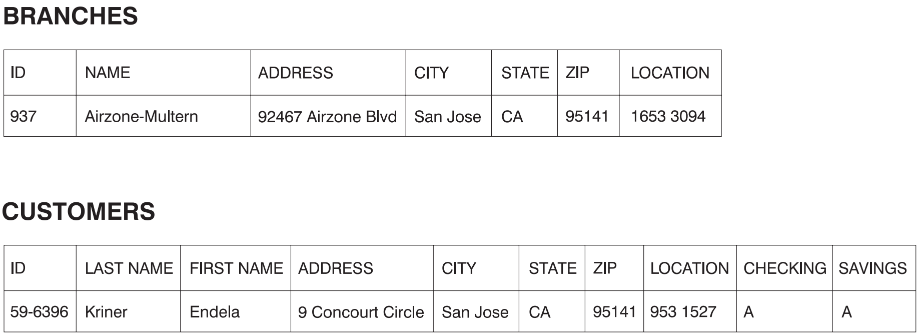
Where Spatial Data Comes From

You can obtain spatial data by:

- Deriving it from attribute data
- Deriving it from other spatial data
- Importing it

Using Attribute Data as Source Data

Deriving spatial data from attribute data (such as addresses) is called *geocoding*. Figure 32 on page 81 shows two columns, one in the BRANCHES table and one in the CUSTOMERS table, designated for spatial data. Imagine that Spatial Extender geocodes the addresses in these tables and places the resulting output (coordinates that correspond to the addresses) into the columns. Figure 33 illustrates this result.



BRANCHES

ID	NAME	ADDRESS	CITY	STATE	ZIP	LOCATION
937	Airzone-Multern	92467 Airzone Blvd	San Jose	CA	95141	1653 3094

CUSTOMERS

ID	LAST NAME	FIRST NAME	ADDRESS	CITY	STATE	ZIP	LOCATION	CHECKING	SAVINGS
59-6396	Kriner	Endela	9 Concourt Circle	San Jose	CA	95141	953 1527	A	A

Figure 33. Tables that include spatial data derived from source data. The LOCATION column in the CUSTOMERS table contains coordinates that a geocoder derived from the address in the ADDRESS, CITY, STATE, and ZIP columns. Similarly, the LOCATION column in the BRANCHES table contains coordinates that the geocoder derived from the address in this table's ADDRESS, CITY, STATE, and ZIP columns.

Spatial Extender uses a function, called a *geocoder*, to geocode attribute data and place the resulting spatial data into columns.

Using Other Spatial Data as Source Data

Spatial data can be generated not only from attribute data, but also from other spatial data. For example, suppose that the bank whose branches are defined in the BRANCHES table wants to know how many customers are located within five miles of each branch. Before the bank can obtain this information from the database, it must supply the database with the definition of a zone that lies within a five-mile radius around each branch. A Spatial Extender function, ST_Buffer, can create such a definition. Using the coordinates of each branch as input, this function can generate the coordinates that demarcate the perimeters of the desired zones. Figure 34 shows the BRANCHES table with information supplied by ST_Buffer.

BRANCHES

ID	NAME	ADDRESS	CITY	STATE	ZIP	LOCATION	SALES_AREA
937	Airzone-Multern	92467 Airzone Blvd	San Jose	CA	95141	1653 3094	1002 2001, 1192 3564, 2502 3415, 1915 3394, 1002 2001

Figure 34. Table that includes new spatial data derived from existing spatial data. The coordinates in the SALES_AREA column were derived by the ST_Buffer function from the coordinates in the LOCATION column.

In addition to ST_Buffer, Spatial Extender provides several other functions that derive new spatial data from existing spatial data.

Importing Spatial Data

A third way to obtain spatial data is to import it from files that are in one of the formats that Spatial Extender supports. These files contain data that is usually applied to maps: census tracks, flood plains, earthquake faults, and so on. By using such data in combination with spatial data that you produce, you can augment the geographic information available to you. For example, if a public works department needs to determine the hazards to which a residential community is vulnerable, it could use ST_Buffer to define a zone around the community, and then import data on flood plains and earthquake faults to see which of these problem areas overlap the zone.

Part 3. Database Design

Chapter 7. Logical Database Design

This section describes the following steps in database design:

- “Decide What Data to Record in the Database”
- “Define Tables for Each Type of Relationship” on page 89
- “Provide Column Definitions for All Tables” on page 91
- “Identify One or More Columns as the Primary Key” on page 94
- “Ensure that Equal Values Represent the Same Entity” on page 97
- “Consider Normalizing Your Tables” on page 98
- “Planning for Constraints Enforcement” on page 103
- “Other Database Design Considerations” on page 110.

Your goal in designing a database is to produce a representation of your environment that is easy to understand and that will serve as a basis for expansion. In addition, you want a database design that will help you maintain consistency and integrity of your data. You can do this by producing a design that will reduce redundancy and eliminate anomalies that can occur during the updating of your database.

These steps are part of *logical* database design. Database design is not a linear process; you will probably have to redo steps as you work out the design.

The *physical* implementation of the database design is described in “Chapter 8. Physical Database Design” on page 113, and in ″Implementing Your Design″ in the *Administration Guide: Implementation*.

Decide What Data to Record in the Database

The first step in developing a database design is to identify the types of data to be stored in database tables. A database includes information about the *entities* in an organization or business, and their relationships to each other. In a relational database, *entities* are represented as *tables*.

An *entity* is a person, object, or concept about which you want to store information. Some of the entities described in the sample tables are employees, departments, and projects. (For a description of the sample database, refer to the *SQL Reference*.)

In the sample employee table, the entity ″employee″ has *attributes*, or properties, such as employee number, name, work department, and salary amount. Those properties appear as the *columns* EMPNO, FIRSTNME, LASTNAME, WORKDEPT, and SALARY.

An *occurrence* of the entity ″employee″ consists of the values in all of the columns for one employee. Each employee has a unique employee number (EMPNO) that can be used to identify an occurrence of the entity ″employee″. Each row in a table represents an occurrence of an entity or relationship. For example, in the following table the values in the first row describe an employee named Haas.

Table 4. Occurrences of Employee Entities and their Attributes

EMPNO	FIRSTNME	LASTNAME	WORKDEPT	JOB
000010	Christine	Haas	A00	President
000020	Michael	Thompson	B01	Manager
000120	Sean	O'Connell	A00	Clerk
000130	Dolores	Quintana	C01	Analyst
000030	Sally	Kwan	C01	Manager
000140	Heather	Nicholls	C01	Analyst
000170	Masatoshi	Yoshimura	D11	Designer

There is a growing need to support non-traditional database applications such as multimedia. You may want to consider attributes to support multimedia objects such as documents, video or mixed media, image, and voice.

Within a table, each column of a row is related in some way to all the other columns of that row. Some of the relationships expressed in the sample tables are:

- Employees are assigned to departments
  - Dolores Quintana is assigned to Department C01
- Employees perform a job
  - Dolores works as an Analyst
- Departments report to other departments
  - Department C01 reports to Department A00
  - Department B01 reports to Department A00
- Employees work on projects
  - Dolores and Heather both work on project IF1000
- Employees manage departments
  - Sally manages department C01.

″Employee″ and ″department″ are entities; Sally Kwan is part of an occurrence of ″employee,″ and C01 is part of an occurrence of ″department″. The same relationship applies to the same columns in every row of a table. For example, one row of a table expresses the relationship that Sally Kwan manages Department C01; another, the relationship that Sean O'Connell is a clerk in Department A00.

The information contained within a table depends on the relationships to be expressed, the amount of flexibility needed, and the data retrieval speed desired.

In addition to identifying the entity relationships within your enterprise, you also need to identify other types of information, such as the business rules that apply to that data.

Define Tables for Each Type of Relationship

Several types of relationships can be defined in a database. Consider the possible relationships between employees and departments. An employee can work in only one department; this relationship is *single-valued* for employees. On the other hand, one department can have many employees; this relationship is *multi-valued* for departments. The relationship between employees (single-valued) and departments (multi-valued) is a *one-to-many* relationship. The following types of relationships are discussed in this section:

- "One-to-Many and Many-to-One Relationships"
- "Many-to-Many Relationships" on page 90
- "One-to-One Relationships" on page 91

One-to-Many and Many-to-One Relationships

To define tables for each one-to-many and each many-to-one relationship:

1. Group all the relationships for which the ″many″ side of the relationship is the same entity.
2. Define a single table for all the relationships in the group.

In the following example, the ″many″ side of the first and second relationships is ″employees″ so we define an employee table, EMPLOYEE.

Table 5. Many-to-One Relationships

Entity	Relationship	Entity
Employees	are assigned to	departments
Employees	work at	jobs
Departments	report to	(administrative) departments

In the third relationship, ″departments″ is on the ″many″ side, so we define a department table, DEPARTMENT.

The following shows how these relationships are represented in tables:

The EMPLOYEE table:

EMPNO	WORKDEPT	JOB
000010	A00	President
000020	B01	Manager
000120	A00	Clerk
000130	C01	Analyst
000030	C01	Manager
000140	C01	Analyst
000170	D11	Designer

The DEPARTMENT table:

DEPTNO	ADMRDEPT
C01	A00
D01	A00
D11	D01

Many-to-Many Relationships

A relationship that is multi-valued in both directions is a many-to-many relationship. An employee can work on more than one project, and a project can have more than one employee. The questions ″What does Dolores Quintana work on?″, and ″Who works on project IF1000?″ both yield multiple answers. A many-to-many relationship can be expressed in a table with a column for each entity (″employees″ and ″projects″), as shown in the following example.

The following shows how a many-to-many relationship (an employee can work on many projects, and a project can have many employees working on it) is represented in a table:

The employee activity (EMP_ACT) table:

EMPNO	PROJNO
000030	IF1000
000030	IF2000
000130	IF1000
000140	IF2000
000250	AD3112

One-to-One Relationships

One-to-one relationships are single-valued in both directions. A manager manages one department; a department has only one manager. The questions, ″Who is the manager of Department C01?″, and ″What department does Sally Kwan manage?″ both have single answers. The relationship can be assigned to either the DEPARTMENT table or the EMPLOYEE table. Because all departments have managers, but not all employees are managers, it is most logical to add the manager to the DEPARTMENT table, as shown in the following example.

The following shows how a one-to-one relationship is represented in a table:

The DEPARTMENT table:

DEPTNO	MGRNO
A00	000010
B01	000020
D11	000060

Provide Column Definitions for All Tables

To define a column in a relational table:

1. Choose a name for the column.

   Each column in a table must have a name that is unique for that table. Selecting column names is described in detail in “Appendix B. Naming Rules” on page 349.

2. State what kind of data is valid for the column.

   The *data type* and *length* specify the type of data and the maximum length that are valid for the column. Data types may be chosen from those provided by the database manager or you may choose to create your own user-defined types. For information about the data types provided by DB2, and about user-defined types, refer to the *SQL Reference*.

Examples of data type categories are: numeric, character string, double-byte (or graphic) character string, date-time, and binary string.

Large object (LOB) data types support multi-media objects such as documents, video, image and voice. These objects are implemented using the following data types:

- A *binary large object* (BLOB) string. Examples of BLOBs are photographs of employees, voice, and video.
- A *character large object* (CLOB) string, where the sequence of characters can be either single- or multi-byte characters, or a combination of both. An example of a CLOB is an employee's resume.
- A *double-byte character large object* (DBCLOB) string, where the sequence of characters is double-byte characters. An example of a DBCLOB is a Japanese resume.

For a better understanding of large object support, refer to the *SQL Reference*.

A *user-defined type* (UDT), is a type that is derived from an existing type. You may need to define types that are derived from and share characteristics with existing types, but that are nevertheless considered to be separate and incompatible.

A *structured type* is a user-defined type whose structure is defined in the database. It contains a sequence of named *attributes*, each of which has a data type. A structured type may be defined as a *subtype* of another structured type, called its *supertype*. A subtype inherits all the attributes of its supertype and may have additional attributes defined. The set of structured types that are related to a common supertype is called a *type hierarchy*, and the supertype that does not have any supertype is called the *root type* of the type hierarchy.

A structured type may be used as the type of a table or a view. The names and data types of the attributes of the structured types, together with the object identifier, become the names and data types of the columns of this *typed table* or *typed view*. Rows of the typed table or typed view can be thought of as a representation of instances of the structured type.

A structured type cannot be used as the data type of a column of a table or a view. There is also no support for retrieving a whole structured type instance into a host variable in an application program.

A *reference type* is a companion type to the structured type. Similar to a distinct type, a reference type is a scalar type that shares a common representation with one of the built-in data types. This same representation is shared for all types in the type hierarchy. The reference type representation is defined when the root type of a type hierarchy is created. When using a reference type, a structured type is specified as a parameter of the type. This parameter is called the *target type* of the reference.

The target of a reference is always a row in a typed table or view. When a reference type is used, it may have a *scope* defined. The scope identifies a table (called the *target table*) or view (called the *target view*) that contains the target row of a reference value. The target table or view must have the same type as the target type of the reference type. An instance of a scoped reference type uniquely identifies a row in a typed table or typed view, called its *target row*.

A *user-defined function* (UDF) can be used for a number of reasons, including invoking routines that allow comparison or conversion between user-defined types. UDFs extend and add to the support provided by built-in SQL functions, and can be used wherever a built-in function can be used. There are two types of UDFs:

- An external function, which is written in a programming language
- A sourced function, which will be used to invoke other UDFs

For example, two numeric data types are European Shoe Size and American Shoe Size. Both types represent shoe size, but they are incompatible, because the measurement base is different and cannot be compared. A user-defined function can be invoked to convert one shoe size to another.

For a better understanding of user-defined types, structured types, reference types, and user-defined functions, refer to the *SQL Reference*.

3. State which columns might need default values.

   Some columns cannot have meaningful values in all rows because:

   - A column value is not applicable to the row.

     For example, a column containing an employee’s middle initial is not applicable to an employee who has no middle initial.
   - A value is applicable, but is not yet known.

     For example, the MGRNO column might not contain a valid manager number because the previous manager of the department has been transferred, and a new manager has not been appointed yet.

   In both situations, you can choose between allowing a NULL value (a special value indicating that the column value is unknown or not applicable), or allowing a non-NULL default value to be assigned by the database manager or by the application.

NULL values and default values are described in detail in the *SQL Reference*.

Identify One or More Columns as the Primary Key

A *key* is a set of columns that can be used to identify or access a particular row or rows. The key is identified in the description of a table, index, or referential constraint. The same column can be part of more than one key.

A *unique key* is a key that is constrained so that no two of its values are equal. The columns of a unique key cannot contain NULL values. For example, an employee number column can be defined as a unique key, because each value in the column identifies only one employee. No two employees can have the same employee number.

The mechanism used to enforce the uniqueness of the key is called a *unique index*. The unique index of a table is a column, or an ordered collection of columns, for which each value identifies (functionally determines) a unique row. A unique index can contain NULL values.

The *primary key* is one of the unique keys defined on a table, but is selected to be the key of first importance. There can be only one primary key on a table.

A *primary index* is automatically created for the primary key. The primary index is used by the database manager for efficient access to table rows, and allows the database manager to enforce the uniqueness of the primary key. (You can also define indexes on non-primary key columns to efficiently access data when processing queries.)

If a table does not have a ″natural″ unique key, or if arrival sequence is the method used to distinguish unique rows, using a time stamp as part of the key can be helpful. (See also “Defining Identity Columns” on page 96.)

Primary keys for some of the sample tables are:

Table	**Key Column**
Employee table	EMPNO
Department table	DEPTNO
Project table	PROJNO

The following example shows part of the PROJECT table, including its primary key column.

Table 6. A Primary Key on the PROJECT Table

PROJNO (Primary Key)	PROJNAME	DEPTNO
MA2100	Weld Line Automation	D01
MA2110	Weld Line Programming	D11

If every column in a table contains duplicate values, you cannot define a primary key with only one column. A key with more than one column is a *composite key*. The combination of column values should define a unique entity. If a composite key cannot be easily defined, you may consider creating a new column that has unique values.

The following example shows a primary key containing more than one column (a composite key):

Table 7. A Composite Primary Key on the EMP_ACT Table

EMPNO (Primary Key)	PROJNO (Primary Key)	ACTNO (Primary Key)	EMPTIME	EMSTDATE (Primary Key)
000250	AD3112	60	1.0	1982-01-01
000250	AD3112	60	.5	1982-02-01
000250	AD3112	70	.5	1982-02-01

Identifying Candidate Key Columns

To identify candidate keys, select the smallest number of columns that define a unique entity. There may be more than one candidate key. In Table 2 on page 22, there appear to be many candidate keys. The EMPNO, the PHONENO, and the LASTNAME columns each uniquely identify the employee.

The criteria for selecting a primary key from a pool of candidate keys should be persistence, uniqueness, and stability:

- Persistence means that a primary key value for each row always exists.
- Uniqueness means that the key value for each row is different from all the others.
- Stability means that primary key values never change.

Of the three candidate keys in the example, only EMPNO satisfies all of these criteria. An employee may not have a phone number when joining a company. Last names can change, and, although they may be unique at one point, are not guaranteed to be so. The employee number column is the best choice for the primary key. An employee is assigned a unique number only once, and that number is generally not updated as long as the employee remains with the company. Since each employee must have a number, values in the employee number column are persistent.

Defining Identity Columns

An *identity column* provides a way for DB2 to automatically generate a unique numeric value for each row in a table. A table can have a single column that is defined with the identity attribute. Examples of an identity column include order number, employee number, stock number, and incident number.

Values for an identity column can be generated always or by default.

- An identity column that is defined as *generated always* is guaranteed to be unique by DB2. Its values are always generated by DB2; applications are not allowed to provide an explicit value.
- An identity column that is defined as *generated by default* gives applications a way to explicitly provide a value for the identity column. If a value is not given, DB2 generates one, but cannot guarantee the uniqueness of the value in this case. DB2 guarantees uniqueness only for the set of values that it generates. Generated by default is meant to be used for data propagation, in which the contents of an existing table are copied, or for the unloading and reloading of a table.

Identity columns are ideally suited to the task of generating unique primary key values. Applications can use identity columns to avoid the concurrency and performance problems that can result when an application generates its own unique counter outside of the database. For example, one common application-level implementation is to maintain a 1-row table containing a counter. Each transaction locks this table, increments the number, and then commits; that is, only one transaction at a time can increment the counter. In contrast, if the counter is maintained through an identity column, much higher levels of concurrency can be achieved because the counter is not locked by transactions. One uncommitted transaction that has incremented the counter will not prevent subsequent transactions from also incrementing the counter.

The counter for the identity column is incremented (or decremented) independently of the transaction. If a given transaction increments an identity counter two times, that transaction may see a gap in the two numbers that are generated because there may be other transactions concurrently incrementing the same identity counter (that is, inserting rows into the same table). If an application must have a consecutive range of numbers, that application should take an exclusive lock on the table that has the identity column. This decision must be weighed against the resulting loss of concurrency. Furthermore, it is possible that a given identity column can appear to have generated gaps in the number, because a transaction that generated a value for the identity column has rolled back, or the database that has cached a range of values has been deactivated before all of the cached values were assigned.

The sequential numbers that are generated by the identity column have the following additional properties:

- The values can be of any exact numeric data type with a scale of zero; that is, SMALLINT, INTEGER, BIGINT, or DECIMAL with a scale of zero. (Single and double precision floating point are considered to be approximate numeric data types.)
- Consecutive values can differ by any specified integer increment. The default increment is 1.
- The counter value for the identity column is recoverable. If a failure occurs, the counter value is reconstructed from the logs, thereby guaranteeing that unique values continue to be generated.
- Identity column values can be cached to give better performance.

Ensure that Equal Values Represent the Same Entity

You can have more than one table describing the attributes of the same set of entities. For example, the EMPLOYEE table shows the number of the department to which an employee is assigned, and the DEPARTMENT table shows which manager is assigned to each department number. To retrieve both sets of attributes simultaneously, you can join the two tables on the matching columns, as shown in the following example. The values in WORKDEPT and DEPTNO represent the same entity, and represent a *join path* between the DEPARTMENT and EMPLOYEE tables.

The DEPARTMENT table:

DEPTNO	**DEPTNAME**	**MGRNO**	**ADMRDEPT**
D21	Administration Support	000070	D01

The EMPLOYEE table:

EMPNO	**FIRSTNAME**	**LASTNAME**	**WORKDEPT**	**JOB**
000250	Daniel	Smith	D21	Clerk

When you retrieve information about an entity from more than one table, ensure that equal values represent the same entity. The connecting columns can have different names (like WORKDEPT and DEPTNO in the previous example), or they can have the same name (like the columns called DEPTNO in the department and project tables).

Consider Normalizing Your Tables

Normalization helps eliminate redundancies and inconsistencies in table data. It is the process of reducing tables to a set of columns where all the non-key columns depend on the primary key column. If this is not the case, the data can become inconsistent during updates.

This section briefly reviews the rules for first, second, third, and fourth normal form. The fifth normal form of a table, which is covered in many books on database design, is not described here.

Form **Description**

First At each row and column position in the table, there exists *one* value, never a set of values (see "First Normal Form").

Second Each column that is not part of the key is dependent upon the key (see "Second Normal Form" on page 99).

Third Each non-key column is independent of other non-key columns, and is dependent only upon the key (see "Third Normal Form" on page 100).

Fourth No row contains two or more independent multi-valued facts about an entity (see "Fourth Normal Form" on page 102).

First Normal Form

A table is in *first normal form* if there is only one value, never a set of values, in each cell. A table that is in first normal form does not necessarily satisfy the criteria for higher normal forms.

For example, the following table violates first normal form because the WAREHOUSE column contains several values for each occurrence of PART.

Table 8. Table Violating First Normal Form

PART (Primary Key)	WAREHOUSE
P0010	Warehouse A, Warehouse B, Warehouse C
P0020	Warehouse B, Warehouse D

The following example shows the same table in first normal form.

Table 9. Table Conforming to First Normal Form

PART (Primary Key)	WAREHOUSE (Primary Key)	QUANTITY
P0010	Warehouse A	400
P0010	Warehouse B	543

Table 9. Table Conforming to First Normal Form (continued)

PART (Primary Key)	WAREHOUSE (Primary Key)	QUANTITY
P0010	Warehouse C	329
P0020	Warehouse B	200
P0020	Warehouse D	278

Second Normal Form

A table is in *second normal form* if each column that is not part of the key is dependent upon the *entire* key.

Second normal form is violated when a non-key column is dependent upon *part* of a composite key, as in the following example:

Table 10. Table Violating Second Normal Form

PART (Primary Key)	WAREHOUSE (Primary Key)	QUANTITY	WAREHOUSE_ADDRESS
P0010	Warehouse A	400	1608 New Field Road
P0010	Warehouse B	543	4141 Greenway Drive
P0010	Warehouse C	329	171 Pine Lane
P0020	Warehouse B	200	4141 Greenway Drive
P0020	Warehouse D	278	800 Massey Street

The primary key is a composite key, consisting of the PART and the WAREHOUSE columns together. Because the WAREHOUSE_ADDRESS column depends only on the value of WAREHOUSE, the table violates the rule for second normal form.

The problems with this design are:

- The warehouse address is repeated in every record for a part stored in that warehouse.
- If the address of a warehouse changes, every row referring to a part stored in that warehouse must be updated.
- Because of this redundancy, the data might become inconsistent, with different records showing different addresses for the same warehouse.
- If at some time there are no parts stored in a warehouse, there might not be a row in which to record the warehouse address.

The solution is to split the table into the following two tables:

Table 11. PART_STOCK Table Conforming to Second Normal Form

PART (Primary Key)	**WAREHOUSE (Primary Key)**	**QUANTITY**
P0010	Warehouse A	400
P0010	Warehouse B	543
P0010	Warehouse C	329
P0020	Warehouse B	200
P0020	Warehouse D	278

Table 12. WAREHOUSE Table Conforms to Second Normal Form

WAREHOUSE (Primary Key)	**WAREHOUSE_ADDRESS**
Warehouse A	1608 New Field Road
Warehouse B	4141 Greenway Drive
Warehouse C	171 Pine Lane
Warehouse D	800 Massey Street

There is a performance consideration in having the two tables in second normal form. Applications that produce reports on the location of parts must join both tables to retrieve the relevant information.

To better understand performance considerations, refer to ″Tuning Application Performance″ in the *Administration Guide: Performance*.

Third Normal Form

A table is in third normal form if each non-key column is independent of other non-key columns, and is dependent only on the key.

The first table in the following example contains the columns EMPNO and WORKDEPT. Suppose a column DEPTNAME is added (see Table 14 on page 101). The new column depends on WORKDEPT, but the primary key is EMPNO. The table now violates third normal form. Changing DEPTNAME for a single employee, John Parker, does not change the department name for other employees in that department. Note that there are now two different department names used for department number E11. The inconsistency that results is shown in the updated version of the table.

Table 13. Unnormalized EMPLOYEE_DEPARTMENT Table Before Update

EMPNO (Primary Key)	FIRSTNAME	LASTNAME	WORKDEPT	DEPTNAME
000290	John	Parker	E11	Operations
000320	Ramlal	Mehta	E21	Software Support
000310	Maude	Setright	E11	Operations

Table 14. Unnormalized EMPLOYEE_DEPARTMENT Table After Update. Information in the table has become inconsistent.

EMPNO (Primary Key)	FIRSTNAME	LASTNAME	WORKDEPT	DEPTNAME
000290	John	Parker	E11	Installation Mgmt
000320	Ramlal	Mehta	E21	Software Support
000310	Maude	Setright	E11	Operations

The table can be normalized by creating a new table, with columns for WORKDEPT and DEPTNAME. An update like changing a department name is now much easier; only the new table needs to be updated.

An SQL query that returns the department name along with the employee name is more complex to write, because it requires joining the two tables. It will probably also take longer to run than a query on a single table. Additional storage space is required, because the WORKDEPT column must appear in both tables.

The following tables are defined as a result of normalization:

Table 15. EMPLOYEE Table After Normalizing the EMPLOYEE_DEPARTMENT Table

EMPNO (Primary Key)	FIRSTNAME	LASTNAME	WORKDEPT
000290	John	Parker	E11
000320	Ramlal	Mehta	E21
000310	Maude	Setright	E11

Table 16. DEPARTMENT Table After Normalizing the EMPLOYEE_DEPARTMENT Table

DEPTNO (Primary Key)	DEPTNAME
E11	Operations
E21	Software Support

Fourth Normal Form

A table is in fourth normal form if no row contains two or more independent multi-valued facts about an entity.

Consider these entities: employees, skills, and languages. An employee can have several skills and know several languages. There are two relationships, one between employees and skills, and one between employees and languages. A table is not in fourth normal form if it represents both relationships, as in the following example:

Table 17. Table Violating Fourth Normal Form

EMPNO (Primary Key)	SKILL (Primary Key)	LANGUAGE (Primary Key)
000130	Data Modelling	English
000130	Database Design	English
000130	Application Design	English
000130	Data Modelling	Spanish
000130	Database Design	Spanish
000130	Application Design	Spanish

Instead, the relationships should be represented in two tables:

Table 18. EMPLOYEE_SKILL Table Conforming to Fourth Normal Form

EMPNO (Primary Key)	SKILL (Primary Key)
000130	Data Modelling
000130	Database Design
000130	Application Design

Table 19. EMPLOYEE_LANGUAGE Table Conforming to Fourth Normal Form

EMPNO (Primary Key)	LANGUAGE (Primary Key)
000130	English
000130	Spanish

If, however, the attributes are interdependent (that is, the employee applies certain languages only to certain skills), the table should *not* be split.

A good strategy when designing a database is to arrange all data in tables that are in fourth normal form, and then to decide whether the results give you an acceptable level of performance. If they do not, you can rearrange the data in tables that are in third normal form, and then reassess performance.

Planning for Constraints Enforcement

A *constraint* is a rule that the database manager enforces. Four types of constraints handling are covered in this section:

Unique Constraints
Ensure that key values in a table are unique. Any changes to the columns that make up the primary key are checked for uniqueness.

Referential Integrity
Enforces referential constraints on insert, update, and delete operations. It is the state of a database in which all values of all foreign keys are valid.

Table Check Constraints
Verify that changed data does not violate conditions specified when a table was created or altered.

Triggers
Define a set of actions that are to be taken when called by an update, delete, or insert operation on a specified table.

Unique Constraints

A *unique constraint* is the rule that ensures that key values are unique within the table. Each column making up the key in a unique constraint must be defined as NOT NULL. Unique constraints are defined in the CREATE TABLE or the ALTER TABLE statement, using the PRIMARY KEY clause or the UNIQUE clause.

A table can have any number of unique constraints; however, you can only define one unique constraint as the primary key for a table. Moreover, a table cannot have more than one unique constraint on the same set of columns.

When a unique constraint is defined, the database manager creates a unique index (if needed), and designates it as either a primary or a unique system-required index. The constraint is enforced through the unique index. Once a unique constraint has been established on a column, the check for uniqueness during multiple row updates is deferred until the end of the update.

A unique constraint can also be used as the parent key in a referential constraint.

Referential Integrity

The database manager maintains referential integrity through *referential constraints*, which require that all values for a given attribute or table column also exist in some other table or column. For example, a referential constraint might require that every employee in the EMPLOYEE table be in a department that exists in the DEPARTMENT table. No employee can be in a department that does not exist.

You can build referential constraints into a database to ensure that referential integrity is maintained, and to allow the optimizer to exploit knowledge of these special relationships to process queries more efficiently. When planning for referential integrity, identify all of the relationships between database tables. You can identify a relationship by defining a primary key and referential constraints.

Consider the following related tables:

Table 20. DEPARTMENT Table

DEPTNO (Primary Key)	DEPTNAME	MGRNO
A00	Spiffy Computer Service Div.	000010
B01	Planning	000020
C01	Information Center	000030
D11	Manufacturing Systems	000060

Table 21. EMPLOYEE Table

EMPNO (Primary Key)	FIRSTNAME	LASTNAME	WORKDEPT (Foreign Key)	PHONENO
000010	Christine	Haas	A00	3978
000030	Sally	Kwan	C01	4738
000060	Irving	Stern	D11	6423
000120	Sean	O’Connell	A00	2167
000140	Heather	Nicholls	C01	1793
000170	Masatoshi	Yoshimura	D11	2890

Many of the following concepts, useful for understanding referential integrity, are discussed in relation to these tables.

A *unique key* is a column or a set of columns where no values in a row are duplicated in any other row. You can define one unique key as the primary key for the table. The unique key may also be known as a *parent key* when referenced by a foreign key.

A *primary key* is a unique key that is part of the definition of the table. Each table can have only one primary key. In the preceding tables, DEPTNO and EMPNO are the primary keys of the DEPARTMENT and EMPLOYEE tables, respectively.

A *foreign key* is a column or a set of columns in a table that refer to a unique key or the primary key of the same or another table. A foreign key is used to establish a relationship with a unique key or the primary key to enforce referential integrity among tables. The column WORKDEPT in the EMPLOYEE table is a foreign key because it refers to the primary key, DEPTNO, in the DEPARTMENT table.

A *composite key* is a key that has more than one column. Both primary and foreign keys can be composite keys. For example, if departments were uniquely identified by the combination of division number and department number, two columns would be needed to create the key for the DEPARTMENT table.

A *parent key* is a primary key or a unique key of a referential constraint. The *primary key constraint* is the default parent key of a referential constraint when a set of parent key columns is not specified.

A *parent table* is a table containing a parent key that is related to at least one foreign key in the same or another table. A table can be a parent in an arbitrary number of relationships. For example, the DEPARTMENT table, which has a primary key of DEPTNO, is a parent of the EMPLOYEE table, which contains the foreign key WORKDEPT.

A *parent row* is a row of a parent table whose parent key value matches at least one foreign key value in a dependent table. A row in a parent table is not necessarily a parent row. The fourth row (D11) of the DEPARTMENT table is the parent row of the third and sixth rows in the EMPLOYEE table. The second row (B01) of the DEPARTMENT table is not the parent of any other row.

A *dependent table* is a table containing one or more foreign keys. A dependent table can also be a parent table. A table can be a dependent table in an arbitrary number of relationships. The EMPLOYEE table contains the foreign key WORKDEPT, and is dependent on the DEPARTMENT table, whose primary key is DEPTNO.

A *dependent row* is a row of a dependent table that has a non-NULL foreign key value that matches a parent key value. The foreign key value represents a reference from the dependent row to the parent row. Since foreign keys can accept NULL values, a row in a dependent table is not necessarily a dependent row.

A table is a *descendent* of a table if it is a dependent table, or if it is a descendent of a dependent table. A descendent table contains a foreign key that can be traced back to the parent key of some table.

A *referential cycle* is a path that connects a table to itself. When a table is directly connected to itself, it is a *self-referencing* table. If the EMPLOYEE table had another column called MGRID that contains the EMPNO of each employee's manager, the EMPLOYEE table would be a self-referencing table. MGRID would be a foreign key for the EMPLOYEE table.

A self-referencing table is both a parent and a dependent in the same relationship. A self-referencing row is a row that is both a parent and a dependent of itself. The constraint that exists in this situation is called a self-referencing constraint. For example, if the value for the foreign key in a row of a self-referencing table matches the value of the unique key in that row, the row is self-referencing.

A *referential constraint* is an assertion that non-NULL values of a designated foreign key are valid only if they also appear as values for a unique key of a designated table. The purpose of referential constraints is to guarantee that database relationships are maintained, and that data entry rules are followed.

Implications for SQL Operations

Enforcement of referential constraints has special implications for some SQL operations that depend on whether the table is a parent or a dependent. This section describes the effects of maintaining referential integrity on SQL INSERT, DELETE, UPDATE, and DROP operations.

DB2 does *not* automatically enforce referential constraints across systems. Consequently, if you want to enforce referential constraints across systems, your applications must contain the necessary logic.

The following topics are discussed:

- "INSERT Rules"
- "DELETE Rules" on page 107
- "UPDATE Rules" on page 108.

INSERT Rules: You can insert a row at any time into a parent table without any action being taken in dependent tables. However, you cannot insert a row into a dependent table, unless there is a row in the parent table with a parent key value equal to the foreign key value of the row that is being inserted, unless that foreign key value is NULL. The value of a composite foreign key is NULL if any component of the value is NULL.

This rule is implicit when a foreign key is specified.

If you try to insert a row into a table that has referential constraints, the INSERT operation is not allowed if any of the non-NULL foreign key values are not present in the parent key. If the INSERT operation fails for one row during an attempt to insert more than one row, none of the rows are inserted.

DELETE Rules: When you delete a row from a parent table, DB2 checks if there are any dependent rows in dependent tables with matching foreign key values. If any dependent rows are found, several actions are possible. You can determine which action will be taken by specifying a *delete* rule when you create the dependent table.

The delete rules for a dependent table (the table containing the foreign key) when a primary key is deleted are:

RESTRICT
Prevents any row in the parent table from being deleted if any dependent rows are found. If you need to remove both parent and dependent rows, delete dependent rows first. Deleting the parent row first violates the referential constraint, and is not allowed.

NO ACTION
Enforces the presence of a parent row for every child after all referential constraints are applied.

CASCADE
Implies that deleting a row in the parent table automatically deletes any related rows in the dependent table. This rule is useful when a row in the dependent table has no significance without a row in the parent table.

Deleting the parent row first automatically deletes the dependent rows referencing a primary key. The dependent rows do not need to be deleted first. If some of these dependent rows have dependents of their own, the delete rule for those relationships is applied. DB2 manages cascading deletions.

SET NULL
Ensures that deletion of a row in the parent table sets the values of the foreign key in any

dependent rows to NULL. Other parts of the row remain unchanged.

If no delete rule is explicitly defined when the table is created, the NO ACTION rule applies.

Any table that can be involved in a delete operation is said to be delete-connected. The following restrictions apply to delete-connected relationships.

- A table cannot be delete-connected to itself in a referential cycle of more than one table.
- When a table is delete-connected to another table through more than one dependent relationship, these relationships must have the same delete rule, either CASCADE or NO ACTION.
- When a self-referencing table is a dependent of another table in a CASCADE relationship, the delete rule for the self-referencing relationship must also be CASCADE.

You can, at any time, delete rows from a dependent table without taking any action against the parent table. In the department-employee relationship, for example, an employee could retire and have his row deleted from the employee table with no effect on the department table. (In the reverse relationship of employee-department, the department manager ID is a foreign key referring to the parent key of the employee table. If a manager retires, there *is* an effect on the department table.)

UPDATE Rules: DB2 prevents the update of a unique key for a parent row. When you update a foreign key in a dependent table, and that foreign key is not NULL, it must match some value of the parent key for the parent table of the relationship. If any referential constraint is violated by an UPDATE operation, an error occurs and no rows are updated.

When a value in a column of the parent key is updated:

- If any row in the dependent table matches the original value of the key, the update is rejected when the update rule is RESTRICT.
- If any row in the dependent table does not have a corresponding parent key when the update statement is completed (except after triggers), the update is rejected when the update rule is NO ACTION.

To update the value of a parent key that is in a parent row, you must first remove the relationship to any child rows in the dependent tables by either:

- Deleting the child rows; or,
- Updating the foreign keys in dependent tables to include another valid key value.

If there is no dependency to the key value in the row, the row is no longer a parent in a referential relationship and can be updated.

If part of a foreign key is being updated and no part of the foreign key value is NULL, the new value of the foreign key must appear as a unique key value in the parent table. If there is no foreign key dependent on a given unique key; that is, the row containing the unique key is *not* a parent row, part of the unique key may be updated. However, no more than one row can be selected for update in this case, because you are working with a unique key, and duplicate rows are not allowed.

Table Check Constraints

Business rules identified in your design can be enforced through table check constraints. *Table check constraints* specify search conditions that are applied to each row of a table. These constraints are automatically activated when an update or insert statement is applied against the table. They are defined through the CREATE TABLE or the ALTER TABLE statement.

A table check constraint can be used for validation. For example, values for a department number must be within the range of 10 to 100; the job title of an employee can only be ″Sales″, ″Manager″, or ″Clerk″; or an employee who has been with the company for more than 8 years must earn more than $40,500.

Refer to the *Data Movement Utilities Guide and Reference* for information about the impact of table check constraints on the IMPORT and LOAD commands.

Triggers

A *trigger* is a defined set of actions that are performed whenever a delete, insert, or update operation is carried out against a specified table. Triggers can be defined to help support business rules. Triggers can also be used to automatically update summary or audit data. Because triggers are stored in the database, you do not have to code the actions in every application program. The trigger is coded once, stored in the database, and automatically called by DB2, as required, when an application uses the database. This ensures that the business rules related to the data are always enforced. If a business rule changes, only the triggers need to be modified.

A user-defined function (UDF) can be called within a triggered SQL statement. This allows the triggered action to perform a non-SQL operation when the trigger is fired. For example, e-mail can be sent as an alert mechanism. For more information about triggers, see ″Creating a Trigger″ in the *Administration Guide: Implementation* and refer to the *Application Development Guide*.

Other Database Design Considerations

When designing a database, it is important to consider which tables users should be able to access. Access to tables is granted or revoked through authorizations. The highest level of authority is system administration authority (SYSADM). A user with SYSADM authority can assign other authorizations, including database administrator authority (DBADM).

There are other issues that you may want to consider in your design, such as *audit activities*, *historical data*, *summary tables*, *security*, *data typing*, and *parallel processing capability*.

For *audit* purposes, you may have to record every update made to your data for a specified period. For example, you may want to update an audit table each time an employee's salary is changed. Updates to this table could be made automatically if an appropriate trigger is defined. Audit activities can also be carried out through the DB2 audit facility. For more information, see ″Auditing DB2 Activities″ in the *Administration Guide: Implementation*.

For performance reasons, you may only want to access a selected amount of data, while maintaining the base data as *history*. You should include within your design, the requirements for maintaining this historical data, such as the number of months or years of data that is required to be available before it can be purged.

You may also want to make use of *summary* information. For example, you may have a table that has all of your employee information in it. However, you would like to have this information divided into separate tables by division or department. In this case, a summary table for each division or department based on the data in the original table would be helpful. For more information about summary tables, see ″Creating a Summary Table″ in the *Administration Guide: Implementation*.

Security implications should also be identified within your design. For example, you may decide to support user access to certain types of data through security tables. You can define access levels to various types of data, and who can access this data. Confidential data, such as employee and payroll data, would have stringent security restrictions. For more information about security and authorizations, see ″Controlling Database Access″ in the *Administration Guide: Implementation*.

You can create tables that have a *structured type* associated with them. With such typed tables, you can establish a hierarchical structure with a defined relationship between those tables called a *type hierarchy*. The type hierarchy is made up of a single root type, supertypes, and subtypes.

A *reference type* representation is defined when the root type of a type hierarchy is created. The target of a reference is always a row in a typed table or view.

For more information about implementing a design that includes typed rows and tables, see "Implementing Your Design" in the *Administration Guide: Implementation*. Refer to the *Data Movement Utilities Guide and Reference* for information about moving data between typed tables that are in a hierarchical structure.

As your business grows, you may need the additional capacity and performance capability provided by DB2 Enterprise - Extended Edition. In this environment, your database is partitioned across several machines or systems, each responsible for the storage and retrieval of a portion of the overall database. Each partition (or node) works in parallel to handle SQL or utility operations.

Issues and considerations relating to parallel operations are included throughout this book.

Chapter 8. Physical Database Design

After you have completed your logical database design (see “Chapter 7. Logical Database Design” on page 87), there are a number of issues you should consider about the physical environment in which your database and tables will reside. These include understanding the files that will be created to support and manage your database, understanding how much space will be required to store your data, and determining how you should use the table spaces that are required to store your data.

The following topics are covered:

- “Database Directories”
- “Estimating Space Requirements for Tables” on page 115
- “Additional Space Requirements” on page 123
- “Designing Nodegroups” on page 124
- “Designing and Choosing Table Spaces” on page 132
- “Federated Database Design Considerations” on page 153

Database Directories

When a database is created, DB2 creates a separate subdirectory to store control files (such as log header files) and to allocate containers to default table spaces. Objects associated with the database are not always stored in the database directory; they can be stored in various locations, including devices.

The database is created in the instance that is defined by the DB2INSTANCE environment variable, or in the instance to which you have explicitly attached (using the ATTACH command). For an introduction to instances, see "Using Multiple Instances of the Database Manager" in the *Administration Guide: Implementation*.

The naming scheme used on UNIX based systems is:

```
specified_path/$DB2INSTANCE/NODEnnnn/SQL00001
```

The naming scheme used on OS/2 and the Windows operating systems:

```
D:\$DB2INSTANCE\NODEnnnn\SQL00001
```

where

- `specified_path` is the optional, user-specified location to install the instance.

- NODEnnnn is the node identifier in a partitioned database environment. The first node is NODE0000.
- ″D:″ is a ″drive letter″ identifying the volume on which the root directory is located.

SQL00001 contains objects associated with the first database created, and subsequent databases are given higher numbers: SQL00002, and so on.

The subdirectories are created in a directory with the same name as the database manager instance to which you are attached when you create the database. (On OS/2 and the Windows operating systems, the subdirectories are created under the root directory for a volume that is identified by a ″drive letter″.) These instance and database subdirectories are created within the path specified on the CREATE DATABASE command, and the database manager maintains them automatically. Depending on your platform, each instance might be owned by an instance owner, who has system administrator (SYSADM) authority over the databases belonging to that instance.

To avoid potential problems, do not create directories that use the same naming scheme, and do not manipulate directories that have already been created by the database manager.

Database Files

The following files are associated with a database:

File Name **Description**

SQLDBCON This file stores the tuning parameters and flags for the database. Refer to *Administration Guide: Performance* for information about changing database configuration parameters.

SQLOGCTL.LFH
This file is used to help track and control all of the database log files.

Syyyyyyy.LOG
Database log files, numbered from 0000000 to 9999999. The number of these files is controlled by the *logprimary* and the *logsecond* database configuration parameters. The size of the individual files is controlled by the *logfilsiz* database configuration parameter.

With circular logging, the files are reused and the same numbers remain. With archive logging, the file numbers increase in sequence as logs are archived and new logs are allocated. When 9999999 is reached, the number wraps.

By default, these log files are stored in a directory called SQLOGDIR. SQLOGDIR is found in the SQL*nnnnn* subdirectory.

SQLINSLK
This file helps to ensure that a database is used by only one instance of the database manager.

SQLTMPLK
This file helps to ensure that a database is used by only one instance of the database manager.

SQLSPCS.1
This file contains the definition and current state of all table spaces in the database.

SQLSPCS.2
This file is a backup copy of `SQLSPCS.1`. Without one of these files, you will not be able to access your database.

SQLBP.1
This file contains the definition of all buffer pools used in the database.

SQLBP.2
This file is a backup copy of `SQLBP.1`. Without one of these files, you will not be able to access your database.

DB2RHIST.ASC
This file is the database history file. It keeps a history of administrative operations on the database, such as backup and restore operations.

DB2RHIST.BAK
This file is a backup copy of `DB2RHIST.ASC`.

Notes:

1. Do *not* make any direct changes to these files. They can only be accessed indirectly using the documented APIs and by tools that implement those APIs, including the command line processor and the Control Center.
2. Do not move these files.
3. Do not remove these files.
4. The only supported means of backing up a database or a table space is through the **sqlubkp** (Backup Database) API, including the command line processor and Control Center implementations of that API.

Estimating Space Requirements for Tables

Estimating the size of database objects is an imprecise undertaking. Overhead caused by disk fragmentation, free space, and the use of variable length columns makes size estimation difficult, because there is such a wide range of possibilities for column types and row lengths. After initially estimating your database size, create a test database and populate it with representative data.

From the Control Center, you can access a number of utilities that are designed to assist you in determining the size requirements of various database objects:

- You can select an object and then use the ″Estimate Size″ utility. This utility can tell you the current size of an existing object, such as a table. You can then change the object, and the utility will calculate new estimated values for the object. The utility will help you approximate storage requirements, taking future growth into account. It gives more than a single estimate of the size of the object. It also provides possible size ranges for the object: both the smallest size, based on current values, and the largest possible size.
- You can determine the relationships between objects by using the ″Show Related″ dialog.
- You can select any database object on the instance and request ″Generate DDL″. This function uses the **db2look** utility to generate data definition statements for the database. For information about this utility, refer to the *Command Reference*.

In each of these cases, either the ″Show SQL″ or the ″Show Command″ button is available to you. You can also save the resulting SQL statements or commands in script files to be used later. All of these utilities have online help to assist you.

Keep these utilities in mind as you work through the planning of your physical database requirements.

When estimating the size of a database, the contribution of the following must be considered:

- "System Catalog Tables" on page 117
- "User Table Data" on page 117
- "Long Field Data" on page 119
- "Large Object (LOB) Data" on page 119
- "Index Space" on page 120

Space requirements related to the following are not discussed:

- The local database directory file
- The system database directory file
- The file management overhead required by the operating system, including:
  - file block size
  - directory control space

System Catalog Tables

System catalog tables are created when a database is created. The system tables grow as database objects and privileges are added to the database. Initially, they use approximately 3.5 MB of disk space.

The amount of space allocated for the catalog tables depends on the type of table space, and the extent size of the table space containing the catalog tables. For example, if a DMS table space with an extent size of 32 is used, the catalog table space will initially be allocated 20 MB of space. For more information, see “Designing and Choosing Table Spaces” on page 132.

Note: For databases with multiple partitions, the catalog tables reside only on the partition from which the CREATE DATABASE command was issued. Disk space for the catalog tables is only required for that partition.

User Table Data

By default, table data is stored on 4 KB pages. Each page (regardless of page size) contains 76 bytes of overhead for the database manager. This leaves 4020 bytes to hold user data (or rows), although no row on a 4 KB page can exceed 4005 bytes in length. A row will *not* span multiple pages. You can have a maximum of 500 columns when using a 4 KB page size.

Table data pages *do not* contain the data for columns defined with LONG VARCHAR, LONG VARGRAPHIC, BLOB, CLOB, or DBCLOB data types. The rows in a table data page do, however, contain a descriptor for these columns. (See “Long Field Data” on page 119 and “Large Object (LOB) Data” on page 119 for information about estimating the space requirements for table objects that do contain these data types.)

Rows are usually inserted into a table in first-fit order. The file is searched (using a free space map) for the first available space that is large enough to hold the new row. When a row is updated, it is updated in place, unless there is insufficient space left on the page to contain it. If this is the case, a record is created in the original row location that points to the new location in the table file of the updated row.

If the ALTER TABLE APPEND ON statement is invoked, data is always appended, and information about any free space on the data pages is not kept. For more information about this statement, refer to the *SQL Reference*.

The number of 4 KB pages for each user table in the database can be estimated by calculating:

```
ROUND DOWN(4020/(average row size + 10)) = records_per_page
```

and then inserting the result into:

```
   (number_of_records/records_per_page) * 1.1 = number_of_pages
```

where the average row size is the sum of the average column sizes, (For information about the size of each column, refer to the CREATE TABLE statement in the *SQL Reference*.), and the factor of ″1.1″ is for overhead.

Note: This formula only provides an estimate. Accuracy of the estimate is reduced if the record length varies because of fragmentation and overflow records.

You also have the option to create buffer pools or table spaces that have an 8 KB, 16 KB, or 32 KB page size. All tables created within a table space of a particular size have a matching page size. A single table or index object can be as large as 512 GB, assuming a 32 KB page size. You can have a maximum of 1012 columns when using an 8 KB, 16 KB, or 32 KB page size. The maximum number of columns is 500 for a 4 KB page size. Maximum row lengths also vary, depending on page size:

- When the page size is 4 KB, the row length can be up to 4005 bytes.
- When the page size is 8 KB, the row length can be up to 8101 bytes.
- When the page size is 16 KB, the row length can be up to 16 293 bytes.
- When the page size is 32 KB, the row length can be up to 32 677 bytes.

Having a larger page size facilitates a reduction in the number of levels in any index. If you are working with OLTP (online transaction processing) applications, which perform random row reads and writes, a smaller page size is better, because it wastes less buffer space with undesired rows. If you are working with DSS (decision support system) applications, which access large numbers of consecutive rows at a time, a larger page size is better, because it reduces the number of I/O requests required to read a specific number of rows. An exception occurs when the row size is smaller than the page size divided by 255. In such a case, there is wasted space on each page. (Recall that there can be a maximum of only 255 rows per page.) To reduce this wasted space, a smaller page size may be more appropriate.

You cannot restore a backup to a different page size.

You cannot import IXF data files that represent more than 755 columns. For more information about importing data into tables, and IXF data files, refer to the *Data Movement Utilities Guide and Reference*.

Declared temporary tables can only be created in their own ″user temporary″ table space type. There is no default user temporary table space. Temporary tables cannot have LONG data. The tables are dropped implicitly when an application disconnects from the database, and estimates of their space requirements should take this into account.

Long Field Data

Long field data is stored in a separate table object that is structured differently from other data types (see “User Table Data” on page 117 and “Large Object (LOB) Data”).

Data is stored in 32 KB areas that are broken up into segments whose sizes are ″powers of two″ times 512 bytes. (Hence these segments can be 512 bytes, 1024 bytes, 2048 bytes, and so on, up to 32 700 bytes.)

Long field data types (LONG VARCHAR or LONG VARGRAPHIC) are stored in a way that enables free space to be reclaimed easily. Allocation and free space information is stored in 4 KB allocation pages, which appear infrequently throughout the object.

The amount of unused space in the object depends on the size of the long field data, and whether this size is relatively constant across all occurrences of the data. For data entries larger than 255 bytes, this unused space can be up to 50 percent of the size of the long field data.

If character data is less than the page size, and it fits into the record along with the rest of the data, the CHAR, GRAPHIC, VARCHAR, or VARGRAPHIC data types should be used instead of LONG VARCHAR or LONG VARGRAPHIC.

Large Object (LOB) Data

Large object (LOB) data is stored in two separate table objects that are structured differently from other data types (see “User Table Data” on page 117 and “Long Field Data”).

To estimate the space required by LOB data, you need to consider the two table objects used to store data defined with these data types:

- **LOB Data Objects**

  Data is stored in 64 MB areas that are broken up into segments whose sizes are ″powers of two″ times 1024 bytes. (Hence these segments can be 1024 bytes, 2048 bytes, 4096 bytes, and so on, up to 64 MB.)

  To reduce the amount of disk space used by LOB data, you can specify the COMPACT option on the *lob-options* clause of the CREATE TABLE and the ALTER TABLE statements. The COMPACT option minimizes the amount of disk space required by allowing the LOB data to be split into smaller segments. This process does not involve data compression, but simply uses the minimum amount of space, to the nearest 1 KB boundary. Using the COMPACT option may result in reduced performance when appending to LOB values.

The amount of free space contained in LOB data objects is influenced by the amount of update and delete activity, as well as the size of the LOB values being inserted.

- **LOB Allocation Objects**

  Allocation and free space information is stored in 4 KB allocation pages that are separated from the actual data. The number of these 4 KB pages is dependent on the amount of data, including unused space, allocated for the large object data. The overhead is calculated as follows: one 4 KB page for every 64 GB, plus one 4 KB page for every 8 MB.

If character data is less than the page size, and it fits into the record along with the rest of the data, the CHAR, GRAPHIC, VARCHAR, or VARGRAPHIC data types should be used instead of BLOB, CLOB, or DBCLOB.

Index Space

For each index, the space needed can be estimated as:

```
(average index key size + 8) * number of rows * 2
```

where:

- The ″average index key size″ is the byte count of each column in the index key. Refer to the CREATE TABLE statement in the *SQL Reference* for information on how to calculate the byte count for columns with different data types. (When estimating the average column size for VARCHAR and VARGRAPHIC columns, use an average of the current data size, plus one byte. Do not use the maximum declared size.)
- The factor of ″2″ is for overhead, such as non-leaf pages and free space.

Note: For every column that allows NULLs, add one extra byte for the null indicator.

Temporary space is required when creating the index. The maximum amount of temporary space required during index creation can be estimated as:

```
(average index key size + 8) * number of rows * 3.2
```

where the factor of ″3.2″ is for index overhead, and space required for sorting during index creation.

Note: In the case of non-unique indexes, only four bytes are required to store duplicate key entries. The estimates shown above assume no duplicates. The space required to store an index may be over-estimated by the formula shown above.

The following two formulas can be used to estimate the number of leaf pages (the second provides a more accurate estimate). The accuracy of these estimates depends largely on how well the averages reflect the actual data.

Note: For SMS, the minimum required space is 12 KB. For DMS, the minimum is an extent.

- A rough estimate of the average number of keys per leaf page is:

  ```
  (.9 * (U - (M*2))) * (D + 1)
  ----------------------------
        K + 6 + (4 * D)
  ```

  where:

  – U, the usable space on a page, is approximately equal to the page size minus 100. For a page size of 4096, U is 3996.

  – M = U / (8 + *minimumKeySize*)

  – D = average number of duplicates per key value

  – K = *averageKeySize*

  Remember that *minimumKeySize* and *averageKeysize* must have an extra byte for each nullable key part, and an extra byte for the length of each variable length key part.

  If there are include columns, they should be accounted for in *minimumKeySize* and *averageKeySize*.

  The `.9` can be replaced by any (100 - pctfree)/100 value, if a percent free value other than the default value of ten percent was specified during index creation.

- A more accurate estimate of the average number of keys per leaf page is:

  ```
  L = number of leaf pages = X / (avg number of keys on leaf page)
  ```

  where X is the total number of rows in the table.

  You can estimate the original size of an index as:

  ```
  (L + 2L/(average number of keys on leaf page)) * pagesize
  ```

  For DMS table spaces, add together the sizes of all indexes on a table, and round up to a multiple of the extent size for the table space on which the index resides.

  You should provide additional space for index growth due to INSERT/UPDATE activity, which may result in page splits.

Use the following calculations to obtain a more accurate estimate of the original index size, as well as an estimate of the number of levels in the index. (This may be of particular interest if include columns are being used in the index definition.) The average number of keys per non-leaf page is roughly:

```
(.9 * (U - (M*2))) * (D + 1)
----------------------------
      K + 12 + (8 * D)
```

where:

- U, the usable space on a page, is approximately equal to the page size minus 100. For a page size of 4096, U is 3996.
- D is the average number of duplicates per key value on non-leaf pages (this will be much smaller than on leaf pages, and you may want to simplify the calculation by setting the value to 0).
- M = U / (8 + *minimumKeySize* for non-leaf pages)
- K = *averageKeySize* for non-leaf pages

The *minimumKeySize* and the *averageKeySize* for non-leaf pages will be the same as for leaf pages, except when there are include columns. Include columns are not stored on non-leaf pages.

You should not replace .9 with (100 - pctfree)/100, unless this value is greater than .9, because a maximum of 10 percent free space will be left on non-leaf pages during index creation.

The number of non-leaf pages can be estimated as follows:

```
if L > 1 then {P++; Z++}
While (Y > 1)
{
   P = P + Y
   Y = Y / N
  Z++
}
```

where:

- P is the number of pages (0 initially).
- L is the number of leaf pages.
- N is the number of keys for each non-leaf page.
- Y = L / N
- Z is the number of levels in the index tree (1 initially).

Total number of pages is:

```
T = (L + P + 2) * 1.0002
```

The additional 0.02 percent is for overhead, including space map pages.

The amount of space required to create the index is estimated as:

```
T * pagesize
```

Additional Space Requirements

Additional space is also required for:

- "Log File Space"
- "Temporary Work Space" on page 124

Log File Space

The amount of space (in bytes) required for log files can range from:

```
( logprimary * (logfilsiz + 2 ) * 4096 ) + 8192
```

to:

```
( (logprimary + logsecond) * (logfilsiz + 2 ) * 4096 ) + 8192
```

where:

- *logprimary* is the number of primary log files, defined in the database configuration file
- *logsecond* is the number of secondary log files, defined in the database configuration file
- *logfilsiz* is the number of pages in each log file, defined in the database configuration file
- 2 is the number of header pages required for each log file
- 4096 is the number of bytes in one page
- 8192 is the size (in bytes) of the log control file.

Refer to the *Administration Guide: Performance* for more information about these configuration parameters.

Note: The total active log space cannot exceed 32 GB.

The upper limit of log file space is dependent on the actual number of secondary log files that the database manager requires at run time. This upper limit may never be required, or may be needed only during occasional periods of high volume activity.

If the database is enabled for roll-forward recovery, special log space requirements should be taken into consideration:

- With the *logretain* configuration parameter enabled, the log files will be archived in the log path directory. The online disk space will eventually fill up, unless you move the log files to a different location.
- With the *userexit* configuration parameter enabled, a user exit program moves the archived log files to a different location. Extra log space is still required to allow for:
  - Online archived logs that are waiting to be moved by the user exit program
  - New log files being formatted for future use.

Temporary Work Space

Some SQL statements require temporary tables for processing (such as a work file for sorting operations that cannot be done in memory). These temporary tables require disk space; the amount of space required is dependent upon the queries, and the size of returned tables, and cannot be estimated.

You can use the database system monitor and the query table space APIs to track the amount of work space being used during the normal course of operations.

Designing Nodegroups

A *nodegroup* is a named set of one or more nodes that are defined as belonging to a database. Each database partition that is part of the database system configuration must already be defined in a *partition configuration file* called `db2nodes.cfg`. A nodegroup can contain as little as one database partition, or as much as the entire set of database partitions defined for the database system.

You create a new nodegroup using the CREATE NODEGROUP statement, and can modify it using the ALTER NODEGROUP statement. You can add or drop one or more database partitions from a nodegroup. The database partitions must be defined in the `db2nodes.cfg` file before modifying the nodegroup. Table spaces reside within nodegroups. Tables reside within table spaces.

When a nodegroup is created or modified, a *partitioning map* is associated with it. A partitioning map, in conjunction with a *partitioning key* and a hashing algorithm, is used by the database manager to determine which database partition in the nodegroup will store a given row of data. For more information about partitioning maps, see “Partitioning Maps” on page 127. For more information about partitioning keys, see “Partitioning Keys” on page 128.

In a non-partitioned database, no partitioning key or partitioning map is required. There are no nodegroup design considerations if you are using a non-partitioned database. A *database partition* is a part of the database, complete with user data, indexes, configuration files, and transaction logs. Default nodegroups that were created when the database was created, are used by the database manager. IBMCATGROUP is the default nodegroup for the table space containing the system catalogs. IBMTEMPGROUP is the default nodegroup for system temporary table spaces. IBMDEFAULTGROUP is the default nodegroup for the table spaces containing the user defined tables that you may choose to put there. A user temporary table space for a declared temporary table can be created in IBMDEFAULTGROUP or any user-created nodegroup, but not in IBMTEMPGROUP.

If you are using a multiple partition nodegroup, consider the following design points:

- In a multiple partition nodegroup, you can only create a unique index if it is a superset of the partitioning key.
- Depending on the number of database partitions in the database, you may have one or more single-partition nodegroups, and one or more multiple partition nodegroups present.
- Each database partition must be assigned a unique partition number. The same database partition may be found in one or more nodegroups.
- To ensure fast recovery of the database partition containing system catalog tables, avoid placing user tables on the same database partition. This is accomplished by placing user tables in nodegroups that do not include the database partition in the IBMCATGROUP nodegroup.

You should place small tables in single-partition nodegroups, except when you want to take advantage of *collocation* with a larger table. Collocation is the placement of rows from different tables that contain related data in the same database partition. Collocated tables allow DB2 to utilize more efficient join strategies. Collocated tables can reside in a single-partition nodegroup. Tables are considered collocated if they reside in a multiple partition nodegroup, have the same number of columns in the partitioning key, and if the data types of the corresponding columns are partition compatible. Rows in collocated tables with the same partitioning key value are placed on the same database partition. Tables can be in separate table spaces in the same nodegroup, and still be considered collocated.

You should avoid extending medium-sized tables across too many database partitions. For example, a 100 MB table may perform better on a 16 partition nodegroup than on a 32 partition nodegroup.

You can use nodegroups to separate online transaction processing (OLTP) tables from decision support (DSS) tables, to ensure that the performance of OLTP transactions is not adversely affected.

Nodegroup Design Considerations

Your logical database design, and the amount of data to be processed, will suggest whether your database needs to be partitioned. This section covers the following topics related to database partitioning:

- “Data Partitioning”
- “Partitioning Maps” on page 127
- “Partitioning Keys” on page 128
- “Table Collocation” on page 130
- “Partition Compatibility” on page 131
- “Replicated Summary Tables” on page 131

Data Partitioning

DB2 supports a partitioned storage model that allows you to store data across several database partitions in the database. This means that the data is physically stored across more than one database partition, and yet can be accessed as though it were located in the same place. Applications and users accessing data in a partitioned database do not need to be aware of the physical location of the data.

The data, while physically split, is used and managed as a logical whole. Users can choose how to partition their data by declaring partitioning keys. Users can also determine across which and how many database partitions their table data can be spread, by selecting the table space and the associated nodegroup in which the data should be stored. In addition, an updatable partitioning map is used with a hashing algorithm to specify the mapping of partitioning key values to database partitions, which determines the placement and retrieval of each row of data. As a result, you can spread the workload across a partitioned database for large tables, while allowing smaller tables to be stored on one or more database partitions. Each database partition has local indexes on the data it stores, resulting in increased performance for local data access.

You are not restricted to having all tables divided across all database partitions in the database. DB2 supports *partial declustering*, which means that you can divide tables and their table spaces across a subset of database partitions in the system (that is, a nodegroup).

An alternative to consider when you want tables to be positioned on each database partition, is to use summary tables and then replicate those tables. You can create a summary table containing the information that you need, and then replicate it to each node. For more information, see “Replicated Summary Tables” on page 131.

Partitioning Maps

In a partitioned database environment, the database manager must have a way of knowing which table rows are stored on which database partition. The database manager must know where to find the data it needs, and uses a map, called a *partitioning map*, to find the data.

A partitioning map is an internally generated array containing either 4 096 entries for multiple partition nodegroups, or a single entry for single-partition nodegroups. For a single-partition nodegroup, the partitioning map has only one entry containing the partition number of the database partition where all the rows of a database table are stored. For multiple partition nodegroups, the partition numbers of the nodegroup are specified in a round-robin fashion. Just as a city map is organized into sections using a grid, the database manager uses a *partitioning key* to determine the location (the database partition) where the data is stored.

For example, assume that you have a database created on four database partitions (numbered 0–3). The partitioning map for the IBMDEFAULTGROUP nodegroup of this database would be:

```
   0 1 2 3 0 1 2 ...
```

If a nodegroup had been created in the database using database partitions 1 and 2, the partitioning map for that nodegroup would be:

```
   1 2 1 2 1 2 1 ...
```

If the partitioning key for a table to be loaded in the database is an integer that has possible values between 1 and 500 000, the partitioning key is hashed to a partition number between 0 and 4 095. That number is used as an index into the partitioning map to select the database partition for that row.

Figure 35 on page 128 shows how the row with the partitioning key value (c1, c2, c3) is mapped to partition 2, which, in turn, references database partition n5.

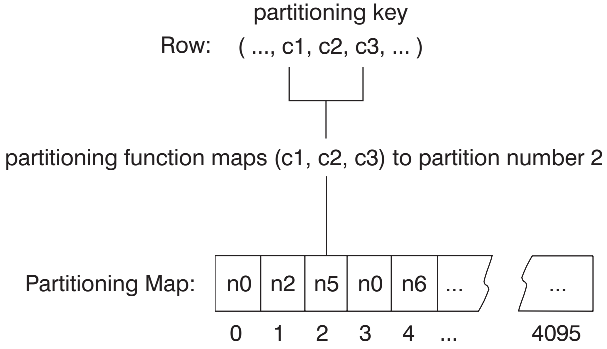

partitioning key
Row: (..., c1, c2, c3, ...)
partitioning function maps (c1, c2, c3) to partition number 2
Partitioning Map:
n0
n2
n5
n0
n6
...
...
0
1
2
3
4
...
4095

Figure 35. Data Distribution Using a Partition Map

A partition map is a flexible way of controlling where data is stored in a partitioned database. If you have a need at some future time to change the data distribution across the database partitions in your database, you can use the data redistribution utility. This utility allows you to rebalance or introduce skew into the data distribution. For more information about this utility, refer to "Redistributing Data Across Database Partitions" in the *Administration Guide: Performance*.

You can use the Get Table Partitioning Information (**sqlugtpi**) API to obtain a copy of a partitioning map that you can view. For more information about this API, refer to the *Administrative API Reference*.

Partitioning Keys

A *partitioning key* is a column (or group of columns) that is used to determine the partition in which a particular row of data is stored. A partitioning key is defined on a table using the CREATE TABLE statement. If a partitioning key is not defined for a table in a table space that is divided across more than one database partition in a nodegroup, one is created by default from the first column of the primary key. If no primary key is specified, the default partitioning key is the first non-long field column defined on that table. (*Long* includes all long data types and all large object (LOB) data types). If you are creating a table in a table space associated with a single-partition nodegroup, and you want to have a partitioning key, you must define the partitioning key explicitly. One is not created by default.

If no columns satisfy the requirement for a default partitioning key, the table is created without one. Tables without a partitioning key are only allowed in single-partition nodegroups. You can add or drop partitioning keys at a later time, using the ALTER TABLE statement. Altering the partition key can only be done to a table whose table space is associated with a single-partition nodegroup.

Choosing a good partitioning key is important. You should take into consideration:

- How tables are to be accessed
- The nature of the query workload
- The join strategies employed by the database system.

If collocation is not a major consideration, a good partitioning key for a table is one that spreads the data evenly across all database partitions in the nodegroup. The partitioning key for each table in a table space that is associated with a nodegroup determines if the tables are collocated. Tables are considered collocated when:

- The tables are placed in table spaces that are in the same nodegroup
- The partition keys in each table have the same number of columns
- The data types of the corresponding columns are partition-compatible.

This ensures that rows of collocated tables with the same partitioning key values are located on the same partition. For more information about partition-compatibility, see "Partition Compatibility" on page 131. For more information about table collocation, see "Table Collocation" on page 130.

An inappropriate partitioning key can cause uneven data distribution. Columns with unevenly distributed data, and columns with a small number of distinct values should not be chosen as a partitioning key. The number of distinct values must be great enough to ensure an even distribution of rows across all database partitions in the nodegroup. The cost of applying the partitioning hash algorithm is proportional to the size of the partitioning key. The partitioning key cannot be more than 16 columns, but fewer columns result in better performance. Unnecessary columns should not be included in the partitioning key.

The following points should be considered when defining partitioning keys:

- Creation of a multiple partition table that contains only long data types (LONG VARCHAR, LONG VARGRAPHIC, BLOB, CLOB, or DBCLOB) is not supported.
- The partition key definition cannot be altered.
- You cannot update the partitioning key column value for a row in the table.
- You can only delete or insert partitioning key column values.
- The partitioning key should include the most frequently joined columns.
- The partitioning key should be made up of columns that often participate in a GROUP BY clause.
- Any unique key or primary key must contain all of the partitioning key columns.

- In an online transaction processing (OLTP) environment, all columns in the partitioning key should participate in the transaction by using equal (=) predicates with constants or host variables. For example, assume you have an employee number, *emp_no*, that is often used in transactions such as:

  ```
  UPDATE emp_table SET ... WHERE
  emp_no = host-variable
  ```

  In this case, the EMP_NO column would make a good single column partitioning key for EMP_TABLE.

Hash partitioning is the method by which the placement of each row in the partitioned table is determined. The method works as follows:

1. The hashing algorithm is applied to the value of the partitioning key, and generates a partition number between zero and 4095.
2. The partitioning map is created when a nodegroup is created. Each of the partition numbers is sequentially repeated in a round-robin fashion to fill the partitioning map. For more information about partitioning maps, see “Partitioning Maps” on page 127.
3. The partition number is used as an index into the partitioning map. The number at that location in the partitioning map is the number of the database partition where the row is stored.

Table Collocation

You may discover that two or more tables frequently contribute data in response to certain queries. In this case, you will want related data from such tables to be located as close together as possible. In an environment where the database is physically divided among two or more database partitions, there must be a way to keep the related pieces of the divided tables as close together as possible. The ability to do this is called *table collocation*.

Tables are collocated when they are stored in the same nodegroup, and when their partitioning keys are compatible. Placing both tables in the same nodegroup ensures a common partitioning map. The tables may be in different table spaces, but the table spaces must be associated with the same nodegroup. The data types of the corresponding columns in each partitioning key must be *partition-compatible*. For information about partition compatibility, see “Partition Compatibility” on page 131.

DB2 has the ability to recognize, when accessing more than one table for a join or a subquery, that the data to be joined is located at the same database partition. When this happens, DB2 can choose to perform the join or subquery at the database partition where the data is stored, instead of having to move data between database partitions. This ability to carry out joins or subqueries at the database partition has significant performance advantages. For more information, refer to ″Collocated Joins″ in the *Administration Guide: Performance*.

Partition Compatibility

The base data types of corresponding columns of partitioning keys are compared and can be declared *partition compatible*. Partition compatible data types have the property that two variables, one of each type, with the same value, are mapped to the same partition number by the same partitioning algorithm.

Partition compatibility has the following characteristics:

- A base data type is compatible with another of the same base data type.
- Internal formats are used for DATE, TIME, and TIMESTAMP data types. They are not compatible with each other, and none are compatible with CHAR.
- Partition compatibility is not affected by columns with NOT NULL or FOR BIT DATA definitions.
- NULL values of compatible data types are treated identically; those of non-compatible data types may not be.
- Base data types of a user defined type are used to analyze partition compatibility.
- Decimals of the same value in the partitioning key are treated identically, even if their scale and precision differ.
- Trailing blanks in character strings (CHAR, VARCHAR, GRAPHIC, or VARGRAPHIC) are ignored by the hashing algorithm.
- BIGINT, SMALLINT, and INTEGER are compatible data types.
- REAL and FLOAT are compatible data types.
- CHAR and VARCHAR of different lengths are compatible data types.
- GRAPHIC and VARGRAPHIC are compatible data types.
- Partition compatibility does not apply to LONG VARCHAR, LONG VARGRAPHIC, CLOB, DBCLOB, and BLOB data types, because they are not supported as partitioning keys.

Replicated Summary Tables

A *summary table* is a table that is defined by a query that is also used to determine the data in the table. Summary tables can be used to improve the performance of queries. If DB2 determines that a portion of a query could be resolved using a summary table, the query may be rewritten by the database manager to use the summary table.

In a partitioned database environment, you can replicate summary tables. You can use *replicated summary tables* to improve query performance. A replicated summary table is based on a table that may have been created in a single-partition nodegroup, but that you want replicated across all of the database partitions in the nodegroup. To create the replicated summary table, invoke the CREATE TABLE statement with the REPLICATED keyword. The

REPLICATED keyword can only be specified for a summary table that is defined with the REFRESH DEFERRED option.

For more information about summary tables, see ″Creating a Summary Table″ in the *Administration Guide: Implementation*.

By using replicated summary tables, you can obtain collocation between tables that are not typically collocated. Replicated summary tables are particularly useful for joins in which you have a large fact table and small dimension tables. To minimize the extra storage required, as well as the impact of having to update every replica, tables that are to be replicated should be small and infrequently updated.

Note: You should also consider replicating larger tables that are infrequently updated: the one-time cost of replication is offset by the performance benefits that can be obtained through collocation.

By specifying a suitable predicate in the subselect clause used to define the replicated table, you can replicate selected columns, selected rows, or both.

For more information about replicated summary tables, refer to the CREATE TABLE statement in the *SQL Reference*. For more information about collocated joins, refer to ″Collocated Joins″ in the *Administration Guide: Implementation*.

Designing and Choosing Table Spaces

A table space is a storage model that provides a level of indirection between a database and the tables stored within that database. Table spaces reside in nodegroups. They allow you to assign the location of database and table data directly onto containers. (A container can be a directory name, a device name, or a file name.) This can provide improved performance, more flexible configuration, and better integrity.

For information about creating or altering a table space, see ″Creating a Table Space″, or ″Altering a Table Space″ in the *Administration Guide: Implementation*.

Since table spaces reside in nodegroups, the table space selected to hold a table defines how the data for that table is distributed across the database partitions in a nodegroup. A single table space can span several containers. It is possible for multiple containers (from one or more table spaces) to be created on the same physical disk (or drive). For improved performance, each container should use a different disk. Figure 36 on page 133 illustrates the relationship between tables and table spaces within a database, and the containers associated with that database.

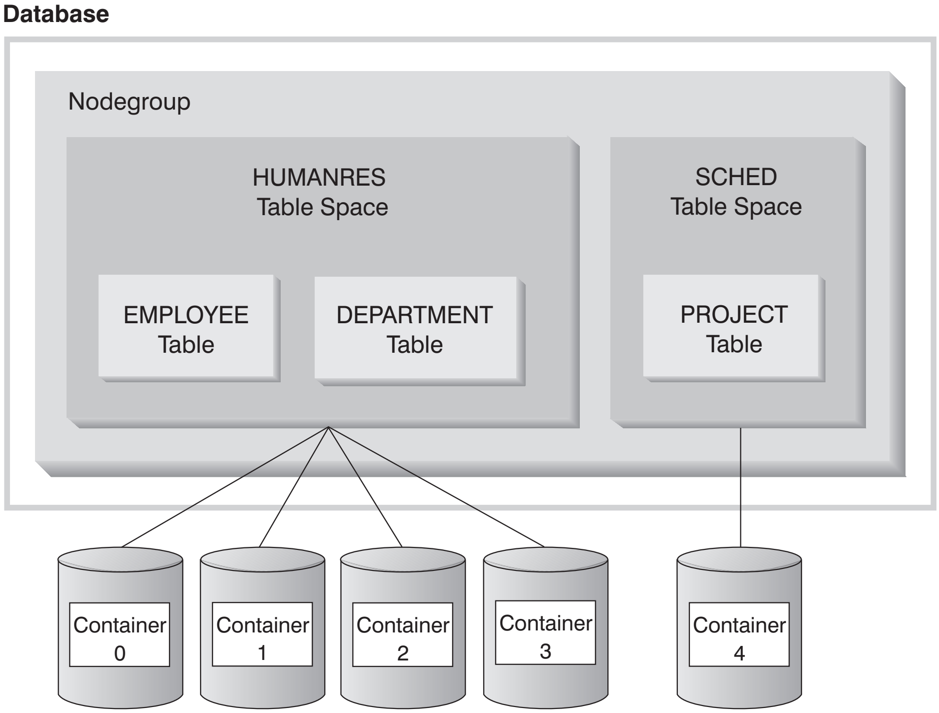


Figure 36. Table Spaces and Tables Within a Database

The EMPLOYEE and DEPARTMENT tables are in the HUMANRES table space, which spans containers 0, 1, 2 and 3. The PROJECT table is in the SCHED table space in container 4. This example shows each container existing on a separate disk.

The database manager attempts to balance the data load across containers. As a result, all containers are used to store data. The number of pages that the database manager writes to a container before using a different container is called the *extent size*. The database manager does not always start storing table data in the first container.

Figure 37 on page 134 shows the HUMANRES table space with an extent size of two 4 KB pages, and four containers, each with a small number of allocated extents. The DEPARTMENT and EMPLOYEE tables both have seven pages, and span all four containers.

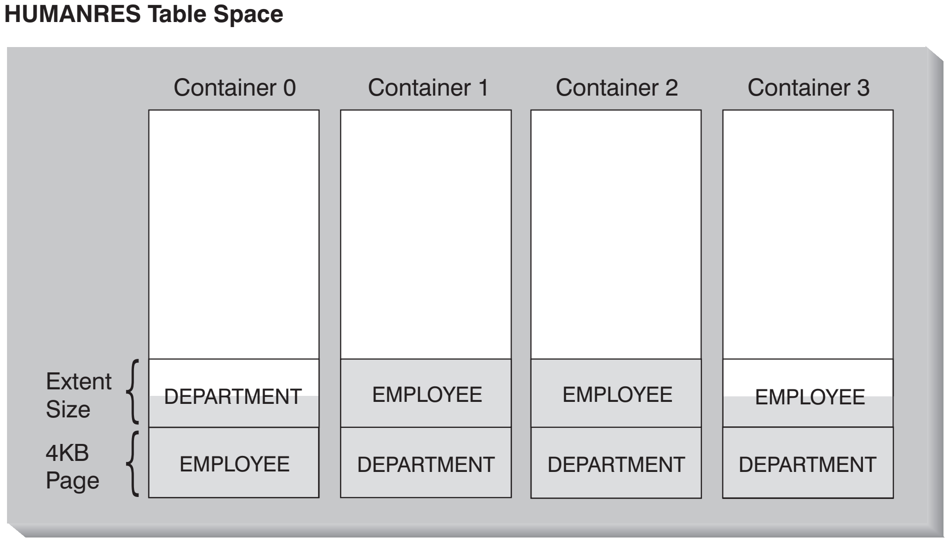


Figure 37. Containers and Extents

A database must contain at least three table spaces:

- One *catalog table space*, which contains all of the system catalog tables for the database. This table space is called SYSCATSPACE, and it cannot be dropped. IBMCATGROUP is the default nodegroup for this table space.
- One or more *user table spaces*, which contain all user defined tables. By default, one table space, USERSPACE1, is created. IBMDEFAULTGROUP is the default nodegroup for this table space.

  You should specify a table space name when you create a table, or the results may not be what you intend. If you do not specify a table space name, the table is placed according to the following rules: If user-created table spaces exist, choose the one with the smallest page size large enough for this table. Otherwise, use USERSPACE1 if it's page size is large enough for the table. If no table spaces with a large enough page size exist, the table is not created.

  A table's page size is determined either by row size, or the number of columns. The maximum allowable length for a row is dependent upon the page size of the table space in which the table is created. Possible values for page size are 4 KB (the default), 8 KB, 16 KB, and 32 KB. You can use a table space with one page size for the base table, and a different table space with a different page size for long or LOB data. (Recall that SMS does not support tables that span table spaces, but that DMS does.) If the number of columns or the row size exceeds the limits for a table space's page size, an error is returned (SQLSTATE 42997).
- One or more *temporary table spaces*, which contain temporary tables. Temporary table spaces can be *system temporary table spaces* or *user temporary*

table spaces. A database must have at least one system temporary table space; by default, one system temporary table space called TEMPSPACE1 is created at database creation time. IBMTEMPGROUP is the default nodegroup for this table space. User temporary table spaces are *not* created by default at database creation time.

If a database uses more than one temporary table space, temporary objects are allocated among the temporary table spaces in a round-robin fashion.

If queries are running against tables in table spaces that are defined with a page size larger than the 4 KB default (for example, an ORDER BY on 1012 columns), some of them may fail. This will occur if there are no temporary table spaces defined with a larger page size. You may need to create a temporary table space with a larger page size (8 KB, 16 KB, or 32 KB). Any DML (Data Manipulation Language) statement could fail unless there exists a temporary table space with the same page size as the largest page size in the user table space.

You should define a single SMS temporary table space with a page size equal to the page size used in the majority of your user table spaces. This should be adequate for typical environments and workloads. See also "Recommendations for Temporary Table Spaces" on page 147.

In a partitioned database environment, the catalog node will contain all three default table spaces, and the other database partitions will each contain only TEMPSPACE1 and USERSPACE1.

There are two types of table space, both of which can be used in a single database:

- "System Managed Space": The operating system's file manager controls the storage space.
- "Database Managed Space Table Space" on page 139: The database manager controls the storage space.

System Managed Space

In an SMS (System Managed Space) table space, the operating system's file system manager allocates and manages the space where the table is stored. The storage model typically consists of many files, representing table objects, stored in the file system space. The user decides on the location of the files, DB2 controls their names, and the file system is responsible for managing them. By controlling the amount of data written to each file, the database manager distributes the data evenly across the table space containers. An SMS table space is the default table space.

Each table has at least one SMS physical file associated with it. See "SMS Physical Files" on page 138 for a list of these files and a description of their contents.

In an SMS table space, a file is extended one page at a time as the object grows. If you need improved insert performance, you can consider enabling multipage file allocation. This allows the system to allocate or extend the file by more than one page at a time. You must run **db2empfa** to enable multipage file allocation. In a partitioned database environment, this utility must be run on each database partition. Once multipage file allocation is enabled, it cannot be disabled. For more information about **db2empfa**, refer to the *Command Reference*.

You should explicitly define SMS table spaces using the MANAGED BY SYSTEM option on the CREATE DATABASE command, or on the CREATE TABLESPACE statement. You must consider two key factors when you design your SMS table spaces:

- Containers for the table space.

  You must specify the number of containers that you want to use for your table space. It is very important to identify all the containers you want to use, because you cannot add or delete containers after an SMS table space is created. In a partitioned database environment, when a new partition is added to the nodegroup for an SMS table space, the ALTER TABLESPACE statement can be used to add containers for the new partition.

  Each container used for an SMS table space identifies an absolute or relative directory name. Each of these directories can be located on a different file system (or physical disk). The maximum size of the table space can be estimated by:

  ```
  number of containers * (maximum file system size
     supported by the operating system)
  ```

  This formula assumes that there is a distinct file system mapped to each container, and that each file system has the maximum amount of space available. In practice, this may not be the case, and the maximum table space size may be much smaller.

  Note: Care must be taken when defining the containers. If there are existing files or directories on the containers, an error (SQL0298N) is returned.

- Extent size for the table space.

  The extent size can only be specified when the table space is created. Because it cannot be changed later, it is important to select an appropriate value for the extent size. For more information, see “Choosing an Extent Size” on page 146.

  If you do not specify the extent size when creating a table space, the database manager will create the table space using the default extent size, defined by the *dft_extent_sz* database configuration parameter (refer to the *Administration Guide: Performance* for more information about this parameter). This configuration parameter is initially set based on information provided when the database is created. If the DFT_EXTENT_SZ parameter is not specified on the CREATE DATABASE command, the default extent size will be set to 32.

To choose appropriate values for the number of containers and the extent size for the table space, you must understand:

- The limitation that your operating system imposes on the size of a logical file system.

  For example, some operating systems have a 2 GB limit. Therefore, if you want a 64 GB table object, you will need at least 32 containers on this type of system.

  When you create the table space, you can specify containers that reside on different file systems and as a result, increase the amount of data that can be stored in the database.
- How the database manager manages the data files and containers associated with a table space.

  The first table data file (`SQL00001.DAT`) is created in the first container specified for the table space, and this file is allowed to grow to the extent size. After it reaches this size, the database manager writes data to `SQL00001.DAT` in the next container. This process continues until all of the containers contain `SQL00001.DAT` files, at which time the database manager returns to the first container. This process (known as *striping*) continues through the container directories until a container becomes full (SQL0289N), or no more space can be allocated from the operating system (disk full error). Striping is also used for index (`SQL`*nnnnn*`.INX`), long field (`SQL`*nnnnn*`.LF`), and LOB (`SQL`*nnnnn*`.LB` and `SQL`*nnnnn*`.LBA`) files.

  Note: The SMS table space is full as soon as any one of its containers is full. Thus, it is important to allocate the same amount of space for each container.

  To help distribute data across the containers more evenly, the database manager determines which container to use first by taking the table identifier (1 in the above example) modulo the number of containers. Containers are numbered sequentially, starting at 0.

  For more information about the files used in an SMS table space, see “SMS Physical Files” on page 138.

SMS Physical Files

The following files are found within an SMS table space directory container:

File Name **Description**

SQLTAG.NAM
There is one of these files in each container subdirectory, and they are used by the database manager when you connect to the database to verify that the database is complete and consistent.

SQLxxxxx.DAT
Table file. All table rows are stored here, with the exception of LONG VARCHAR, LONG VARGRAPHIC, BLOB, CLOB, or DBCLOB data.

SQLxxxxx.LF File containing LONG VARCHAR or LONG VARGRAPHIC data (also called ″long field data″). This file is only created if LONG VARCHAR or LONG VARGRAPHIC columns exist in the table.

SQLxxxxx.LB Files containing BLOB, CLOB, or DBCLOB data (also called ″LOB data″). These files are only created if BLOB, CLOB, or DBCLOB columns exist in the table.

SQLxxxxx.LBA
Files containing allocation and free space information about the `SQL`*`xxxxx`*`.LB` files.

SQLxxxxx.INX
Index file for a table. All indexes for the corresponding table are stored in this single file. It is only created if indexes have been defined.

Note: When an index is dropped, the space is not physically freed from the index (.INX) file until the index file is deleted. The index file will be deleted if all the indexes on the table are dropped (and committed), or if the table is reorganized. If the index file is not deleted, the space will be marked free once the drop has been committed, and will be reused for future index creation or index maintenance.

SQLxxxxx.DTR
Temporary data file for the reorganization of a DAT file. When reorganizing a table, the reorg utility (through the REORG TABLE command) creates a table in one of the system temporary table spaces. These temporary table spaces can be defined to use containers different from those used for the user defined tables.

SQLxxxxx.LFR

Temporary data file for the reorganization of an LF file. When reorganizing a table, the reorg utility (through the REORG TABLE command) creates a table in one of the system temporary table spaces. These temporary table spaces can be defined to use containers different from those used for the user defined tables.

SQLxxxxx.RLB

Temporary data file for the reorganization of an LB file. When reorganizing a table, the reorg utility (through the REORG TABLE command) creates a table in one of the system temporary table spaces. These temporary table spaces can be defined to use containers different from those used for the user defined tables.

SQLxxxxx.RBA

Temporary data file for the reorganization of an LBA file. When reorganizing a table, the reorg utility (through the REORG TABLE command) creates a table in one of the system temporary table spaces. These temporary table spaces can be defined to use containers different from those used for the user defined tables.

Notes:

1. Do *not* make any direct changes to these files. They can only be accessed indirectly using the documented APIs and by tools that implement those APIs, including the command line processor and the Control Center.
2. Do not move these files.
3. Do not remove these files.
4. The only supported means of backing up a database or a table space is through the **sqlubkp** (Backup Database) API, including the command line processor and Control Center implementations of that API.

Database Managed Space Table Space

In a DMS (Database Managed Space) table space, the database manager controls the storage space. The storage model consists of a limited number of devices whose space is managed by DB2. The Administrator decides which devices to use, and DB2 manages the space on those devices. The table space is essentially an implementation of a special purpose file system designed to best meet the needs of the database manager. The table space definition includes a list of the devices or files that belong to the table space, and in which data can be stored.

A DMS table space containing user defined tables and data can be defined as:

- A *regular* table space to store normal table and index data

- A *long* table space to store long field or LOB data.

When designing your DMS table spaces and containers, you should consider the following:

- The database manager uses striping to ensure an even distribution of data across all containers.
- The maximum size of regular table spaces is 64 GB for 4 KB pages; 128 GB for 8 KB pages; 256 GB for 16 KB pages; and 512 GB for 32 KB pages. The maximum size of long table spaces is 2 TB.
- Unlike SMS table spaces, the containers that make up a DMS table space do not need to be the same size; however, this is not normally recommended, because it results in uneven striping across the containers, and sub-optimal performance. If any container is full, DMS table spaces use available free space from other containers.
- Because space is pre-allocated, it must be available before the table space can be created. When using device containers, the device must also exist with enough space for the definition of the container. Each device can have only one container defined on it. To avoid wasted space, the size of the device and the size of the container should be equivalent. If, for example, the device is allocated with 5 000 pages, and the device container is defined to allocate 3 000 pages, 2 000 pages on the device will not be usable.
- One page in every container is reserved for overhead, and the remaining pages will be used one extent at a time. Only full extents are used, so for optimal space management, you can use the following formula to determine an appropriate size to use when allocating a container:

  ```
  (extent_size * n) + 1
  ```

  where *extent_size* is the size of each extent in the table space, and *n* is the number of extents that you want to store in the container.
- Three extents in the table space are reserved for overhead.
- At least two extents are required to store any user table data. (These extents are required for the regular data for one table, and not for any index, long field or large object data, which require their own extents.)
- Device containers must use logical volumes with a ″character special interface″, not physical volumes.
- You can use files instead of devices with DMS table spaces. No operational difference exists between a file and a device; however, a file can be less efficient because of the run-time overhead associated with the file system. Files are useful when:
  - Devices are not directly supported
  - A device is not available
  - Maximum performance is not required
  - You do not want to set up devices.

- If your workload involves LOBs or LONG VARCHAR data, you may derive performance benefits from file system caching. Note that LOBs and LONG VARCHARs are not buffered by DB2's buffer pool.
- Some operating systems allow you to have physical devices greater than 2 GB in size. You should consider partitioning the physical device into multiple logical devices, so that no container is larger than the size allowed by the operating system.

Adding Containers to DMS Table Spaces

The ALTER TABLESPACE statement lets you add a container to an existing table space to increase its storage capacity. The contents of the table space are then rebalanced across all containers. Access to the table space is not restricted during rebalancing. If you need to add more than one container, you should add them at the same time, either in one ALTER TABLESPACE statement, or within the same transaction, to prevent the database manager from having to rebalance the containers more than once.

You should check how full the containers for a table space are by using the LIST TABLESPACE CONTAINERS or the LIST TABLESPACES command. Adding new containers should be done before the existing containers are almost or completely full. The new space across all containers is not available until rebalancing is complete.

Adding a container which is smaller than existing containers results in a uneven distribution of data. This can cause parallel I/O operations, such as prefetching data, to perform less efficiently than they otherwise could on containers of equal size.

Table Space Design Considerations

This section covers the following topics:

- "Considerations for Table Space Input and Output (I/O)" on page 142
- "Mapping Table Spaces to Buffer Pools" on page 143
- "Mapping Table Spaces to Nodegroups" on page 144
- "Mapping Tables to Table Spaces" on page 144
- "Choosing an Extent Size" on page 146
- "Recommendations for Temporary Table Spaces" on page 147
- "Recommendations for Catalog Table Spaces" on page 148
- "Workload Considerations" on page 149
- "Choosing an SMS or DMS Table Space" on page 150
- "Optimizing Performance When Data is Placed on RAID Devices" on page 151.

Considerations for Table Space Input and Output (I/O)

The type and design of your table space determines the efficiency of the I/O performed against that table space. Following are concepts that you should understand before considering further the issues surrounding table space design and use:

Big-block reads
A read where several pages (usually an extent) are retrieved in a single request. Reading several pages at once is more efficient than reading each page separately.

Prefetching
The reading of pages in advance of those pages being referenced by a query. The overall objective is to reduce response time. This can be achieved if the prefetching of pages can occur asynchronously to the execution of the query. The best response time is achieved when either the CPU or the I/O subsystem is operating at maximum capacity.

Page cleaning
As pages are read and modified, they accumulate in the database buffer pool. When a page is read in, it is read into a buffer pool page. If the buffer pool is full of modified pages, one of these modified pages must be written out to the disk before the new page can be read in. To prevent the buffer pool from becoming full, page cleaner agents write out modified pages to guarantee the availability of buffer pool pages for future read requests.

Whenever it is advantageous to do so, DB2 performs big-block reads. This typically occurs when retrieving data that is sequential or partially sequential in nature. The amount of data read in one read operation depends on the extent size — the bigger the extent size, the more pages can be read at one time.

How the extent is stored on disk affects I/O efficiency. In a DMS table space using device containers, the data tends to be contiguous on disk, and can be read with a minimum of seek time and disk latency. If files are being used, however, the data may have been broken up by the file system and stored in more than one location on disk. This occurs most often when using SMS table spaces, where files are extended one page at a time, making fragmentation more likely. A large file that has been pre-allocated for use by a DMS table space tends to be contiguous on disk, especially if the file was allocated in a clean file space.

You can control the degree of prefetching by tuning the PREFETCHSIZE parameter on the CREATE TABLESPACE statement. (The default value for all table spaces in the database is set by the *dft_prefetch_sz* database configuration parameter.) The PREFETCHSIZE parameter tells DB2 how many pages to read whenever a prefetch is triggered. By setting PREFETCHSIZE to be a multiple of the EXTENTSIZE parameter on the CREATE TABLESPACE statement, you can cause multiple extents to be read in parallel. (The default value for all table spaces in the database is set by the *dft_extent_sz* database configuration parameter.) The EXTENTSIZE parameter specifies the number of 4 KB pages that will be written to a container before skipping to the next container.

For example, suppose you had a table space that used three devices. If you set the PREFETCHSIZE to be three times the EXTENTSIZE, DB2 can do a big-block read from each device in parallel, thereby significantly increasing I/O throughput. This assumes that each device is a separate physical device, and that the controller has sufficient bandwidth to handle the data stream from each device. Note that DB2 may have to dynamically adjust the prefetch parameters at run time based on query speed, buffer pool utilization, and other factors.

Some file systems use their own prefetching method (such as the Journaled File System on AIX). In some cases, file system prefetching is set to be more aggressive than DB2 prefetching. This may cause prefetching for SMS and DMS table spaces with file containers to appear to outperform prefetching for DMS table spaces with devices. This is misleading, because it is likely the result of the additional level of prefetching that is occurring in the file system. DMS table spaces should be able to outperform any equivalent configuration.

For prefetching (or even reading) to be efficient, a sufficient number of clean buffer pool pages must exist. For example, there could be a parallel prefetch request that reads three extents from a table space, and for each page being read, one modified page is written out from the buffer pool. The prefetch request may be slowed down to the point where it cannot keep up with the query. Page cleaners should be configured in sufficient numbers to satisfy the prefetch request. At least one page cleaner should be defined for each real disk used by the database. For more information about these topics, refer to the *Administration Guide: Performance*.

Mapping Table Spaces to Buffer Pools

Each table space is associated with a specific buffer pool. The default buffer pool is IBMDEFAULTBP. If another buffer pool is to be associated with a table space, the buffer pool must exist (it is defined with the CREATE BUFFERPOOL statement), and the association is defined when the table space is created (using the CREATE TABLESPACE statement). The association between the table space and the buffer pool can be changed using the ALTER TABLESPACE statement.

Having more than one buffer pool allows you to configure the memory used by the database to improve overall performance. For table spaces with one or more large tables that are accessed randomly by users, the size of the buffer pool can be limited, because caching the data pages might not be beneficial.

The table space for an online transaction application might be associated with a larger buffer pool, so that the data pages used by the application can be cached longer, resulting in faster response times. Care must be taken in configuring new buffer pools. For more information on this topic, refer to ″Managing the Database Buffer Pool″ in the *Administration Guide: Performance*.

Note: If you have determined that a page size of 8 KB, 16 KB, or 32 KB is required by your database, each table space with one of these page sizes must be mapped to a buffer pool with the same page size.

The storage required for all the buffer pools must be available to the database manager when the database is started. If DB2 is unable to obtain the required storage, the database manager will start up with default buffer pools (one each of 4 KB, 8 KB, 16 KB, and 32 KB page sizes), and issue a warning.

In a partitioned database environment, you can create a buffer pool of the same size for all partitions in the database. You can also create buffer pools of different sizes on different partitions. For more information about the CREATE BUFFERPOOL statement, refer to the *SQL Reference*.

Mapping Table Spaces to Nodegroups

In a partitioned database environment, each table space is associated with a specific nodegroup. This allows the characteristics of the table space to be applied to each node in the nodegroup. The nodegroup must exist (it is defined with the CREATE NODEGROUP statement), and the association between the table space and the nodegroup is defined when the table space is created using the CREATE TABLESPACE statement.

You cannot change the association between table space and nodegroup using the ALTER TABLESPACE statement. You can only change the table space specification for individual partitions within the nodegroup. In a single-partition environment, each table space is associated with the default nodegroup. The default nodegroup, when defining a table space, is IBMDEFAULTGROUP, unless a system temporary table space is being defined; then IBMTEMPGROUP is used. For more information about the CREATE NODEGROUP statement, refer to the *SQL Reference*. For more information about nodegroups and physical database design, see “Designing Nodegroups” on page 124.

Mapping Tables to Table Spaces

When determining how to map tables to table spaces, you should consider:

- The partitioning of your tables.

  At a minimum, you should ensure that the table space you choose is in a nodegroup with the partitioning you want.
- The amount of data in the table.

If you plan to store many small tables in a table space, consider using SMS for that table space. The DMS advantages with I/O and space management efficiency are not as important with small tables. The SMS advantages of allocating space one page at a time, and only when needed, are more attractive with smaller tables. If one of your tables is larger, or you need faster access to the data in the tables, a DMS table space with a small extent size should be considered.

You may wish to use a separate table space for each very large table, and group all small tables together in a single table space. This separation also allows you to select an appropriate extent size based on the table space usage. (See "Choosing an Extent Size" on page 146 for additional information.)

- The type of data in the table.

  You may, for example, have tables containing historical data that is used infrequently; the end-user may be willing to accept a longer response time for queries executed against this data. In this situation, you could use a different table space for the historical tables, and assign this table space to less expensive physical devices that have slower access rates.

  Alternatively, you may be able to identify some essential tables which require high availability and fast response time. You may want to put these tables into a table space assigned to a fast physical device that can help support these important data requirements.

  Using DMS table spaces, you can also distribute your table data across three different table spaces: one for index data; one for LOB and long field data; and one for regular table data. This allows you to choose the table space characteristics and the physical devices supporting those table spaces to best suit the data. For example, you could put your index data on the fastest devices you have available, and as a result, obtain significant performance improvements. If you split a table across DMS table spaces, you should consider backing up and restoring all parts of the table together if roll-forward recovery is enabled. SMS table spaces do not support this type of data distribution across table spaces.

- Administrative issues.

  Some administrative functions can be performed at the table space level instead of the database or table level. For example, taking a backup of a table space instead of a database can help you make better use of your time and resources. It allows you to frequently back up table spaces with large volumes of changes, while only occasionally backing up tables spaces with very low volumes of changes.

  You can restore a database or a table space. If unrelated tables do not share table spaces, you have the option to restore a smaller portion of your database and reduce costs.

A good approach is to group related tables in a set of table spaces. These tables could be related through referential constraints, or through other defined business constraints.

If you need to drop and redefine a particular table often, you may want to define the table in its own table space, because it is more efficient to drop a DMS table space than it is to drop a table.

Choosing an Extent Size

The extent size for a table space represents the number of pages of table data that will be written to a container before data will be written to the next container. When selecting an extent size, you should consider:

- The size and type of tables in the table space.

  Space in DMS table spaces is allocated to a table one extent at a time. As the table is populated and an extent becomes full, a new extent is allocated.

  A table is made up of the following separate table objects:

  - A data object. This is where the regular column data is stored.
  - An index object. This is where all indexes defined on the table are stored.
  - A long field object. This is where long field data, if your table has one or more LONG columns, is stored.
  - Two LOB objects. If your table has one or more LOB columns, they are stored in these two table objects:
    - One table object for the LOB data
    - A second table object for meta-data describing the LOB data.

  Each table object is stored separately, and each object allocates new extents as needed. Each table object is also paired with a meta-data object called an *extent map*, which describes all of the extents in the table space that belong to the table object. Space for extent maps is also allocated one extent at a time.

  The initial allocation of space for a table, therefore, is two extents for each table object. If you have many small tables in a table space, you may have a relatively large amount of space allocated to store a relatively small amount of data. In such a case, you should specify a small extent size, or use an SMS table space, which allocates pages one at a time.

  If, on the other hand, you have a very large table that has a high growth rate, and you are using a DMS table space with a small extent size, you could have unnecessary overhead related to the frequent allocation of additional extents.

- The type of access to the tables.

  If access to the tables includes many queries or transactions that process large quantities of data, prefetching data from the tables may provide significant performance benefits. (Refer to *Administration Guide: Performance* for information about data prefetching and its relationship to the extent size.)

- The minimum number of extents required.

  If there is not enough space in the containers for five extents of the table space, the table space will not be created.

Recommendations for Temporary Table Spaces

It is recommended that you define a single SMS temporary table space with a page size equal to the page size used in the majority of your regular table spaces. This should be suitable for typical environments and workloads. However, it can be advantageous to experiment with different temporary table space configurations and workloads. The following points should be considered:

- Temporary tables are in most cases accessed in batches and sequentially. That is, a batch of rows is inserted, or a batch of sequential rows is fetched. Therefore, a larger page size typically results in better performance, because fewer logical or physical page I/O requests are required to read a given amount of data. This is not always the case when the average temporary table row size is smaller than the page size divided by 255. A maximum of 255 rows can exist on any page, regardless of the page size. For example, a query that requires a temporary table with 15-byte rows would be better served by a 4 KB temporary table space page size, because 255 such rows can all be contained within a 4 KB page. An 8 KB (or larger) page size would result in at least 4 KB (or more) bytes of wasted space on each temporary table page, and would not reduce the number of required I/O requests.
- If more than fifty percent of the regular table spaces in your database use the same page size, it can be advantageous to define your temporary table spaces with the same page size. The reason for this is that this arrangement enables your temporary table space to share the same buffer pool space with most or all of your regular table spaces. This, in turn, simplifies buffer pool tuning.
- When reorganizing a table using a temporary table space, the page size of the temporary table space must match that of the table. For this reason, you should ensure that there are temporary table spaces defined for each different page size used by existing tables that you may reorganize using a temporary table space.

  You can also reorganize without a temporary table space by reorganizing the table ″inplace″; that is, directly in the target table space. Of course, this ″inplace″ reorganization requires that there be extra space in the target table space for the reorganization process. For additional information about table reorganization, refer to *Administration Guide: Performance*.

- In general, when temporary table spaces of differing page sizes exist, the optimizer will most often choose the temporary table space with the largest buffer pool. In such cases, it is often wise to assign an ample buffer pool to one of the temporary table spaces, and leave any others with a smaller buffer pool. Such a buffer pool assignment will help ensure efficient utilization of main memory. For example, if your catalog table space uses 4 KB pages, and the remaining table spaces use 8 KB pages, the best temporary table space configuration may be a single 8 KB temporary table space with a large buffer pool, and a single 4 KB table space with a small buffer pool.

  Note: Catalog table spaces are restricted to using the 4 KB page size. As such, the database manager always enforces the existence of a 4 KB temporary table space to enable catalog table reorganizations.
- There is generally no advantage to defining more than one temporary table space of any single page size.
- SMS is almost always a better choice than DMS for temporary table spaces because:
  - Disk space is allocated on demand in SMS, whereas it must be pre-allocated in DMS. Pre-allocation can be difficult: Temporary table spaces hold transient data that can have a very large peak storage requirement, and a much smaller average storage requirement. With DMS, the peak storage requirement must be pre-allocated, whereas with SMS, the extra disk space can be used for other purposes during off-peak hours.
  - The database manager attempts to keep temporary table pages in memory, rather than writing them out to disk. As a result, the performance advantages of DMS are less significant.
  - SMS containers can take advantage of file system buffering; DMS containers cannot.

Recommendations for Catalog Table Spaces

An SMS table space is recommended for database catalogs, for the following reasons:

- The database catalog consists of many tables of varying sizes. When using a DMS table space, a minimum of two extents are allocated for each table object. Depending on the extent size chosen, a significant amount of allocated and unused space may result. When using a DMS table space, a small extent size (two to four pages) should be chosen; otherwise, an SMS table space should be used.
- There are large object (LOB) columns in the catalog tables. LOB data is not kept in the buffer pool with other data, but is read from disk each time it is needed. Reading LOBs from disk reduces performance. Since a file system usually has its own place for storing (or caching) data, using an SMS table

space, or a DMS table space built on file containers, makes avoidance of I/O possible if the LOB has previously been referenced.

Given these considerations, an SMS table space is a somewhat better choice for the catalogs.

Another factor to consider is whether you will need to enlarge the catalog table space in the future. While some platforms have support for enlarging the underlying storage for SMS containers, and while you can use redirected restore to enlarge an SMS table space, the use of a DMS table space facilitates the addition of new containers.

Workload Considerations

The primary type of workload being managed by DB2 in your environment can affect your choice of what table space type to use, and what page size to specify. An online transaction processing (OLTP) workload is characterized by transactions that need random access to data and that usually return small sets of data. Given that the access is random, and involves one or a few pages, prefetching is not possible.

DMS table spaces using device containers perform best in this situation. DMS table spaces with file containers, or SMS table spaces, are also reasonable choices for OLTP workloads if maximum performance is not required. With little or no sequential I/O expected, the settings for the EXTENTSIZE and the PREFETCHSIZE parameters on the CREATE TABLESPACE statement are not important for I/O efficiency.

A query workload is characterized by transactions that need sequential or partially sequential access to data, and that usually return large sets of data. A DMS table space using multiple device containers (where each container is on a separate disk) offers the greatest potential for efficient parallel prefetching. The value of the PREFETCHSIZE parameter on the CREATE TABLESPACE statement should be set to the value of the EXTENTSIZE parameter, multiplied by the number of device containers. This allows DB2 to prefetch from all containers in parallel.

A reasonable alternative for a query workload is to use files, if the file system has its own prefetching. The files can be either of DMS type using file containers, or of SMS type. Note that if you use SMS, you need to have the directory containers map to separate physical disks to achieve I/O parallelism.

Your goal for a mixed workload is to make single I/O requests as efficient as possible for OLTP workloads, and to maximize the efficiency of parallel I/O for query workloads.

The considerations for determining the page size for a table space are as follows:

- For OLTP applications that perform random row read and write operations, a smaller page size is usually preferable, because it wastes less buffer pool space with unwanted rows.
- For DSS applications that access large numbers of consecutive rows at a time, a larger page size is usually better, because it reduces the number of I/O requests that are required to read a specific number of rows. There is, however, an exception to this. If your row size is smaller than:

  ```
  pagesize / 255
  ```

  there will be wasted space on each page (there is a maximum of 255 rows per page). In this situation, a smaller page size may be more appropriate.
- Larger page sizes may allow you to reduce the number of levels in the index.
- Larger pages support rows of greater length.
- On default 4 KB pages, tables are restricted to 500 columns, while the larger page sizes (8 KB, 16 KB, and 32 KB) support 1012 columns.
- The maximum size of the table space is proportional to the page size of the table space. The limits are documented in the *SQL Reference*.

Choosing an SMS or DMS Table Space

There are a number of trade-offs to consider when determining which type of table space you should use to store your data.

Advantages of an SMS Table Space:

- Space is not allocated by the system until it is required.
- Creating a database requires less initial work, because you do not have to predefine the containers.

Advantages of a DMS Table Space:

- The size of a table space can be increased by adding containers, using the ALTER TABLESPACE statement. Existing data is automatically rebalanced across the new set of containers to retain optimal I/O efficiency.
- A table can be split across multiple table spaces, based on the type of data being stored:
  - Long field and LOB data
  - Indexes
  - Regular table data

  You might want to separate your table data for performance reasons, or to increase the amount of data stored for a table. For example, you could have a table with 64 GB of regular table data, 64 GB of index data and 2 TB of long data. If you are using 8 KB pages, the table data and the index data can be as much as 128 GB. If you are using 16 KB pages, it can be as much as 256 GB. If you are using 32 KB pages, the table data and the index data can be as much as 512 GB.

- The location of the data on the disk can be controlled, if this is allowed by the operating system.
- If all table data is in a single table space, a table space can be dropped and redefined with less overhead than dropping and redefining a table.
- In general, a well-tuned set of DMS table spaces will outperform SMS table spaces.

Note: On Solaris and PTX (IBM NUMA-Q), DMS table spaces with raw devices is strongly recommend for performance-critical workloads.

In general, small personal databases are easiest to manage with SMS table spaces. On the other hand, for large, growing databases you will probably only want to use SMS table spaces for the temporary table spaces, and separate DMS table spaces, with multiple containers, for each table. In addition, you will probably want to store long field data and indexes on their own table spaces.

If you choose to use DMS table spaces with device containers, you must be willing to tune and administer your environment. For more information, refer to "Performance Considerations for DMS Devices" in the *Administration Guide: Performance*.

Optimizing Performance When Data is Placed on RAID Devices

This section describes how to optimize performance when data is placed on Redundant Array of Independent Disks (RAID) devices. In general, you should do the following for each table space that uses a RAID device:

- Define a single container for the table space (using the RAID device).
- Make the EXTENTSIZE of the table space equal to, or a multiple of, the RAID stripe size.
- Ensure that the PREFETCHSIZE of the table space is:
  – the RAID stripe size multiplied by the number of RAID parallel devices (or a whole multiple of this product), and
  – a multiple of the EXTENTSIZE.
- Use the DB2_PARALLEL_IO registry variable (described below) to enable parallel I/O for the table space.
- Use the DB2_STRIPED_CONTAINERS registry variable (described below) to ensure extent boundaries are aligned in the table space.

DB2_PARALLEL_IO

When reading data from, or writing data to table space containers, DB2 may use parallel I/O if the number of containers in the database is greater than 1. However, there are situations when it would be beneficial to have parallel I/O enabled for single container table spaces. For example, if the container is created on a single RAID device that is composed of more than one physical disk, you may want to issue parallel read and write calls.

To force parallel I/O for a table space that has a single container, you can use the DB2_PARALLEL_IO registry variable. This variable can be set to ″*″ (asterisk), meaning every table space, or it can be set to a list of table space IDs separated by commas. For example:

```
db2set DB2_PARALLEL_IO=*        {turn parallel I/O on for all table spaces}
db2set DB2_PARALLEL_IO=1,2,4,8  {turn parallel I/O on for table spaces 1, 2,
                                 4, and 8}
```

After setting the registry variable, DB2 must be stopped (**db2stop**), and then restarted (**db2start**), for the changes to take effect.

DB2_STRIPED_CONTAINERS

Currently, when creating a DMS table space container (device or file), a one-page tag is stored at the beginning of the container. The remaining pages are available for data storage by DB2, and are grouped into extent-sized blocks.

When using RAID devices for table space containers, it is suggested that the table space be created with an extent size that is equal to, or a multiple of, the RAID stripe size. However, because of the one-page container tag, the extents will not line up with the RAID stripes, and it may be necessary during an I/O request to access more physical disks than would be optimal.

DMS table space containers can now be created in such a way that the tag exists in its own (full) extent. This avoids the problem described above, but it requires an extra extent of overhead within the container. To create containers in this fashion, you must set the DB2 registry variable DB2_STRIPED_CONTAINERS to ″ON″, and then stop and restart your instance:

```
db2set DB2_STRIPED_CONTAINERS=ON
db2stop
db2start
```

Any DMS container that is created (with the CREATE TABLESPACE or the ALTER TABLESPACE statement) will have tags taking up a full extent. Existing containers will remain unchanged.

To stop creating containers with this attribute, reset the variable, and then stop and restart your instance:

```
db2set DB2_STRIPED_CONTAINERS=
db2stop
db2start
```

The Control Center and the LIST TABLESPACE CONTAINERS command do not show whether a container has been created as a striped container. They use the label ″file″ or ″device″, depending on how the container was created. To verify that a container was created as a striped container, you can use the /DTSF option of DB2DART to dump table space and container information, and then look at the type field for the container in question. The query container APIs (**sqlbftcq** and **sqlbtcq**), can be used to create a simple application that will display the type.

Federated Database Design Considerations

When designing a federated database, consider the following design topics:

- Space requirements
- Network prioritization.

Typically, the data accessible from a federated database is not stored at that database. References to data source tables and views are stored within the system catalog, but the actual data is located at the data source. As such, a federated database might need less storage space than a conventional database. This general rule might not apply if your queries, due to collating system differences or lack of function at a data source, must be executed locally. In this case, tables are materialized at DB2 for processing.

Because the majority of federated system data is typically located at one or more data sources located across a network, consider changing the resources assigned to DB2 and your network system. You might see performance increases after allocating more resources to the network at the DB2 system, than to the database manager itself.

Chapter 9. Designing Distributed Databases

A transaction is commonly referred to in DB2 as a *unit of work*. A unit of work is a recoverable sequence of operations within an application process. It is used by the database manager to ensure that a database is in a consistent state. Any reading from or writing to the database is done within a unit of work.

For example, a bank transaction might involve the transfer of funds from a savings account to a checking account. After the application subtracts an amount from the savings account, the two accounts are inconsistent, and remain so until the amount is added to the checking account. When *both* steps are completed, a point of consistency is reached. The changes can be committed and made available to other applications.

A unit of work starts when the first SQL statement is issued against the database. The application must end the unit of work by issuing either a COMMIT or a ROLLBACK statement. The COMMIT statement makes permanent all changes made within a unit of work. The ROLLBACK statement removes these changes from the database. If the application ends normally without either of these statements being explicitly issued, the unit of work is automatically committed. If it ends abnormally in the middle of a unit of work, the unit of work is automatically rolled back. Once issued, a COMMIT or a ROLLBACK cannot be stopped. With some multi-threaded applications, or some operating systems (such as Windows), if the application ends normally without either of these statements being explicitly issued, the unit of work is automatically rolled back. It is recommended that your applications always explicitly commit or roll back complete units of work. If part of a unit of work does not complete successfully, the updates are rolled back, leaving the participating tables as they were before the transaction began. This ensures that requests are neither lost nor duplicated.

The following topics provide additional information:

- “Using a Single Database in a Transaction” on page 156
- “Using Multiple Databases in a Single Transaction” on page 157
- “Other Configuration Considerations” on page 162
- “Understanding the Two-Phase Commit Process” on page 165
- “Recovering from Problems During Two-Phase Commit” on page 168.

For information about creating applications that use distributed databases, refer to the *Application Development Guide* and the *CLI Guide and Reference*.

Using a Single Database in a Transaction

The simplest form of transaction is to read from and write to only one database within a single unit of work. This type of database access is called a *remote unit of work*.

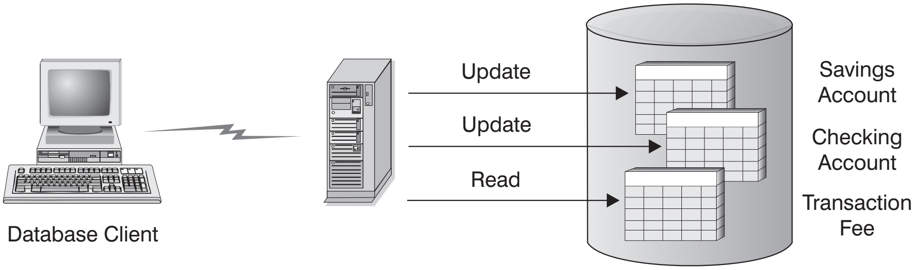


Figure 38. Using a Single Database in a Transaction

Figure 38 shows a database client running a funds transfer application that accesses a database containing checking and savings account tables, as well as a banking fee schedule. The application must:

- Accept the amount to transfer from the user interface
- Subtract the amount from the savings account, and determine the new balance
- Read the fee schedule to determine the transaction fee for a savings account with the given balance
- Subtract the transaction fee from the savings account
- Add the amount of the transfer to the checking account
- Commit the transaction (unit of work).

To set up such an application, you must:

1. Create the tables for the savings account, checking account and banking fee schedule in the same database (see ″Implementing Your Design″ in the *Administration Guide: Implementation*)
2. If physically remote, set up the database server to use the appropriate communications protocol, as described in the *Installation and Configuration Supplement*
3. If physically remote, catalog the node and the database to identify the database on the database server, as described in the *Quick Beginnings* books
4. Precompile your application program to specify a type 1 connection; that is, specify CONNECT 1 (the default) on the PRECOMPILE PROGRAM command, as described in the *Application Development Guide*.

Using Multiple Databases in a Single Transaction

When using multiple databases in a single transaction, the requirements for setting up and administering your environment are different, depending on the number of databases that are being updated in the transaction.

Updating a Single Database

If your data is distributed across multiple databases, you may wish to update one database while reading from one or more other databases. This type of access, which is called *multisite update*, or *two-phase commit*, can be performed within a single unit of work (transaction). See "Updating Multiple Databases" on page 158 for another example of a multisite update.

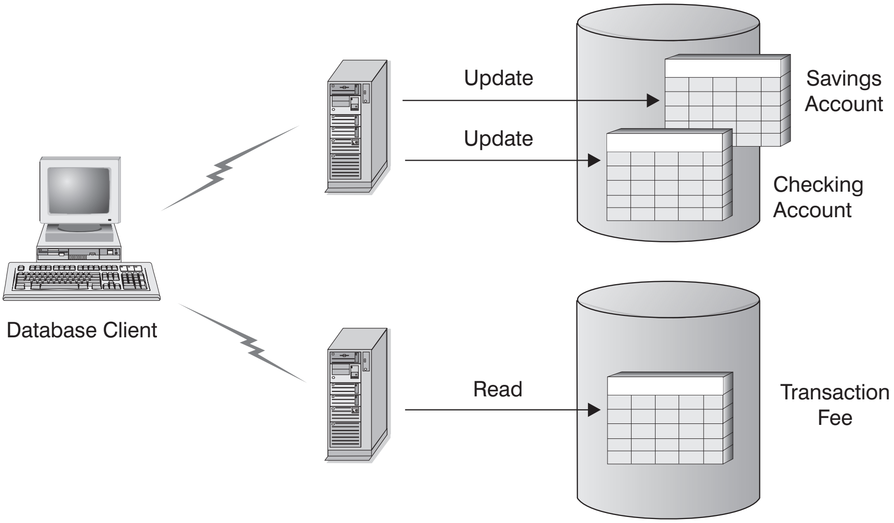


Figure 39. Using Multiple Databases in a Single Transaction

Figure 39 shows a database client running a funds transfer application that accesses two database servers: one containing the checking and savings accounts, and another containing the banking fee schedule. This example is similar to the one shown in Figure 38 on page 156, except for the number of databases, and the location of the tables.

To set up a funds transfer application for this environment, you must:

1. Create the necessary tables in the appropriate databases (see "Implementing Your Design" in the *Administration Guide: Implementation*)
2. If physically remote, set up the database servers to use the appropriate communications protocols, as described in the *Installation and Configuration Supplement*

3. If physically remote, catalog the nodes and the databases to identify the databases on the database servers, as described in the *Quick Beginnings* books
4. Precompile your application program to specify a type 2 connection (that is, specify CONNECT 2 on the PRECOMPILE PROGRAM command), and one-phase commit (that is, specify SYNCPOINT ONEPHASE on the PRECOMPILE PROGRAM command), as described in the *Application Development Guide*.

If databases are located on a host or AS/400 database server, you require DB2 Connect for connectivity to these servers. For information about setup, refer to one of the DB2 Connect *Quick Beginnings* books. For information about using DB2 Connect, refer to the *DB2 Connect User's Guide*.

Updating Multiple Databases

If your data is distributed across multiple databases, you may want to read and update several databases in a single transaction. This type of database access is called a *multisite update*.

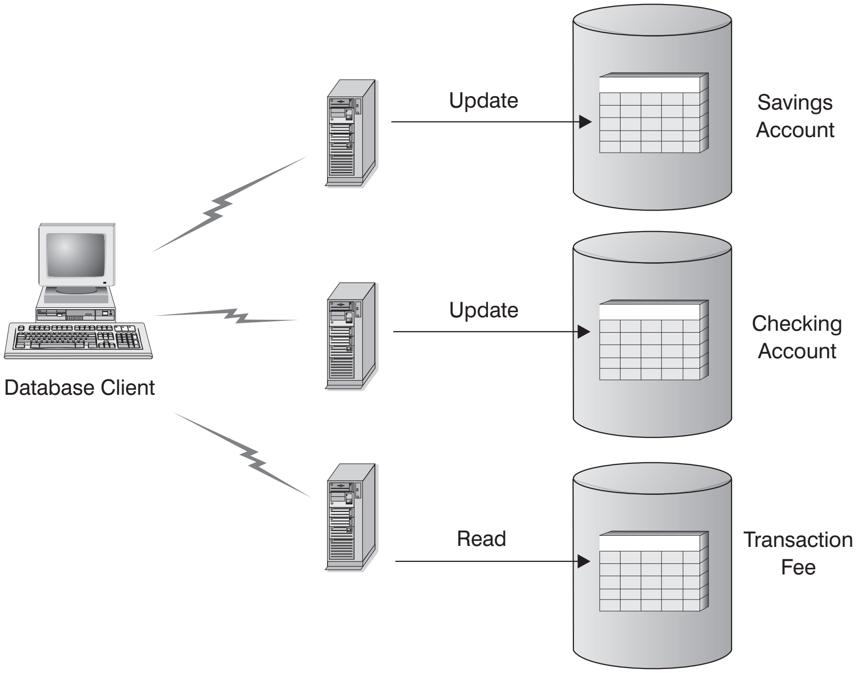


Figure 40. Updating Multiple Databases in a Single Transaction

Figure 40 on page 158 shows a database client running a funds transfer application that accesses three database servers: one containing the checking account, another containing the savings account, and the third containing the banking fee schedule.

To set up a funds transfer application for this environment, you must:

1. Create the necessary tables in the appropriate databases (see "Implementing Your Design" in the *Administration Guide: Implementation*)
2. If physically remote, set up the database servers to use the appropriate communications protocols, as described in the *Installation and Configuration Supplement*
3. If physically remote, catalog the nodes and the databases to identify the databases on the database servers, as described in the *Quick Beginnings* books
4. Precompile your application program to specify a type 2 connection (that is, specify CONNECT 2 on the PRECOMPILE PROGRAM command), and one-phase commit (that is, specify SYNCPOINT ONEPHASE on the PRECOMPILE PROGRAM command), as described in the *Application Development Guide*.
5. Configure the DB2 transaction manager (TM), as described in "Using the DB2 Transaction Manager".

Using the DB2 Transaction Manager

The database manager provides transaction manager functions that can be used to coordinate the updating of several databases within a single unit of work. The database client automatically coordinates the unit of work, and uses a *transaction manager database* to register each transaction and track its completion status.

If you are using an XA-compliant transaction manager, such as IBM TXSeries, BEA Tuxedo, or Microsoft Transaction Server, see "Chapter 10. Designing for Transaction Managers" on page 171 for integration instructions.

When using DB2 UDB for UNIX based systems, Windows operating systems, or OS/2 to coordinate your transactions, you must fulfill certain configuration requirements. If you use TCP/IP exclusively for communications, and DB2 UDB and DB2 for OS/390 are the only database servers involved in your transactions, configuration is straightforward.

DB2 UDB and DB2 for OS/390 Using TCP/IP Connectivity: If each of the following statements is true for your environment, the configuration steps for multisite update are straightforward.

- All communications with remote database servers (including DB2 UDB for OS/390) use TCP/IP exclusively.

- DB2 UDB for UNIX based systems, Windows operating systems, OS/2, or OS/390 are the only database servers involved in the transaction.
- The DB2 Connect sync point manager (SPM) is not configured.

  The DB2 Connect sync point manager is configured automatically at DB2 instance creation time, and is required when:

  - SNA connectivity is used with host or AS/400 database servers for multisite updates.
  - An XA-compliant transaction manager (such as IBM TXSeries CICS) is coordinating the two-phase commit.

    This applies to both SNA and TCP/IP connectivity with host or AS/400 database servers. For detailed information, see "Chapter 10. Designing for Transaction Managers" on page 171. If your environment does not require the DB2 Connect sync point manager, you can turn it off by issuing the command `db2 update dbm cfg using spm_name NULL` at the DB2 Connect server. Then stop and restart DB2.

The database that will be used as the transaction manager database is determined at the database client by the database manager configuration parameter *tm_database*. For more information about this configuration parameter, see "Configuring DB2" in the *Administration Guide: Performance*. Consider the following factors when setting this configuration parameter:

- The transaction manager database can be:
  - A DB2 UDB for UNIX based systems, Windows operating systems, or OS/2 database
  - A DB2 for OS/390 Version 5 or later database.

    This is the recommended database server to use as the transaction manager database. OS/390 systems are, generally, more secure than workstation servers, reducing the possibility of accidental power downs, reboots, and so on. Therefore the recovery logs, used in the event of resynchronization, are more secure.
- If a value of `1ST_CONN` is specified for the *tm_database* configuration parameter, the first database to which an application connects is used as the transaction manager database.

  Care must be taken when using `1ST_CONN`. You should only use this configuration if it is easy to ensure that all involved databases are cataloged correctly; that is, if:

  - The database client initiating the transaction is in the same instance that contains the participating databases, including the transaction manager database.
  - You are using DCE directory services to catalog and manage access to your databases.

Note that if your application attempts to disconnect from the database being used as the transaction manager database, you will receive a warning message, and the connection will be held until the unit of work is committed.

Other Environments: If, in your environment:

- TCP/IP is not used exclusively for communications with remote database servers (for example, NETBIOS is used)
- DB2 for MVS Version 3 or Version 4, DB2 for AS/400, or DB2 for VM&VSE is accessed
- DB2 for OS/390 is accessed using SNA
- The DB2 Connect sync point manager is used to access host or AS/400 database servers

the configuration steps for multisite update are more involved.

The database that will be used as the transaction manager database is determined at the database client by the database manager configuration parameter *tm_database*. For more information about this configuration parameter, see "Configuring DB2" in the *Administration Guide: Performance*. Consider the following factors when setting this configuration parameter:

- The transaction manager database can be a DB2 UDB for UNIX based systems, Windows operating systems, or OS/2 database.
- If a value of `1ST_CONN` is specified for the *tm_database* configuration parameter, the first database to which an application connects is used as the transaction manager database.

  Care must be taken when using `1ST_CONN`. You should only use this configuration if it is easy to ensure that all involved databases are cataloged correctly; that is, if:

  - The database client initiating the transaction is in the same instance that contains the participating databases, including the transaction manager database.
  - You are using DCE directory services to catalog and manage access to your databases.

  Note that if your application attempts to disconnect from the database being used as the transaction manager database, you will receive a warning message, and the connection will be held until the unit of work is committed.

Other Configuration Considerations

You should consider the following configuration parameters when you are setting up your environment. For additional information about setting these parameters, refer to the *DB2 Connect User's Guide*.

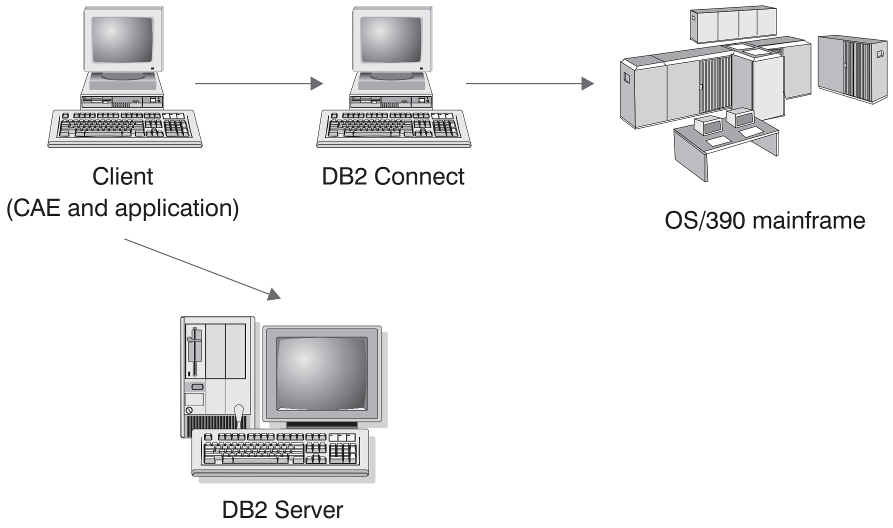


Figure 41. Configuration Considerations

Database Manager Configuration Parameters

- *tm_database*

  This parameter identifies the name of the Transaction Manager (TM) database for each DB2 instance.
- *spm_name*

  This parameter identifies the name of the DB2 Connect sync point manager instance to the database manager. For resynchronization to be successful, the name must be unique across your network.
- *resync_interval*

  This parameter identifies the time interval (in seconds) after which the DB2 Transaction Manager, the DB2 server database manager, and the DB2 Connect sync point manager or the DB2 UDB sync point manager should retry the recovery of any outstanding indoubt transactions.
- *spm_log_file_sz*

  This parameter specifies the size (in 4 KB pages) of the SPM log file.
- *spm_max_resync*

  This parameter identifies the number of agents that can simultaneously perform resynchronization operations.

- *spm_log_path*

  This parameter identifies the log path for the SPM log files.

Database Configuration Parameters

- *maxappls*

  This parameter specifies the maximum permitted number of active applications. Its value must be equal to or greater than the sum of the connected applications, plus the number of these applications that may be concurrently in the process of completing a two-phase commit or rollback, plus the anticipated number of indoubt transactions that might exist at any one time. For more information about indoubt transactions, see "Recovering from Problems During Two-Phase Commit" on page 168.
- *autorestart*

  This database configuration parameter specifies whether the RESTART DATABASE routine will be invoked automatically when needed. The default value is YES (that is, enabled).

  A database containing indoubt transactions requires a restart database operation to start up. If *autorestart* is not enabled when the last connection to the database is dropped, the next connection will fail and require an explicit RESTART DATABASE invocation. This condition will exist until the indoubt transactions have been removed, either by the transaction manager's resynchronization operation, or through a heuristic operation initiated by the administrator. When the RESTART DATABASE command is issued, a message is returned if there are any indoubt transactions in the database. The administrator can then use the LIST INDOUBT TRANSACTIONS command and other command line processor commands to find get information about those indoubt transactions.

For more information about these configuration parameters, refer to the *Administration Guide: Performance*.

Host or AS/400 Applications Accessing a LAN Based DB2 Universal Database Server in a Multisite Update

DB2 Universal Database does not support multisite update from the host or AS/400 database clients using TCP/IP connectivity. In this situation, only SNA (Systems Network Architecture) connectivity is supported. The DB2 sync point manager is required for multisite update. DB2 Connect is not used in this scenario.

The database server that is being accessed from the host or the AS/400 database client does not have to be local to the workstation with the DB2 sync point manager. The host or AS/400 database client could connect to a DB2 UDB server using the DB2 sync point manager workstation as an interim gateway. This allows you to isolate the DB2 sync point manager workstation in a secure environment, while the actual DB2 UDB servers are remote in your organization. This also permits a DB2 common server Version 2 database to be involved in multisite updates originating from the host or AS/400 database clients.

The steps are as follows:

- On the workstation that will be directly accessed by the host or AS/400 application:
  1. Install DB2 Universal Database Enterprise Edition, or Enterprise - Extended Edition to provide multisite update support with the host or AS/400 database clients.
  2. Create a database instance on the same system. For example, you can use the default instance, `DB2`, or use the following command to create a new instance:

     ```
     db2icrt myinstance
     ```
  3. Supply licensing information, as required.
  4. Ensure that the registry value DB2COMM includes the value `APPC`.
  5. Configure SNA communications, as required. When the supported IBM SNA products are used, the SNA profiles required for the DB2 sync point manager are created automatically, based on the value of the *spm_name* database manager configuration parameter. Any other supported SNA stack will require manual configuration. For details, refer to the *Installation and Configuration Supplement*.
  6. Determine the value to be specified for the *spm_name* database manager configuration parameter. This parameter is pre-configured at DB2 instance creation time with a derivative of the TCP/IP host name for the machine. If this is acceptable and unique within your environment, do not change it.
  7. If necessary, update *spm_name* on the DB2 Universal Database server, using the UPDATE DATABASE MANAGER CONFIGURATION command.
  8. Configure communications required for this DB2 workstation to connect to remote DB2 UDB servers, if any.
  9. Configure communications required for remote DB2 UDB servers to connect to this DB2 server.
  10. Stop and restart the database manager on the DB2 Universal Database server to start the SPM for the first time.

      You should be able to connect to the remote DB2 UDB servers from this DB2 UDB workstation.
- On each remote DB2 UDB server that will be accessed by the host or AS/400 database client:

1. Configure communications required for the remote DB2 UDB workstation with the DB2 sync point manager to connect to this DB2 UDB server.
2. Stop and restart the database manager.

Understanding the Two-Phase Commit Process

Figure 42 on page 166 illustrates the steps involved in a multisite update. Understanding how a transaction is managed will help you to resolve the problem if an error occurs during the two-phase commit process.

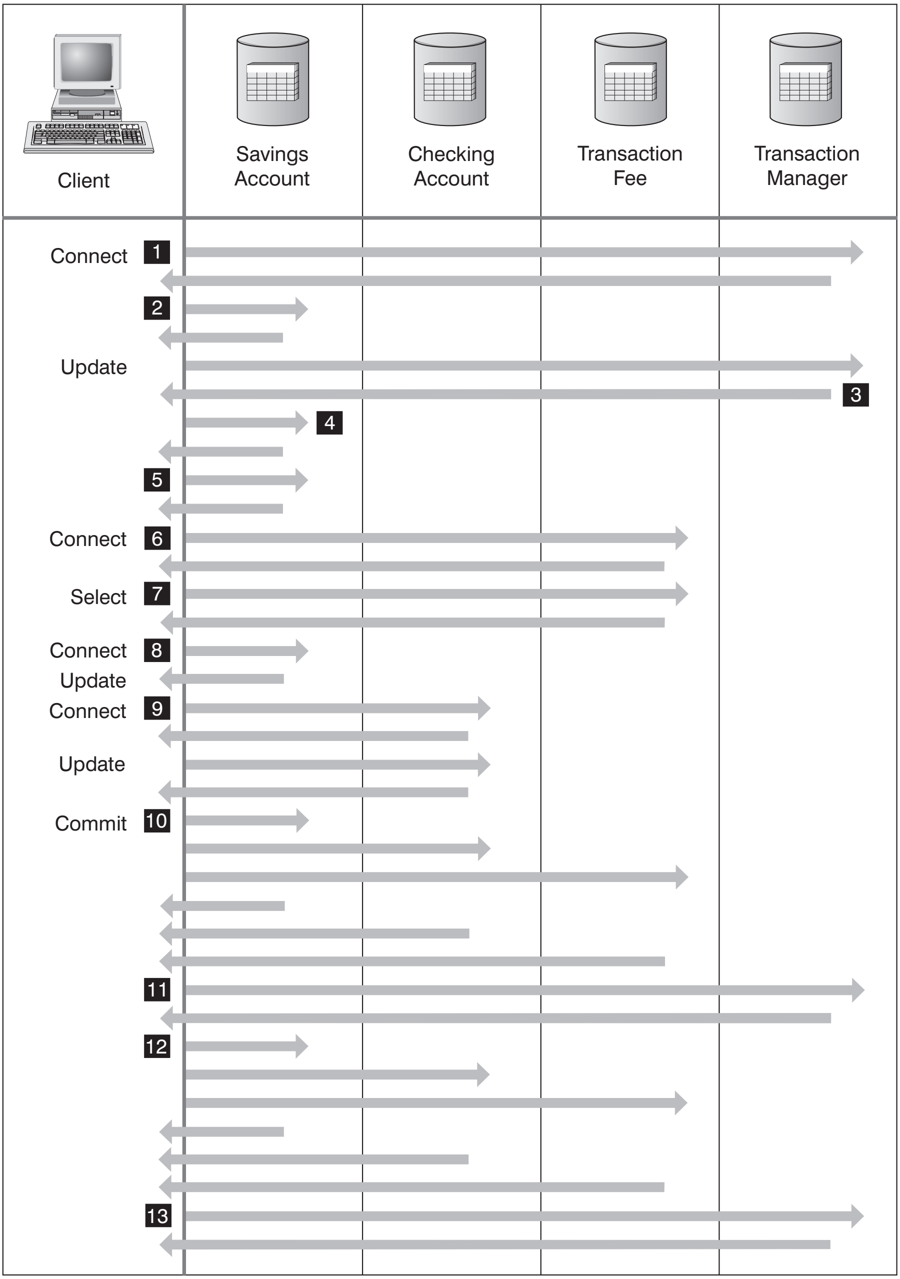


Figure 42. Updating Multiple Databases

0 The application is prepared for two-phase commit. This can be accomplished through precompilation options (refer to the *Application*

Development Guide for details). This can also be accomplished through DB2 CLI (Call Level Interface) configuration (refer to the *CLI Guide and Reference* for details).

1 When the database client wants to connect to the SAVINGS_DB database, it first internally connects to the transaction manager (TM) database. The TM database returns an acknowledgment to the database client. If the database manager configuration parameter *tm_database* is set to `1ST_CONN`, SAVINGS_DB becomes the transaction manager database for the duration of this application instance.

2 The connection to the SAVINGS_DB database takes place and is acknowledged.

3 The database client begins the update to the SAVINGS_ACCOUNT table. This begins the unit of work. The TM database responds to the database client, providing a transaction ID for the unit of work. Note that the registration of a unit of work occurs when the first SQL statement in the unit of work is run, not during the establishment of a connection.

4 After receiving the transaction ID, the database client registers the unit of work with the database containing the SAVINGS_ACCOUNT table. A response is sent back to the client to indicate that the unit of work has been registered successfully.

5 SQL statements issued against the SAVINGS_DB database are handled in the normal manner. The response to each statement is returned in the SQLCA when working with SQL statements embedded in a program. (The SQLCA is described in the *Application Development Guide* and in the *SQL Reference*.)

6 The transaction ID is registered at the FEE_DB database containing the TRANSACTION_FEE table, during the first access to that database within the unit of work.

7 Any SQL statements against the FEE_DB database are handled in the normal way.

8 Additional SQL statements can be run against the SAVINGS_DB database by setting the connection, as appropriate. Since the unit of work has already been registered with the SAVINGS_DB database **4**, the database client does not need to perform the registration step again.

9 Connecting to, and using the CHECKING_DB database follows the same rules described in **6** and **7**.

10 When the database client requests that the unit of work be committed, a *prepare* message is sent to all databases participating in the unit of work. Each database writes a ″PREPARED″ record to its log files, and replies to the database client.

11 After the database client receives a positive response from all of the databases, it sends a message to the transaction manager database, informing it that the unit of work is now ready to be committed (PREPARED). The transaction manager database writes a ″PREPARED″ record to its log file, and sends a reply to inform the client that the second phase of the commit process can be started.

12 During the second phase of the commit process, the database client sends a message to all participating databases to tell them to commit. Each database writes a ″COMMITTED″ record to its log file, and releases the locks that were held for this unit of work. When the database has completed committing the changes, it sends a reply to the client.

13 After the database client receives a positive response from all participating databases, it sends a message to the transaction manager database, informing it that the unit of work has been completed. The transaction manager database then writes a ″COMMITTED″ record to its log file, indicating that the unit of work is complete, and replies to the client, indicating that it has finished.

Recovering from Problems During Two-Phase Commit

Recovering from error conditions is a normal task associated with application programming, system administration, database administration and system operation. Distributing databases over several remote servers increases the potential for error resulting from network or communications failures. To ensure data integrity, the database manager provides the two-phase commit process, which is illustrated in "Understanding the Two-Phase Commit Process" on page 165 (points **10** , **11** , and **12** . The following explains how the database manager handles errors during the two-phase commit process:

- **First Phase Error**

  If a database communicates that it has failed to prepare to commit the unit of work, the database client will roll back the unit of work during the second phase of the commit process. A prepare message will *not* be sent to the transaction manager database in this case.

  During the second phase, the client sends a rollback message to all participating databases that successfully prepared to commit during the first phase. Each database then writes an ″ABORT″ record to its log file, and releases the locks that were held for this unit of work.

- **Second Phase Error**

  Error handling at this stage is dependent upon whether the second phase will commit or roll back the transaction. The second phase will only roll back the transaction if the first phase encountered an error.

  If one of the participating databases fails to commit the unit of work (possibly due to a communications failure), the transaction manager database will retry the commit on the failed database. The application, however, will be informed that the commit was successful through the SQLCA. DB2 will ensure that the uncommitted transaction in the database server is committed. The database manager configuration parameter *resync_interval* (see ″Configuring DB2″ in the *Administration Guide: Performance*) is used to specify how long the transaction manager database should wait between attempts to commit the unit of work. All locks are held at the database server until the unit of work is committed.

  If the transaction manager database fails, it will resynchronize the unit of work when it is restarted. The resynchronization process will attempt to complete all *indoubt transactions*; that is, those transactions that have finished the first phase, but have not completed the second phase of the commit process. The database manager associated with the transaction manager database performs the resynchronization by:

  1. Connecting to the databases that indicated they were ″PREPARED″ to commit during the first phase of the commit process.
  2. Attempting to commit the indoubt transactions at those databases. (If the indoubt transactions cannot be found, the database manager assumes that the database successfully committed the transactions during the second phase of the commit process.)
  3. Committing the indoubt transactions in the transaction manager database, after all indoubt transactions have been committed in the participating databases.

  If one of the participating databases fails and is restarted, the database manager for this database will query the transaction manager database for the status of this transaction, to determine whether the transaction should be rolled back. If the transaction is not found in the log, the database manager assumes that the transaction was rolled back, and will roll back the indoubt transaction in this database. Otherwise, the database waits for a commit request from the transaction manager database.

  If the transaction was coordinated by a transaction processing monitor (XA-compliant transaction manager), the database will always depend on the TP monitor to initiate the resynchronization.

If, for some reason, you cannot wait for the transaction manager to automatically resolve indoubt transactions, there are actions you can take to manually resolve them. This manual process is sometimes referred to as

″making a heuristic decision″. For more information about manual recovery of indoubt transactions, see “Making a Heuristic Decision” on page 182.

Resynchronizing Indoubt Transactions if AUTORESTART=OFF

For configuration considerations in the DB2 Universal Database two-phase commit environment, see “Other Configuration Considerations” on page 162.

In particular, if the *autorestart* database configuration parameter is set to `OFF`, and there are indoubt transactions in either the TM or RM databases, the RESTART DATABASE command is required to start the resynchronization process. When issuing the RESTART DATABASE command from the command line processor, use different sessions. If you restart a different database from the same session, the connection established by the previous invocation will be dropped, and must be restarted once again. Issue the TERMINATE command to drop the connection after no more indoubt transactions are returned by the LIST INDOUBT TRANSACTIONS command.

Chapter 10. Designing for Transaction Managers

You may want to use your databases with an XA-compliant transaction manager if you have resources other than DB2 databases that you want to participate in a two-phase commit transaction. If your transactions only access DB2 databases, you should use the DB2 transaction manager, described in "Updating Multiple Databases" on page 158.

The following topics will assist you in using the database manager with an XA-compliant transaction manager, such as IBM TXSeries CICS, IBM TXSeries Encina, BEA Tuxedo, or Microsoft Transaction Server:

- "X/Open Distributed Transaction Processing Model" on page 172
- "Setting Up a Database as a Resource Manager" on page 176
- "xa_open and xa_close Strings Usage" on page 176
- "New xa_open String Format for DB2 Version 7" on page 176
- "TPM and TP_MON_NAME Values" on page 178
- "xa_open String Format for Earlier Versions of DB2" on page 181
- "Updating Host or AS/400 Database Servers" on page 181
- "Database Connection Considerations" on page 182
- "Making a Heuristic Decision" on page 182
- "Security Considerations" on page 185
- "Configuration Considerations" on page 186
- "XA Function Supported" on page 187
- "XA Interface Problem Determination" on page 189
- "Configuring XA Transaction Managers to Use DB2 UDB" on page 190
- "Configuring Microsoft Transaction Server" on page 195.

If you are using an XA-compliant transaction manager, or are implementing one, more information is available from our technical support web site:

```
http://www.ibm.com/software/data/db2/library/
```

Once there, choose "DB2 Universal Database", then search the web site using the keyword "XA" for the latest available information on XA-compliant transaction managers.

X/Open Distributed Transaction Processing Model

The X/Open Distributed Transaction Processing (DTP) model includes three interrelated components:

- "Application Program (AP)"
- "Transaction Manager (TM)" on page 174
- "Resource Managers (RM)" on page 175.

Figure 43 illustrates this model, and shows the relationship among these components.

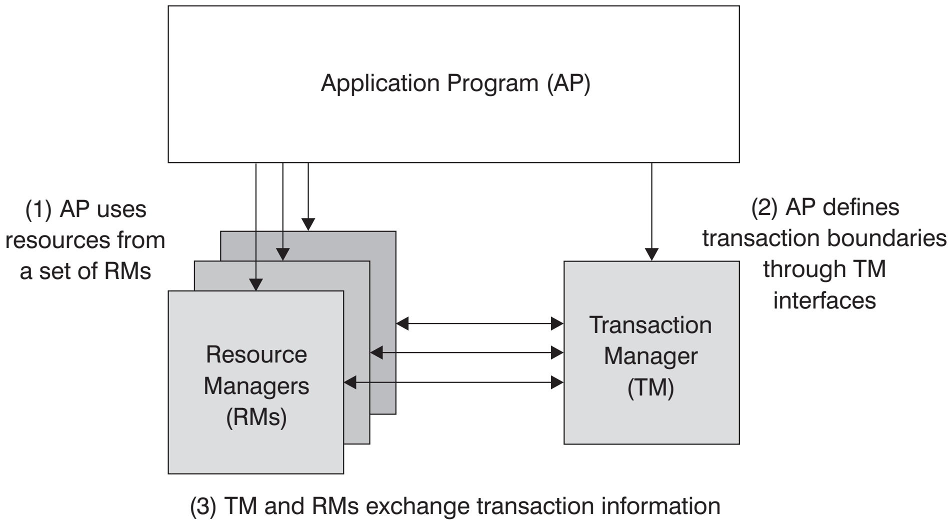

Application Program (AP)
(1) AP uses resources from a set of RMs
(2) AP defines transaction boundaries through TM interfaces
Resource Managers (RMs)
Transaction Manager (TM)
(3) TM and RMs exchange transaction information

Figure 43. X/Open Distributed Transaction Processing (DTP) Model

Application Program (AP)

The application program (AP) defines transaction boundaries, and defines the application-specific actions that make up the transaction.

For example, a CICS* application program might want to access resource managers (RMs), such as a database and a CICS Transient Data Queue, and use programming logic to manipulate the data. Each access request is passed to the appropriate resource managers through function calls specific to that RM. In the case of DB2, these could be function calls generated by the DB2 precompiler for each SQL statement, or database calls coded directly by the programmer using the APIs.

A transaction manager (TM) product usually includes a transaction processing (TP) monitor to run the user application. The TP monitor provides APIs to allow an application to start and end a transaction, and to perform application scheduling and load balancing among the many users who want to run the application. The application program in a distributed transaction processing (DTP) environment is really a combination of the user application and the TP monitor.

To facilitate an efficient online transaction processing (OLTP) environment, the TP monitor pre-allocates a number of server processes at startup, and then schedules and reuses them among the many user transactions. This conserves system resources, by allowing more concurrent users to be supported with a smaller number of server processes and their corresponding RM processes. Reusing these processes also avoids the overhead of starting up a process in the TM and RMs for each user transaction or program. (A program invokes one or more transactions.) This also means that the server processes are the actual ″user processes″ to the TM and the RMs. This has implications for security administration and application programming. For details, see "Security Considerations" on page 185.

The following types of transactions are possible from a TP monitor:

- Non-XA transactions

  These transactions involve RMs that are not defined to the TM, and are therefore not coordinated under the two-phase commit protocol of the TM. This might be necessary if the application needs to access an RM that does not support the XA interface. The TP monitor simply provides efficient scheduling of applications and load balancing. Since the TM does not explicitly ″open″ the RM for XA processing, the RM treats this application as any other application that runs in a non-DTP environment.
- Global transactions

  These transactions involve RMs that are defined to the TM, and are under the TM's two-phase commit control. A global transaction is a unit of work that could involve one or more RMs. A *transaction branch* is the part of work between a TM and an RM that supports the global transaction. A global transaction could have multiple transaction branches when multiple RMs are accessed through one or more application processes that are coordinated by the TM.

  Loosely coupled global transactions exist when each of a number of application processes accesses the RMs as if they are in a separate global transaction, but those applications are under the coordination of the TM. Each application process will have its own transaction branch within an RM. When a commit or rollback is requested by any one of the APs, TM, or RMs, the transaction branches are completed altogether. It is the application's responsibility to ensure that resource deadlock does not occur among the branches. (Note that the transaction coordination performed by the DB2 transaction manager for applications prepared with the

SYNCPOINT(TWOPHASE) option is roughly equivalent to these loosely coupled global transactions. See “Updating Multiple Databases” on page 158.)

Tightly coupled global transactions exist when multiple application processes take turns to do work under the same transaction branch in an RM. To the RM, the two application processes are a single entity. The RM must ensure that resource deadlock does not occur within the transaction branch.

Transaction Manager (TM)

The transaction manager (TM) assigns identifiers to transactions, monitors their progress, and takes responsibility for transaction completion and failure. The transaction branch identifiers (known as XIDs) are assigned by the TM to identify both the global transaction, and the specific branch within an RM. This is the correlation token between the log in a TM and the log in an RM. The XID is needed for two-phase commit, or rollback, to perform the *resynchronization* operation (also known as a *resync*) on system startup, or to let the administrator perform a *heuristic* operation (also known as *manual intervention*), if necessary.

After a TP monitor is started, it asks the TM to open all the RMs that a set of application servers have defined. The TM passes **xa_open** calls to the RMs, so that they can be initialized for DTP processing. As part of this startup procedure, the TM performs a resync to recover all *indoubt transactions*. An indoubt transaction is a global transaction that was left in an uncertain state. This occurs when the TM (or at least one RM) becomes unavailable after successfully completing the first phase (that is, the prepare phase) of the two-phase commit protocol. The RM will not know whether to commit or roll back its branch of the transaction until the TM can reconcile its own log with the RM logs when they become available again. To perform the resync operation, the TM issues a **xa_recover** call one or more times to each of the RMs to identify all the indoubt transactions. The TM compares the replies with the information in its own log to determine whether it should inform the RMs to **xa_commit** or **xa_rollback** those transactions. If an RM has already committed or rolled back its branch of an indoubt transaction through a heuristic operation by its administrator, the TM issues an **xa_forget** call to that RM to complete the resync operation.

When a user application requests a commit or a rollback, it must use the API provided by the TP monitor or TM, so that the TM can coordinate the commit and rollback among all the RMs involved. For example, when a CICS application issues the CICS SYNCPOINT request to commit a transaction, the CICS XA TM (implemented in the Encina Server) will in turn issue XA calls, such as **xa_end**, **xa_prepare**, **xa_commit**, or **xa_rollback** to request the RM to commit or roll back the transaction. The TM could choose to use one-phase instead of two-phase commit if only one RM is involved, or if an RM replies that its branch is read-only.

Resource Managers (RM)

A resource manager (RM) provides access to shared resources, such as databases.

DB2, as resource manager of a database, can participate in a *global transaction* that is being coordinated by an XA-compliant TM. As required by the XA interface, the database manager provides a *db2xa_switch* external C variable of type xa_switch_t to return the XA switch structure to the TM. This data structure contains the addresses of the various XA routines to be invoked by the TM, and the operating characteristics of the RM. For more information about XA functions supported by the database manager, see "XA Function Supported" on page 187.

There are two methods by which the RM can register its participation in each global transaction: *static registration* and *dynamic registration*:

- Static registration requires the TM to issue (for every transaction) the **xa_start**, **xa_end**, and **xa_prepare** series of calls to all the RMs defined for the server application, regardless of whether a given RM is used by the transaction. This is inefficient if not every RM is involved in every transaction, and the degree of inefficiency is proportional to the number of defined RMs.
- Dynamic registration (used by DB2) is flexible and efficient. An RM registers with the TM using an **ax_reg** call only when the RM receives a request for its resource. Note that there is no performance disadvantage with this method, even when there is only one RM defined, or when every RM is used by every transaction, because the **ax_reg** and the **xa_start** calls have similar paths in the TM.

The XA interface provides two-way communication between a TM and an RM. It is a system-level interface between the two DTP software components, not an ordinary application program interface to which an application developer codes. However, application developers should be familiar with the programming restrictions that the DTP software components impose. For information about X/Open XA interface programming considerations, refer to the *Application Development Guide*.

Although the XA interface is invariant, each XA-compliant TM may have product-specific ways of integrating an RM. For information about integrating your DB2 product as a resource manager with a specific transaction manager, see the appropriate TM product documentation. Integration information regarding the most popular TP monitors is provided in "Configuring XA Transaction Managers to Use DB2 UDB" on page 190.

Setting Up a Database as a Resource Manager

Each database is defined as a separate resource manager (RM) to the transaction manager (TM), and the database must be identified with an xa_open string. For a description of DB2's xa_open string format, see "xa_open and xa_close Strings Usage".

xa_open and xa_close Strings Usage

The database manager xa_open string has two accepted formats. One format is new to DB2 Version 7. The second format is used by earlier versions of DB2, and remains for back-level compatibility. New implementations should use the new format, and older implementations should be migrated to the new format when possible. Future versions of DB2 may not support the older xa_open string format. For information about the original xa_open string format, see "xa_open String Format for Earlier Versions of DB2" on page 181.

When setting up a database as a resource manager, you do not need the xa_close string. If provided, this string will be ignored by the database manager.

New xa_open String Format for DB2 Version 7

The following xa_open string format is new to DB2 Version 7:

```
parm_id1 = <parm value>,parm_id2 = <parm value>, ...
```

It does not matter in what order these parameters are specified. Valid values for *parm_id* are described in the following table.

Table 22. Valid Values for parm_id

Parameter Name	Value	Mandatory?	Case Sensitive?	Default Value
DB	Database alias	Yes	No	None
Database alias used by the application to access the database.				
UID	User ID	No	Yes	None
User ID that has authority to connect to the database. Required if a password is specified.				
PWD	Password	No	Yes	None
A password that is associated with the user ID. Required if a user ID is specified.				
TPM	Transaction processing monitor name	No	No	None
Name of the TP monitor being used. For supported values, see "TPM and TP_MON_NAME Values" on page 178. This parameter can be specified to allow multiple TP monitors to use a single DB2 instance. The specified value will override the value specified in the *tp_mon_name* database manager configuration parameter.				

Table 22. Valid Values for parm_id (continued)

<table>
<tr><th>Parameter Name</th><th>Value</th><th>Mandatory?</th><th>Case Sensitive?</th><th>Default Value</th></tr>
<tr><td>AXLIB</td><td>Library that contains the TP monitor's <b>ax_reg</b> and <b>ax_unreg</b> functions.</td><td>No</td><td>Yes</td><td>None</td></tr>
<tr><td colspan="5">This value is used by DB2 to obtain the addresses of the required <b>ax_reg</b> and <b>ax_unreg</b> functions. It can be used to override assumed values based on the TPM parameter, or it can be used by TP monitors that do not appear on the list for TPM.</td></tr>
<tr><td>CHAIN_END</td><td>xa_end chaining flag. Valid values are T, F, or no value.</td><td>No</td><td>No</td><td>F</td></tr>
<tr><td colspan="5">XA_END chaining is an optimization that can be used by DB2 to reduce network flows. If the TP monitor environment is such that it can be guaranteed that <b>xa_prepare</b> will be invoked within the same thread or process immediately following the call to <b>xa_end</b>, and if CHAIN_END is on, the xa_end flag will be chained with the <b>xa_prepare</b> command, thus elimintaing one network flow. A value of T means that CHAIN_END is on; a value of F means that CHAIN_END is off; no specified value means that CHAIN_END is on. This parameter can be used to override the setting derived from a specified TPM value.</td></tr>
<tr><td>SUSPEND_ CURSOR</td><td>Specifies whether cursors are to be kept when a transaction thread of control is suspended. Valid values are T, F, or no value.</td><td>No</td><td>No</td><td>F</td></tr>
<tr><td colspan="5">TP monitors that suspend a transaction branch can reuse the suspended thread or process for other transactions. In these situations, cursors must be closed so that the new transaction does not inherit them. When the suspended transaction is resumed, the application must obtain the cursors again. If SUSPEND_CURSOR is on, any open cursors are not closed, but the thread or process cannot be reused for other transactions. Only the resumption of the suspended transaction is permitted. A value of T means that SUSPEND_CURSOR is on; a value of F means that SUSPEND_CURSOR is off; no specified value means that SUSPEND_CURSOR is on. This parameter can be used to override the setting derived from a specified TPM value.</td></tr>
</table>

Table 22. Valid Values for parm_id (continued)

Parameter Name	Value	Mandatory?	Case Sensitive?	Default Value
HOLD_CURSOR	Specifies whether cursors are held across transaction commits. Valid values are `T`, `F`, or no value.	No	No	F
TP monitors typically reuse threads or processes for multiple applications. To ensure that a newly loaded application does not inherit cursors opened by a previous application, cursors are closed after a commit. If HOLD_CURSOR is on, cursors are held across transaction commits. A value of `T` means that HOLD_CURSOR is on; a value of `F` means that HOLD_CURSOR is off; no specified value means that HOLD_CURSOR is on. This parameter can be used to override the setting derived from a specified TPM value.				

TPM and TP_MON_NAME Values

The xa_open string TPM parameter and the *tp_mon_name* database manager configuration parameter are used to indicate to DB2 which TP monitor is being used. The *tp_mon_name* value applies to the entire DB2 instance. The TPM parameter applies only to the specific XA resource manager. The TPM value overrides the *tp_mon_name* parameter. Valid values for the TPM and *tp_mon_name* parameters are as follows:

Table 23. Valid Values for TPM and tp_mon_name

TPM Value	TP Monitor Product	Internal Settings

CICS	IBM TxSeries CICS	`AXLIB=libEncServer (for Windows)`
		`=/usr/lpp/encina/lib/libEncServer`
		`(for UNIX based systems)`
		`HOLD_CURSOR=T`
		`CHAIN_END=T`
		`SUSPEND_CURSOR=F`

ENCINA	IBM TxSeries Encina Monitor	`AXLIB=libEncServer (for Windows)`
		`=/usr/lpp/encina/lib/libEncServer`
		`(for UNIX based systems)`
		`HOLD_CURSOR=F`
		`CHAIN_END=T`
		`SUSPEND_CURSOR=F`

Table 23. Valid Values for TPM and tp_mon_name (continued)

TPM Value	TP Monitor Product	Internal Settings

MQ	IBM MQSeries	AXLIB=mqmax (for Windows)
		=/usr/mqm/lib/libmqmax.a
		(for AIX)
		=/opt/mqm/lib/libmqmax.a
		(for Solaris)
		HOLD_CURSOR=F
		CHAIN_END=F
		SUSPEND_CURSOR=F

CB	IBM Component Broker	AXLIB=somtrx1i (for Windows)
		=libsomtrx1
		(for UNIX based systems)
		HOLD_CURSOR=F
		CHAIN_END=T
		SUSPEND_CURSOR=F

SF	IBM San Francisco	AXLIB=ibmsfDB2
		HOLD_CURSOR=F
		CHAIN_END=T
		SUSPEND_CURSOR=F

TUXEDO	BEA Tuxedo	AXLIB=libtux
		HOLD_CURSOR=F
		CHAIN_END=F
		SUSPEND_CURSOR=F

MTS	Microsoft Transaction Server	It is not necessary to configure DB2 for MTS. MTS is automatically detected by DB2's ODBC driver.
JTA	Java Transaction API	It is not necessary to configure DB2 for Enterprise Java Servers (EJS) such as IBM WebSphere. DB2's JDBC driver automatically detects this environment.

Examples

1. You are using IBM TxSeries CICS on WIndows NT. The TxSeries documentation indicates that you need to configure *tp_mon_name* with a value of `libEncServer:C`. This is still an acceptable format; however, with DB2 UDB or DB2 Connect Version 7, you have the option of:

- Specifying a *tp_mon_name* of CICS (recommended for this scenario):

  ```
  db2 update dbm cfg using tp_mon_name CICS
  ```

  For each database defined to CICS in the Region-> Resources-> Product-> XAD-> Resource manager initialization string, specify:

  ```
  db=dbalias,uid=userid,pwd=password
  ```
- For each database defined to CICS in the Region-> Resources-> Product-> XAD-> Resource manager initialization string, specify:

  ```
  db=dbalias,uid=userid,pwd=password,tpm=cics
  ```

2. You are using IBM MQSeries on Windows NT. The MQSeries documentation indicates that you need to configure *tp_mon_name* with a value of mqmax. This is still an acceptable format; however, with DB2 UDB or DB2 Connect Version 7, you have the option of:
   - Specifying a *tp_mon_name* of MQ (recommended for this scenario):

     ```
     db2 update dbm cfg using tp_mon_name MQ
     ```

     For each database defined to CICS in the Region-> Resources-> Product-> XAD-> Resource manager initialization string, specify:

     ```
     uid=userid,db=dbalias,pwd=password
     ```
   - For each database defined to CICS in the Region-> Resources-> Product-> XAD-> Resource manager initialization string, specify:

     ```
     uid=userid,db=dbalias,pwd=password,tpm=mq
     ```
3. You are using both IBM TxSeries CICS and IBM MQSeries on WIndows NT. A single DB2 instance is being used. In this scenario, you would configure as follows:
   a. For each database defined to CICS in the Region-> Resources-> Product-> XAD-> Resource manager initialization string, specify:

      ```
      pwd=password,uid=userid,tpm=cics,db=dbalias
      ```

   b. For each database defined as a resource in the queue manager properties, specify an XaOpenString as:

      ```
      db=dbalias,uid=userid,pwd=password,tpm=mq
      ```
4. You are developing your own XA-compliant transaction manager (XA TM) on Windows NT, and you want to tell DB2 that library ″myaxlib″ has the required functions **ax_reg** and **ax_unreg**. Library ″myaxlib″ is in a directory specified in the PATH statement. You have the option of:
   - Specifying a *tp_mon_name* of myaxlib:

     ```
     db2 update dbm cfg using tp_mon_name myaxlib
     ```

     and, for each database defined to the XA TM, specifying an xa_open string:

     ```
     db=dbalias,uid=userid,pwd=password
     ```
   - For each database defined to the XA TM, specifying an xa_open string:

```
db=dbalias,uid=userid,pwd=password,axlib=myaxlib
```

5. You are developing your own XA-compliant transaction manager (XA TM) on Windows NT, and you want to tell DB2 that library ″myaxlib″ has the required functions **ax_reg** and **ax_unreg**. Library ″myaxlib″ is in a directory specified in the PATH statement. You also want to enable XA END chaining. You have the option of:
   - For each database defined to the XA TM, specifying an xa_open string:
     ```
     db=dbalias,uid=userid,pwd=password,axlib=myaxlib,chain_end=T
     ```
   - For each database defined to the XA TM, specifying an xa_open string:
     ```
     db=dbalias,uid=userid,pwd=password,axlib=myaxlib,chain_end
     ```

xa_open String Format for Earlier Versions of DB2

Earlier versions of DB2 used the xa_open string format described here. This format is still supported for compatibility reasons. Applications should be migrated to the new format (see "New xa_open String Format for DB2 Version 7" on page 176) when possible.

Each database is defined as a separate resource manager (RM) to the transaction manager (TM), and the database must be identified with an xa_open string that has the following syntax:

```
"database_alias<,userid,password>"
```

The *database_alias* is required to specify the alias name of the database. The alias name is the same as the database name unless you have explicitly cataloged an alias name after you created the database. The user name and password are optional and, depending on the authentication method, are used to provide authentication information to the database.

When setting up a database as a resource manager, you do not need the xa_close string. If provided, this string will be ignored by the database manager.

Updating Host or AS/400 Database Servers

Host and AS/400 database servers may be updatable depending upon the architecture of the XA Transaction Manager. To support commit sequences from different processes, the DB2 Connect concentrator must be enabled. To enable the DB2 Connect EE concentrator, set the database manager configuration parameter *max_logicagents* to a value greater then *maxagents*. Note that the DB2 Connect EE concentrator requires a DB2 Version 7.1 client to support XA commit sequences from different processes. For information about the SQL statements that are allowed in this environment, refer to the *Application Development Guide*. For information about the concentrator, refer to the *DB2 Connect User's Guide*.

If you will be updating host or AS/400 database servers, you will require DB2 Connect with the DB2 sync point manager (SPM) configured. Refer to one of the *Quick Beginnings* books for instructions.

Database Connection Considerations

The following topics are covered in this section:

- “RELEASE Statement”
- “Transactions Accessing Partitioned Databases”

RELEASE Statement

If a RELEASE statement is used to release a connection to a database, a CONNECT statement, rather than SET CONNECTION, should be used to reconnect to that database.

Transactions Accessing Partitioned Databases

In a partitioned database environment, user data may be distributed across database partitions. An application accessing the database connects and sends requests to one of the database partitions (the coordinator node). Different applications can connect to different database partitions, and the same application can choose different database partitions for different connections.

For transactions against a database in a partitioned database environment, all access must be through the *same* database partition. That is, the same database partition must be used from the start of the transaction until (and including) the time that the transaction is committed.

Any transaction against the partitioned database must be committed before disconnecting.

Making a Heuristic Decision

An XA-compliant transaction manager (Transaction Processing Monitor) uses a two-phase commit process similar to that used by the DB2 transaction manager, described in “Understanding the Two-Phase Commit Process” on page 165. The principal difference between the two environments is that the TP monitor provides the function of logging and controlling the transaction, instead of the DB2 transaction manager and the transaction manager database.

Errors similar to those discussed for the DB2 transaction manager (see “Recovering from Problems During Two-Phase Commit” on page 168) can occur when using an XA-compliant transaction manager. Similar to the DB2 transaction manager, an XA-compliant transaction manager will attempt to resynchronize indoubt transactions.

If, for some reason, you cannot wait for the transaction manager to automatically resolve indoubt transactions, there are actions you can take to manually resolve them. This manual process is sometimes referred to as ″making a heuristic decision″.

The LIST INDOUBT TRANSACTIONS command (using the WITH PROMPTING option), or the related set of APIs, allows you to query, commit, and roll back indoubt transactions. In addition, it also allows you to ″forget″ transactions that have been heuristically committed or rolled back, by removing the log records and releasing the log space. To obtain indoubt transaction information from DB2 UDB on UNIX based systems, the Windows operating system, or OS/2, connect to the database and issue the LIST INDOUBT TRANSACTIONS WITH PROMPTING command, or the equivalent API. For information about this command or the related administrative APIs, refer to the *Command Reference* or the *Administrative API Reference*.

For indoubt transaction information with respect to host or AS/400 database servers, you have two choices:

- You can obtain indoubt information directly from the host or AS/400 server.

  To obtain indoubt information directly from DB2 for OS/390, invoke the DISPLAY THREAD TYPE(INDOUBT) command. Use the RECOVER command to make a heuristic decision. To obtain indoubt information directly from DB2 for OS/400, invoke the **wrkcmtdfn** command.
- You can obtain indoubt information from the DB2 Connect server used to access the host or AS/400 database server.

  To obtain indoubt information from the DB2 Connect server, first connect to the DB2 sync point manager by connecting to the DB2 instance represented by the value of the *spm_name* database manager configuration parameter. Then issue the LIST DRDA INDOUBT TRANSACTIONS WITH PROMPTING command to display indoubt transactions and to make heuristic decisions.

Use these commands (or related APIs) with *extreme caution*, and only as a last resort. The best strategy is to wait for the transaction manager to drive the resynchronization process. You could experience data integrity problems if you manually commit or roll back a transaction in one of the participating databases, and the opposite action is taken against another participating database. Recovering from data integrity problems requires you to understand the application logic, to identify the data that was changed or rolled back, and then to perform a point-in-time recovery of the database, or manually undo or reapply the changes.

If you cannot wait for the transaction manager to initiate the resynchronization process, and you must release the resources tied up by an indoubt transaction, heuristic operations are necessary. This situation could occur if the transaction manager will not be available for an extended period of time to perform the resynchronization, and the indoubt transaction is tying up resources that are urgently needed. An indoubt transaction ties up the resources that were associated with this transaction before the transaction manager or resource managers became unavailable. For the database manager, these resources include locks on tables and indexes, log space, and storage taken up by the transaction. Each indoubt transaction also decreases (by one) the maximum number of concurrent transactions that can be handled by the database.

Although there is no foolproof way to perform heuristic operations, the following provides some general guidelines:

1. Connect to the database for which you require all transactions to be complete.
2. Use the LIST INDOUBT TRANSACTIONS command to display the indoubt transactions. The *xid* represents the global transaction ID, and is identical to the *xid* used by the transaction manager and by other resource managers participating in the transaction.
3. For each indoubt transaction, use your knowledge about the application and the operating environment to determine the other participating resource managers.
4. Determine if the transaction manager is available:
   - If the transaction manager is available, and the indoubt transaction in a resource manager was caused by the resource manager not being available in the second commit phase, or for an earlier resynchronization process, you should check the transaction manager's log to determine what action has been taken against the other resource managers. You should then take the same action against the database; that is, using the LIST INDOUBT TRANSACTIONS command, either heuristically commit or heuristically roll back the transaction.
   - If the transaction manager is *not* available, you will need to use the status of the transaction in the other participating resource managers to determine what action you should take:
     - If at least one of the other resource managers has committed the transaction, you should heuristically commit the transaction in all the resource managers.
     - If at least one of the other resource managers has rolled back the transaction, you should heuristically roll back the transaction.
     - If the transaction is in ″prepared″ (indoubt) state in all of the participating resource managers, you should heuristically roll back the transaction.

- If one or more of the other resource managers is not available, you should heuristically roll back the transaction.

Do not perform the heuristic forget function unless a heuristically committed or rolled back transaction causes a log full condition, indicated in output from the LIST INDOUBT TRANSACTIONS command. The heuristic forget function releases the log space occupied by an indoubt transaction. The implication is that if a transaction manager eventually performs a resynchronization operation for this indoubt transaction, it could potentially make the wrong decision to commit or roll back other resource managers, because there is no log record for the transaction in this resource manager. In general a ″missing″ log record implies that the resource manager has rolled back the transaction.

Security Considerations

The TP monitor pre-allocates a set of server processes and runs the transactions from different users under the IDs of the server processes. To the database, each server process appears as a big application that has many units of work, all being run under the same ID associated with the server process.

For example, in an AIX environment using CICS, when a TXSeries CICS region is started, it is associated with the AIX user name under which it is defined. All the CICS Application Server processes are also being run under this TXSeries CICS ″master″ ID, which is usually defined as ″cics″. CICS users can invoke CICS transactions under their DCE login ID, and while in CICS, they can also change their ID using the CESN signon transaction. In either case, the end user's ID is not available to the RM. Consequently, a CICS Application Process might be running transactions on behalf of many users, but they appear to the RM as a single program with many units of work from the same ″cics″ ID. Optionally, you can specify a user ID and password on the xa_open string, and that user ID will be used, instead of the ″cics″ ID, to connect to the database.

There is not much impact on static SQL statements, because the binder's privileges, not the end user's privileges, are used to access the database. This does mean, however, that the EXECUTE privilege of the database packages must be granted to the server ID, and not to the end user ID.

For dynamic statements, which have their access authentication done at run time, access privileges to the database objects must be granted to the server ID and not to the actual user of those objects. Instead of relying on the database to control the access of specific users, you must rely on the TP monitor system to determine which users can run which programs. The server ID must be granted all privileges that its SQL users require.

To determine who has accessed a database table or view, you can perform the following steps:

1. From the SYSCAT.PACKAGEDEP catalog view, obtain a list of all packages that depend on the table or view.
2. Determine the names of the server programs (for example, CICS programs) that correspond to these packages through the naming convention used in your installation.
3. Determine the client programs (for example, CICS transaction IDs) that could invoke these programs, and then use the TP monitor's log (for example, the CICS log) to determine who has run these transactions or programs, and when.

Configuration Considerations

You should consider the following configuration parameters when you are setting up your TP monitor environment:

- *tp_mon_name*

  This database manager configuration parameter identifies the name of the TP monitor product being used ("CICS", or "ENCINA", for example).
- *tpname*

  This database manager configuration parameter identifies the name of the remote transaction program that the database client must use when issuing an allocate request to the database server, using the APPC communications protocol. The value is set in the configuration file at the server, and must be the same as the transaction processor (TP) name configured in the SNA transaction program. Refer to the *Quick Beginnings* manuals for more information.
- *tm_database*

  Because DB2 does *not* coordinate transactions in the XA environment, this database manager configuration parameter is not used for XA-coordinated transactions.
- *maxappls*

  This database configuration parameter specifies the maximum number of active applications allowed. The value of this parameter must be equal to or greater than the sum of the connected applications, plus the number of these applications that may be concurrently in the process of completing a two-phase commit or rollback. This sum should then be increased by the anticipated number of indoubt transactions that might exist at any one time. For more information about indoubt transactions, see "Recovering from Problems During Two-Phase Commit" on page 168.

  For a TP monitor environment (for example, TXSeries CICS), you may need to increase the value of the *maxappls* parameter. This would help to ensure that all TP monitor processes can be accommodated.

- *autorestart*

  This database configuration parameter specifies whether the RESTART DATABASE routine will be invoked automatically when needed. The default value is YES (that is, enabled).

  A database containing indoubt transactions requires a restart database operation to start up. If *autorestart* is not enabled when the last connection to the database is dropped, the next connection will fail and require an explicit RESTART DATABASE invocation. This condition will exist until the indoubt transactions have been removed, either by the transaction manager's resync operation, or through a heuristic operation initiated by the administrator. When the RESTART DATABASE command is issued, a message is returned if there are any indoubt transactions in the database. The administrator can then use the LIST INDOUBT TRANSACTIONS command and other command line processor commands to find get information about those indoubt transactions.

XA Function Supported

DB2 Universal Database supports the XA91 specification defined in *X/Open CAE Specification Distributed Transaction Processing: The XA Specification*, with the following exceptions:

- Asynchronous services

  The XA specification allows the interface to use asynchronous services, so that the result of a request can be checked at a later time. The database manager requires that the requests be invoked in synchronous mode.

- Static registration

  The XA interface allows two ways to register an RM: static registration and dynamic registration. DB2 Universal Database supports only dynamic registration, which is more advanced and efficient. For more information about these two methods, see "Resource Managers (RM)" on page 175.

- Association Migration

  DB2 Universal Database does not support transaction migration between threads of control.

For information about xa_open and xa_close strings usage, see "xa_open and xa_close Strings Usage" on page 176.

XA Switch Usage and Location

As required by the XA interface, the database manager provides a *db2xa_switch* external C variable of type xa_switch_t to return the XA switch structure to the TM. Other than the addresses of various XA functions, the following fields are returned:

Field **Value**

name The product name of the database manager. For example, DB2 for AIX.

flags TMREGISTER | TMNOMIGRATE

Explicitly states that DB2 Universal Database uses dynamic registration, and that the TM should not use association migration. Implicitly states that asynchronous operation is not supported.

version Must be zero.

Using the DB2 Universal Database XA Switch

The XA architecture requires that a Resource Manager (RM) provide a *switch* that gives the XA Transaction Manager (TM) access to the RM's **xa_** routines. An RM switch uses a structure called xa_switch_t. The switch contains the RM's name, non-NULL pointers to the RM's XA entry points, a flag, and a version number.

UNIX Based Systems and OS/2: DB2 UDB's switch can be obtained through either of the following two ways:

- Through one additional level of indirection. In a C program, this can be accomplished by defining the macro:

  ```
  #define db2xa_switch (*db2xa_switch)
  ```

  prior to using *db2xa_switch*.
- By calling **db2xacic**

  DB2 UDB provides this API, which returns the address of the *db2xa_switch* structure. This function is prototyped as:

  ```
  struct xa_switch_t * SQL_API_FN  db2xacic( )
  ```

With either method, you must link your application with `libdb2` (on UNIX based system) or `db2api.lib` (on OS/2).

Windows NT: The pointer to the *xa_switch* structure, *db2xa_switch*, is exported as DLL data. This implies that a Windows NT application using this structure must reference it in one of three ways:

- Through one additional level of indirection. In a C program, this can be accomplished by defining the macro:

  ```
  #define db2xa_switch (*db2xa_switch)
  ```

  prior to using *db2xa_switch*.
- If using the Microsoft Visual C++ compiler, *db2xa_switch* can be defined as:

  ```
  extern __declspec(dllimport) struct xa_switch_t db2xa_switch
  ```
- By calling **db2xacic**

  DB2 UDB provides this API, which returns the address of the *db2xa_switch* structure. This function is prototyped as:

  ```
  struct xa_switch_t * SQL_API_FN  db2xacic( )
  ```

With any of these methods, you must link your application with db2api.lib.

Example C Code: The following code illustrates the different ways in which the *db2xa_switch* can be accessed via a C program on any DB2 UDB platform. Be sure to link your application with the appropriate library.

```
   #include <stdio.h>
   #include <xa.h>

   struct xa_switch_t * SQL_API_FN  db2xacic( );

   #ifdef DECLSPEC_DEFN
   extern __declspec(dllimport) struct xa_switch_t db2xa_switch;
   #else
   #define db2xa_switch (*db2xa_switch)
   extern struct xa_switch_t db2xa_switch;
   #endif
main( )
   {
      struct xa_switch_t *foo;
      printf ( "%s \n", db2xa_switch.name );
      foo = db2xacic();
      printf ( "%s \n", foo->name );
      return ;
   }
```

XA Interface Problem Determination

When an error is detected during an XA request from the TM, the application program may not be able to get the error code from the TM. If your program abends, or gets a cryptic return code from the TP monitor or the TM, you should check the First Failure Service Log, which reports XA error information when diagnostic level 3 or greater is in effect. For more information about the First Failure Service Log, refer to the *Troubleshooting Guide*.

You should also consult the console message, TM error file, or other product-specific information about the external transaction processing software that you are using.

The database manager writes all XA-specific errors to the First Failure Service Log with SQLCODE -998 (transaction or heuristic errors) and the appropriate reason codes. Following are some of the more common errors:

- Invalid syntax in the xa_open string.
- Failure to connect to the database specified in the open string as a result of one of the following:
  - The database has not been cataloged.
  - The database has not been started.
  - The server application’s user name or password is not authorized to connect to the database.

- Communications error.

Following is an example of an error log for an xa_open error (due to a missing xa_open string) generated on AIX:

```
Tue Apr  4 15:59:08 1995
toop pid(83378) process (xatest) XA DTP Support      sqlxa_open Probe:101
DIA4701E Database "" could not be opened for distributed transaction
processing.
String Title : XA Interface SQLCA  pid(83378)
SQLCODE = -998  REASON CODE: 4  SUBCODE: 1
Dump File : /u/toop/diagnostics/83378.dmp Data : SQLCA
```

Configuring XA Transaction Managers to Use DB2 UDB

Note that the information in this section supercedes the similar section in the *Administration Guide: Performance*.

The sections that follow describe how to configure specific products to use DB2 as a resource manager. You can use any of the following:

- "Configuring IBM TXSeries CICS"
- "Configuring IBM TXSeries Encina"
- "Configuring BEA Tuxedo" on page 193
- "Configuring Microsoft Transaction Server" on page 195.

Configuring IBM TXSeries CICS

For information about how to configure IBM TXSeries CICS to use DB2 as a resource manager, refer to your *IBM TXSeries CICS Administration Guide*. TXSeries documentation can be viewed online at http://www.transarc.com/Library/documentation/websphere/WAS-EE/en_US/html/.

Host and AS/400 database servers can participate in CICS-coordinated transactions.

Configuring IBM TXSeries Encina

Following are the various APIs and configuration parameters required for the integration of Encina Monitor and DB2 Universal Database servers, or DB2 for MVS, DB2 for OS/390, DB2 for AS/400, or DB2 for VSE&VM when accessed through DB2 Connect. TXSeries documentation can be viewed online at http://www.transarc.com/Library/documentation/websphere/WAS-EE/en_US/html/.

Host and AS/400 database servers can participate in Encina-coordinated transactions.

Configuring DB2

To configure DB2:

1. Each database name must be defined in the DB2 database directory. If the database is a remote database, a node directory entry must also be defined. You can perform the configuration using the Client Configuration Assistant (CCA), or the DB2 command line processor (CLP). For example:

   ```
   DB2 CATALOG DATABASE inventdb AS inventdb AT NODE host1 AUTH SERVER
   DB2 CATALOG TCPIP NODE host1 REMOTE hostname1 SERVER svcname1
   ```

2. The DB2 client can optimize its internal processing for Encina if it knows that it is dealing with Encina. You can specify this by setting the *tp_mon_name* database manager configuration parameter to ENCINA. The default behavior is no special optimization. If *tp_mon_name* is set, the application must ensure that the thread that performs the unit of work also immediately commits the work after ending it. No other unit of work may be started. If this is *not* your environment, ensure that the *tp_mon_name* value is NONE (or, through the CLP, that the value is set to NULL). The parameter can be updated through the Control Center or the CLP. The CLP command is:

   ```
   db2 update dbm cfg using tp_mon_name ENCINA
   ```

Configuring Encina for Each Resource Manager

To configure Encina for each resource manager (RM), an administrator must define the Open String, Close String, and Thread of Control Agreement for each DB2 database as a resource manager before the resource manager can be registered for transactions in an application. The configuration can be performed using the Enconcole full screen interface, or the Encina command line interface. For example:

```
monadmin create rm inventdb -open "db=inventdb,uid=user1,pwd=password1"
```

There is one resource manager configuration for each DB2 database, and each resource manager configuration must have an rm name (″logical RM name″). To simplify the situation, you should make it identical to the database name.

The xa_open string contains information that is required to establish a connection to the database. The content of the string is RM-specific. The xa_open string of DB2 UDB contains the alias name of the database to be opened, and optionally, a user ID and password to be associated with the connection. Note that the database name defined here must also be cataloged into the regular database directory required for all database access. For information about DB2's xa_open string, see "Setting Up a Database as a Resource Manager" on page 176.

The xa_close string is not used by DB2.

The Thread of Control Agreement determines if an application agent thread can handle more than one transaction at a time. DB2 UDB supports the default of TMXA_SERIALIZE_ALL_OPERATIONS, where a thread can be reused only after a transaction has completed.

If you are accessing DB2 for OS/390, DB2 for MVS, DB2 for AS/400, or DB2 for VSE&VM, you must use the DB2 Syncpoint Manager. Refer to the *DB2 Connect Enterprise Edition for OS/2 and Windows Quick Beginnings* manual for configuration instructions.

Referencing a DB2 Database from an Encina Application

To reference a DB2 database from an Encina application:

1. Use the Encina Scheduling Policy API to specify how many application agents can be run from a single TP monitor application process. For example:

   ```
   rc = mon_SetSchedulingPolicy (MON_EXCLUSIVE)
   ```

   For DB2 (DB2 Universal Database, host, or AS/400 database servers), you should use the default setting of MON_EXCLUSIVE. This ensures that:

   - The application process is locked during the lifetime of the transaction.
   - The application acts single-threaded.

   Note: If you are using the ODBC or DB2 Call Level Interface, you must disable multithread support. You can do this by setting the CLI configuration parameter `DISABLEMULTITHREAD = 1` (disables multithreading). The default for DB2 Universal Database is `DISABLEMULTITHREAD = 0` (enables multithreading). Refer to the *CLI Guide and Reference* for more information.

2. Use the Encina RM Registration API to provide the XA switch and the logical RM name to be used by Encina when referencing the RM in an application process. For example:

   ```
   rc = mon_RegisterRmi ( &db2xa_switch,   /* xa switch */
                          "inventdb",      /* logical RM name */
                          &rmiId );        /* internal RM ID */
   ```

   The XA switch contains the addresses of the XA routines in the RM that the TM can call, and it also specifies the functionality that is provided by the RM. The XA switch of DB2 Universal Database is `db2xa_switch`, and it resides in the DB2 client library (`db2app.dll` on Windows operating systems and OS/2, and `libdb2` on UNIX based systems).

   The logical RM name is the one used by Encina, and is not the actual database name that is used by the SQL application that runs under Encina.

The actual database name is specified in the xa_open string in the Encina RM Registration API. The logical RM name is set to be the same as the database name in this example.

The third parameter returns an internal identifier or handle that is used by the TM to reference this connection.

Note: When using Encina for transaction processing with DB2 through the TM-XA interface, note that Encina-nested transactions are not currently supported by the DB2 XA interface. Avoid using these transactions, if possible. If you cannot, ensure that SQL work is done in only one member of the Encina transaction family.

Configuring BEA Tuxedo

To configure Tuxedo to use DB2 as a resource manager, perform the following steps:

1. Install Tuxedo as specified in the documentation for that product. Ensure that you perform all basic Tuxedo configuration, including the log files and environment variables.

   You also require a compiler and the DB2 Application Development Client. Install these if necessary.
2. At the Tuxedo server ID, set the DB2INSTANCE environment variable to reference the instance that contains the databases that you want Tuxedo to use. Set the PATH variable to include the DB2 program directories. Confirm that the Tuxedo server ID can connect to the DB2 databases.
3. Update the *tp_mon_name* database manager configuration parameter with the value `TUXEDO`.
4. Add a definition for DB2 to the Tuxedo resource manager definition file. In the examples that follow, `UDB_XA` is the locally-defined Tuxedo resource manager name for DB2, and *db2xa_switch* is the DB2-defined name for a structure of type `xa_switch_t`:
   - For AIX. In the file `${TUXDIR}/udataobj/RM`, add the definition:

     ```
     # DB2 UDB
     UDB_XA:db2xa_switch:-L${DB2DIR} /lib -ldb2
     ```

     where `{TUXDIR}` is the directory where you installed Tuxedo, and `{DB2DIR}` is the DB2 instance directory.
   - For Windows NT. In the file `%TUXDIR%\udataobj\rm`, add the definition:

     ```
     # DB2 UDB
     UDB_XA;db2xa_switch;%DB2DIR%\lib\db2api.lib
     ```

     where `%TUXDIR%` is the directory where you installed Tuxedo, and `%DB2DIR%` is the DB2 instance directory.
5. Build the Tuxedo transaction monitor server program for DB2:

- For AIX:

  ```
  ${TUXDIR}/bin/buildtms -r UDB_XA -o ${TUXDIR}/bin/TMS_UDB
  ```

  where {TUXDIR} is the directory where you installed Tuxedo.
- For Windows NT:

  ```
  %TUXDIR%\bin\buildtms -r UDB_XA -o %TUXDIR%\bin\TMS_UDB
  ```

6. Build the application servers. In the examples that follow, the -r option specifies the resource manager name, the -f option (used one or more times) specifies the files that contain the application services, the -s option specifies the application service names for this server, and the -o option specifies the output server file name:
   - For AIX:

     ```
     ${TUXDIR}/bin/buildserver -r UDB_XA -f svcfile.o -s SVC1,SVC2
        -o UDBserver
     ```

     where {TUXDIR} is the directory where you installed Tuxedo.
   - For Windows NT:

     ```
     %TUXDIR%\bin\buildserver -r UDB_XA -f svcfile.o -s SVC1,SVC2
        -o UDBserver
     ```

     where %TUXDIR% is the directory where you installed Tuxedo.
7. Set up the Tuxedo configuration file to reference the DB2 server. In the *GROUPS section of the UDBCONFIG file, add an entry similar to:

   ```
   UDB_GRP   LMID=simp GRPNO=3
     TMSNAME=TMS_UDB TMSCOUNT=2
     OPENINFO="UDB_XA:db=sample,uid=db2_user,pwd=db2_user_pwd"
   ```

   where the TMSNAME parameter specifies the transaction monitor server program that you built previously, and the OPENINFO parameter specifies the resource manager name. This is followed by the database name, and the DB2 user and password, which are used for authentication.

   The application servers that you built previously are referenced in the *SERVERS section of the Tuxedo configuration file.
8. If the application is accessing data residing on DB2 for OS/390, DB2 for OS/400, or DB2 for VM&VSE, the DB2 Connect XA concentrator will be required. For configuration details and limitations, refer to the *DB2 Connect User's Guide*.
9. Start Tuxedo:

   ```
   tmboot -y
   ```

   After the command completes, Tuxedo messages should indicate that the servers are started. In addition, if you issue the DB2 command LIST

APPLICATIONS ALL, you should see two connections (in this situation, specified by the TMSCOUNT parameter in the UDB group in the Tuxedo configuration file, `UDBCONFIG`.

Configuring Microsoft Transaction Server

DB2 UDB V5.2 and later can be fully integrated with Microsoft Transaction Server (MTS) Version 2.0. Applications running under MTS on Windows 32-bit operating systems can use MTS to coordinate two-phase commit with multiple DB2 UDB, host, and AS/400 database servers, as well as with other MTS-compliant resource managers.

Enabling MTS Support in DB2

Microsoft Transaction Server support is automatically enabled. While you can set the *tp_mon_name* database manager configuration parameter to `MTS`, it is not necessary and will be ignored.

Note: Additional technical information may be provided on the IBM web site to assist you with installation and configuration of DB2 MTS support. Set your URL to http://www.ibm.com/software/data/db2/library/, and search for a DB2 Universal Database ″Technote″ with the keyword ″MTS″.

MTS Software Prerequisites

MTS support requires the DB2 Client Application Enabler (CAE) Version 5.2, or later, and MTS must be at Version 2.0 with Hotfix 0772 or later releases.

The installation of the DB2 ODBC driver on Windows 32-bit operating systems will automatically add a new keyword into the registry:

```
HKEY_LOCAL_MACHINE\software\ODBC\odbcinit.ini\IBM DB2 ODBC Driver:
Keyword Value Name: CPTimeout
Data Type: REG_SZ
Value: 60
```

Installation and Configuration

Following is a summary of installation and configuration considerations for MTS. To use DB2’s MTS support, you must:

1. Install MTS and the DB2 client on the same machine where the MTS application runs.
2. If host or AS/400 database servers are to be involved in a multisite update:
   a. Install DB2 Connect Enterprise Edition (EE), either on your local machine or on a remote machine. DB2 Connect EE allows host or AS/400 database servers to participate in a multisite update transaction.

b. Ensure that your DB2 Connect EE server is enabled for multisite update. For information about enabling DB2 Connect for multisite updates, refer to the DB2 Connect Enterprise Edition Quick Beginnings manual for your platform.

When running DB2 CLI/ODBC applications, the following configuration keywords (as set in the `db2cli.ini` file) must not be changed from their default values:

- CONNECTYPE keyword (default 1)
- MULTICONNECT keyword (default 1)
- DISABLEMULTITHREAD keyword (default 1)
- CONNECTIONPOOLING keyword (default 0)
- KEEPCONNECTION keyword (default 0)

DB2 CLI applications written to make use of MTS support must not change the attribute values corresponding to the above keywords. In addition, the applications must not change the default values of the following attributes:

- SQL_ATTR_CONNECT_TYPE attribute (default SQL_CONCURRENT_TRANS)
- SQL_ATTR_CONNECTON_POOLING attribute (default SQL_CP_OFF)

Note: Additional technical information may be provided on the IBM web site to assist you with installation and configuration of DB2 MTS support. Set your URL to http://www.ibm.com/software/data/db2/library/, and search for a DB2 Universal Database ″Technote″ with the keyword ″MTS″.

Verifying the Installation

1. Configure your DB2 client and DB2 Connect EE to access your DB2 UDB, host, or AS/400 server.
2. Verify the connection from the DB2 CAE machine to the DB2 UDB database servers.
3. Verify the connection from the DB2 Connect machine to your host or AS/400 database server with DB2 CLP, and issue a few queries.
4. Verify the connection from the DB2 CAE machine through the DB2 Connect gateway to your host or AS/400 database server, and issue a few queries.

Supported DB2 Database Servers

The following servers are supported for multisite update using MTS-coordinated transactions:

- DB2 Universal Database Enterprise Edition Version 5.2
- DB2 Enterprise - Extended Edition Version 5.2

- DB2 for OS/390
- DB2 for MVS
- DB2 for AS/400
- DB2 for VM&VSE
- DB2 Common Server for SCO, Version 2
- DB2 Universal Database for AIX with PTF U453782
- DB2 Universal Database for HP-UX with PTF U453784
- DB2 Universal Database Enterprise Edition for OS/2 with PTF WR09033
- DB2 Universal Database for SOLARIS with PTF U453783
- DB2 Universal Database Enterprise Edition for Windows NT with PTF WR09034
- DB2 Universal Database Extended Enterprise Edition for UNIX or Windows NT.

MTS Transaction Time-Out and DB2 Connection Behavior

You can set the transaction time-out value in the MTS Explorer tool. For more information, refer to the online *MTS Administrator Guide*.

If a transaction takes longer than the transaction time-out value (default value is 60 seconds), MTS will asynchronously issue an abort to all Resource Managers involved, and the whole transaction is aborted.

For the connection to a DB2 server, the abort is translated into a DB2 rollback request. Like any other database request, the rollback request is serialized on the connection to guarantee the integrity of the data on the database server.

As a result:

- If the connection is idle, the rollback is executed immediately.
- If a long-running SQL statement is processing, the rollback request waits until the SQL statement finishes.

Connection Pooling

Connection pooling enables an application to use a connection from a pool of connections, so that the connection does not need to be re-established for each use. Once a connection has been created and placed in a pool, an application can reuse that connection without performing a complete connection process. The connection is pooled when the application disconnects from the ODBC data source, and will be given to a new connection whose attributes are the same.

Connection pooling has been a feature of ODBC driver Manager 2.x. With the latest ODBC driver manager (version 3.5) that was shipped with MTS, connection pooling has some configuration changes and new behavior for

ODBC connections of transactional MTS COM objects (see “Reusing ODBC Connections Between COM Objects Participating in the Same Transaction” on page 199).

ODBC driver Manager 3.5 requires that the ODBC driver register a new keyword in the registry before it allows connection pooling to be activated. The keyword is:

```
Key Name: SOFTWARE\ODBC\ODBCINST.INI\IBM DB2 ODBC DRIVER
Name: CPTimeout
Type: REG_SZ
Data: 60
```

The DB2 ODBC driver Version 6 and later for the 32-bit Windows operating system fully supports connection pooling; therefore, this keyword is registered. Version 5.2 clients must install FixPack 3 (WR09024) or later.

The default value of 60 means that the connection will be pooled for 60 seconds before it is disconnected.

In a busy environment, it is better to increase the CPTimeout value to a large number (Microsoft sometimes suggests 10 minutes for certain environments) to prevent too many physical connects and disconnects, because these consume large amounts of system resource, including system memory and communications stack resources.

In addition, to ensure that the same connection is used between objects in the same transaction in a multiple processor machine, you must turn off ″multiple pool per processor″ support. To do this, copy the following registry setting into a file called `odbcpool.reg`, save it as a plain text file, and issue the command **odbcpool.reg**. The Windows operating system will import these registry settings.

```
REGEDIT4

[HKEY_LOCAL_MACHINE\SOFTWARE\ODBC\ODBCINST.INI\ODBC Connection Pooling]
"NumberOfPools"="1"
```

Without this keyword set to 1, MTS may pool connections in different pools, and hence will not reuse the same connection.

MTS Connection Pooling using ADO 2.1 and Later

If the MTS COM objects use ADO to access the database, you must turn off the OLEDB resource pooling so that the Microsoft OLEDB provider for ODBC (MSDASQL) will not interfere with ODBC connection pooling. This feature was initialized to `OFF` in ADO 2.0, but is initialized to `ON` in ADO 2.1. To turn OLEDB resource polling off, copy the following lines into a file called `oledb.reg`, save it as a plain text file, and issue the command **oledb.reg**. The Windows operating system will import this registry setting.

```
REGEDIT4

[HKEY_CLASSES_ROOT\CLSID\{c8b522cb-5cf3-11ce-ade5-00aa0044773d}]
@="MSDASQL"
"OLEDB_SERVICES"=dword:fffffffc
```

Reusing ODBC Connections Between COM Objects Participating in the Same Transaction

ODBC connections in MTS COM objects have connection pooling turned on automatically (whether or not the COM object is transactional).

For multiple MTS COM objects participating in the same transaction, the connection can be reused between two or more COM objects in the following manner.

Suppose that there are two COM objects, COM1 and COM2, that connect to the same ODBC data source and participate in the same transaction.

After COM1 connects and does its work, it disconnects, and the connection is pooled. However, this connection will be reserved for the use of other COM objects of the same transaction. It will be available to other transactions only after the current transaction ends.

When COM2 is invoked in the same transaction, it is given the pooled connection. MTS will ensure that the connection can only be given to the COM objects that are participating in the same transaction.

On the other hand, if COM1 does not explicitly disconnect, it will tie up the connection until the transaction ends. When COM2 is invoked in the same transaction, a separate connection will be acquired. Subsequently, this transaction ties up two connections instead of one.

This reuse of connection feature for COM objects participating in the same transaction is preferable for the following reasons:

- It uses fewer resources in both the client and the server. Only one connection is needed.
- It eliminates the possibility that two connections participating in the same transaction (accessing the same database server and accessing the same data) can lock one another, because DB2 servers treat different connections from MTS COM objects as separate transactions.

Tuning TCP/IP Communications

If a small CPTimeout value is used in a high-workload environment where too many physical connects and disconnects occur at the same time, the TCP/IP stack may encounter resource constraints.

To alleviate this problem, use the TCP/IP Registry Entries. These are described in the *Windows NT Resource Guide*, Volume 1. The registry key values are located in HKEY_LOCAL_MACHINE—> SYSTEM—> CurrentControlSet—> Services—> TCPIP—> Parameters.

The default values and suggested settings are as follows:

Name	Default Value	Suggested Value
KeepAlive time	7200000 (2 hours)	Same
KeepAlive interval	1000 (1 second)	10000 (10 seconds)
TcpKeepCnt	120 (2 minutes)	240 (4 minutes)
TcpKeepTries	20 (20 re-tries)	Same
TcpMaxConnectAttempts	3	6
TcpMaxConnectRetransmission	3	6
TcpMaxDataRetransmission	5	8
TcpMaxRetransmissionAttempts	7	10
If the registry value is not defined, create it.		

Testing DB2 With The MTS ″BANK″ Sample Application

You can use the ″BANK″ sample program that is shipped with MTS to test the setup of the client products and MTS.

Follow these steps:

1. Change the file `\Program Files\Common Files\ODBC\Data Sources\MTSSamples.dsn` so that it looks like this:

   ```
   [ODBC]
   DRIVER=IBM DB2 ODBC DRIVER
   UID=your_user_id
   PWD=your_password
   DSN=your_database_alias
   Description=MTS Samples
   ```

   where:

   - *your_user_id*and *your_password* are the user ID and password used to connect to the host.
   - *your_database_alias* is the database alias used to connect to the database server.
2. Go to ODBC Administration in the Control Panel, select the **System DSN** tab, and then add the data source:
   a. Select IBM ODBC Driver, and then select **Finish**.

b. When presented with the list of database aliases, choose the one that was specified previously.

c. Select **OK**.

3. Use DB2 CLP to connect to a DB2 database under the ID *your_user_id*, as above.

a. Bind the db2cli.lst file:

```
db2 bind @C:\sqllib\bnd\db2cli.lst blocking all grant public
```

b. Bind the utilities.

If the server is a DRDA host server, bind ddcsmvs.lst, ddcs400.lst, or ddcsvm.lst, depending on the host that you are connecting to (OS/390, AS/400, or VSE&VM). For example:

```
db2 bind @C:\sqllib\bnd\@ddcsmvs.lst blocking all grant public
```

Otherwise, bind the db2ubind.lst file:

```
db2 bind @C:\sqllib\bnd\@db2ubind.lst blocking all grant public
```

c. Create the sample table and data for the MTS sample application, as follows:

```
db2 create table account (accountno int, balance int)
db2 insert into account values(1, 1)
```

4. On the DB2 client, ensure that the database manager configuration parameter *tp_mon_name* is set to MTS.

5. Run the ″BANK″ application. Select the **Account** button and the **Visual C++** option, then submit the request. Other options may use SQL that is specific to SQL Server, and may not work.

Chapter 11. Designing for High Availability

DB2 Universal Database provides *high availability failover support* on many platforms. Failover capability allows for the automatic transfer of workload from one processor to another when there is hardware failure. For example, on AIX, DB2 UDB supports failover through the capabilities of IBM High Availability Cluster Multi-Processing (HACMP). Throughout this section, examples from AIX are used to introduce the concepts associated with high availability.

HACMP provides increased availability through clusters of processors that share resources such as disks or network access. If a processor fails, another processor in the cluster substitutes for it.

There are three modes of failover support:

Hot Standby

In this mode, one processor is used to run your DB2 instance, and the second processor is in standby mode, ready to take over the instance if there is an operating system or hardware failure involving the first processor.

Mutual Takeover

In this mode:

- Both processors are used to run separate DB2 instances.
- One processor is used to run a DB2 instance, while the other one is used to run DB2 applications.

If there is an operating system or hardware failure on one of the processors, the other processor takes over the tasks of the failed processor, eventually doing the work of both processors.

Concurrent Access

In this mode, multiple processors are used to scale to a single database instance using the DB2 Universal Database Enterprise - Extended Edition (EEE) product. This is done using a shared-nothing model, partitioning the data such that one or more partitions are running on each processor in the cluster. If an operating system or hardware failure occurs on one of the processors, the other processor takes over the partitions of the failing processor. DB2 UDB EEE does not require a Concurrent Resource Manager to provide redundancy. Redundancy is managed by using the hot standby or the mutual takeover mode. The capabilities of the concurrent access mode are only required by database managers with a shared architecture.

Each of these configurations can be used to failover one or more partitions of a partitioned database. In addition, each can failover a complete instance of a single partition installation.

Hot Standby

The *hot standby* capability can be used to failover the entire instance of a single partition database or a partition of a partitioned database configuration. If one processor fails, another processor in the cluster can substitute for the failed processor by automatically transferring the instance. The database instance and the actual database must be accessible to both the primary and the failover processor.

- The DB2 installation path can be either on a path shared by both systems, or on a non-shared file system. If using a non-shared file system, the installation levels must be identical.
- The DB2 instance path can be either on a shared file system, or on a manually mirrored file system.
- The database and associated containers must be on file systems (or devices) accessible to both systems.
- During failover of a partition in a partitioned database configuration, the partition is restarted on the second processor: the failover script changes the `db2nodes.cfg` file to point to this partition on the new processor, and starts the partition on that processor.
- When a failover occurs, the external communications addresses for supported communications protocols are transparently transferred as part of the failover procedure.

For detailed information about the actual installation requirements, and about instance creation, refer to *HACMP for AIX, Version 4.2: Installation Guide*, SC23-1940.

Examples

Each of the following examples has a sample script that is stored (on DB2 for AIX installations) in `sqllib/samples/hacmp`.

Instance Failover

The following hot standby failover scenario consists of a two-processor HACMP cluster running a single-partition database instance (Figure 44 on page 205). For information about configuring your HACMP cluster, see “Resources” on page 210.

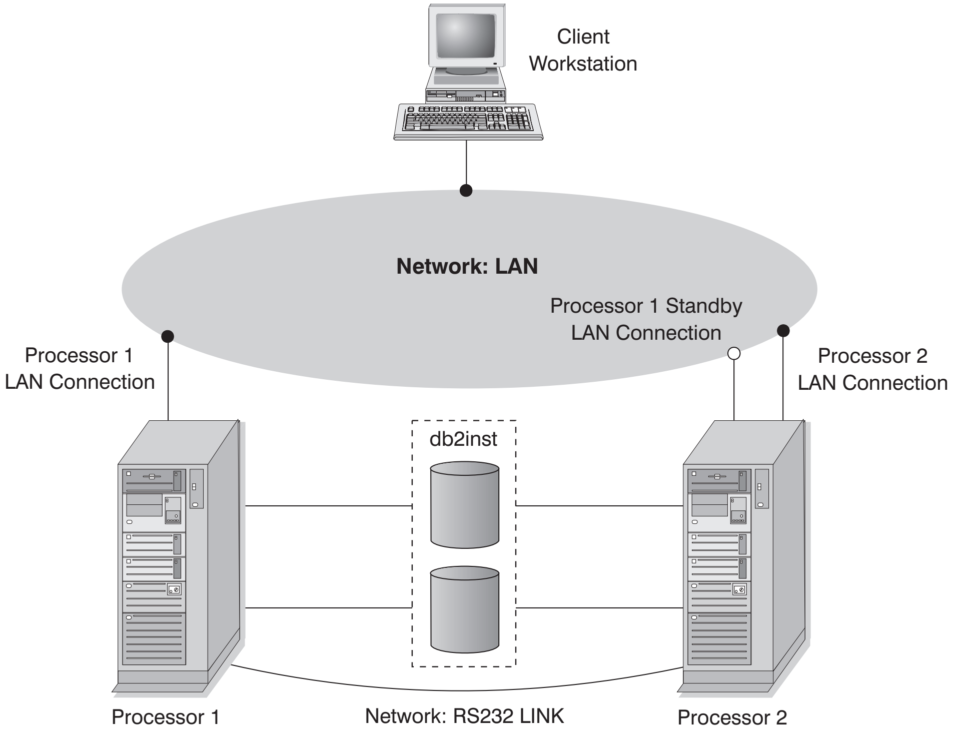


Figure 44. Example of a Hot Standby Failover Configuration

Both processors have access to the installation directory, the instance directory, and the database directory. The database instance ″db2inst″ is being actively executed on processor 1. Processor 2 is not active, and is being used as a hot standby. A failure occurs on processor 1, and the instance is taken over by processor 2. Once the failover is complete, both remote and local applications can access the database within instance ″db2inst″. The database will have to be manually restarted or, if AUTORESTART is on, the first connection to the database will initiate a restart operation. In the sample script provided, it is assumed that AUTORESTART is off, and that the failover script performs the restart for the database. For more information about AUTORESTART, see ″Overview of Recovery″ in the *Administration Guide: Implementation*.

Sample script:

```
hacmp-s1.sh
```

Partition Failover

In the following hot standby failover scenario, we are using an instance partition instead of the entire instance. The scenario includes a two processor HACMP cluster as in the previous example, but the machine represents one of the partitions of a partitioned database server. Processor 1 is running a single partition of the overall configuration, and processor 2 is being used as the failover processor. When processor 1 fails, the partition is restarted on the second processor. The failover updates the `db2nodes.cfg` file, pointing to processor 2's host name and net name, and then restarts the partition on the new processor.

Following is a portion of the `db2nodes.cfg` file, both before and after the failover. In this example, node number 2 is running on processor 1 of the HACMP machine, which has both a host name and a net name of ″node201″. After the failover, node number 2 is running on processor 2 of the HACMP machine, which has both a host name and a net name of ″node202″.

```
Before:
        1 node101 0 node101
        2 node201 0 node201    <= HACMP
        3 node301 0 node301

   db2start nodenum 2 restart hostname node202 port 0 netname node202

After:
        1 node101 0 node101
        2 node202 0 node202    <= HACMP
        3 node301 0 node301
```

Sample script:

```
   hacmp-s2.sh
```

Multiple Logical Node Failover

A more complex variation on the previous example involves the failover of multiple logical nodes from one processor to another. Again, we are using the same two processor HACMP cluster configuration as above. However, in this scenario, processor 1 is running 3 logical partitions. The setup is the same as that for the simple partition failover scenario, but in this case when processor 1 fails, each of the logical partitions must be started on processor 2. Each logical partition must be started in the order that is defined in the `db2nodes.cfg` file: the logical partition with port number 0 must always be started first.

Following is a portion of the `db2nodes.cfg` file, both before and after the failover. In this example, there are 3 logical partitions defined on processor 1 of a two-processor HACMP cluster.

```
Before:
        1 node101 0 node101
        2 node201 0 node201    <= HACMP
        3 node201 1 node201    <= HACMP
        4 node201 2 node201    <= HACMP
        5 node301 0 node301

   db2start nodenum 2 restart hostname node202 port 0 netname node202
   db2start nodenum 3 restart hostname node202 port 1 netname node202
   db2start nodenum 4 restart hostname node202 port 2 netname node202
```

```
After:
        1 node101 0 node101
        2 node202 0 node202    <= HACMP
        3 node202 1 node202    <= HACMP
        4 node202 2 node202    <= HACMP
        5 node301 0 node301
```

Sample script:

```
   hacmp-s3.sh
```

Mutual Takeover

In *mutual takeover* mode, one processor can failover the single-partition database instance, or the partitions of a partitioned database, while running another instance or other partitions of a partitioned database configuration. As with the hot standby configuration, the installation path, the instance directory, and the database must be accessible to each processor that may be involved in failover processing. The installation and instance paths can either be on a shared file system, or mirrored on separate file systems.

When using the mutual takeover strategy for instance failover, the instances must be defined in such a manner that both instances can be run simultaneously on the same processor. For detailed information about the actual installation requirements, and about instance creation, refer to *HACMP for AIX, Version 4.2: Installation Guide*, SC23-1940.

Examples

Each of the following examples has a sample script that is stored (on DB2 for AIX installations) in `sqllib/samples/hacmp`.

Mutual DB2 Instance Failover

The following mutual instance failover scenario consists of an HACMP system with two processors known as ″node10″ and ″node20″.

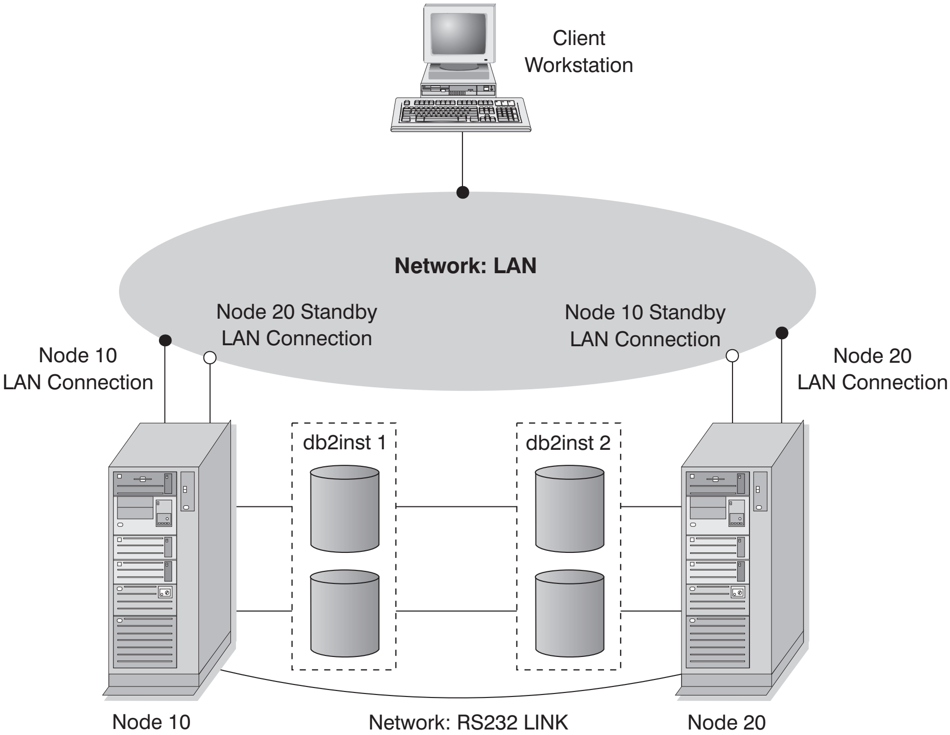


Figure 45. Example of a Mutual Instance Failover Configuration

Two instances, ″db2inst1″ and ″db2inst2″, are created from a single installation path on a shared file system. Instance ″db2inst1″ is created on `/u/db2inst1`, and instance ″db2inst2″ is created on `/u/db2inst2`. Both of these paths are on a shared file system that is accessible to both processors. Each instance has a single database, with a unique path, that is also on a shared resource accessible to both processors.

Both instances are accessed via remote clients over the TCP/IP protocol: ″db2inst1″ uses the service name ″db2inst1_port″ (port number 5500), and ″db2inst2″ uses the service name ″db2inst2_port″ (port number 5550). Remote clients accessing the ″db2inst1″ instance have this instance cataloged in their node directory using ″node10″ as the host name. Remote clients accessing the ″db2inst2″ instance have this instance cataloged in their node directory using ″node20″ as the host name. Under normal operating conditions, ″db2inst1″ is running on ″node10″, and ″db2inst2″ is running on ″node20″. If ″node10″ were to fail, the failover script will start ″db2inst1″ on ″node20″, and the external IP address associated with ″node10″ will be switched over to ″node20″. Once the instance has been started by the failover script, and the database has been restarted, the remote clients can connect to the database within this instance as if it were running on ″node10″.

Sample script:

```
hacmp-s4.sh
```

Mutual DB2 Partition Failover

Mutual failover of partitions in a partitioned database server environment requires that the failover of the partition occur as a logical node on the failover processor. For example, if we have two partitions of a partitioned database server running on separate processors of a two-processor HACMP cluster configured for mutual takeover, the partitions must failover as logical nodes. The default partition at each node must be defined as logical node 0, meaning that when a partition fails over from one processor to another, it will start as a logical node having no direct remote communication protocol listeners. Such a partition cannot be used as a coordinator node.

One other important consideration when configuring a system for mutual partition takeover pertains to the local partition database path. When a database is created in a partitioned database environment, it is created on a root path that is not shared across the partitioned database servers. For example, consider the following command:

```
CREATE DATABASE db_a1 ON /dbpath
```

This command is run under instance ″db2inst″, and creates the database db_a1 on `/dbpath`. Each database partition is created on its local file system under `/dbpath/db2inst/node`*xxxx*, where *xxxx* represents the node number. HACMP failover will attempt to mount the `/dbpath` file system, which is already being used by the other processor. Therefore, the failover script must mount the file system under a different logical point, and set up a symbolic link from that file system to the appropriate `/dpath/db2inst/node`*xxxx* path.

Following is a portion of the `db2nodes.cfg` file, both before and after the failover. In this example, node number 2 is running on processor 1 of the HACMP machine, which has both a host name and a net name of ″node201″. Node number 3 is running on processor 2 of the HACMP machine, which has both a host name and a net name of ″node202″.

```
Before:
        1 node101 0 node101
        2 node201 0 node201    <= HACMP
        3 node202 0 node202    <= HACMP
        4 node301 0 node301

   db2start nodenum 2 restart hostname node202 port 1 netname node202

After:
        1 node101 0 node101
        2 node202 1 node202    <= HACMP
        3 node202 0 node202    <= HACMP
        4 node301 0 node301
```

After the failover, any remote clients trying to directly access node number 2 as the coordinator will have to recatalog the node entry for the database to point to the failover node. Using a mutual failover scenario for coordinator nodes is not recommended. If you require redundancy for your coordinator node, use the hot standby configuration.

Sample script:

```
hacmp-s5.sh
```

Reconnecting after a Failover

If a client uses the SET CLIENT statement to connect to a specific node, and that node moves to a different host during failover, the next connect request from the client will fail. Issue **db2stop**, followed by **db2start** *nodenum* on the node where the SET CLIENT statement was run, and then reissue the statement so that both client and server detect the new physical location of the target node.

Resources

For detailed information about HACMP concepts, installation, and configuration, refer to the following books:

- *HACMP for AIX, Version 4.2: Concepts and Facilities*, SC23-1938
- *HACMP for AIX, Version 4.2: Installation Guide*, SC23-1940
- *HACMP for AIX, Version 4.2: Planning Guide*, SC23-1939.

Part 4. High Availability

Chapter 12. High Availability Cluster Multi-processing, Enhanced Scalability (HACMP ES) for AIX

Enhanced scalability (ES) is a feature of High Availability Cluster Multi-processing (HACMP) for AIX Version 4.2.2, which currently runs only on RS/6000 SP nodes.

This feature provides the same failover recovery as HACMP, and has the same event structure as previous HACMP versions (see *HACMP for AIX, V4.2.2, Enhanced Scalability Installation and Administration Guide*). Enhanced scalability also provides:

- Larger HACMP clusters, with scalability up to 16 nodes per cluster.
- Additional error coverage through *user-defined events*. Monitored areas can trigger user-defined events, which can be as diverse as the death of a process, or the fact that paging space is nearing capacity. Such events include pre- and post-events that can be added to the failover recovery process, if needed. Extra functions that are specific to the different implementations can be placed within the HACMP pre- and post-event streams.

  A *rules file* (`/usr/sbin/cluster/events/rules.hacmprd`) contains the HACMP events. User-defined events are added to this file. The script files that are to be run when events occur are part of this definition.

  For more information about user-defined events and the rules file, see “HACMP ES Event Monitoring and User-defined Events” on page 233.
- HACMP client utilities for monitoring and detecting status changes (in one or more clusters) from AIX physical nodes outside of the HACMP cluster.

The nodes in HACMP ES clusters exchange messages called *heartbeats*, or *keepalive packets*, by which each node informs the other nodes about its availability. A node that has stopped responding causes the remaining nodes in the cluster to invoke recovery. The recovery process is called a *node_down event* and may also be referred to as *failover*. The completion of the recovery process is followed by the re-integration of the node into the cluster. This is called a *node_up event*.

There are two types of events: standard events that are anticipated within the operations of HACMP ES, and user-defined events that are associated with the monitoring of parameters in hardware and software components.

One of the standard events is the node_down event. When planning what should be done as part of the recovery process, HACMP allows two failover options: hot (or idle) standby, and mutual takeover.

Cluster Configuration

In a *hot standby* configuration, the AIX processor node that is the takeover node *is not* running any other workload. In a *mutual takeover* configuration, the AIX processor node that is the takeover node *is* running other workloads.

Generally, DB2 Universal Database Enterprise - Extended Edition (UDB EEE) runs in mutual takeover mode with partitions on each node. One exception is a scenario in which the catalog node is part of a hot standby configuration.

When planning a large DB2 installation on an RS/6000 SP using HACMP ES, you need to consider how to divide the nodes of the cluster within or between the RS/6000 SP frames. Having a node and its backup in different SP frames allows takeover in the event one frame goes down (that is, the frame power/switch board fails). However, such failures are expected to be exceedingly rare, because there are *N*+1 power supplies in each SP frame, and each SP switch has redundant paths, along with *N*+1 fans and power. In the case of a frame failure, manual intervention may be required to recover the remaining frames. This recovery procedure is documented in the SP Administration Guide. HACMP ES provides for recovery of SP node failures; recovery of frame failures is dependent on the proper layout of clusters within one or more SP frames.

Another planning consideration is how to manage big clusters. It is easier to manage a small cluster than a big one; however, it is also easier to manage one big cluster than many smaller ones. When planning, consider how your applications will be used in your cluster environment. If there is a single, large, homogeneous application running, for example, on 16 nodes, it is probably easier to manage the configuration as a single cluster rather than as eight two-node clusters. If the same 16 nodes contain many different applications with different networks, disks, and node relationships, it is probably better to group the nodes into smaller clusters. Keep in mind that nodes integrate into an HACMP cluster one at a time; it will be faster to start a configuration of multiple clusters rather than one large cluster. HACMP ES supports both single and multiple clusters, as long as a node and its backup are in the same cluster.

HACMP ES failover recovery allows pre-defined (also known as *cascading*) assignment of a resource group to a physical node. The failover recovery procedure also allows floating (or *rotating*) assignment of a resource group to a physical node. IP addresses, and external disk volume groups, or file systems, or NFS file systems, and application servers within each resource group specify either an application or an application component, which can be manipulated by HACMP ES between physical nodes by failover and reintegration. Failover and reintegration behavior is specified by the type of resource group created, and by the number of nodes placed in the resource group.

For example, consider a DB2 database partition (logical node). If its log and table space containers were placed on external disks, and other nodes were linked to those disks, it would be possible for those other nodes to access these disks and to restart the database partition (on a takeover node). It is this type of operation that is automated by HACMP. HACMP ES can also be used to recover NFS file systems used by DB2 instance main user directories.

Read the HACMP ES documentation thoroughly as part of your planning for recovery with DB2 UDB EEE. You should read the Concepts, Planning, Installation, and Administration guides, then build the recovery architecture for your environment. For each subsystem that you have identified for recovery, based on known points of failure, identify the HACMP clusters that you need, as well as the recovery nodes (either hot standby or mutual takeover). This is a starting point for completing the HACMP worksheets that are included in the documentation.

It is strongly recommended that both disks and adapters be mirrored in your external disk configuration. For DB2 physical nodes that are configured for HACMP, care is required to ensure that nodes on the volume group can vary from the shared external disks. In a mutual takeover configuration, this arrangement requires some additional planning, so that the paired nodes can access each other's volume groups without conflicts. For DB2 UDB EEE, this means that all container names must be unique across all databases.

One way to achieve uniqueness is to include the partition number as part of the name. You can specify a node expression for container string syntax when creating either SMS or DMS containers. When you specify the expression, the node number can be part of the container name or, if you specify additional arguments, the results of those arguments can be part of the container name. Use the argument ″ $N″ ([blank]$N) to indicate the node expression. The argument must occur at the end of the container string, and can only be used in one of the following forms:

Table 24. Arguments for Creating Containers. The node number is assumed to be five.

Syntax	Example	Value
[blank]$N	″ $N″	5
[blank]$N+[number]	″ $N+1011″	1016
[blank]$N%[number]	″ $N%3″	2
[blank]$N+[number]%[number]	″ $N+12%13″	4
[blank]$N%[number]+[number]	″ $N%3+20″	22

Notes:

1. % is modulus.
2. In all cases, the operators are evaluated from left to right.

Following are some examples of how to create containers using this special argument:

- Creating containers for use on a two-node system.

  ```
  CREATE TABLESPACE TS1 MANAGED BY DATABASE USING
     (device '/dev/rcont $N' 20000)
  ```

  The following containers would be used:

  ```
  /dev/rcont0   - on Node 0
  /dev/rcont1   - on Node 1
  ```

- Creating containers for use on a four-node system.

  ```
  CREATE TABLESPACE TS2 MANAGED BY DATABASE USING
     (file '/DB2/containers/TS2/container $N+100' 10000)
  ```

  The following containers would be used:

  ```
  /DB2/containers/TS2/container100   - on Node 0
  /DB2/containers/TS2/container101   - on Node 1
  /DB2/containers/TS2/container102   - on Node 2
  /DB2/containers/TS2/container103   - on Node 3
  ```

- Creating containers for use on a two-node system.

  ```
  CREATE TABLESPACE TS3 MANAGED BY SYSTEM USING
     ('/TS3/cont $N%2, '/TS3/cont $N%2+2')
  ```

  The following containers would be used:

  ```
  /TS3/cont0   - on Node 0
  /TS3/cont2   - on Node 0
  /TS3/cont1   - on Node 1
  /TS3/cont3   - on Node 1
  ```

Figure 46 on page 217 and Figure 47 on page 218 show an example of a DB2 SSA I/O subsystem configuration, and some of the planning necessary to ensure both a highly available external disk configuration, and the ability to access all volume groups without conflict.

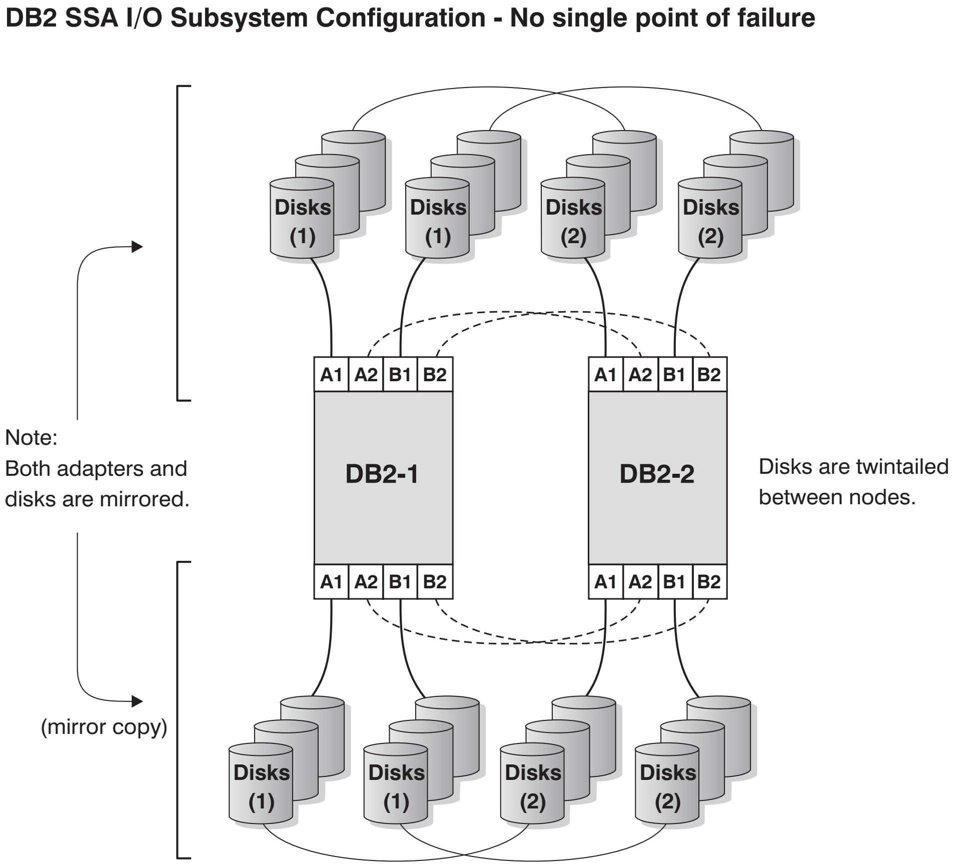

DB2 SSA I/O Subsystem Configuration - No single point of failure
Disks (1)
Disks (1)
Disks (2)
Disks (2)
A1
A2
B1
B2
DB2-1
DB2-2
Note:
Both adapters and disks are mirrored.
Disks are twintailed between nodes.
(mirror copy)

Figure 46. No Single Point of Failure

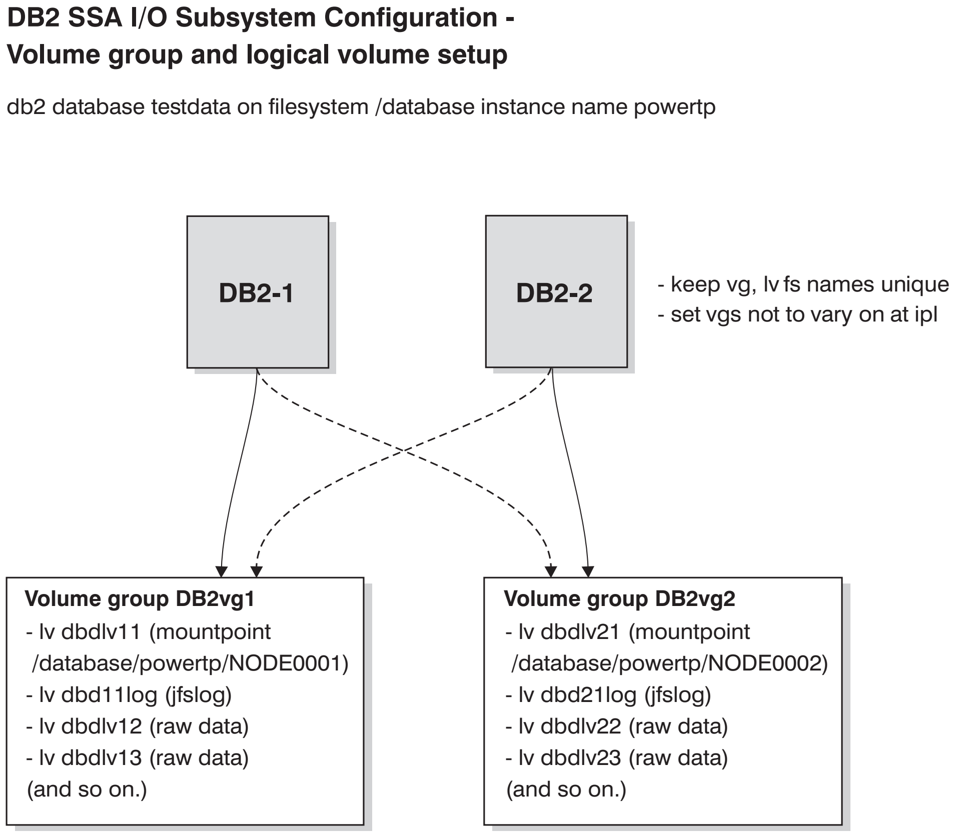

DB2 SSA I/O Subsystem Configuration -
Volume group and logical volume setup
db2 database testdata on filesystem /database instance name powertp
DB2-1
DB2-2
- keep vg, lv fs names unique
- set vgs not to vary on at ipl
Volume group DB2vg1
- lv dbdlv11 (mountpoint /database/powertp/NODE0001)
- lv dbd11log (jfslog)
- lv dbdlv12 (raw data)
- lv dbdlv13 (raw data)
(and so on.)
Volume group DB2vg2
- lv dbdlv21 (mountpoint /database/powertp/NODE0002)
- lv dbd21log (jfslog)
- lv dbdlv22 (raw data)
- lv dbdlv23 (raw data)
(and so on.)

Figure 47. Volume Group and Logical Volume Setup

Configuring a DB2 Database Partition

Once configured, each database partition in an instance is started by HACMP ES, one physical node at a time. Multiple clusters are recommended for starting parallel DB2 configurations that are larger than four nodes. Note that in a 64-node parallel DB2 configuration, it is faster to start 32 two-node HACMP clusters in parallel, than four 16-node clusters.

A script file, `rc.db2pe`, is packaged with DB2 UDB EEE (and installed on each node in `/usr/bin`) to assist in configuring for HACMP ES failover or recovery in either hot standby or mutual takeover nodes. In addition, DB2 buffer pool sizes can be customized during failover in mutual takeover configurations from within `rc.db2pe`. (Buffer pool sizes need to be configured to ensure proper performance when two database partitions run on one physical node.)

When you create an application server in an HACMP configuration of a DB2 database partition, specify `rc.db2pe` as a start and stop script as follows:

```
/usr/bin/rc.db2pe <instance> <dpn> <secondary dpn> start <use switch>
/usr/bin/rc.db2pe <instance> <dpn> <secondary dpn> stop <use switch>

where:

<instance> is the instance name.
<dpn> is the database partition number.
<secondary dpn> is the "companion" database partition number in
   mutual takeover configurations only; in hot standby configurations,
   it is the same as <dpn>.
<use switch> is usually blank; when blank, it indicates that
   the SP switch network is used for the hostname field
   in the db2nodes.cfg file (all traffic for DB2 is routed over the SP switch);
   if not blank, the name used is the host name of the SP node to be used.
```

The DB2 command LIST DATABASE DIRECTORY is used from within `rc.db2pe` to find all databases configured for this database partition. The script file then looks for the `/usr/bin/reg.parms.`*DATABASE* file and the `/usr/bin/failover.parms.`*DATABASE* file, where *DATABASE* is each of the databases configured for this database partition. In a mutual takeover configuration, it is recommended that you create the parameter files `reg.parms.`*xxx* and `failover.parms.`*xxx*. In the `failover.parms.`*xxx* file, the settings for BUFFPAGE, DBHEAP, and any others affecting buffer pools, should be adjusted to account for the possibility of more than one buffer pool. Sample files `reg.parms.SAMPLE` and `failover.parms.SAMPLE` are provided for your use.

One of the important parameters in this environment is the *start_stop_time* database manager configuration parameter, which has a default value of 10 minutes. However, `rc.db2pe` sets this parameter to 2 minutes. You should set this parameter through `rc.db2pe` to a value of 10 minutes, or slightly more. In this context, the specified duration is the time interval between the failure of the partition, and its recovery. If applications running on a partition are issuing frequent COMMITs, 10 minutes following failure on a database partition should be sufficient time to roll back uncommitted transactions and to reach a point of consistency in the database on that partition. If your workload is heavy, or you have many partitions, you may need to increase the duration to decrease the probability of timeouts occurring before the rollback operation completes.

Following is an example of a hot standby configuration and a mutual takeover configuration. In both examples, the resource groups contain a Service IP switch alias address. This switch alias address is used for:

- NFS access to a file server for the DB2 instance owner file systems

- Other client access that needs to be maintained in the case of a failover, TSM (Tivoli Storage Manager, formerly ADSM) connection, or other similar operation.

If your implementation does not require these aliases, they can be removed. If removed, be sure to set the *MOUNT_NFS* parameter to `NO` in the `rc.db2pe` script file.

Example of a Hot Standby Configuration

The assumption in this example is that a hot standby configuration exists between physical nodes 1 and 2, and that the DB2 instance name is POWERTP. The database partition is 1, and the database is TESTDATA, residing on file system `/database`.

```
Resource group name: db2_dp_1
Node Relationship: cascading
Participating nodenames: node1_eth, node2_eth
Service_IP_label: nfs_switch_1     (<<< this is the switch alias address)
Filesystems: /database/powertp/NODE0001
Volume Groups: DB2vg1
Application Servers: db2_dp1_app
Application Server Start Script: /usr/bin/rc.db2pe powertp 1 1 start
Application Server Stop Script: /usr/bin/rc.db2pe powertp 1 1 stop
```

Example of a Mutual Takeover Configuration

The assumption in this example is that a mutual takeover configuration exists between physical nodes 1 and 2, and that the DB2 instance name is POWERTP. The database partitions are 1 and 2, and the database is TESTDATA, residing on file system `/database`.

```
Resource group name: db2_dp_1
Node Relationship: cascading
Participating nodenames: node1_eth, node2_eth
Service_IP_label: nfs_switch_1     (<<< this is the switch alias address)
Filesystems: /database/powertp/NODE0001
Volume Groups: DB2vg1
Application Servers: db2_dp1_app
Application Server Start Script: /usr/bin/rc.db2pe powertp 1 2 start
Application Server Stop Script: /usr/bin/rc.db2pe powertp 1 2 stop

Resource group name: db2_pd_2
Node Relationship: cascading
Participating nodenames: node2_eth, node1_eth
Service_IP_label: nfs_switch_2     (<<< this is the switch alias address)
Filesystems: /database/powertp/NODE0002
Volume Groups: DB2vg2
Application Servers: db2_dp2_app
Application Server Start Script: /usr/bin/rc.db2pe powertp 2 1 start
Application Server Stop Script: /usr/bin/rc.db2pe powertp 2 1 stop
```

Configuration of an NFS Server Node

The `rc.db2pe` script can also be used to make available NFS-mounted directories of DB2 parallel instance user directories. This can be accomplished by setting the *MOUNT_NFS* parameter to YES in the rc.db2pe script file, and configuring the NFS failover server pair as follows:

- Configure the home directory and export it as ″root″ using /etc/exports and the **exportfs** command to the IP address used on the nodes in the same subnet as the NFS server's IP address. Include both the HACMP boot and service addresses. The NFS server's IP address is the same address as the service address in HACMP, and which can be taken over by a backup. The home directory of the DB2 instance owner should be NFS-mounted directly, not automounted. (The use of the automounter is not supported by the scripts as a DB2 instance owner home directory.)
- Using SMIT or a bottom-line configuration, create a separate /etc/filesystems entry for this file system, so that all nodes in the DB2 parallel grouping, including the file server, can mount using the NFS file system command.

  For example, an /nfshome JFS file system can be exported to all nodes as /dbhome. Each node creates an NFS file system /dbname, which is nfs_server:/nfshome. Therefore, the home directory of the DB2 instance owner would be /dbhome/powertp if the instance name is ″powertp″.

  Ensure that the NFS parameters for the mount in /etc/filesystems are ″hard″, ″bg″, ″intr″, and ″rw″.
- Ensure that the DB2 instance owner definitions associated with the home directory /dbhome/powertp in /etc/passwd are the same on all nodes.

  The user definitions in an SP environment are typically created on the control workstation, and ″supper″ or ″pcp″ is used to distribute /etc/passwd, /etc/security/passwd, /etc/security/user, and /etc/security/group to all nodes.
- Do *not* configure the ″nfs_filesystems to export″ in HACMP resource groups for the volume group and the file system that is exported. Instead, configure it normally to NFS. The scripts for the NFS server will control the exporting of the file systems.
- Ensure that the major number of the volume group where the file system resides is the same on both the primary node and the takeover node. This is accomplished by using **importvg** with the -V option.
- Verify that the *MOUNT_NFS* parameter is set to YES in the rc.db2pe script file, and that each node has the NFS file system to mount in /etc/filesystems. If this is not the case, rc.db2pe will not be able to mount the file system and start DB2.
- If the DB2 instance owner was already created, and you are copying the user's directory structure to the file system you are creating, ensure that you **tar** (-cvf) the directory. This ensures the preservation of symbolic links.
- Do not forget to mirror both the adapters and the disks for the logical volumes, and the file system logs of the file system you are creating.

Example of an NFS Server Takeover Configuration

The assumption in this example is that there is an NFS server file system /nfshome in the volume group nfsvg over the IP address ″nfs_server″. The DB2 instance name is POWERTP, and the home directory is /dbhome/powertp.

```
   Resource group name: nfs_server
   Node Relationship: cascading
   Participating nodenames: node1_eth, node2_eth
   Service_IP_label: nfs_server     (<<< this is the switch alias address)
   Filesystems: /nfshome
   Volume Groups: nfsvg
   Application Servers: nfs_server_app
   Application Server Start Script: /usr/bin/rc.db2pe powertp NFS SERVER start
  Application Server Stop Script: /usr/bin/rc.db2pe powertp NFS SERVER stop
```

In this example:

- /etc/filesystems on all nodes would contain an entry for /dbhome as mounting nfs_server:/nfshome. nfs_server is a Service IP switch alias address.
- /etc/exports on the nfs_server node and the backup node would include the boot and service addresses, and contain an entry for /nfsfs -root=nfs_switch_1, nfs_switch_2, ...

Considerations When Configuring the SP Switch

When implementing HACMP ES with the SP switch, consider the following:

- There are ″base″ and ″alias″ addresses on the SP switch. The base addresses are those defined in the SP System Data Repository (SDR), and are configured by rc.switch when the system is ″booted″. The alias addresses are IP addresses configured, in addition to the base address, into the css0 interface through the **ifconfig** command with an alias attribute. For example:

  ```
  ifconfig css0 inet alias sw_alias_1 up
  ```
- When configuring the DB2 db2nodes.cfg file, SP switch ″base″ IP address names should be used for both the ″hostname″ and the ″netname″ field. Switch IP address aliases are *only* used to maintain NFS connectivity. DB2 failover is achieved by restarting DB2 with the **db2start** (RESTART) command (which updates db2nodes.cfg).
- Do not confuse the switch addresses with the etc/hosts aliases. Both the SP switch addresses and the SP switch alias addresses are real in either etc/hosts or DNS. The switch alias addresses are not another name for the SP switch base address; each has its own separate address.
- The SP switch base addresses are always present on a node when it is up. HACMP ES does not configure or move these addresses between nodes.
- If you intend to use SP switch alias addresses, configure these to HACMP as boot and service addresses for ″heartbeating″ and IP address takeover. If you do not intend to use SP switch alias addresses, configure the base SP

switch address to HACMP as a service address for ″heartbeating″ *only* (no IP address takeover). Do not, in any configuration, configure alias addresses *and* the switch base address; this configuration is not supported by HACMP ES.

- Only the SP switch alias addresses (and not the SP switch base addresses) are moved between nodes for an IP takeover configuration.
- The need for SP switch aliases arises because there can only be one SP switch adapter per node. Using alias addresses allows a node to take over another node's switch alias IP address without adding another switch adapter. This is useful in nodes that are ″slot-constrained″. For more information about handling recovery from SP switch adapter failures, see the network failure section under "HACMP ES Script Files" on page 237.
- If you configure the SP switch for IP address takeover, you will need to create two extra alias IP addresses per node: one as a boot address and one as a service address.
- Do not forget to use ″HPS″ in the HACMP ES network name definition for an SP switch base IP address or an SP switch alias IP address.
- `rc.cluster` in HACMP automatically **ifconfig**s in the SP switch boot address when HACMP is started. No additional configuration is required, other than creating the IP address and name, and defining them to HACMP.
- The Eprimary node of the SP switch is the server that implements the Estart, Efence, and Eunfence commands. The HACMP scripts attempt to Eunfence or to Estart a node when HACMP is started, and make the switch available if it is defined as one of its networks. For this reason, ensure that the Eprimary node is available when you start HACMP. The HACMP code waits up to 12 minutes for an Eprimary failover to complete before it exits with an error.
- The Eprimary node of the SP switch is moved between nodes by the SP Parallel System Support Program (PSSP), and not HACMP. If an Eprimary node goes offline, the PSSP automatically has a backup node assume responsibility as the Eprimary node. The switch network is unaffected by this change and remains up.

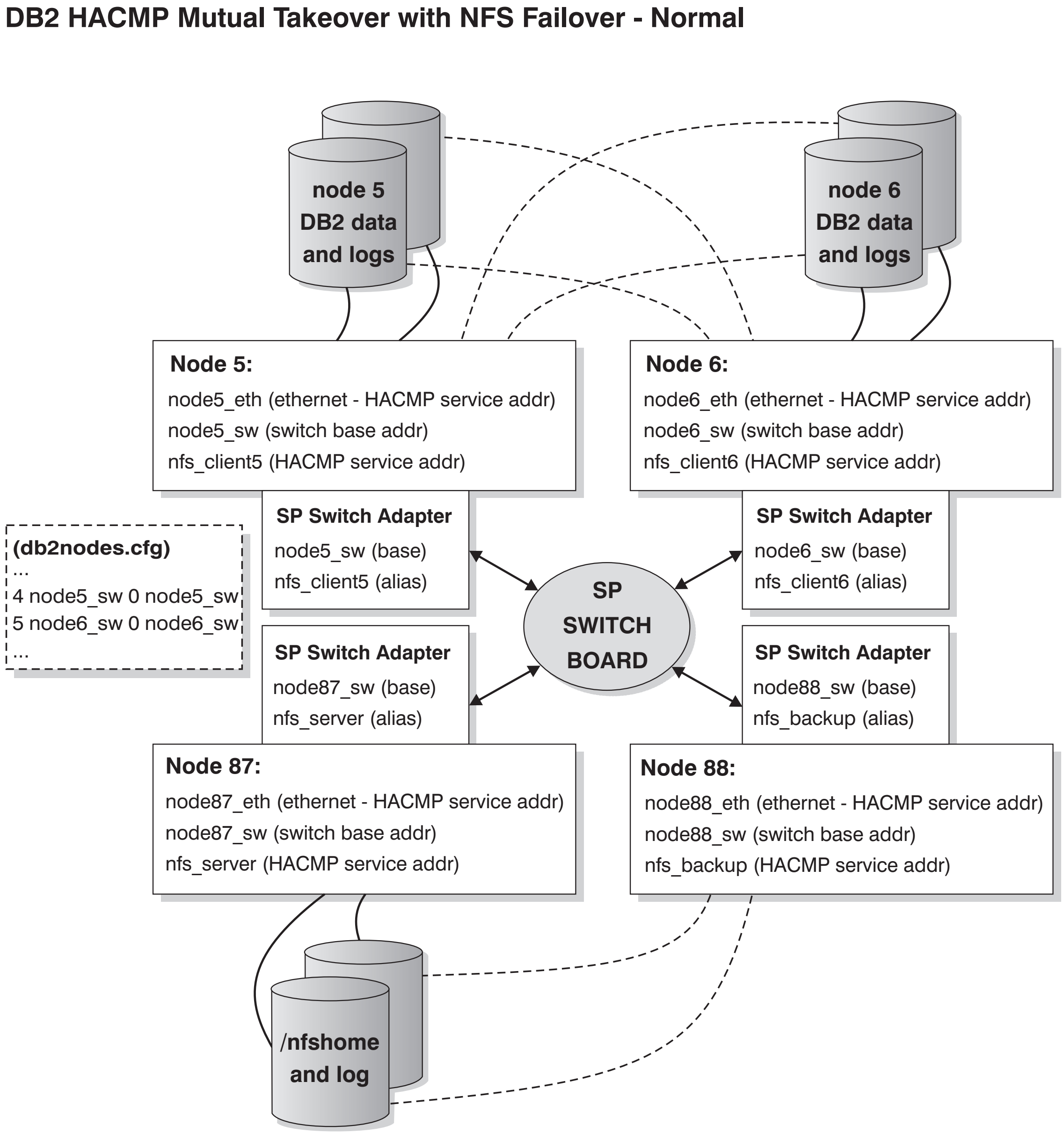
DB2 HACMP Configuration Examples

The following examples illustrate different failover support configurations and show what happens when failure occurs.

In the case of DB2 HACMP mutual takeover configurations (Figure 48 on page 225, Figure 49 on page 226, and Figure 50 on page 227):

- HACMP adapters are defined for ethernet, and SP switch alias boot and service aliases — base addresses are untouched. Remember to use an ″HPS″ string in the HACMP network name.

- The `NFS_server/nfshome` is mounted as `/dbhome` on all nodes through switch aliases.
- The `db2nodes.cfg` file contains SP switch base addresses. The `db2nodes.cfg` file is changed by the **db2start** (RESTART) command after a DB2 database partition (logical node) failover.
- The SP switch alias boot addresses are not shown.
- Nodes can be in different SP frames.


DB2 HACMP Mutual Takeover with NFS Failover - Normal
node 5
DB2 data
and logs
node 6
DB2 data
and logs
Node 5:
node5_eth (ethernet - HACMP service addr)
node5_sw (switch base addr)
nfs_client5 (HACMP service addr)
Node 6:
node6_eth (ethernet - HACMP service addr)
node6_sw (switch base addr)
nfs_client6 (HACMP service addr)
SP Switch Adapter
node5_sw (base)
nfs_client5 (alias)
SP Switch Adapter
node6_sw (base)
nfs_client6 (alias)
SP
SWITCH
BOARD
SP Switch Adapter
node87_sw (base)
nfs_server (alias)
SP Switch Adapter
node88_sw (base)
nfs_backup (alias)
Node 87:
node87_eth (ethernet - HACMP service addr)
node87_sw (switch base addr)
nfs_server (HACMP service addr)
Node 88:
node88_eth (ethernet - HACMP service addr)
node88_sw (switch base addr)
nfs_backup (HACMP service addr)
/nfshome
and log

```
(db2nodes.cfg)
...
4 node5_sw 0 node5_sw
5 node6_sw 0 node6_sw
...
```

Figure 48. Mutual Takeover with NFS Failover - Normal

DB2 HACMP Mutual Takeover with NFS Failover - NFS failover

- nfs_server SP Switch alias IP addr and nfs mounted /nfshome moved from node 87 to 88.
- SP switch arp code has functionality to update all switch arp caches with this move.

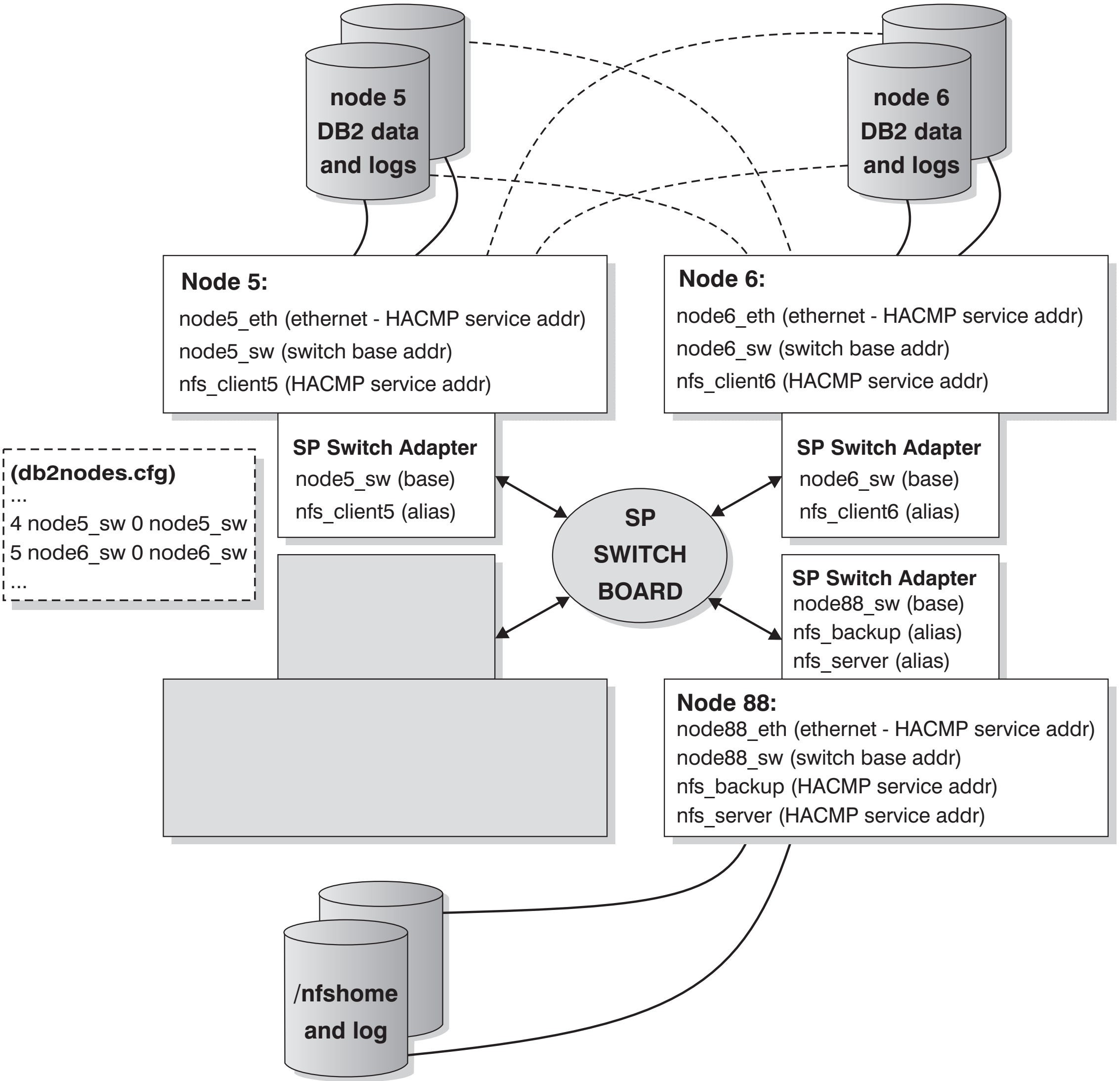
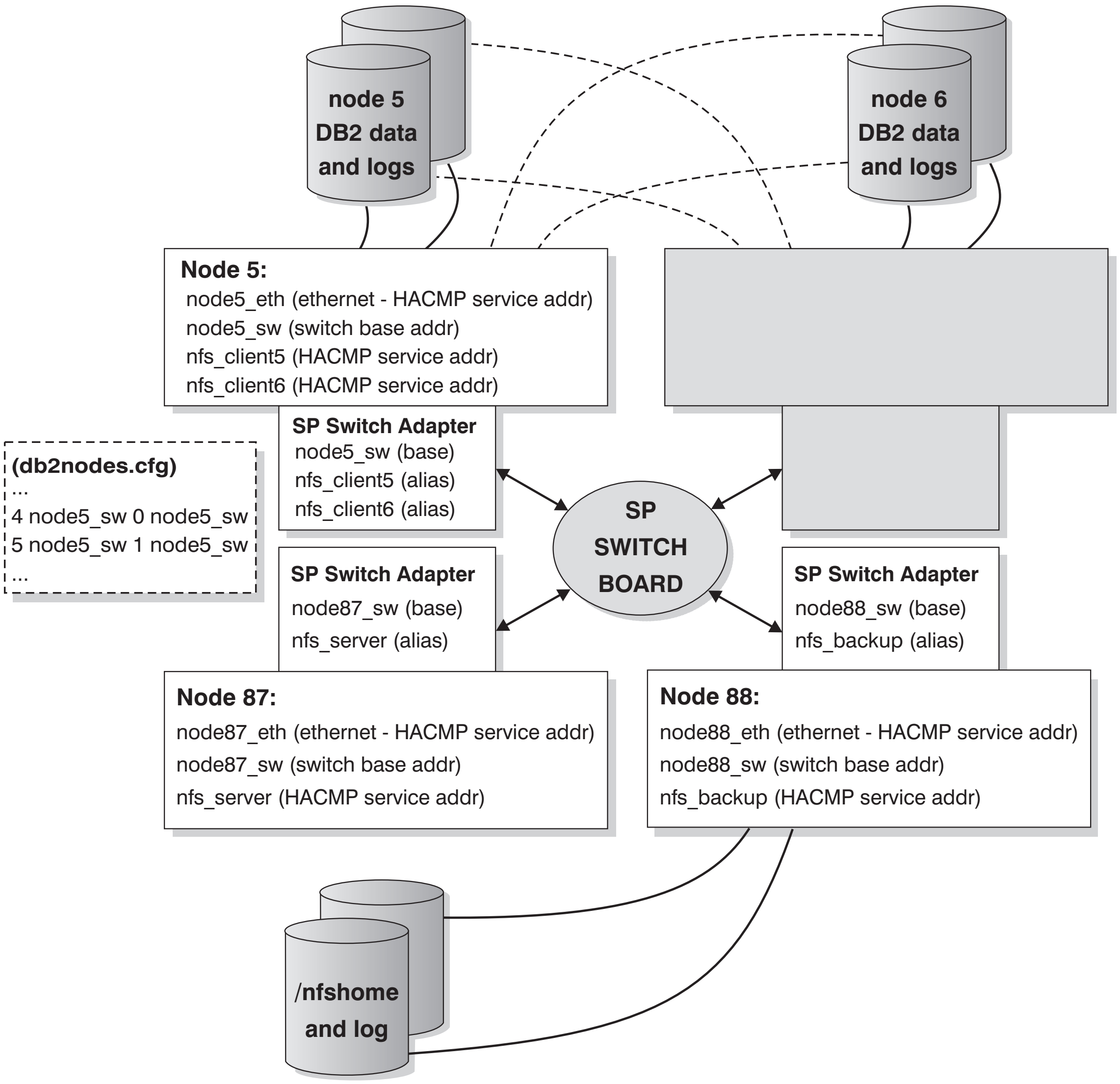

node 5
DB2 data
and logs
node 6
DB2 data
and logs
Node 5:
node5_eth (ethernet - HACMP service addr)
node5_sw (switch base addr)
nfs_client5 (HACMP service addr)
Node 6:
node6_eth (ethernet - HACMP service addr)
node6_sw (switch base addr)
nfs_client6 (HACMP service addr)
SP Switch Adapter
node5_sw (base)
nfs_client5 (alias)
SP Switch Adapter
node6_sw (base)
nfs_client6 (alias)
(db2nodes.cfg)
...
4 node5_sw 0 node5_sw
5 node6_sw 0 node6_sw
...
SP
SWITCH
BOARD
SP Switch Adapter
node88_sw (base)
nfs_backup (alias)
nfs_server (alias)
Node 88:
node88_eth (ethernet - HACMP service addr)
node88_sw (switch base addr)
nfs_backup (HACMP service addr)
nfs_server (HACMP service addr)
/nfshome
and log

Figure 49. Mutual Takeover with NFS Failover - NFS Failover

DB2 HACMP Mutual Takeover with NFS Failover - DB2 failover

- switch IP address takeover allows other servers (like ADSM) to retain connectivity.
- Node 5 runs 2 logical nodes of DB2.


node 5 DB2 data and logs
node 6 DB2 data and logs
Node 5:
node5_eth (ethernet - HACMP service addr)
node5_sw (switch base addr)
nfs_client5 (HACMP service addr)
nfs_client6 (HACMP service addr)
SP Switch Adapter
node5_sw (base)
nfs_client5 (alias)
nfs_client6 (alias)
SP SWITCH BOARD
SP Switch Adapter
node87_sw (base)
nfs_server (alias)
SP Switch Adapter
node88_sw (base)
nfs_backup (alias)
Node 87:
node87_eth (ethernet - HACMP service addr)
node87_sw (switch base addr)
nfs_server (HACMP service addr)
Node 88:
node88_eth (ethernet - HACMP service addr)
node88_sw (switch base addr)
nfs_backup (HACMP service addr)
/nfshome and log

```
(db2nodes.cfg)
...
4 node5_sw 0 node5_sw
5 node5_sw 1 node5_sw
...
```

Figure 50. Mutual Takeover with NFS Failover - DB2 Failover

In the case of DB2 HACMP hot standby configurations (Figure 51 on page 229 and Figure 52 on page 230):

- HACMP adapters are defined for ethernet, and SP switch alias boot and service aliases — base addresses are untouched. Remember to use an ″HPS″ string in the HACMP network name.
- The `NFS_server/nfshome` is mounted as `/dbhome` on all nodes through switch aliases.
- The `db2nodes.cfg` file contains SP switch base addresses. The `db2nodes.cfg` file is changed by the **db2start** (RESTART) command after a DB2 database partition (logical node) failover.
- The SP switch alias boot addresses are not shown.

DB2 HACMP Hot Standby with NFS Failover - Normal

Note: Hot Standby node can back up more than one node, depending on disk cabling.

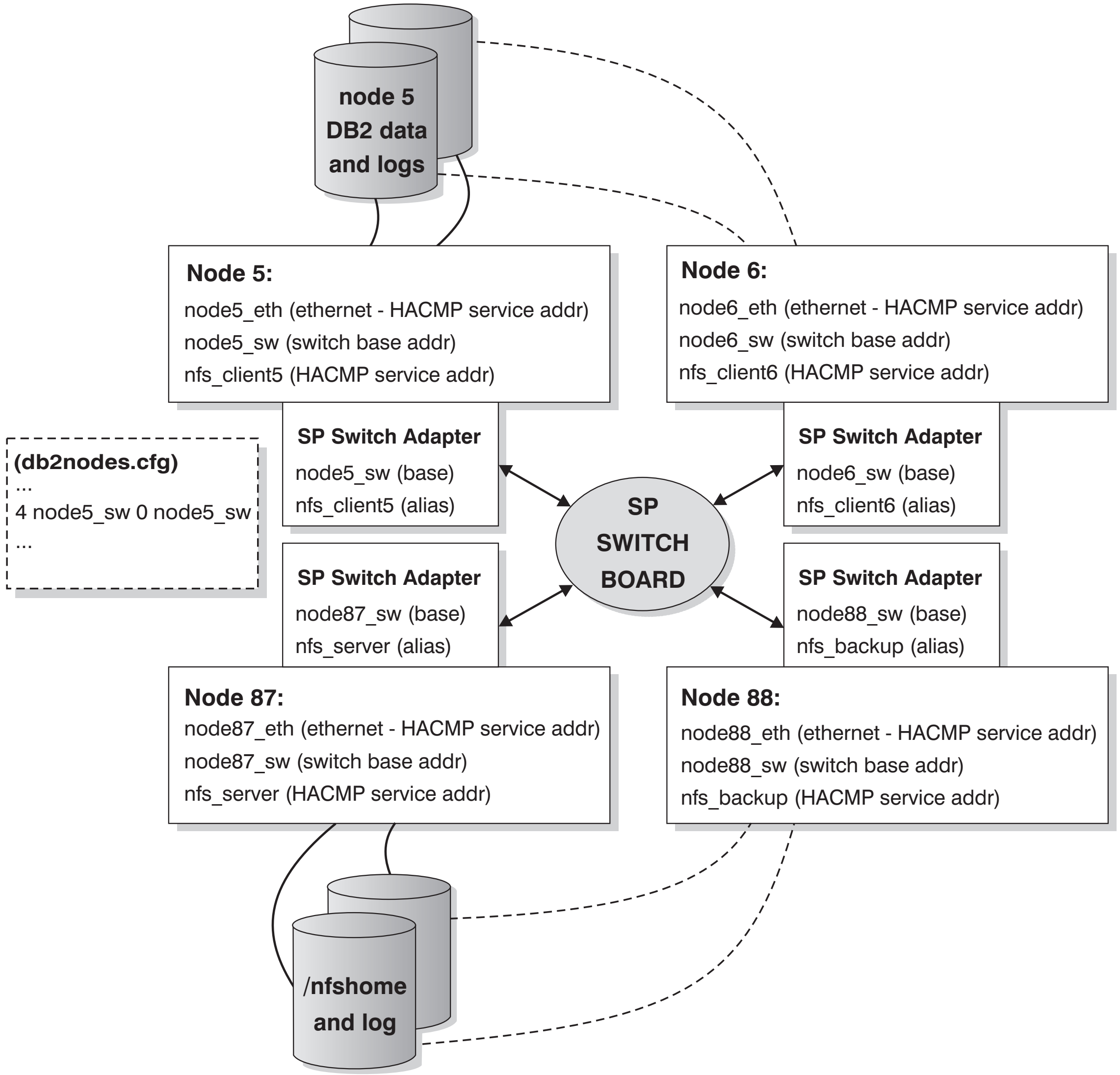
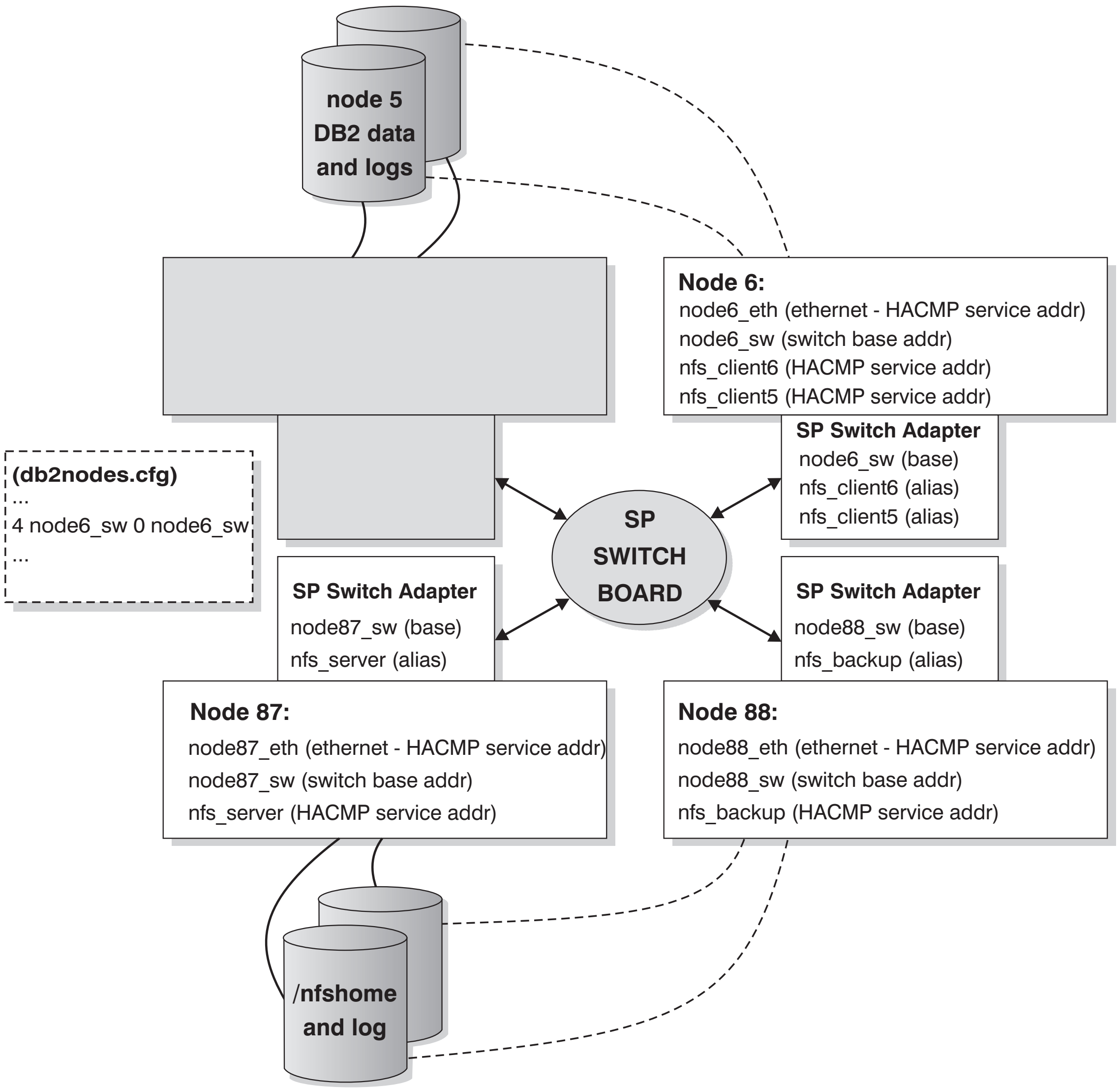

node 5
DB2 data
and logs
Node 5:
node5_eth (ethernet - HACMP service addr)
node5_sw (switch base addr)
nfs_client5 (HACMP service addr)
Node 6:
node6_eth (ethernet - HACMP service addr)
node6_sw (switch base addr)
nfs_client6 (HACMP service addr)
(db2nodes.cfg)
...
4 node5_sw 0 node5_sw
...
SP Switch Adapter
node5_sw (base)
nfs_client5 (alias)
SP Switch Adapter
node6_sw (base)
nfs_client6 (alias)
SP
SWITCH
BOARD
SP Switch Adapter
node87_sw (base)
nfs_server (alias)
SP Switch Adapter
node88_sw (base)
nfs_backup (alias)
Node 87:
node87_eth (ethernet - HACMP service addr)
node87_sw (switch base addr)
nfs_server (HACMP service addr)
Node 88:
node88_eth (ethernet - HACMP service addr)
node88_sw (switch base addr)
nfs_backup (HACMP service addr)
/nfshome
and log

Figure 51. Hot Standby with NFS Failover - Normal

DB2 HACMP Hot Standby with NFS Failover- DB2 Failover

Note: Hot Standby node can back up more than one node, depending on disk cabling.



node 5
DB2 data
and logs
Node 6:
node6_eth (ethernet - HACMP service addr)
node6_sw (switch base addr)
nfs_client6 (HACMP service addr)
nfs_client5 (HACMP service addr)
SP Switch Adapter
node6_sw (base)
nfs_client6 (alias)
nfs_client5 (alias)
(db2nodes.cfg)
...
4 node6_sw 0 node6_sw
...
SP
SWITCH
BOARD
SP Switch Adapter
node87_sw (base)
nfs_server (alias)
SP Switch Adapter
node88_sw (base)
nfs_backup (alias)
Node 87:
node87_eth (ethernet - HACMP service addr)
node87_sw (switch base addr)
nfs_server (HACMP service addr)
Node 88:
node88_eth (ethernet - HACMP service addr)
node88_sw (switch base addr)
nfs_backup (HACMP service addr)
/nfshome
and log

Figure 52. Hot Standby with NFS Failover - DB2 Failover

In the case of DB2 HACMP mutual takeover without NFS failover configurations (Figure 53 on page 231 and Figure 54 on page 232):

- HACMP adapters are defined for ethernet, and SP switch base addresses. Remember that when base addresses are configured to HACMP as service addresses, there is no boot address (only a ″heartbeat″). Do not forget to use an ″HPS″ string in the HACMP network name for the SP switch.

- The `db2nodes.cfg` file contains SP switch base addresses. The `db2nodes.cfg` file is changed by the **db2start** (RESTART) command after a DB2 database partition (logical node) failover.
- No NFS failover functions are shown.
- Nodes can be in different SP frames.

DB2 HACMP Mutual Takeover without NFS Failover - Normal

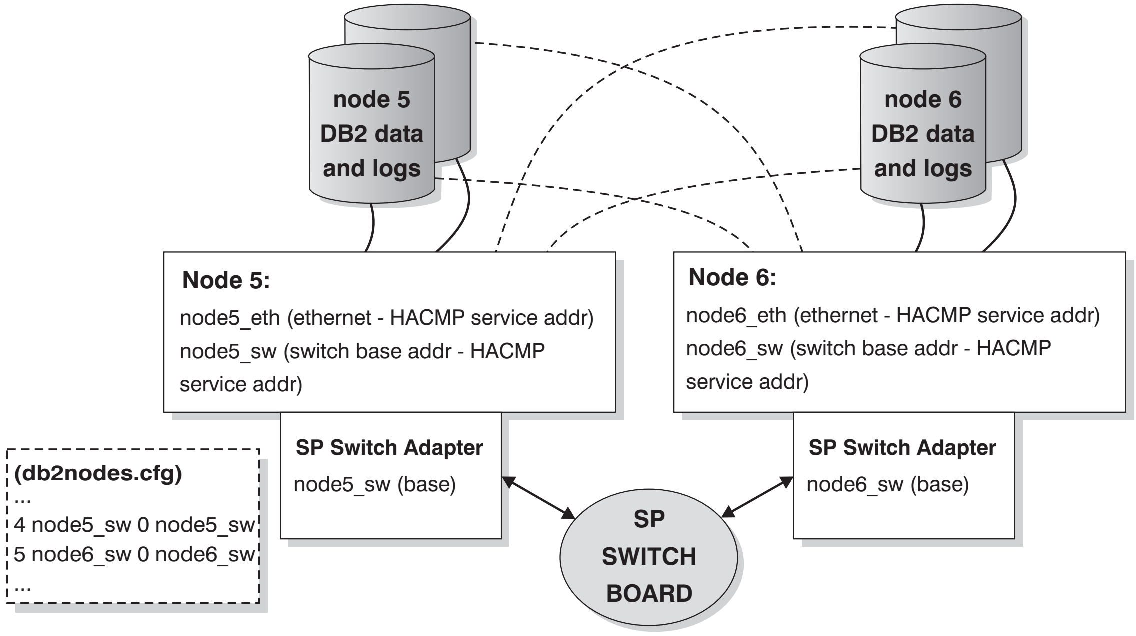
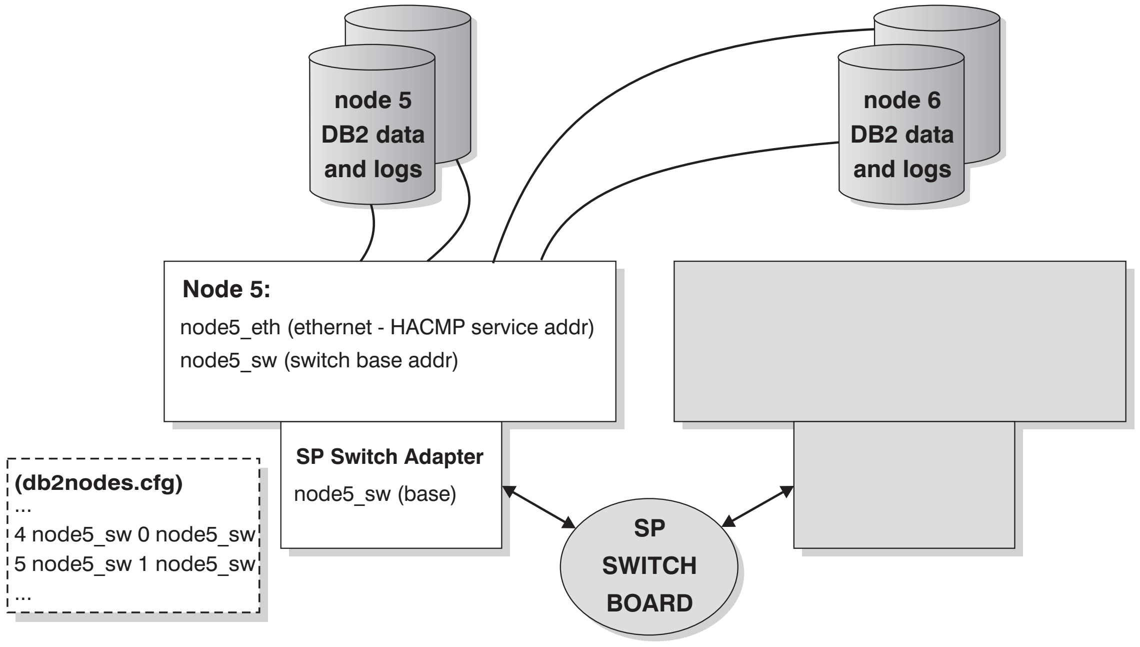


Figure 53. Mutual Takeover without NFS Failover - Normal

DB2 HACMP Mutual Takeover without NFS Failover - DB2 failover

- Node 5 runs 2 logical nodes of DB2.



```
(db2nodes.cfg)
...
4 node5_sw 0 node5_sw
5 node5_sw 1 node5_sw
...
```

Figure 54. Mutual Takeover without NFS Failover - DB2 Failover

DB2 HACMP Startup Recommendations

It is recommended that you do not specify that HACMP is to be started at boot time in `/etc/inittab`. HACMP should be started manually after the nodes are booted. This allows for non-disruptive maintenance of a failed node.

As an example of ″disruptive maintenance″, consider the case in which a node has a hardware failure and crashes. Failover is initiated automatically by HACMP, and recovery completes successfully. However, the failed node needs to be fixed. If HACMP was configured in `/etc/inittab` to start on reboot, this node will try to reintegrate after boot completion, which is not desirable in this case.

For ″non-disruptive maintenance″, consider manually starting HACMP on each node. In this way, failed nodes can be fixed and reintegrated without affecting the other nodes. The `ha_cmd` script is provided for controlling HACMP start and stop commands from the control workstation.

Note: When creating a DB2 instance for the first time, the following entry is appended to the /etc/inittab file:

```
rcdb2:2:once:/etc/rc.db2 > /dev/console 2>&1 # Autostart DB2 Services
```

If HACMP or HACMP ES is enabled, update the /etc/inittab file by placing the above line before the HACMP entry. Following is a sample HACMP entry in the /etc/inittab file:

```
clinit:a:wait:touch /usr/sbin/cluster/.telinit # HACMP for AIX
```

The entry must be the last entry in the /etc/inittab file.

HACMP ES Event Monitoring and User-defined Events

Shutting down DB2 database partitions on an AIX physical node when paging space reaches a certain percentage of fullness, and restarting a DB2 database partition, or initiating a failover operation if a process dies on a given node, are two examples of user-defined events. Examples that illustrate user-defined events, such as shutting down a database partition and forcing a transaction abort to free paging space, can be found in the samples subdirectory.

A rules file, /user/sbin/cluster/events/rules.hacmprd, contains HACMP events. Each event description in this file has the following nine components:

- Event name, which must be unique.
- State, or qualifier for the event. The event name and state are the rule triggers. HACMP ES Cluster Manager initiates recovery only if it finds a rule with a trigger corresponding to the event name and state.
- Resource program path, a full-path specification of the *xxx*.rp file containing the recovery program.
- Recovery type. This is reserved for future use.
- Recovery level. This is reserved for future use.
- Resource variable name, which is used for Event Manager events.
- Instance vector, which is used for Event Manager events. This is a set of elements of the form ″name=value″. The values uniquely identify the copy of the resource in the system and, by extension, the copy of the resource variable.
- Predicate, which is used for Event Manager events. This is a relational expression between a resource variable and other elements. When this expression is true, the Event Management subsystem generates an event to notify the Cluster Manager and the appropriate application.
- Rearm predicate, which is used for Event Manager events. This is a predicate used to generate an event that alters the status of the primary

predicate. This predicate is typically the inverse of the primary predicate. It can also be used with the event predicate to establish an upper and a lower boundary for a condition of interest.

Each object requires one line in the event definition, even if the line is not used. If these lines are removed, HACMP ES Cluster Manager cannot parse the event definition properly, and this may cause the system to hang. Any line beginning with ″#″ is treated as a comment line.

Note: The rules file requires exactly nine lines for each event definition, not counting any comment lines. When adding a user-defined event at the bottom of the rules file, it is important to remove the unnecessary empty line at the end of the file, or the node will hang.

Following is an example of an event definition for `node_up`:

```
##### Beginning of the Event Definition: node_up
#
TE_JOIN_NODE
0
/usr/sbin/cluster/events/node_up.rp
2
0
# 6) Resource variable - only used for event management events

# 7) Instance vector - only used for event management events

# 8) Predicate  - only used for event management events

# 9) Rearm predicate - only used for event management events

###### End of the Event Definition: node_up
```

This example is just one of the event definitions that can be found in the `rules.hacmprd` file. In this example, the recovery program `/usr/sbin/cluster/events/node_up.rp` is invoked when the `node_up` event occurs. Values are specified for the state, recovery type, and recovery level. There are four empty lines for resource variable, instance variable, predicate, and rearm predicate.

You can define other events to react to non-standard HACMP ES events. For example, to define the event that the `/tmp` file system is over 90 per cent full, the `rules.hacmprd` file must be modified.

Many events are predefined in the IBM Parallel System Support Program (PSSP). These events can be exploited (when used within user-defined events) as follows:

1. Stop the cluster.

2. Edit the `rules.hacmprd` file. Back up the file before modifying it. Add the predefined PSSP event manually. If you need synchronizing points across all nodes in the cluster, use the **barrier** command in the recovery program. (Read more about the barrier command, and synchronization of recovery programs in the HACMP Concepts, Installation, and Administration Guides.)
3. Restart the cluster. The `rules.hacmprd` file is stored in memory when Cluster Manager is started. To accurately implement the changes, restart all the clusters. There should not be any inconsistent rules in a cluster.
4. Cluster Manager uses all events in the `rules.hacmprd` file.

HACMP ES uses PSSP event detection to treat user-defined events. The PSSP Event Management subsystem provides comprehensive event detection by monitoring various hardware and software resources.

Resource states are represented by resource variables. Resource conditions are represented as expressions called predicates.

Event Management receives resource variables from the Resource Monitor, which observes the state of specific system resources and transforms this state into several resource variables. These variables are periodically passed to Event Management. Event Management applies predicates that are specified by the HACMP ES Cluster Manager in `rules.hacmprd` for each resource variable. When the predicate is evaluated as being true, an event is generated and sent to the Cluster Manager. Cluster Manager initiates the voting protocol, and the recovery program file (`xxx.rp`) is run (according to event priority) on a set of nodes specified by ″node sets″ in the recovery program.

The recovery program file (*`xxx`*`.rp`) is made up of one or more recovery program lines. Each line is declared in the following format:

```
relationship     command_to_run     expected_status     NULL
```

There must be at least one space between each value in the line. ″Relationship″ is a value used to decide which program should run on which kind of node. Three types of relationship are supported:

- All. The specified command or program is run on all nodes of the current HACMP cluster.
- Event. The specified command or program is run only on the nodes on which the event occurred.
- Other. The specified command or program is run on all nodes on which the event did *not* occur.

″Command_to_run″ is a quotation mark-delimited string, with or without a full-path specification to an executable program. Only HACMP-delivered event scripts can use a relative-path definition. Other scripts or programs must use the full-path specification, even if they are located in the same directory as the HACMP event scripts.

″Expected_states″ is the return code of the specified command or program. It is either an integer value, or an ″x″. If ″x″ is used, Cluster Manager does not care about the return code. All other codes must be equal to the expected return code, otherwise Cluster Manager detects the event failure. The handling of this event ″hangs″ the process until recovery (through manual intervention) occurs. Without manual intervention, the node does not synchronize with the other nodes. Synchronization across all nodes is required for the Cluster Manager to control all the nodes.

″NULL″ is a field reserved for future use. The word ″NULL″ must appear at the end of each line except the barrier line. If you specify multiple recovery commands between two barrier commands, or before the first one, the recovery commands are run in parallel on the node itself, and between the nodes.

The barrier command is used to synchronize all the commands across all the cluster nodes. When a node hits the barrier statement in the recovery program, Cluster Manager initiates the barrier protocol on this node. Since the barrier protocol is a two-phase protocol, all nodes are notified that both phases have completed when all of the nodes have met the barrier in the recovery program, and ″voted″ to approve the protocol.

The process can be summarized as follows:

1. Either Group Services/ES (for predefined events) or Event Management (for user-defined events) notifies HACMP ES Cluster Manager of the event.
2. Cluster Manager reads the `rules.hacmprd` file, and determines the recovery program that is mapped to the event.
3. Cluster Manager runs the recovery program, which consists of a sequence of recovery commands.
4. The recovery program executes the recovery commands, which may be shell scripts or binary commands. (In HACMP for AIX, the recovery commands are the same as the HACMP event scripts.)
5. Cluster Manager receives the return status from the recovery commands. An unexpected status ″hangs″ the cluster until manual intervention (using `smit cm_rec_aids` or the `/usr/sbin/cluster/utilities/clruncmd` command) is carried out.

HACMP ES Script Files

The following sample scripts for failover recovery and user-defined events are included with DB2 UDB EEE. The script files are located in the $INSTNAME/sqllib/samples/hacmp/es directory. The scripts will work ″as is″, or you can customize the recovery action.

- DB2 database partition recovery script rc.db2pe. This is the script file used to start and stop the HACMP configuration on a database partition. It also works as an HACMP start and stop script for an NFS server of the DB2 instance owner.
- DB2-specific user-defined events for HACMP ES. Six default events are included: one for process recovery, two for paging space, and three for NFS and automounter recovery.
- DB2 instance NFS file server failover. This script provides failover recovery of the file system server for a DB2 instance to a backup.
- Network failover. The scripts network_up_complete, network_back, network_down_complete, and network_down allow SP DB2 database partitions to failover if their SP switch adapter fails.
- Scripts to define monitoring events for the SP GUI Perspectives. Monitoring of failover and user-defined recovery is possible through the Event and Hardware Perspectives. Read the documentation for PSSP Administration to find out more about Perspectives.
- Installation scripts to install and remove core scripts and events on the HACMP ES nodes.
- Script files to create and remove the SP Perspectives problem management (pman) resources for monitoring the HACMP and DB2 configuration.

The recovery scripts must be installed on each node that will run recovery operations. The script files can be centrally installed from the SP control workstation or other designated SP node:

1. Copy the scripts from the $INSTNAME/sqllib/samples/hacmp/es directory to one of either the SP control workstation or another SP node that can run the **pcp** and **pexec** commands. These commands are required for the install operation.
2. Customize the reg.parms.SAMPLE and failover.parms.SAMPLE files for your environment by setting key parameters (such as BUFFPAGE) for failover configurations. Typically, for mutual takeover configurations, your failure settings will be adjusted lower to one-half the size of your regular settings or less. Also, you will use a copy of these files renamed with your own name (instead of ″SAMPLE″).
3. Customize (as necessary) the five parameters NFS_RETRIES, START_RETRIES, MOUNT_NFS, STOP_RETRIES, and FAILOVER in the rc.db2pe file. The retry and failover settings should be adequate for most implementations. The MOUNT_NFS setting should be configured, depending on whether you will be using the package for NFS server

availability. You should specify this setting if you want rc.db2pe to mount and verify the NFS home directory of the DB2 instance owner for you. Setting the FAILOVER parameter to ″YES″ will invoke db2_proc_restart and launch an attempt to restart a DB2 database partition. If the restart operation is unsuccessful, HACMP will shut down with a failover.

4. Customize db2_paging_action, db2_proc_recovery, and nfs_auto_recovery in the event file. Edit pwq to change this to the DB2 instance owner. Customize db2_paging_action to specify which action is to be taken if paging space becomes more that ninety percent full. (If this does occur, the DB2 database partition is stopped.) Modify the script if additional recovery actions are required.
5. Use db2_inst_ha to install the scripts and events on the nodes you specify. (HACMP ES must be pre-installed on these nodes before you begin.) The syntax of db2_inst_ha is:

   ```
   db2_inst_ha $INSTNAME/sqllib/samples/hacmp/es <nodelist> <DATABASENAME>

   where

      $INSTNAME/sqllib/samples/hacmp/es is the directory in which the
         scripts and the event are located
      <nodelist> is the pcp or pexec style of the nodes; for example,
         1-16 or 1,2,3,4
      <DATABASENAME> is the name of the database for regular and
         failover parameter files.
   ```

   The reg.parms.SAMPLE and failover.parms.SAMPLE files will be copied to each node and renamed reg.parms.DATABASENAME. db2_inst_ha copies files to each node in /usr/bin, and updates the HACMP event files:

   ```
   /usr/sbin/cluster/events/rules.hacmprd
   /usr/sbin/cluster/events/network_up_complete
   /usr/sbin/cluster/events/network_down_complete
   ```
6. Configure your system and scripts with HACMP.
7. Use the **create_db2_events** command to install the monitoring events for problem management resources (pman) and the SP GUI Perspectives. Additional configuration and customization in Perspectives is needed. For more information about Perspectives, read the PSSP Administration Guide.
8. Use the ha_db2stop command to shutdown the database partitions without HACMP ES failover recovery taking place. To use this command, copy the file to the database user’s home directory and make sure permissions and ownership are set for that user. To stop the database without failover recovery, then as that user, type:

   ```
   ha_db2stop
   ```

Note: You must wait for the command to return. Exiting by using a `ctrl-C` interrupt, or by killing the process, may re-enable failover recovery prematurely, and some database partitions may not be stopped.

DB2 Recovery Script Operations with HACMP ES

HACMP ES invokes the DB2 recovery scripts in the following way:

- `node_up_local` (starting a node)

  HACMP runs the `node_up` sequence, acquiring volume groups, logical volumes, file systems, and IP addresses specified in resource groups that are owned (through cascading) or assigned (through rotating) to this node.

  When `node_up_local_complete` is run, the application server definition that contains `rc.db2pe` is initiated to start the database partition specified in the application server definitions on this physical node.

  Note: `rc.db2pe`, when running in start mode, adjusts the DB2 parameters specified in `reg.parms.DATABASE` for each DATABASE in the database directory that matches a parameter (parms) file.

  Each node follows this sequence when starting. If you have multiple HACMP clusters and start them in parallel, multiple nodes are brought up at once.

- `node_down_remote` (failover)

  HACMP acquires volume groups, logical volumes, file systems, and IP addresses that are specified in the resource group on the designated takeover node.

  When `node_down_remote_complete` is run, HACMP will run `rc.db2pe` as the application server start script specified in the resource group for this database partition.

  Note: `rc.db2pe`, when running in mutual takeover mode, stops the DB2 database partition running on it, adjusts the DB2 parameters specified in `failover.parms.DATABASE` for each DATABASE in the database directory that matches a parameter (parms) file, and then starts both database partitions on the physical takeover node.

- `node_up_remote` (reintegration of a failed node - cascading mutual takeover resource group)

  When `node_up_remote` is run on the old takeover node, the application server definition causes `rc.db2pe` to be run in stop mode.

  Note: `rc.db2pe`, when running in a reintegration mode (mutual takeover), stops both of the database partitions running on it, adjust the DB2 parameters specified in `reg.parms.DATABASE` for each DATABASE in the database directory that matches a parameter (parms) file, and then starts just the database partition to be kept on this physical takeover node.

The old takeover node releases volume groups, logical volumes, file systems, and IP addresses specified in resource groups that are to be owned by the reintegrating node.

HACMP re-acquires volume groups, logical volumes, file systems, and IP addresses specified in the resource group that is now owned by the reintegrating node.

When `node_up_local_complete` is run, the application server definition that contains `rc.db2pe` is initiated to start the DB2 database partition specified in the application server definition on this reintegrating physical node.

Note: `rc.db2pe`, when running in start mode, adjusts the DB2 parameters specified in `reg.parms.DATABASE` for each DATABASE in the database directory that matches a parameter (parms) file.

- `node_down_local` (node stop or stop with takeover)

  When `node_down_local` is run on the stopping node, the application server definition causes `rc.db2pe` to be run in stop mode.

  Note: `rc.db2pe`, when running in stop mode, adjusts the DB2 parameters specified in `failover.parms.DATABASE` for each DATABASE in the database directory that matches a parameter (parms) file, and then stops the DB2 database partition (this is for takeover).

  HACMP releases volume groups, logical volumes, file systems, and IP addresses specified in resource groups that are now owned by the node.

- `db2_proc_recovery` (db2 process death)

  All nodes run the `db2_proc_restart` script. The node on which the failure occurred restarts the correct DB2 database partition.

- `db2_paging_recovery` (paging space recovery)

  All nodes run the `db2_paging_action` script. If a node has more than 70 percent of paging space filled, a wall command is issued. If a node has more than 90 percent of paging space filled, DB2 database partitions on this physical node are stopped and then restarted.

- `nfs_auto_recovery` (nfs or automount process failure)

  All nodes run the `rc.db2pe` script in NFS mode. If an NFS process stops running, it is restarted. Similarly, if the automount process stops running, it is restarted.

- `network_down_complete` (network failure - SP switch)

The net_down script is called. This verifies the network as the SP switch network, and verifies that it is down. If that is the case, it waits a user-defined time interval. The default time interval is 100 seconds.

If the SP switch network comes back, as indicated by an network_up_complete event, no recovery is effected.

If the time limit is reached, HACMP is stopped with failover.

Note: All events can be monitored through SP problem management and the SP Perspectives GUI.

Other Script Utilities

Other script utilities are available for your use, including:

- ha_cmd, a command provided to start HACMP on SP nodes from the control workstation. The syntax is:

```
ha_cmd <noderange> <START|STOP|TAKE|FORCE>

where

   <noderange> is a pcp or pexec style of SP noderange.
      For example, "ha_cmd 3-6 START" would start HACMP on nodes 3,4,5,6.
                   "ha_cmd 5 TAKE" would shut down HACMP on node 5
                      for mutual takeover.
```

- ha_mon, a command for monitoring HACMP hacmp_out files from the SP control workstation. The syntax is:

```
ha_mon <node>

where

   <node> is the SP node to be monitored.

   ha_mon will "tail -f" the /tmp/hacmp.out file on the node you specify.
```

- db2_turnoff_recov, a command for temporarily disabling all HACMP (non-failover) recovery, and designed for extremely rare situations. No DB2 process, paging, NFS, or automounter recovery is initiated. This function removes the event stanzas for that recovery from the HACMP rules file. HACMP must be stopped and restarted. The syntax is:

```
db2_turnoff_recov <nodelist>
```

- db2_turnon_recov, a command to re-enable HACMP (non-failover) recovery. This command is used after db2_turnoff_recov to restore HACMP rules files, so that user-defined event recovery can occur. HACMP must be stopped and restarted. The syntax is:

```
db2_turnon_recov <nodelist>
```

Monitoring HACMP Clusters

Scripts are provided for creating SP problem management (pman) events to monitor the DB2 HACMP ES configuration, in addition to those monitoring utilities already present in HACMP ES. To monitor HACMP status from the SP control workstation:

- Install the HACMP client code on the control workstation.
- Edit the `/usr/sbin/cluster/etc/clhosts` file, and include the SP ethernet IP addresses of the nodes that you want to monitor.
- Invoke the command `startsrc -s clinfo` to start monitoring the clusters.

HACMP supplies an interface for monitoring the clusters (`/usr/sbin/cluster/clstat`.

To use the problem management monitoring with SP Perspectives GUI for HACMP RS and user-defined events:

1. Invoke `create_db2_events <nodelist>`, where *nodelist* contains pcp or pexec style nodes. This script creates five pman events for monitoring by Perspectives.

   Note: The resource variables `PSSP.pm.User_state12-16` are used in the creation of these events. If these resource variables are already being used for some other purpose, `create_db2_events` and `update_db2_events` must be updated to use different resource variables.
2. Start Perspectives on the control workstation. From the launch pad, choose the event perspective. You should see five events: `db2_hacmp_recovery`, `db2_process_recovery`, `db2_paging_err`, `db2_nfs_err`, and `Errlog_PERM_entry`.
3. Double-click on each event. On the screen that appears, register (within the Definition Table) a condition for the event. Click next to the down arrow by `Name: "unnamed"`, and select the same name as the event you specify as the condition. Select the `"Response Options"` tab. Click on the button at the top of the display (″Send Message to Perspectives event session″). You can specify commands, errlog entries, as well as SNMP traps for these event occurrences. The event log displays are maintained only across Perspective sessions; therefore, you might want to create AIX error log entries for each. Select **OK**, and close the window.
4. From the Perspectives launch pad, select the hardware Perspective.
5. When the hardware frame GUI appears, select ″View″ and then ″Monitor″. You are provided with a list of events that can be monitored for your SP. Scrolling to the bottom of the list, you will find two additional events: one for HACMP DB2 recovery (`db2_ha_ind`), and the other for SP node PERM errors (`Errlog_PERM_mon`. Select those that you want to monitor. (When an event occurs, the node displays a red ″X″. If all monitored conditions are

fine, the node display is green.) host_responds, switch_responds, and node_power_LED are typically used. You can also monitor the DB2 HACMP recovery, as well as PERM errors, on the node.

Note: The db2_hacmp_mon and db2_hacmp_recovery variables for pman and Perspectives do not reflect HACMP cluster status. Rather, these variables reflect the status of the rc.db2pe operation to start or stop DB2. The "real" HACMP status is shown in the HACMP clstat monitor, and reflects the HACMP cluster state. If you want db2_hacmp_ind to reflect monitoring similar to HACMP status, add the following line to your /etc/inittab file:

```
haind:2:wait:/usr/bin/db2_update_events HAIND OFF 2>&1 >/dev/null
```

If you are planning to use NetView for your implementation, consider using HAVIEW (which is part of HACMP) for monitoring your configuration. For information about configuring that product, refer to the NetView documentation.

DB2 SP HACMP ES Installation

To help you plan for the installation of HACMP ES with DB2 Universal Database, following is a step-by-step overview of the installation and migration processes.

DB2 SP HACMP ES New Installation

To install HACMP ES:

1. Install the AIX operating system on each SP node, (refer to the SP Installation and Administration Guides). Ensure that proper paging space is available on both the control workstation, and each of the SP nodes. Ensure that switch configuration has been considered and implemented, along with any other modifiable configuration parameters. Put in place the SP monitoring (Perspectives) that you want to use. Ensure that the SP dsh, pcp, and pexec commands work.
2. Design your database layout. This should, at a minimum, include the number of nodes to be used, the mapping of DB2 database partitions to physical nodes, the disk requirements per node or partition, and table space considerations. You should also consider who the main DB2 instance owner will be, and what access authorization this and other users will require.
3. Plan your external SSA disk configuration, including redundant adapters, mirrored disks, and the twin-tailing of disks.
4. Using your database layout and SSA configuration, complete the HACMP worksheets located in the HACMP Planning, Installation, and Administration Guides.

5. Implement your external SSA disk configuration. Ensure that microcode levels are consistent across all drives, and use the Maymap utility to validate and fill in any gaps in your worksheets.
6. Install DB2 UDB EEE on each SP node.
7. Install HACMP ES on each SP node.
8. Install the DB2 UDB EEE HACMP ES on SP Package, using the **db2_inst_ha** command.
9. Create the main DB2 instance user, and ensure that it can access all nodes. This is not a highly available user at this point. This can be temporarily an SP user on the SP control workstation.
10. Create your DB2 instance and database. Ensure that it is operational by invoking **db2start**, and then **db2stop**, before proceeding to the next step.
11. If you want to load the database before adding HACMP, do this now.
12. Configure HACMP ES on the SP nodes topology and resource groups according to the HACMP worksheets and the information in this document.
13. Beginning with your NFS server node for the main DB2 instance user, change this user (by modifying `/etc/security/user` and `/etc/passwd` on all nodes, in accordance with what is specified in this document. This user will become a highly available NFS user; and this node and its backup will update `/etc/exports`. All nodes will be able to mount this directory using NFS (with an entry in `/etc/filesystems` on each node) through the switch alias IP addresses.
14. ″Tar″ the home directory of the main instance user and ″un-tar″ the home directory in the new location.
15. Create an NFS file system on each of the SP nodes to mount a new main instance home directory.
16. Start HACMP on the NFS server node. Verify that it comes up successfully by investigating `/tmp/hacmp.out`. The **ha_mon** command can be used to monitor this file as it is written.
17. Bring up the other nodes one at a time, verifying each successful completion by investigating `/tmp/hacmp.out`. The **ha_mon** command can be used to monitor this file as it is written.
18. Set up the optional monitoring through Perspectives and Problem Management.
19. Validate failover functionality on each node by simulating a concurrent maintenance action on each node. The **ha_cmd** command (specifying the TAKE option) can be used to stop HACMP gracefully with takeover. Verify that the takeovers and the reintegrations are successful by interrogating `/tmp/hacmp.out` and using your monitoring tools.

DB2 SP HACMP ES Migration

If you are migrating from a non-HACMP installation to one with HACMP, consider the following overview:

1. Convert your existing external disks to a highly available, twin-tailed, mirrored configuration. Add any extra hardware and disks to achieve this configuration, remembering that names of different logical volumes on different nodes *must* be unique when they are twin-tailed. This applies to volume groups, logical volumes, and file systems.
2. Complete the HACMP planning and the related worksheets, including the worksheets in this document.
3. Implement your external SSA disk configuration changes. Ensure that microcode levels are consistent across all drives, and use the Maymap utility to validate and eliminate any gaps in the worksheets.

   Note: SSA disks in a RAID5 configuration is supported. Two SSA adapters in the same RAID loop is the only configuration permitted. For an HACMP configuration with the RAID disks twin-tailed, only one adapter per node is supported. In this configuration, the adapter is a single point of failure for access to the disks, and extra configuration is recommended to detect the adapter outage and promote this to an HACMP failover event. AIX error notification is the simplest way to configure a node for failover, should the SSA adapter fail. Refer to *HACMP for AIX, V4.2.2, Enhanced Scalability Installation and Administration Guide* for more information about AIX error notification.
4. Install HACMP ES on each SP node.
5. Install the DB2 UDB EEE HACMP ES on SP Package, using the **db2_inst_ha** command.
6. Configure HACMP ES on the SP nodes topology and resource groups according to the HACMP worksheets and the information in this document.
7. Beginning with your NFS server node for the main DB2 instance user, change this user (by modifying `/etc/security/user` and `/etc/passwd` on all nodes, in accordance with what is specified in this document. This user will become a highly available NFS user; and this node and its backup will update `/etc/exports`. All nodes will be able to mount this directory using NFS (with an entry in `/etc/filesystems` on each node) through the switch alias IP addresses.
8. ″Tar″ the home directory of the main instance user and ″un-tar″ the home directory in the new location.
9. Create an NFS file system on each of the SP nodes to mount a new main instance home directory.

10. Start HACMP on the NFS server node. Verify that it comes up successfully by investigating /tmp/hacmp.out. The **ha_mon** command can be used to monitor this file as it is written.
11. Bring up the other nodes one at a time, verifying each successful completion by investigating /tmp/hacmp.out. The **ha_mon** command can be used to monitor this file as it is written.
12. Set up the optional monitoring through Perspectives and Problem Management.
13. Validate failover functionality on each node by simulating a concurrent maintenance action on each node. The **ha_cmd** command (specifying the TAKE option) can be used to stop HACMP gracefully with takeover. Verify that the takeovers and the reintegrations are successful by interrogating /tmp/hacmp.out and using your monitoring tools.

DB2 SP HACMP ES Worksheets

The following worksheets are designed to be used with HACMP worksheets that should be completed in preparation for your external SSA disk configuration (and that are located in the HACMP Planning, Installation, and Administration Guides). In each case, both a completed example, and a blank worksheet, are provided.

The database configuration on external disks documented in the first sample worksheet is shown in the following figure. The statement used to create the database is:

```
db2 create database pwq on /newdata
```

Both SSA external adapters and external SSA disks are mirrored and twin-tailed for logical volumes with no single point of failure. The diagram depicts a configuration that is similar to output from the **maymap** command. Maymap is a utility (available through AIXTOOLS) that shows the external SSA disk configuration, and should be used when planning your setup.

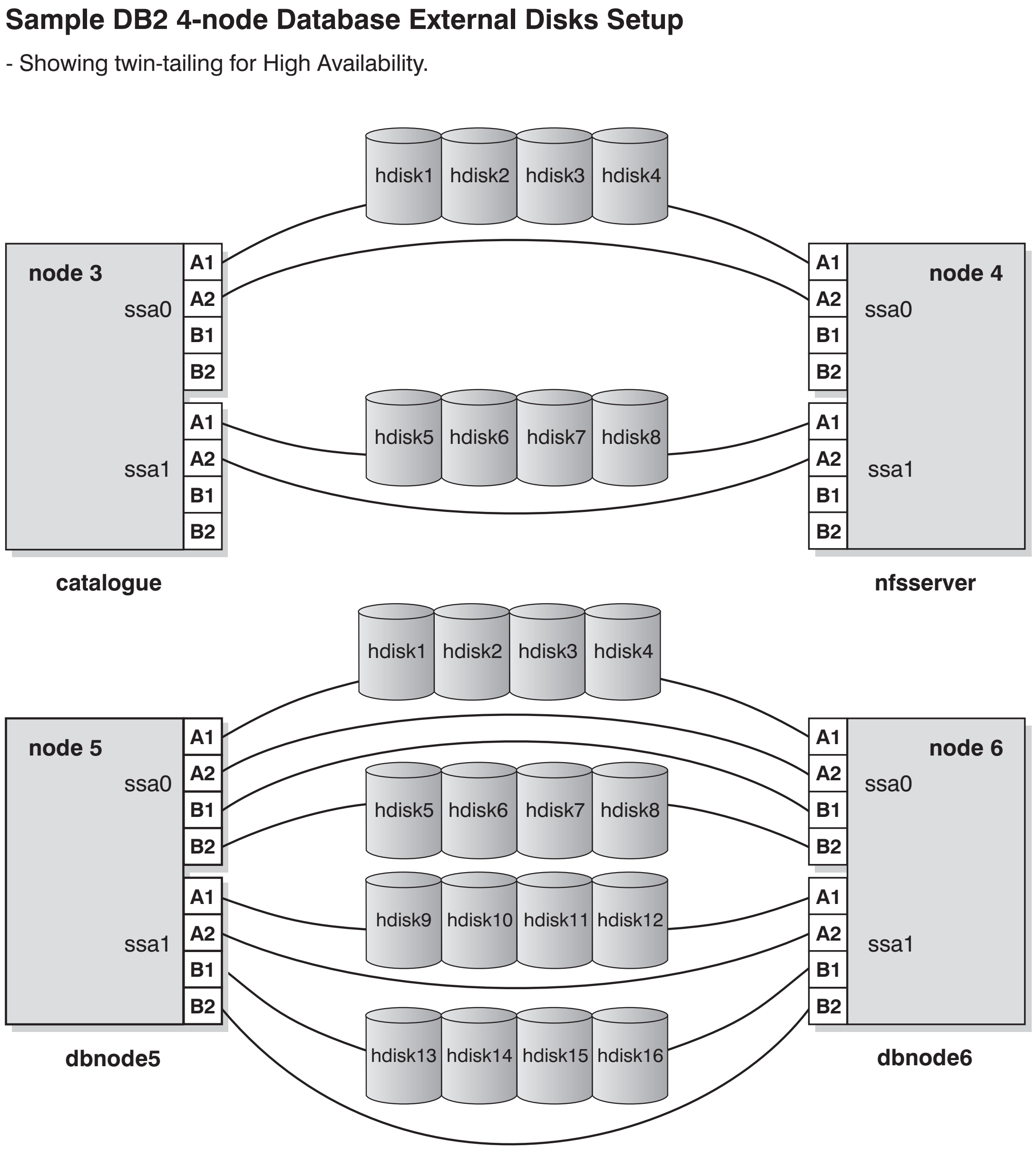


Figure 55. Sample DB2 4-node Database External Disks Setup

Before you review the following table, you should read the HACMP documentation regarding the quorum settings on volume groups, and mirrored write consistency settings on logical volumes. The settings used for both will directly affect your availability and performance. Ensure that you review these settings and understand their implications. The typical setting for both ″quorum″ and ″mirrored write consistency″ is ″off″.

Table 25. HACMP Volume Groups, Logical Volumes, and File Systems

SP Node	Volume Group Name	PP Size (MB)	Logical Volume Name	# of PPs	Copies	hdisk List	File System Mount Point (MB)	File System Log Logical Volume	Node Description and Backup	User Owner of /dev Logical Device
3	havg3	8	hlv300	10	2	hdisk1 hdisk5	/newdata /pwq /NODE0003	hlog301	Catalognode mount point; node 4	root *
3	havg3	8	hlog301	1	2	hdisk1 hdisk5	N/A	N/A	Catalognode jfslog; node 4	root *
3	havg3	8	hlv301	10	2	hdisk2 hdisk6	N/A	N/A	Catalognode rawtemp space; node 4	pwq **
4	havg4	8	hlv400	10	2	hdisk3 hdisk7	/dbmnt	hlog401	nfsserver pwq home; node 3	root *
4	havg4	8	hlog401	1	2	hdisk3 hdisk7	N/A	N/A	nfsserver jfslog; node 3	root *
5	havg5	8	hlv500	10	2	hdisk1 hdisk9	/newdata/ pwq/ NODE0005	HLOG501	Dbnode5 mount point; node 6	root *
5	havg5	8	hlog501	1	2	hdisk1 hdisk9	N/A	N/A	Dbnode5 jfslog; node 6	root *
5	havg5	8	hlv501	10	2	hdisk2 hdisk10	N/A	N/A	Dbnode5 raw temp space; node 6	pwq **
5	havg5	8	hlv502	100	2	hdisk2 hdisk10	N/A	N/A	Dbnode5 raw table space; node 6	pwq **
5	havg5	8	halv503	100	2	hdisk3 hdisk11	N/A	N/A	Dbnode5 raw table space; node 6	pwq **

Table 25. HACMP Volume Groups, Logical Volumes, and File Systems (continued)

SP Node	Volume Group Name	PP Size (MB)	Logical Volume Name	# of PPs	Copies	hdisk List	File System Mount Point (MB)	File System Log Logical Volume	Node Description and Backup	User Owner of /dev Logical Device
5	havg5	8	halv504	100	2	hdisk3 hdisk11	N/A	N/A	Dbnode5 raw table space; node 6	pwq **
5	havg5	8	halv505	100	2	hdisk4 hdisk12	/dbdata5	hlog501	Dbnode6 system table space; node 6	root *
6	havg6	8	hlv600	10	2	hdisk5 hdisk13	/newdata/pwq/NODE0006	hlog601	Dbnode6 mount point; node 5	root *
6	havg6	8	hlog601	1	2	hdisk5 hdisk13	N/A	N/A	Dbnode6 jfslog; node 5	root *
6	havg6	8	hlv601	10	2	hdisk6 hdisk14	N/A	N/A	Dbnode6 raw temp space; node 5	pwq **
6	havg6	8	hlv602	100	2	hdisk6 hdisk14	N/A	N/A	Dbnode6 raw table space; node 5	pwq **
6	havg6	8	hlv603	100	2	hdisk7 hdisk15	N/A	N/A	Dbnode6 raw table space; node 5	pwq **
6	havg6	8	hlv604	100	2	hdisk7 hdisk15	N/A	N/A	Dbnode6 raw table space; node 5	pwq **
6	havg6	8	hlv605	100	2	hdisk8 hdisk16	/dbdata6	hlog601	Dbnode6 system table space; node 5	root *

Table 25. HACMP Volume Groups, Logical Volumes, and File Systems (continued)

SP Node	Volume Group Name	PP Size (MB)	Logical Volume Name	# of PPs	Copies	hdisk List	File System Mount Point (MB)	File System Log Logical Volume	Node Description and Backup	User Owner of /dev Logical Device

Notes:

1. * jfs file system logical volumes and logs keep root permissions.
2. ** raw database spaces get database user permissions on /dev raw file entries (/dev/rxxxx).

Table 26. HACMP Volume Groups, Logical Volumes, and File Systems - Blank

SP Node	Volume Group Name	PP Size (MB)	Logical Volume Name	# of PPs	Copies	hdisk List	File System Mount Point (MB)	File System Log Logical Volume	Node Description and Backup	User Owner of /dev Logical Device

Table 26. HACMP Volume Groups, Logical Volumes, and File Systems - Blank (continued)

SP Node	Volume Group Name	PP Size (MB)	Logical Volume Name	# of PPs	Copies	hdisk List	File System Mount Point (MB)	File System Log Logical Volume	Node Description and Backup	User Owner of /dev Logical Device

Table 26. HACMP Volume Groups, Logical Volumes, and File Systems - Blank (continued)

SP Node	Volume Group Name	PP Size (MB)	Logical Volume Name	# of PPs	Copies	hdisk List	File System Mount Point (MB)	File System Log Logical Volume	Node Description and Backup	User Owner of /dev Logical Device

Table 27. Planning HACMP NFS Server

SP Node	External File System	Backup Node	SP Switch Boot and Service IP Alias Pairs	File System to Mount (/etc/filesystems)	File System to Specify as Database Home Directory	Addresses to which File System is to be Exported (/etc/exports)
3	/dbmnt	4	nfs_boot_3 nfs_client_3	nfs_server:/ dbmnt as /dbi	/dbi/pwq	nfs_boot_3 nfs_client_3 nfs_server_boot nfs_server nfs_boot_5 nfs_client_5 nfs_boot_6 nfs_client_6

Table 27. Planning HACMP NFS Server (continued)

SP Node	External File System	Backup Node	SP Switch Boot and Service IP Alias Pairs	File System to Mount (/etc/filesystems)	File System to Specify as Database Home Directory	Addresses to which File System is to be Exported (/etc/exports)
4	/dbmnt	3	nfs_server_boot nfs_server	nfs_server:/ dbmnt as /dbi	/dbi/pwq	nfs_boot_3 nfs_client_3 nfs_server_boot nfs_server nfs_boot_5 nfs_client_5 nfs_boot_6 nfs_client_6
5	N/A	N/A	nfs_boot_5 nfs_client_5	nfs_server:/ dbmnt as /dbi	/dbi/pwq	N/A
6	N/A	N/A	nfs_boot_6 nfs_client_6	nfs_server:/ dbmnt as /dbi	/dbi/pwq	N/A

Notes:

1. `/etc/passwd` must be the same on all nodes. This can be synchronized from the control workstation.
2. Ensure that the external file system has the permission of the database instance owner.
3. The `/etc/filesystems` must have the mount parameters: hard, bg, intr, and rw.
4. The `/etc/exports` will have

   ```
   -root=ip1:ip2:ip3
   ```

   only on the server and its backup.

Table 28. Planning HACMP NFS Server - Blank

SP Node	External File System	Backup Node	SP Switch Boot and Service IP Alias Pairs	File System to Mount (/etc/filesystems)	File System to Specify as Database Home Directory	Addresses to which File System is to be Exported (/etc/exports)

Table 28. Planning HACMP NFS Server - Blank (continued)

SP Node	External File System	Backup Node	SP Switch Boot and Service IP Alias Pairs	File System to Mount (/etc/filesystems)	File System to Specify as Database Home Directory	Addresses to which File System is to be Exported (/etc/exports)

Table 28. Planning HACMP NFS Server - Blank (continued)

SP Node	External File System	Backup Node	SP Switch Boot and Service IP Alias Pairs	File System to Mount (/etc/filesystems)	File System to Specify as Database Home Directory	Addresses to which File System is to be Exported (/etc/exports)

Chapter 13. High Availability in the Windows NT Environment

You can set up your database system so that if a machine fails, the database server on the failed machine can run on another machine. On Windows NT, failover support can be implemented with Microsoft Cluster Server (MSCS). To use MSCS, you require Windows NT Version 4.0 Enterprise Edition with the MSCS feature installed.

MSCS can perform both failure detection and the restarting of resources in a clustered environment, such as failover support for physical disks and IP addresses. (When the failed machine is online again, resources will not automatically fall back to it, unless you previously configure them to do so. For more information, see "Fallback Considerations" on page 269.)

Before you enable DB2 instances for failover support, perform the following planning steps:

1. Decide which disks you want to use for data storage. Each database server should be assigned at least one disk for its own use. The disk that you use to store data must be attached to a shared disk subsystem, and must be configured as an MSCS disk resource.
2. Ensure that you have one IP address for each database server that you want to use to support remote requests.

When you set up failover support, it can be for an existing instance, or you can create a new instance when you implement the failover support.

To enable failover support, perform the following steps:

1. Create an input file for the DB2MSCS utility.
2. Invoke the **db2mscs** command.
3. If you are using a partitioned database system, register database drive mapping to enable mutual takeover. See "Registering Database Drive Mapping for Mutual Takeover Configurations in a Partitioned Database Environment" on page 269.

After you finish enabling the instance for failover support, your configuration will resemble Figure 56 on page 258.

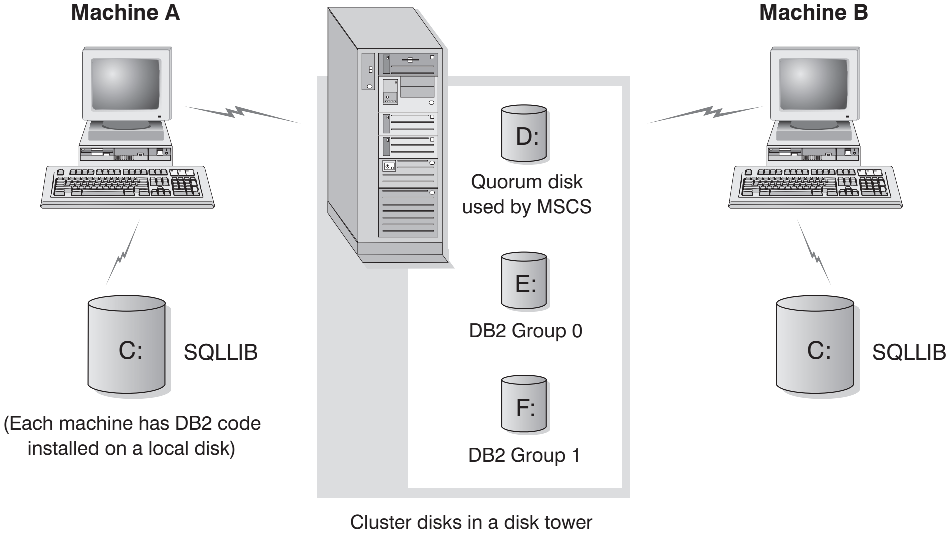


Figure 56. Example MSCS Configuration

The following sections describe the different types of failover support, and how to implement them. Before performing any of the steps described below, you must already have the MSCS software installed on every machine that you want to use in an MSCS cluster. In addition, you must also have DB2 installed on every machine.

Failover Configurations

Two types of configuration are available:

- Hot standby
- Mutual takeover

Currently, MSCS supports clusters of two machines.

In a partitioned database environment, the clusters do not all have to have the same type of configuration. You can have some clusters that are set up to use hot standby, and others that are set up for mutual takeover. For example, if your DB2 instance consists of five workstations, you can have two machines set up to use a mutual takeover configuration, two to use a hot standby configuration, and one machine not configured for failover support.

Hot Standby Configuration

In a hot standby configuration, one machine in the MSCS cluster provides dedicated failover support, and the other machine participates in the database system. If the machine participating in the database system fails, the database server on it will be started on the failover machine. If, in a partitioned database system, you are running multiple logical nodes on a machine and it fails, the logical nodes will be started on the failover machine. Figure 57 shows an example of a hot standby configuration.

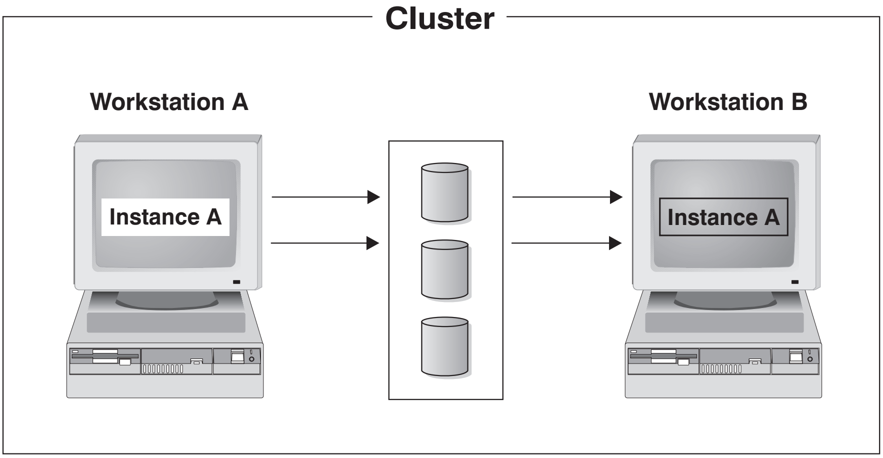


Figure 57. Hot Standby Configuration

Mutual Takeover Configuration

In a mutual takeover configuration, both workstations participate in the database system (that is, each machine has at least one database server running on it). If one of the workstations in the MSCS cluster fails, the database server on the failing machine will be started to run on the other machine. In a mutual takeover configuration, a database server on one machine can fail independently of the database server on another machine. Any database server can be active on any machine at any given point in time. Figure 58 on page 260 shows an example of a mutual takeover configuration.

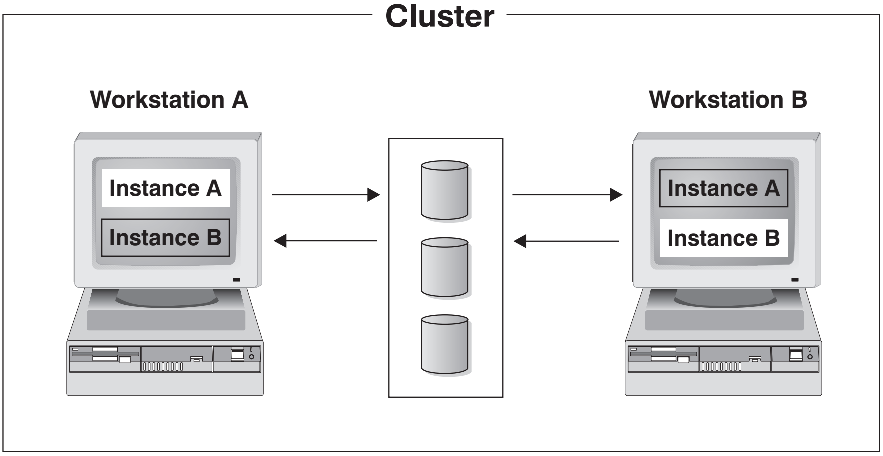


Figure 58. Mutual Takeover Configuration

Using the DB2MSCS Utility

Use the DB2MSCS utility to create the infrastructure for DB2 to support failover in the Windows NT environment using MSCS support. You can use this utility to enable failover in both single-partition and partitioned database environments.

Invoke the **db2mscs** command once for each instance on its instance-owning machine. If there is only one DB2 instance running on one machine in the MSCS cluster, this sets up a hot-standby configuration. If you have an instance running on each machine in the MSCS cluster, you would run DB2MSCS once on each instance-owing machine to set up a mutual takeover configuration.

The DB2MSCS utility:

1. Reads the required MSCS and DB2 parameters from an input file called `DB2MSCS.CFG`. See “Specifying the DB2MSCS.CFG File” on page 261 for information about the full set of input parameters.
2. Validates the parameters in the input file.
3. Registers the DB2 resource type.
4. Creates the MSCS group (or groups) to contain the MSCS and DB2 resources.
5. Creates the IP resource.
6. Creates the Network Name resource.
7. Moves MSCS disks to the group.

8. Creates the DB2 resource (or resources).
9. Adds all required dependencies for the DB2 resource.
10. Converts the non-clustered DB2 instance into a clustered instance.
11. Brings all resources online.

The command syntax is as follows:

```
►►──db2mscs──┬─────────────────┬─────────────────────────────────────────►◄
             └─-f:──input_file─┘
```

Where:

-f:*input_file*

Specifies the `DB2MSCS.CFG` input file to be used by the MSCS utility. If this parameter is not specified, the DB2MSCS utility reads the `DB2MSCS.CFG` file that is in the current directory.

Specifying the DB2MSCS.CFG File

The `DB2MSCS.CFG` file is an ASCII text file that contains parameters that are read by the DB2MSCS utility. You specify each input parameter on a separate line using the following format: PARAMETER_KEYWORD=*parameter_value*. For example:

```
CLUSTER_NAME=WOLFPACK
GROUP_NAME=DB2 Group
IP_ADDRESS=9.21.22.89
```

Two example configuration files are in the `/CFG` subdirectory of the `/SQLLIB` directory. The first, `DB2MSCS.EE`, is an example for single-partition database environments. The second, `DB2MSCS.EEE`, is an example for partitioned database environments.

The parameters for the `DB2MSCS.CFG` file are as follows:

DB2_INSTANCE

The name of the DB2 instance. If the instance name is *not* specified, the default instance (the value of the DB2INSTANCE environment variable) is used.

This parameter has a global scope, and you specify it only once in the `DB2MSCS.CFG` file.

This parameter is optional.

Example:

```
DB2_INSTANCE=DB2
```

The instance must already exist. For information about creating instances, refer to the *DB2 Enterprise - Extended Edition for Windows Quick Beginnings* book.

DB2_LOGON_USERNAME

The name of the logon account for the DB2 service.

This parameter has a global scope, and you specify it only once in the DB2MSCS.CFG file.

This parameter is only required for DB2 Enterprise - Extended Edition instances.

Example:

```
DB2_LOGON_USERNAME=db2user
```

DB2_LOGON_PASSWORD

The password of the logon account for the DB2 service. If the DB2_LOGON_USERNAME parameter is provided, but the DB2_LOGON_PASSWORD parameter is not, the DB2MSCS utility prompts for the password. The password is not displayed when it is typed at the command line.

This parameter has a global scope, and you specify it only once in the DB2MSCS.CFG file.

This parameter is only required for DB2 Enterprise - Extended Edition instances.

Example:

```
DB2_LOGON_PASSWORD=xxxxxx
```

CLUSTER_NAME

The name of the MSCS cluster. All the resources specified following this line are created in this cluster until another CLUSTER_NAME tag is specified.

Specify this parameter once for each cluster.

This parameter is optional. If not specified, the name of the MSCS cluster on the local machine is used.

Example:

```
CLUSTER_NAME=WOLFPACK
```

GROUP_NAME

The name of the MSCS group. If this parameter is specified, a new MSCS group is created if one does not exist. If the group already exists, it is used as the target group. Any MSCS resource created following this line is created in this group until another GROUP_NAME keyword is specified.

Specify this parameter once for each group.

This parameter is required.

Example:

```
GROUP_NAME=DB2 Group
```

DB2_NODE

The node number of the database partition server (node) to be included in the current MSCS group. If multiple logical nodes exist on the same machine, each node requires a separate DB2_NODE keyword.

You specify this parameter after the GROUP_NAME parameter, so that the DB2 resources are created in the correct MSCS group.

This parameter is only required for DB2 Enterprise - Extended Edition instances.

Example:

```
DB2_NODE=0
```

IP_NAME

The name of the IP Address resource. The value for IP_NAME is arbitrary, but must be unique. When this parameter is specified, an MSCS resource of type IP Address is created.

This parameter is required for remote TCP/IP connections. You must specify this parameter for the instance-owning machine in a partitioned database environment. This parameter is optional in single-partition database environments.

Example:

```
IP_NAME=IP Address for DB2
```

Note: DB2 clients should use the TCP/IP address of this IP resource to catalog the TCP/IP node entry. By using the MSCS IP address, when the database server fails over to the other machine, DB2 clients can still connect to the database server, because the IP address is available on the failover machine.

The attributes of the IP resource are as follows:

IP_ADDRESS

The TCP/IP address of the IP resource. Specify this keyword to set the TCP/IP address for the preceding IP resource.

This parameter is required if the IP_NAME parameter is specified.

Example:

```
IP_ADDRESS=9.21.22.34
```

IP_SUBNET

The subnet mask for the preceding IP resource.

This parameter is required if the IP_NAME parameter is specified.

Example:

```
IP_SUBNET=255.255.255.0
```

IP_NETWORK

The name of the MSCS network to which the preceding IP resource belongs. If this parameter is not specified, the first MSCS network detected by the system is used.

This parameter is optional.

Example:

```
IP_NETWORK=Token Ring
```

NETNAME_NAME

The name of the Network Name resource. Specify this parameter to create the Network Name resource.

This parameter is optional for single-partition database environments. It is required for partitioned database environments.

Example:

```
NETNAME_NAME=Network name for DB2
```

The attributes of the Network Name resource are as follows:

NETNAME_VALUE

The value for the Network Name.

This parameter is required if the NETNAME_NAME parameter is specified.

Example:

```
NETNAME_VALUE=DB2SRV
```

NETNAME_DEPENDENCY

The dependency list for the Network Name resource. Each Network Name resource must have a dependency on an IP Address resource. If this parameter is not specified, the Network Name resource has a dependency on the first IP resource in the group.

This parameter is optional.

Example:

```
NETNAME_DEPENDENCY=IP Address for DB2
```

DISK_NAME

The name of the physical disk resources to be moved to the current groups. Specify as many disk resources as you need.

Notes:

1. The disk resources must already exist.
2. When the DB2MSCS utility configures the DB2 instance for MSCS support, the instance directory is copied to the *first* MSCS disk in the group. To specify a different MSCS disk for the instance directory, use the INSTPROF_DISK parameter.

Example:

```
DISK_NAME=Disk E:
DISK_NAME=Disk F:
```

INSTPROF_DISK

An optional parameter to specify an MSCS disk to contain the DB2 instance directory. If this parameter is *not* specified, the DB2MSCS utility uses the *first* MSCS disk that belongs to the same group as the instance directory.

The DB2 instance directory is created on the MSCS disk under the `X:\DB2PROFS` directory (where X is the MSCS disk drive letter).

Example:

```
INSTPROF_DISK=Disk E:
```

TARGET_DRVMAP_DISK

An optional parameter to specify the target MSCS disk for database drive mapping. If a database is created on an MSCS disk that does not belong to the same group as the node, the target drive map disk is used to contain the database partition. If this parameter is not specified, the database drive mapping must be manually registered using the DB2DRVMP utility.

Example:

```
TARGET_DRVMAP_DISK = Disk E:
```

Setting up Failover for a Single-Partition Database System

When you run the DB2MSCS utility against a single-partition database system, one MSCS group contains DB2 and all the dependent MSCS resources (the IP address, Network Name, and disks). For example, the contents of the `DB2MSCS.CFG` file for a single-partition database system will look like the following:

```
#
#  DB2MSCS.CFG for a single-partition database system
#
DB2_INSTANCE=DB2
CLUSTER_NAME=MSCS
GROUP_NAME=DB2 Group
IP_NAME=...
IP_ADDRESS=...
IP_SUBNET=...
IP_NETWORK=...
NETNAME_NAME=...
NETNAME_VALUE=...
DISK_NAME=Disk E:
```

Setting up a Mutual Takeover Configuration for Two Single-Partition Database Systems

You can set up two single-partition database systems, each on a separate machine, so that if the database system on any one machine fails, it is restarted on the other MSCS node.

To set up failover support for this configuration, you need to run the DB2MSCS utility once on each instance-owning machine. You must tailor the configuration file for each database system.

Assume that the DB2 instances are called DB2A and DB2B. The `DB2MSCS.CFG` file for the DB2A instance would be as follows:

```
#
#  DB2MSCS.CFG for first single-partition database system
#
DB2_INSTANCE=DB2A
CLUSTER_NAME=MSCS
GROUP_NAME=DB2A Group
IP_NAME=...
IP_ADDRESS=...
IP_SUBNET=...
IP_NETWORK=...
NETNAME_NAME=...
NETNAME_VALUE=...
DISK_NAME=Disk E:
```

The `DB2MSCS.CFG` file for the DB2A instance would be as follows:

```
#
#  DB2MSCS.CFG for second single-partition database system
#
DB2_INSTANCE=DB2B
CLUSTER_NAME=MSCS
GROUP_NAME=DB2B Group
IP_NAME=...
IP_ADDRESS=...
IP_SUBNET=...
```

```
IP_NETWORK=...
NETNAME_NAME=...
NETNAME_VALUE=...
DISK_NAME=Disk F:
```

For a full example, see “Example - Setting up Two Single-Partition Instances for Mutual Takeover” on page 272.

Setting up Multiple MSCS Clusters for a Partitioned Database System

When you run the DB2MSCS utility against a multi-partition database system, one MSCS group is created for each physical machine that participates in the system. The DB2MSCS.CFG file must contain multiple sections, and each section must have a different value for the GROUP_NAME parameter, and for all the required dependent resources for that group.

In addition, you must specify the DB2_NODE parameter for each database partition server in each MSCS group. If you have multiple logical nodes, each logical node requires a separate DB2_NODE keyword.

For example, assume that you have a multi-partition database system that consists of four database partition servers on four machines, and you want to configure two MSCS clusters using mutual takeover configuration. You would set up the DB2MSCS.CFG configuration file as follows:

```
#
#  DB2MSCS.CFG for one partitioned database system with
#  multiple clusters
DB2_INSTANCE=DB2MPP
DB2_LOGON_USERNAME=db2user
DB2_LOGON_PASSWORD=xxxxxx
CLUSTER_NAME=MSCS1
# Group 1
GROUP_NAME=DB2 Group 1
DB2_NODE=0
IP_NAME=...

...
# Group 2
GROUP_NAME=DB2 Group 2
DB2_NODE=1
IP_NAME=...

...

CLUSTER_NAME=MSCS2
# Group 3
GROUP_NAME=DB2 Group 3
DB2_NODE=2
IP_NAME=...

...
# Group 4
```

```
GROUP_NAME=DB2 Group 4
DB2_NODE=3
IP_NAME=...

...
```

For a full example, see “Example - Setting up a Four-Node Partitioned Database System for Mutual Takeover” on page 274.

Maintaining the MSCS System

When you run the DB2MSCS utility, it creates the infrastructure for failover support for all machines in the MSCS cluster. To remove support from a machine, use the **db2iclus** command with the ″drop″ option. To re-enable support for a machine, use the ″add″ option.

The command syntax is as follows:

```
►►─db2iclus──┬─add──┬──┬──────────────────────┬─/u:─account_name,password──────────►
             └─drop─┘  └─/i:─instance_name─┘

►─┬────────────────────┬─┬────────────────────┬───────────────────────────────────►◄
  └─/m:─machine_name─┘ └─/c:─cluster_name─┘
```

Where:

add
Enables failover support on the machine by adding it to an MSCS cluster. The DB2 resource (database server) can then fail over to this machine.

drop
Removes failover support from the machine by dropping it from an MSCS cluster.

/i: *instance_name*
The name of the instance. (This parameter overrides the setting of the DB2INSTANCE environment variable.)

/u: *account_name*, *password*
The domain account used as the logon account name of the DB2 Service. For example:

```
/u:domainA\db2nt,password
```

This parameter is only required with the ″add″ option.

/m:*machine_name*
The computer name of the machine that you want to add to, or drop from, an MSCS

cluster. You must specify this option if you run the command from a machine other than the one for which you are modifying failover support.

/c: *cluster_name* The name of the MSCS cluster as it is known on the LAN. This name is specified when the MSCS cluster is first created.

Fallback Considerations

By default, groups are set not to fall back to the original (failed) machine. Unless you manually configure a DB2 group to fall back after failing over, it continues to run on the alternative MSCS node after the cause of the failover has been resolved.

If you configure a DB2 group to automatically fall back to the original machine, all the resources in the DB2 group including the DB2 resource will fall back as soon as the original machine is available. If, during the fall back, a database connection exists, the DB2 resource cannot be brought offline, and the fallback processing will fail.

If you want to force all database connections off the database during fallback processing, set the DB2_FALLBACK registry variable to ON. This variable must be set as follows:

```
db2set DB2_FALLBACK=ON
```

You do not have to reboot or restart the cluster service after setting this registry variable.

Registering Database Drive Mapping for Mutual Takeover Configurations in a Partitioned Database Environment

When you create a database in a partitioned database environment, you can specify a drive letter to indicate where the database is to be created.

Note: You do not set database drive mapping for single-partition database environments.

When the CREATE DATABASE command runs, it expects that the drive that you specify will be simultaneously available to all of the machines that participate in the instance. Because this is not possible, DB2 uses database drive mapping to assign the same drive a different name for each machine.

For example, assume that a DB2 instance called DB2 contains two database partition servers:

```
NODE0 is active on machine WOLF_NODE_0
NODE1 is active on machine WOLF_NODE_1
```

Assume also that the share disk E: belongs to the same group as NODE0, and that the share disk F: belongs to the same group as NODE1.

To create a database on the share disk E:

```
db2 create database mppdb on e:
```

For the command to be successful, drive E: must be available to both machines. In a mutual takeover configuration, each database partition server may be active on a different machine, and the cluster disk E: is only available to one machine. In this situation, the CREATE DATABASE command will always fail.

To resolve this problem, the database drive should be mapped as follows:

```
For NODE0, the mapping is from drive F: to drive E:
For NODE1, the mapping is from drive E: to drive F:
```

Any database access for NODE0 to drive F: is then mapped to drive E:, and any database access for NODE1 to drive E: is mapped to drive F:. Using drive mapping, the CREATE DATABASE command will create database files on drive E: for NODE0 and drive F: for NODE1.

Use the **db2drvmp** command to set up the drive mapping. The command syntax is as follows:

```
►►──db2drvmp──┬─add───────┬──node_number──from_drive──to_drive──────────────────────►◄
              ├─drop──────┤
              ├─query─────┤
              └─reconcile─┘
```

The parameters are as follows:

add Assigns a new database drive map.

drop Removes an existing database drive map.

query Queries a database map.

reconcile Repairs a database map drive when the registry contents are damaged. See “Reconciling the Database Drive Mapping” on page 271 for more information.

node_number The node number. This parameter is required for add and drop operations.

from_drive The drive letter from which to map. This parameter is required for add and drop operations.

to_drive The drive letter to which to map. This parameter is required for add operations. It is not applicable to other operations.

If you wanted to set up database drive mapping from F: to E: for NODE0, you would use the following command:

```
db2drvmp add 0 F E
```

Note: Database drive mapping does not apply to table spaces, containers, or any other database storage objects.

Similarly, to set up database drive mapping from E: to F: for NODE1, you would issue the following command:

```
db2drvmp add 1 E F
```

Note: Any setup of, or changes to, the database drive mapping does not take effect immediately. To activate the database drive mapping, use the Cluster Administrator tool to bring the DB2 resource offline, then online.

Using the TARGET_DRVMAP_DISK keyword in the `DB2MSCS.CFG` file will enable the drive mapping to be done automatically.

Reconciling the Database Drive Mapping

When a database is created on a machine that has database drive mapping in effect, the map is saved on the drive in a hidden file. This is to prevent the database drive from being removed after the database is created. You will want to *reconcile* the database drive mapping if, for example, you accidentally drop the database drive map. To reconcile the map, run the **db2drvmp reconcile** command for each database partition server that contains the database. The command syntax is as follows:

```
►►──db2drvmp reconcile──┬─────────────────────┬──────────────────────────►◄
                        └─node_number──drive──┘
```

The parameters are as follows:

node_number The node number of the node to be repaired. If *node_number* is not specified, the command reconciles the mapping for all nodes.

drive The drive to reconcile. If a drive is not specified, the command reconciles the mapping for all drives.

The **db2drvmp** command scans all drives on the machine for database partitions that are managed by the database partition server, and reapplies the database drive mapping to the registry as required.

Example - Setting up Two Single-Partition Instances for Mutual Takeover

The objective for this example is to set up two single-partition database instances with failover support in a mutual takeover configuration. In this example, four servers are configured into two MSCS clusters. By using the mutual takeover configuration, when any of the machine fails, the database server configured for that machine will fail over to the alternative machine, as configured using the MSCS software, and run on the alternative machine.

There are two MSCS clusters in the resulting configuration. Each cluster has:

- Two servers, each with 64 MB of memory and one local SCSI disk of 2 GB
- One SCSI disk tower that has three shared SCSI disks of 2 GB each.

In addition, each machine has one 100X Ethernet Adapter card installed.

Each machine has the following software installed:

- Windows NT Version 4.0 Enterprise Edition with the MSCS feature installed
- DB2 Universal Database Enterprise Edition Version 7.

The resulting network configuration is as follows:

Server 1:	Server 2:
• Machine name:db2test1	• Machine name:db2test2
• TCP/IP hostname:db2test1	• TCP/IP hostname:db2test2
• IP Address: 9.9.9.1	• IP Address: 9.9.9.2
(subnet mask: 255.255.255.0	(subnet mask: 255.255.255.0
• MSCS cluster name: ClusterA	• MSCS cluster name: ClusterA

Both machines in the network are configured with TCP/IP and connected to a private LAN using an Ethernet 100 T-base Hub. In the absence of a Domain Name Server (DNS), all machines have a local TCP/IP `hosts` file, which contains the following entries:

```
9.9.9.1 db2test1 # for Server 1
9.9.9.2 db2test2 # for Server 2
9.9.9.3 ClusterA # for MSCS ClusterA
9.9.9.4 db2tcp1 # for DB2 remote client connection to Server 1
9.9.9.5 db2tcp2 # for DB2 remote client connection to Server 2
```

Preliminary Tasks

Before you perform the following tasks, it is assumed that both machines belong to the same domain, called `DB2NTD`:

1. Create a domain account for DB2 that is a member of the local Administrators group on those machines where DB2 is going to run. Use the account for performing all tasks:
   - Set the user name to db2nt.
   - Set the password to db2nt.
2. Install the MSCS feature on machines db2test1 and db2test2:
   - Name the MSCS cluster ClusterA.
   - The cluster IP Address is 9.9.9.3.
   - Share disk D: will be used by the MSCS software.
   - Share disks E: and F: will be used by DB2.
3. Install DB2 Universal Database Enterprise Edition Version 7 on machine db2test1. Install the software on C:\SQLLIB, which is a local drive.
4. Install DB2 Universal Database Enterprise Edition Version 7 on machine db2test2. Install the software on C:\SQLLIB, which is a local drive.

The next step is to set up the DB2MSCS.CFG file for each instance, and run the DB2MSCS utility for each instance.

Run the DB2MSCS Utility

To set up the db2test1 machine, perform the following tasks:

1. On machine db2test1, log on as user db2nt. The password is db2nt.
2. Create the DB2 instance DB2A, if it does not already exist. The command to create the instance is:
   ```
   db2icrt DB2A
   ```
3. Set up the DB2MSCS.CFG file for the DB2 instance on machine db2test1:
   ```
   #
   #  DB2MSCS.CFG for database system
   #  on machine db2test1
   DB2_INSTANCE=DB2A
   CLUSTER_NAME=ClusterA
   #
   # Group 1
   GROUP_NAME=DB2A Group
   IP_NAME=IP Address for DB2A
   IP_ADDRESS=9.9.9.4
   IP_SUBNET=255.255.255.0
   IP_NETWORK=ClusterA
   NETNAME_NAME=Network name for DB2A
   NETNAME_VALUE=DB2SRV1
   NETNAME_DEPENDENCY=IP Address for DB2A
   DISK_NAME=Disk E:
   INSTPROF_DISK=Disk E:
   ```
4. Run the DB2MSCS utility as follows:
   ```
   db2mscs -f:DB2MSCS.CFG
   ```
5. Log out from the db2nt account.

6. On machine db2test2, log on as user db2nt, which belongs to the local Administrators group. The password is db2nt.
7. Create the DB2 instance DB2B, if it does not already exist. The command to create the instance is:

   ```
   db2icrt DB2B
   ```
8. Set up the DB2MSCS.CFG file for the DB2 instance on machine db2test2:

   ```
   #
   #  DB2MSCS.CFG for database system
   #  on machine db2test2
   DB2_INSTANCE=DB2B
   CLUSTER_NAME=ClusterA
   #
   # Group 1
   GROUP_NAME=DB2B Group
   IP_NAME=IP Address for DB2B
   IP_ADDRESS=9.9.9.5
   IP_SUBNET=255.255.255.0
   IP_NETWORK=ClusterA
   NETNAME_NAME=Network name for DB2B
   NETNAME_VALUE=DB2SRV2
   NETNAME_DEPENDENCY=IP Address for DB2B
   DISK_NAME=Disk F:
   INSTPROF_DISK=Disk F:
   ```
9. Run the DB2MSCS utility as follows:

   ```
   db2mscs -f:DB2MSCS.CFG
   ```
10. Log out from the db2nt account.

Example - Setting up a Four-Node Partitioned Database System for Mutual Takeover

The objective for this example is to set up a four-node partitioned database system with failover support in a mutual takeover configuration. In this example, four servers are configured into two MSCS clusters. By using the mutual takeover configuration, if any machine fails, the database partition servers configured for that machine will fail over to the alternative machine, as configured using the MSCS software, and run as a logical node on the alternative machine.

There are two MSCS clusters in the resulting configuration. Each cluster has:

- Two servers, each with 64 MB of memory and one local SCSI disk of 2 GB
- One SCSI disk tower that has three shared SCSI disks of 2 GB each.

In addition, each machine has one 100X Ethernet Adapter card installed.

Each machine has the following software installed:

- Windows NT Version 4.0 Enterprise Edition with the MSCS feature installed

- DB2 Universal Database Extended Enterprise Edition Version 7.

The resulting network configuration is as follows:

Server 1: • Machine name:db2test1 • TCP/IP hostname:db2test1 • IP Address: 9.9.9.1 (subnet mask: 255.255.255.0 • MSCS cluster name: ClusterA	Server 2: • Machine name:db2test2 • TCP/IP hostname:db2test2 • IP Address: 9.9.9.2 (subnet mask: 255.255.255.0 • MSCS cluster name: ClusterA
Server 3: • Machine name:db2test3 • TCP/IP hostname:db2test3 • IP Address: 9.9.9.3 (subnet mask: 255.255.255.0 • MSCS cluster name: ClusterB	Server 4: • Machine name:db2test4 • TCP/IP hostname:db2test4 • IP Address: 9.9.9.4 (subnet mask: 255.255.255.0 • MSCS cluster name: ClusterB

All machines in the network are configured with TCP/IP and connected to a private LAN using an Ethernet 100 T-base Hub. In the absence of a Domain Name Server (DNS), all machines have a local TCP/IP hosts file, which contains the following entries:

```
9.9.9.1 db2test1 # for Server 1
9.9.9.2 db2test2 # for Server 2
9.9.9.3 db2test3 # for Server 3
9.9.9.4 db2test4 # for Server 4
9.9.9.5 ClusterA # for MSCS Cluster 1
9.9.9.6 ClusterB # for MSCS Cluster 2
9.9.9.7 db2tcp # for DB2 remote client connection
```

Preliminary Tasks

Before you perform the following tasks, it is assumed that all four machines belong to the same domain, called DB2NTD:

1. Create a domain account for DB2 that is a member of the local Administrators group on those machines where DB2 is going to run. Use the account for performing all tasks:
   - Set the user name to db2nt.
   - Set the password to db2nt.
2. Create a second domain account with the *"password never expires"* characteristic. This account will be associated with DB2 services:
   - Set the user name to db2mpp.
   - Set the password to db2mpp.
3. Install the MSCS feature on machines db2test1 and db2test2:

- Name the MSCS cluster ClusterA.
- The cluster IP Address is 9.9.9.5.
- Share disk D: will be used by the MSCS software.
- Share disks E: and F: will be used by DB2.

4. Install the MSCS feature on machines db2test3 and db2test4:
   - Name the MSCS cluster ClusterB.
   - The cluster IP Address is 9.9.9.6.
   - Share disk D: will be used by the MSCS software
   - Share disks E: and F: will be used by DB2.
5. Install DB2 Enterprise - Extended Edition on machine db2test1:
   - Select the *"This machine will be the instance-owning database partition server"* option.
   - The account for the DB2 service is db2mpp. The password is db2mpp.
   - Install the software on C:\SQLLIB, which is a local drive.
6. Install DB2 Enterprise - Extended Edition on machines db2test2, db2test3, and db2test4:
   - Select the *"This machine will be a new node on an existing partitioned database system"* option.
   - Select db2test1 as the instance-owning machine.
   - The account for the DB2 service is db2mpp. The password is db2mpp.
   - Install the software on C:\SQLLIB, which is a local drive.

The next step is to set up the DB2MSCS.CFG file and run the DB2MSCS utility.

Run the DB2MSCS Utility

To set up the db2test1 machine, perform the following tasks:

1. Log on as user db2nt, which belongs to the local Administrators group. The password is db2nt.
2. Set up the DB2MSCS.CFG file:

```
#
#  DB2MSCS.CFG for one partitioned database system with
#  multiple MSCS clusters
DB2_INSTANCE=DB2MPP
CLUSTER_NAME=ClusterA
DB2_LOGON_USERNAME=db2mpp
DB2_LOGON_PASSWORD=db2mpp
# Group 1
# for DB2 node 0
GROUP_NAME=DB2NODE0
DB2_NODE=0
IP_NAME=IP Address for DB2
IP_ADDRESS=9.9.9.7
IP_SUBNET=255.255.255.0
IP_NETWORK=Ethernet
```

```
NETNAME_NAME=Network name for DB2
NETNAME_VALUE=DB2WOLF
NETNAME_DEPENDENCY=IP Address for DB2
DISK_NAME=Disk E:
INSTPROF_DISK=Disk E:
#

# Group 2
# for DB2 node 1
GROUP_NAME=DB2NODE1
DB2_NODE=1
DISK_NAME=Disk F:
#

CLUSTER_NAME=ClusterB
# Group 3
# for DB2 node 2
GROUP_NAME=DB2NODE2
DB2_NODE=2
DISK_NAME=Disk E:

#
# Group 4
# for DB2 node 3
GROUP_NAME=DB2NODE3
DB2_NODE=3
DISK_NAME=Disk F:
```

3. Run the DB2MSCS utility as follows:

   ```
   db2mscs -f:DB2MSCS.CFG
   ```

4. Log out from the `db2nt` account.

The final steps are to register the database drive mapping for the two MSCS clusters.

Register the Database Drive Mapping for ClusterA

To register the database drive mapping for MSCS cluster `ClusterA`, perform the following tasks:

1. On machine `db2test1`, log on as user `db2mpp`, which is the account associated with DB2 services. The password is `db2mpp`.
2. To register the database drive mapping, enter the following commands:

   ```
   db2drvmp add 0 F E

   db2drvmp add 1 E F
   ```

3. Bring all DB2 resources offline, then bring them online.

Register the Database Drive Mapping for ClusterB

To register the database drive mapping for MSCS cluster `ClusterB`, perform the following tasks:

1. On machine `db2test3`, log on as user `db2mpp`, which is the account associated with DB2 services. The password is `db2mpp`.
2. To register the database drive mapping, enter the following commands:

```
db2drvmp add 2 F E
db2drvmp add 3 E F
```

3. Bring all DB2 resources offline, then bring them online.

Administering DB2 in an MSCS Environment

If you are using MSCS clusters, your DB2 instance requires additional planning with regards to daily operation, database deployment, and database configuration. For DB2 to execute transparently on any MSCS node, additional administrative tasks must be performed. All DB2 dependent operating system resources must be available on all MSCS nodes. Some of these operating system resources fall outside the scope of MSCS. That is, they cannot be defined as an MSCS resource. You must ensure that each system is configured such that the same operating system resources are available on all MSCS nodes. The sections that follow describe the additional work that must be done.

Starting and Stopping DB2 Resources

You must start and stop DB2 resources from the Cluster Administrator tool. Several mechanisms are available to start a DB2 instance, such as the **db2start** command, and the **Services** option from the Control Panel. However, if DB2 is not started from the Cluster Administrator, the MSCS software will not be aware of the state of the DB2 instance. If a DB2 instance is started using the Cluster Administrator, and stopped using the **db2stop** command, the MSCS software will interpret the **db2stop** command as a software failure, and attempt to restart DB2. (The current MSCS interfaces do not support notification of a *resource state*.)

Similarly, if you use **db2start** to start a DB2 instance, MSCS cannot detect that the resource is online; if a database server fails, MSCS will not bring the DB2 resource online on the failover machine in the cluster.

Three operations can be applied to a DB2 instance:

Online

This operation is equivalent to using the **db2start** command. If DB2 is already active, this operation can be used simply to notify MSCS that DB2 is active. Any errors during this operation will be written to the Windows NT Event Log.

Offline

This operation is equivalent to using the **db2stop** command. If there are any active attachments to an instance, this operation will fail. This is consistent with the behavior of **db2stop**.

Fail resource

This operation is equivalent to using the **db2stop** command with the

force option specified. DB2 will disconnect all applications from the DB2 system, and stop all database servers.

Running Scripts

You can run scripts both before and after a DB2 resource is brought online. These scripts *must* reside in the instance profile directory that is specified for the DB2INSTPROF environment variable. This directory is the directory path that is specified by the ″-p″ parameter of the **db2icrt** command. You can obtain this value by issuing the following command:

```
   db2set -i:instance_name DB2INSTPROF
```

This file path must be on a clustered disk, so that the instance directory is available on all cluster nodes.

These script files are not required, and are only run if they are found in the instance directory. They are launched by the MSCS Cluster Service in the background. The script files must redirect standard output to capture any information returned from commands within the script files. The output is not displayed to the screen.

In a partitioned database environment, by default, the same script will be used by every database partition server in the instance. If you need to distinguish among the different database partition servers in the instance, use different assignments of the DB2NODE environment variable to target specific node numbers (for example, use the IF statement in the `db2cpre.bat` and `db2cpost.bat` files).

Running Scripts Before Bringing DB2 Resources Online

If you want to run a script *before* you bring a DB2 resource online, the script *must* be named `db2cpre.bat`. DB2 calls functions that will launch this batch file from the Windows NT command line processor (CLP) and wait for the CLP to complete execution before the DB2 resource is brought online. You can use this batch file for tasks such as modifying the DB2 database manager configuration. You may want to change some database manager parameter values if the failover system is constrained, and you must reduce the system resources consumed by DB2.

The commands placed in the `db2cpre.bat` script should execute synchronously. Otherwise, the DB2 resource may be brought online before all tasks in the script are completed, which may result in unexpected behavior. Specifically, **db2cmd** should not be invoked in the `db2cpre.bat` script, because it, in turn, launches another command processor, which will run DB2 commands asynchronously to the **db2cmd** program.

If you want to use DB2 CLP commands in the `db2cpre.bat` script, the commands should be placed in a file and run as a CLP batch file from within a program that initializes the DB2 environment for the DB2 command line processor, and then waits for the completion of the DB2 command line processor. For example:

```
#include <windows.h>

int WINAPI DB2SetCLPEnv_api(DWORD pid);

void main ( int argc, char *argv [ ] )
{
      STARTUPINFO         startInfo   = {0};
      PROCESS_INFORMATION pidInfo     = {0};
      char     title  [32]   = "Run Synchronously";
      char     runCmd [64]  =
                             "DB2 -z c:\\run.out -tvf c:\\run.clp";
/* Invoke API to set up a CLP Environment */
      if ( DB2SetCLPEnv_api (GetCurrentProcessId ()) == 0 )  1
      {
         startInfo.cb          = sizeof(STARTUPINFO);
         startInfo.lpReserved  = NULL;
         startInfo.lpTitle     = title;
         startInfo.lpDesktop   = NULL;
         startInfo.dwX         = 0;
         startInfo.dwY         = 0;
         startInfo.dwXSize     = 0;
         startInfo.dwYSize     = 0;
         startInfo.dwFlags     = 0L;
         startInfo.wShowWindow = SW_HIDE;
         startInfo.lpReserved2 = NULL;
         startInfo.cbReserved2 = 0;
               if ( CreateProcessA( NULL,
                              runCmd,  2
                              NULL,
                              NULL,
                              FALSE,
                              NORMAL_PRIORITY_CLASS CREATE_NEW_CONSOLE,
                              NULL,
                              NULL,
                              &startInfo,
                              &pidInfo ) )
         {
            WaitForSingleObject (pidInfo.hProcess, INFINITE);
            CloseHandle (pidInfo.hProcess);
            CloseHandle (pidInfo.hThread);
         }
      }
      return;
}
```

1 The API DB2SetCLPEnv_api is resolved by the import library DB2API.LIB. This API sets an environment that allows CLP commands to be invoked. If this program is invoked from the db2cpre.bat script, the command processor will wait for the CLP commands to complete.

2 runCmd is the name of the script file that contains the DB2 CLP commands.

A sample program called db2clpex.exe can be found in the MISC subdirectory of the DB2 install path. This executable is similar to the example provided, but accepts the DB2 CLP command as a command line argument. If you want to use this sample program, copy it to the BIN subdirectory. You can use this executable in the db2cpre.bat script as follows (INSTHOME is the instance directory):

```
db2clpex "DB2 -Z INSTHOME\pre.log -tvf INSTHOME\pre.clp"
```

All DB2 ATTACH commands or CONNECT statements should explicitly specify a user; otherwise, they will be executed under the user account associated with the cluster service. CLP scripts should also complete with the TERMINATE command to end the CLP background process.

Following is an example of a db2cpre.bat file:

```
db2cpre.bat : 1
------------------------
db2clpex "db2 -z INSTHOME\pre-%DB2NODE%.log -tvf INSTHOME\pre.clp" 2 - 5
------------------------

PRE.CLP 6
------------------------
update dbm cfg using MAXAGENTS 200;
get dbm cfg;
terminate;
------------------------
```

1 The db2cpre.bat script executes under the user account associated with the Cluster Service. If DB2 actions are required, the user account associated with the Cluster Service must be a valid SQL identifier, as defined by DB2.

2 INSTHOME is the instance directory.

3 The name of the log file must be different for each node to avoid file contention when both logical nodes are brought online at the same time.

4 db2clpex.exe is a sample program that uses a command line argument to specify the CLP command that is to be invoked.

5 The db2clpex.exe sample program must be made available on all MSCS cluster nodes.

6 The CLP commands in this example set a limit on the number of agents.

Running Scripts After Bringing DB2 Resources Online

If you want to run a script *after* you bring a DB2 resource online, it *must* be named `db2cpost.bat`. The script will be run asynchronously from MSCS after the DB2 resource has been successfully brought online. The **db2cmd** command can be used in this script to execute DB2 CLP script files. Use the ″-c″ parameter of the **db2cmd** command to specify that the utility should close all windows on completion of the task. For example:

```
db2cmd -c db2 -tvf mycmds.clp
```

The ″-c″ parameter must be the first argument to the **db2cmd** command, because it prevents orphaned command processors in the background.

The `db2cpost.bat` script is useful if you want to perform database activities immediately after the DB2 resource fails over and becomes active. For example, you can restart or activate databases in the instance so that they are primed for user access.

Following is an example of a `db2cpost.bat` script:

```
db2cpost.bat 1
------------------------
db2cmd -c db2 -z INSTHOME\post-%DB2NODE%.log -tvf INSTHOME\post.clp 2 - 4
------------------------

POST.CLP 5
------------------------
restart database SAMPLE;
connect reset;
activate database SAMPLE;
terminate;
------------------------
```

1 The `db2cpost.bat` script runs under the user account associated with the Cluster Service. If DB2 actions are required, the user account associated with the Cluster Service must be a valid SQL identifier, as defined by DB2.

2 INSTHOME is the instance directory.

3 The name of the log file must be different for each node to avoid file contention when both logical nodes are brought online at the same time.

4 The **db2cmd** command can be used, because the `db2cpost.bat` script can run asynchronously. The ″-c″ parameter must be used to terminate the command processor.

5 The CLP script in this example contains commands to restart and activate the database. This script returns the database to an active state immediately after the database manager is started. In a partitioned database system, you should remove the ACTIVATE

DATABASE command, because multiple DB2 resources are brought online at the same time. The RESTART DATABASE command may fail, because another node is activating the database. If this occurs, rerun the script to ensure that the database is restarted correctly.

Database Considerations

When you create a database, ensure that the database path refers to a share disk. This allows the database to be seen on all MSCS nodes. All logs and other database files must also refer to clustered disks for DB2 to failover successfully. If you do not perform these steps, a DB2 system failure will occur, because it will seem to DB2 that files have been deleted or are unavailable.

Ensure also that the database manager and database configuration parameters are set so that the amount of system resources consumed by DB2 is supported on either MSCS node. The *autorestart* database configuration parameter should be set to `ON`, so that the first database connection on failover will bring the database to a consistent state. (The default setting for *autorestart* is `ON`.) The database can also be brought to a ready state by using the `db2cpost.bat` script to restart and activate the database. This method is preferred, because there will be no dependency on *autorestart*, and the database is brought to a ready state independent of a user connection request.

User and Group Support

DB2 relies on Windows NT for user authentication and group support. For a DB2 instance to fail over from one MSCS node to another in a seamless fashion, each MSCS node must have access to the same Windows NT security databases. You can achieve this by using Windows NT Domain Security.

Define all DB2 users and groups in a Domain Security database. The MSCS nodes must be members of this Domain, or the Domain must be a Trusted Domain. DB2 will then use the Domain Security database for authentication and group support, independent of the MSCS node on which DB2 is running.

If you are using local accounts, the accounts must be replicated on each MSCS node. This approach is not recommended, because it is error prone and requires dual maintenance.

DCE Security is also a supported authentication mode, if all MSCS nodes are clients in the same DCE cell.

You should associate the MSCS service with a user account that follows DB2 naming conventions. This allows the MSCS service to perform actions against DB2 that may be required in the `db2cpre.bat` and `db2cpost.bat` scripts.

For more information about Windows NT user and group support, see ″User Authentication with DB2 for Windows NT″ in the *Administration Guide: Implementation*.

Communications Considerations

DB2 supports two LAN protocols in an MSCS Environment:

- TCP/IP
- NetBIOS

TCP/IP is supported because it is a supported cluster resource type. To enable DB2 to use TCP/IP as a communications protocol for a partitioned database system, create an IP Address resource and place it in the same group as the DB2 resource that represents the database partition server that you want to use as a coordinator node for remote applications. Then create a dependency, using the Cluster Administrator tool, to ensure that the IP resource is online before the DB2 resource is started. DB2 clients can then catalog TCP/IP node directory entries to use this TCP/IP address.

The TCP/IP port associated with the *svcename* database manager configuration parameter must be reserved for use by the DB2 instance on all machines that participate in the instance. The service name associated with the port number must also be the same in the `services` file on all machines.

Although NetBIOS is not a supported cluster resource, you can use NetBIOS as a LAN protocol, because the protocol ensures that NetBIOS names are unique on the LAN. When DB2 registers a NetBIOS name, NetBIOS ensures that the name is not in use on the LAN. In a failover scenario, when DB2 is moved from one system to another, the *nname* used by DB2 will be deregistered from one partner machine in the MSCS cluster and registered on the other machine.

DB2 NetBIOS support uses NetBIOS Frames (NBF). This protocol stack can be associated with different logical adapter numbers (LANA). To ensure consistent NetBIOS access to the server, the LANA associated with the NBF protocol stack should be the same on all clustered nodes. You can configure this by using the **Networks** option from the Control Panel. You should associate NBF with LANA 0, because this is the default setting expected by DB2.

System Time Considerations

DB2 uses the system time to time stamp certain operations. All MSCS nodes that participate in DB2 failover must have the system time zone and system time synchronized to ensure that DB2 behaves consistently on all machines.

Set the system time zone using the **Date/Time** option from the Control Panel dialog. MSCS has a time service that synchronizes the date and time when the

MSCS nodes join to form a cluster. The time service, however, only synchronizes the time every 12 hours, which may result in problems if the time is changed on one system, and DB2 fails over before the time is synchronized.

If the time is changed on one of the MSCS cluster nodes, it should be manually synchronized on the other cluster nodes using the command:

```
net time /set /y \\remote_node
```

Where *remote_node* is the machine name of the cluster node.

Administration Server and Control Center Considerations in a Partitioned Database Environment

The DB2 Administration Server is (optionally) created during the installation of DB2 Universal Database. It is not a partitioned database system. The Control Center uses the services provided by the Administration Server to administer DB2 instances and databases.

In a partitioned database system, a DB2 instance can reside on multiple MSCS nodes. This implies that a DB2 instance must be cataloged on multiple systems under the Control Center so that the instance remains accessible, regardless of the MSCS node on which the DB2 instance is active.

The Administration Server instance directory is not shared. You must mirror all user-defined files in the Administration Server directory to all MSCS nodes to provide the same level of administration to all MSCS nodes. Specifically, you must make user scripts and scheduled executables available on all nodes. You must also ensure that scheduled activities are scheduled on all machines in an MSCS cluster.

Alternatively, instead of duplicating the Administration Server on all machines, you may want to have the Administration Server fail over. For the purposes of the following example, assume that you have two MSCS nodes in the cluster, and that they are called MACH0 and MACH1. MACH0 has access to a cluster disk that will be used by the Administration Server. Assume also that both MACH0 and MACH1 have an Administration Server. You would perform the following steps to make the Administration Server highly available:

1. Stop the Administration Server on both machines by invoking the **db2admin stop** command on each machine.
2. On all administration client machines, uncatalog all references to the Administration Servers on MACH0 and MACH1 using the UNCATALOG NODE command. (You can use the LIST NODE DIRECTORY command on the client machine to determine if any references to the Administration Server exist.)

3. Drop the Administration Server from MACH1 by invoking the **db2admin drop** command from MACH1. (You would only perform this step if you had an Administration Server on both machines.)
4. Determine the name of the Administration Server by issuing the **db2admin** command from MACH0. (The default name is `DB2DAS00`.)
5. Use the DB2MSCS utility to set up failover support for the Administration Server. This entails creating a DB2 resource on MSCS named `DB2DAS00` that has dependencies on the IP and disk resources. (If you have a mutual takeover configuration, you would put the resource in the group that holds the DB2 resource for NODE0.) This resource will be used as the MSCS resource that supports the Administration Server. The `DB2MSCS.ADMIN` file would be as follows:

   ```
   #
   # db2mscs.admin for Administration Server
   # run db2mscs -f:db2mscs.admin
   #
   DB2_INSTANCE=DB2DAS00
   CLUSTER_NAME=CLUSTERA
   DB2_LOGON_USERNAME=db2admin
   DB2_LOGON_PASSWORD=db2admin
   # put Administration server in the same group as DB2 Node 0
   GROUP_NAME=DB2NODE0  1
   DISK_NAME=DISK E:
   INSTPROF_DISK=DISK E:
   IP_NAME= IP Address for Administration Server
   IP_ADDRESS=9.9.9.8
   IP_SUBNET=255.255.255.0
   IP_NETWORK=Ethernet
   ```

   1 This group can be the same as the existing group. This way, you do not require an additional disk for the instance profile directory.
6. On MACH1, invoke the following command to set DB2DAS00 as the Administration Server:

   ```
   db2set -g db2adminserver=DB2DAS00
   ```
7. On MACH0, modify the start-up properties of DB2DAS00 through the Services program so that it is brought up manually and not automatically, because DB2DAS00 is now controlled by MSCS.

When the Administration Server is enabled for failover, all remote access should use an MSCS IP resource for communicating with the Administration Server. The Administration Server will now have the following properties:

- The Administration Server instance directory will fail over with the Administration Server.
- Clients will only catalog a single node to communicate with the Administration Server, regardless of the MSCS node on which it is active.
- Jobs only need to be scheduled once against the Administration Server.

- Local instances can only be controlled by the Administration Server when the Administration Server is active on the same MSCS node as the local instance.
- The Administration Server is not accessible if the Cluster Service is not active.

Limitations and Restrictions

When you run DB2 in an MSCS environment:

- You cannot use physical I/O on shared disks, unless the shared disks have the same physical disk number across both MSCS nodes. You can use logical I/O, because the disk is accessed using a partition identifier.
- You must configure all DB2 resource for MSCS support. If you do not, system errors will occur during DB2 run time (DB2 cannot properly operate in the absence of system resources). For example, if the database logs are not on an MSCS shared disk, DB2 cannot restart the database.
- You must manage a DB2 instance from the Cluster Administrator tool. MSCS will view other mechanisms that are used to start and stop the database manager as software inconsistencies. For example, if you use MSCS to start DB2, and the **db2stop** command to stop DB2, MSCS will detect this as a software failure, and will restart the instance. This also means that you should not use the Control Center to start and stop DB2.
- To uninstall DB2, you must first stop MSCS.

Chapter 14. DB2 and High Availability on Sun Cluster 2.2

This chapter describes in detail how DB2 works with Sun Cluster 2.x (SC2.x) to achieve high availability, and includes a description of the high availability agent, which acts as a mediator between the two software products (see Figure 59).

DB2
HA Agent
SC2.x

Figure 59. DB2, Sun Cluster 2.x, and High Availability

High Availability

The computer systems that host data services contain many distinct components, and each component has a ″mean time before failure″ (MTBF) associated with it. The MTBF is the average time that a component will remain usable. The MTBF for a quality hard drive is in the order of one million hours (approximately 114 years). While this seems like a long time, one out of 200 disks is likely to fail within a 6-month period.

Although there are a number of methods to increase availability for a data service, the most common is an HA cluster. A cluster, when used for high availability, consists of two or more machines, a set of private network interfaces, one or more public network interfaces, and some shared disks. This special configuration allows a data service to be moved from one machine to another. By moving the data service to another machine in the cluster, it should be able to continue providing access to its data. Moving a data service from one machine to another is called a *failover*, as illustrated in Figure 60 on page 290.

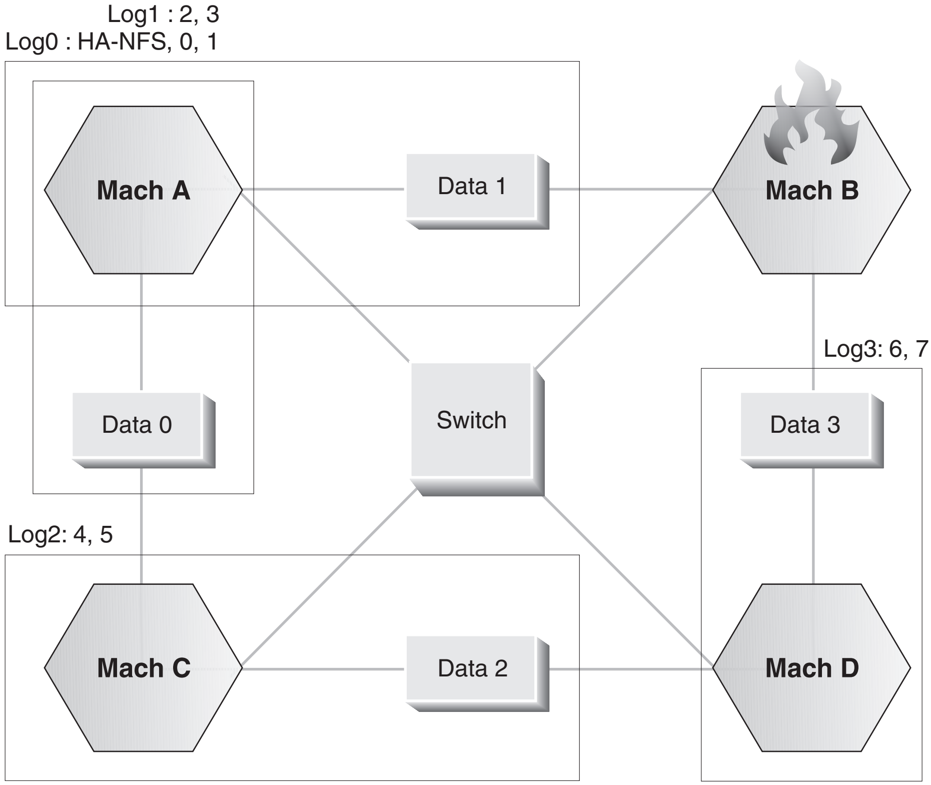


Figure 60. Failover

The private network interfaces are used to send *heartbeat* messages, as well as control messages, among the machines in the cluster. The public network interfaces are used to communicate directly with clients of the HA cluster. The disks in an HA cluster are connected to two or more machines in the cluster, so that if one machine fails, another machine has access to them.

A data service running on an HA cluster has one or more logical public network interfaces and a set of disks associated with it. The clients of an HA data service connect via TCP/IP to the logical network interfaces of the data service only. If a failover occurs, the data service, along with its logical network interfaces and set of disks, are moved to another machine.

One of the benefits of an HA cluster is that a data service can recover without the aid of support staff, and it can do so at any time. Another benefit is redundancy. All of the parts in the cluster should be redundant, including the machines themselves. The cluster should be able to survive any single point of failure.

Even though highly available data services can be very different in nature, they have some common requirements. Clients of a highly available data service expect the network address and host name of the data service to remain the same, and expect to be able to make requests in the same way, regardless of which machine the data service is on.

Consider a Web browser that is accessing a highly available Web server. The request is issued with a URL (Uniform Resource Locator), which contains both a host name, and the path to a file on the Web server. The browser expects both the host name and the path to remain the same after a failover ofáthe Web server. If the browser is downloading a file from the Web server, and the server is failed over, the browser will need to reissue the request.

Availability of a data service is measured by the amount of time the data service is available to its users. The most common unit of measurement for availability is the percentage of ″up time″; this is often referred to as the number of ″nines″:

```
99.99% => service is down for (at most) 52.6 minutes / yr
99.999% => service is down for (at most) 5.26 minutes / yr
99.9999% => service is down for (at most) 31.5 seconds / yr
```

When designing and testing an HA cluster:

1. Ensure that the administrator of the cluster is familiar with the system and what should happen when a failover occurs.
2. Ensure that each part of the cluster is truly redundant and can be replaced quickly if it fails.
3. Force a test system to fail in a controlled environment, and make sure that it fails over correctly each time.
4. Keep track of the reasons for each failover. Although this should not happen often, it is important to address any issues that make the cluster unstable. For example, if one piece of the cluster caused a failover five times in one month, find out why and fix it.
5. Ensure that the support staff for the cluster is notified when a failover occurs.
6. Do not overload the cluster. Ensure that the remaining systems can still handle the workload at an acceptable level after a failover.
7. Check failure-prone components (such as disks) often, so that they can be replaced before problems occur.

Fault Tolerance and Continuous Availability

Another way to increase the availability of a data service is fault tolerance. A *fault tolerant* machine has all of its redundancy built in, and should be able to withstand a single failure of any part, including CPU and memory. Fault tolerant machines are most often used in niche markets, and are usually expensive to implement. An HA cluster with machines in different geographical locations has the added advantage of being able to recover from a disaster affecting only a subset of those locations.

Continuous availability is a step above high availability. It shelters its clients from both planned and unplanned down time. With a continuous availability configuration, the client is completely unaffected if one of the machines hosting the data service fails or is brought down for maintenance. Continuous availability configurations are complex and more expensive to implement.

An HA cluster is the most common solution to increase availability because it is scalable, easy to use, and relatively inexpensive to implement.

Sun Cluster 2.2

Sun Cluster 2.2 (SC2.2) is Sun Microsystems' clustering and high availability product. SC2.2 supports up to four machines in a single cluster. Using four Ultra Enterprise 10000s, a cluster can have up to 256 CPUs and 256 GB of RAM.

Supported Systems

System	UltraSPARC	Memory Capacity	I/O
Ultra Enterprise 1	1	64MB-1GB	3 SBus
Ultra Enterprise 2	1-2	64MB-2GB	4 SBus
Ultra Enterprise 450	1-4	32MB-4GB	10 PCI
Ultra Enterprise 3000	1-6	64MB-6GB	9 SBus
Ultra Enterprise 4000	1-14	64MB-14GB	21 SBus
Ultra Enterprise 5000	1-14	64MB-14GB	21 SBus
Ultra Enterprise 6000	1-30	64MB-30GB	45 SBus
Ultra Enterprise 10000	1-64	512MB-64GB	64 SBus

Agents

The Sun Cluster software includes a number of high availability agents that are supported and shipped with the SC2.2 product. Other HA agents, such as the one for DB2, are developed outside of Sun, and are not shipped with the Sun Cluster software. The HA agent for DB2 is shipped with DB2, and supported by IBM.

The Sun Cluster software works with highly available data services by providing an opportunity to register methods (scripts or programs) that correspond to various components of the Sun Cluster software. Utilizing these methods, the SC2.2 software can control a data service without having intimate knowledge of it. These methods include:

START
Used to start portions of the data service before the logical network interfaces are online.

START_NET
Used to start portions of the data service after the logical network interfaces are online.

STOP Used to stop portions of the data service after the logical network interfaces are offline.

STOP_NET
Used to stop portions of the data service before the logical network interfaces are offline.

ABORT
Like the STOP method, except it is run just before a machine is brought down by the cluster software. In this case, the machine's ″health″ is in question, and a data service may want to execute ″last wish″ requests before the machine is brought down. Run after the logical network interfaces are offline.

ABORT_NET
Like the ABORT method, except it is run before the logical network interfaces are offline.

FM_INIT
Used to initialize fault monitors.

FM_START
Used to start the fault monitors.

FM_STOP
Used to stop the fault monitors.

FM_CHECK
Called by the **hactl** command. Returns the current status of the corresponding data service.

The DB2 agent consists of the following scripts: START_NET, STOP_NET, FM_START, and FM_STOP. The following scripts are not run during cluster reconfiguration: ABORT, ABORT_NET, and FM_CHECK.

A high availability agent consists of one or more of these methods. The methods are registered with SC2.2 through the **hareg** command. Once registered, the Sun Cluster software will call the corresponding method to control the data service.

It is important to remember that the ABORT and STOP methods of a service may not be called. These methods are intended for the controlled shutdown of a data service, and the data service must be able to recover if a machine fails without calling them.

For more information, refer to the Sun Cluster documentation.

Logical Hosts

The SC2.2 software uses the concept of a logical host. A *logical host* consists of a set of disks and one or more logical public network interfaces. A highly available data service is associated with a logical host, and requires the disks that are in the disk groups of the logical host. Logical hosts can be hosted by different machines in the cluster, and "borrow" the CPUs and memory of the machine on which they are running.

Logical Network Interfaces

As with other UNIX based operating systems, Solaris has the ability to have extra IP addresses, in addition to the primary one for a network interface. The extra IP addresses reside on a logical interface in the same way that the primary IP address resides on the physical network interface. Following is an example of the logical interfaces on two machines in a cluster. There are two logical hosts, and both are currently on the machine "thrash".

```
scadmin@crackle(202)# netstat -in
Name Mtu Net/Dest Address Ipkts Ierrs Opkts Oerrs Collis Queue
lo0 8232 127.0.0.0 127.0.0.1 289966 0 289966 0 0 0
hme0 1500 9.21.55.0 9.21.55.98 121657 6098 764122 0 0 0
scid0 16321 204.152.65.0 204.152.65.1 489307 0 476479 0 0 0
scid0:1 16321 204.152.65.32 204.152.65.33 0 0 0 0 0 0
scid1 16321 204.152.65.16 204.152.65.17 347317 0 348073 0 0 0

  1. lo0 is the loopback interface
  2. hme0 is the public network interface (ethernet)
  3. scid0 is the first private network interface (SCI or Scalable
     Coherent Interface)
  4. scid0:1 is a logical network interface that the Sun Cluster software
     uses internally
  5. scid1 is the second private network interface

scadmin@thrash(203)# netstat -in
Name Mtu Net/Dest Address Ipkts Ierrs Opkts Oerrs Collis Queue
lo0 8232 127.0.0.0 127.0.0.1 1128780 0 118780 0 0 0
hme0 1500 9.21.55.0 9.21.55.92 1741422 5692 757127 0 0 0
hme0:1 1500 9.21.55.0 9.21.55.109 0 0 0 0 0 0
hme0:2 1500 9.21.55.0 9.21.55.110 0 0 0 0 0 0
scid0 16321 204.152.65.0 204.152.65.2 476641 0 489476 0 0 0
```

```
scid0:1 16321 204.152.65.32 204.152.65.34 0 0 0 0 0 0
scid1 16321 204.152.65.16 204.152.65.18 348199 0 347444 0 0 0

  1. hme0:1 is a logical network interface for a logical host
  2. hme0:2 is a logical network interface for another logical host
```

A logical host can have one or more logical interfaces associated with it. These logical interfaces move with the logical host from machine to machine, and are used to access the data service that is associated with the logical host. Because these logical interfaces move with the logical hosts, clients can access the data service independently of the machine on which it resides.

A highly available data service should bind to the TCP/IP address INADDR_ANY. This ensures that each IP address on the system can accept connections for the data service. If a data service binds to a specific IP address instead, it must bind the logical interface associated with the logical host that is hosting the data service. Binding to INADDR_ANY also removes the need to rebind to a new IP address if one arrives on the system that is needed by the data service.

Note: Clients of an HA instance should catalog the database using the host name for the logical IP address of a logical host. They should never use the primary host name for a machine, because there is no guarantee that DB2 will be running on that machine.

Disk Groups and File Systems

Disks for a data service are associated with a logical host in groups (or sets). If the cluster is running Sun StorEdge Volume Manager (Veritas), the Sun Cluster software uses the Veritas ″vxdg″ utility to import and deport the disk groups for each logical host. Following is an example of the disk groups for two logical hosts, ″log0″ and ″log1″, which are being hosted by a machine called ″thrash″. The machine called ″crackle″ is not currently hosting any logical hosts.

```
scadmin@crackle(206)# vxdg list
NAME STATE ID
rootdg enabled 899825206.1025.crackle

scadmin@thrash(205)# vxdg list
NAME STATE ID
rootdg enabled 924176206.1025.thrash
data0 enabled 925142028.1157.crackle=
data1 enabled 899826248.1108.crackle
```

The disk groups ″data0″ and ″data1″ correspond to the logical hosts ″log0″ and ″log1″, respectively. The disk group ″data0″ can be deported from ″thrash″ by running

```
vxdg deport data0
```

and imported to "crackle" by running

```
vxdg import data1
```

This is done automatically by the Sun Cluster software, and should not be done manually on a live cluster.

Each disk group contains a number of disks that can be shared between two or more machines in the cluster. A logical host can only be moved to another machine that has physical access to the disks in the disk groups that belong to it.

There are two files that control the file systems for each logical host:

```
/etc/opt/SUNWcluster/conf/hanfs/vfstab.<logical_host>
/etc/opt/SUNWcluster/conf/hanfs/dfstab.<logical_host>
```

where *logical_host* is the name of the associated logical host name.

The vfstab file is similar to the /etc/vfstab file, except that it contains entries for the file systems to be mounted after the disk groups have been imported for a logical host. The dfstab file is similar to the /etc/dfs/dfstab file, except that is contains entries for file systems to export through HA-NFS for a logical host. Each machine has its own copy of these files, and care should be taken to ensure that they have the same content on each machine in the cluster.

Note: The paths for the vfstab and dfstab files of a logical host are misleading, because they contain the directory hanfs. Only the dfstab file for a logical host is used for HA-NFS. The vfstab file is used, even if HA-NFS is not configured.

Following are examples from a cluster running DB2 Universal Database Enterprise - Extended Edition (EEE) in a mutual takeover configuration:

```
scadmin@thrash(217)# ls -l /etc/opt/SUNWcluster/conf/hanfs
total 8
-rw-r--r-- 1 root build 173 Apr 14 15:01 dfstab.log0
-rw-r--r-- 1 root build 316 Apr 26 12:07 vfstab.log0
-rw-r--r-- 1 root build 389 Apr 13 21:04 vfstab.log1

scadmin@thrash(218)# cat dfstab.log0
share -F nfs -o root=crackle:thrash:\
jolt:bump:crackle.torolab.ibm.com:thrash.torolab.ibm.com:\
jolt.torolab.ibm.com:bump.torolab.ibm.com /log0/home
```

The hosts, which are given permission to mount the file system, /log0/home, are from all of the network interfaces (logical and physical) on each machine in the cluster. The file systems are exported with root permissions.

```
scadmin@thrash(220)# cat vfstab.log0
#device to mount             device  to fsck              mount
#                                                         point

/dev/vx/dsk/data0/data1-stat /dev/vx/rdsk/data0/data1-stat /log0
/dev/vx/dsk/data0/vol01      /dev/vx/rdsk/data0/vol01      /log0/home
/dev/vx/dsk/data0/vol02      /dev/vx/rdsk/data0/vol02      /log0/data

scadmin@thrash(221)# cat vfstab.log1
#device to mount             device  to fsck              mount
#                                                         point
/dev/vx/dsk/data1/data1-stat /dev/vx/rdsk/data1/data1-stat /log1
/dev/vx/dsk/data1/vol01      /dev/vx/rdsk/data1/vol01      /log1/home
/dev/vx/dsk/data1/vol02      /dev/vx/rdsk/data1/vol02      /log1/data
/dev/vx/dsk/data1/vol03      /dev/vx/rdsk/data1/vol03      /log1/data1
```

```
FS   fsck mount   options
type pass at boot

ufs  2    no      -
ufs  2    no      -
ufs  2    no      -


FS   fsck mount   options
type pass at boot

ufs  2    no      -
ufs  2    no      -
ufs  2    no      -
ufs  2    no      -
```

The vfstab.log0 file contains three valid entries for file systems under the /log0 directory. Notice that the file systems for the logical host log0 use logical volume devices, which are part of the disk group data0 that is associated with the logical host.

The file systems in the vfstab files are mounted in order from top to bottom, so it is important to ensure that the file systems are listed in the correct order. File systems that are mounted underneath a particular file system should be listed below it. The actual file systems that are needed for a logical host depend on the needs of the data service, and will vary considerably from these examples.

During a failover, the SC2.2 software is responsible for ensuring that the disk groups and logical interfaces associated with a logical host follow it around the cluster from machine to machine. The highly available data service expects to have at least these resources available on a new system after a failover. In fact, many data services are not even aware that they are highly available, and must have these resources ″appear″ to be exactly the same after a failover.

Control Methods

The control methods are registered using

```
hareg(1m)
```

Once an HA service is registered, SC2.2 is responsible for calling the methods that were registered for the HA service at appropriate times during a cluster reconfiguration or failover.

The following actions take place (in the given order) during a cluster reconfiguration (controlled failover). Actions preceding step 5c will not be taken if a machine crashes. (For more information about cluster reconfiguration, refer to the SC2.2 documentation.)

```
 1. FM_STOP method is run.
 2. STOP_NET method is run.
 3. Logical interfaces for the logical host are brought offline.
      - ifconfig hme0:1 0.0.0.0 down
 4. STOP method is run.
 5. Disk groups and file systems are moved.
      a. Unmount logical host file systems.
      b. vxdg deport disk groups on one machine.

    - - Only the steps below are run if a machine crashes - -

      c. vxdg import disk groups on the other machine.
      d. fsck logical host file systems.
      e. Mount logical host file systems.
 6. START method is run.
 7. Logical interfaces for the logical host are brought online.
      - ifconfig hme0:1 <ip address> up
 8. START_NET method is run.
 9. FM_INIT method is run.
10. FM_START method is run.
```

The control methods are run with the following command line arguments:

```
METHOD <logical hosts being hosted> <logical hosts not being hosted> <time-out>
```

The first argument is a comma delimited list of logical hosts that are currently being hosted, and the second is a comma delimited list of logical hosts that are not being hosted. The last argument is the time-out for the method, the amount of time that the method is allowed to run before the SC2.2 software aborts it.

Disk and File System Configuration

SC2.2 supports two volume managers: Sun StorEdge Volume Manager (Veritas) and Solstice Disk Suite. Although both work well, the StorEdge Volume Manager has some advantages in a clustered environment. In some cluster configurations, the controller number for a disk enclosure can be different for each machine in the cluster. If the controller number is different, the paths for the disk devices for the controller will also be different. Because Disk Suite works directly with the disk device paths, it will not work well in this situation. The StorEdge Volume Manager works with the disks themselves, regardless of the controller number, and is not affected if the controller numbers are different.

Since the goal of HA is to increase availability for a data service, it is important to ensure that all file systems and disk devices are mirrored, or in a RAID configuration. This will prevent failovers due to a failed disk, and increase the stability of the cluster.

HA-NFS

DB2 UDB EEE requires a shared file system when an instance is configured across multiple machines. A typical DB2 UDB EEE configuration has the home directory exported from one machine through NFS, and mounted on all of the machines participating in the EEE instance. For a mutual takeover configuration, DB2 UDB EEE depends on HA-NFS to provide a shared, highly available file system. One of the logical hosts exports a file system through HA-NFS, and each machine in the cluster then mounts the file system as the home directory of the EEE instance. For more information about HA-NFS, refer to the Sun Cluster documentation.

The cconsole and ctelnet Utilities

Two useful utilities that come with SC2.2 are `cconsole` and `ctelnet`. These utilities can be used to issue a single command to several machines in a cluster simultaneously. Editing a configuration file with these utilities ensures that it will remain identical on all of the machines in the cluster. These utilities can also be used to install software in exactly the same way on each machine. For more information about these utilities, refer to the Sun Cluster documentation.

Campus Clustering and Continental Clustering

A cluster is called a *campus cluster* when its machines are not in the same building. A campus cluster is useful for removing the building itself as the single point of failure. For example, if the machines in the cluster are all in the same building, and it burns down, the entire cluster is affected. However, if the machines are in different buildings, and one of the buildings burns down, the cluster survives.

A *continental cluster* is a cluster whose machines are distributed among different cities. In this case, the goal is to remove the geographic region as the single point of failure. This type of cluster provides protection against catastrophic events, such as earthquakes and tidal waves.

Currently, a Sun Cluster can support machines as far apart as 10 km, or about 6 miles. This makes campus clustering a viable option for those who need high speed connections between two different sites. A cluster requires two private interconnects, and a number of fiber optic cables for the shared disks. The cost of high speed connections between two sites may offset the benefits.

Common Problems

The SC2.2 software uses the Cluster Configuration Database, or CCD(4), to provide a single cluster-wide repository for the cluster configuration. The CCD has a private API and is stored under the `/etc/opt/SUNWcluster/conf` directory. In rare cases, the CCD can go out of synchrony, and may need to be repaired. The best way to repair the CCD in this situation is to restore it from a backup copy.

To back up the CCD, shut down the cluster software on all machines in the cluster, ″tar″ up the `/etc/opt/SUNWcluster/conf` directory, and store the tar file in a safe place. If the cluster software is not shut down when the backup is made, you may have trouble restoring the CCD. Ensure that the backup copy is kept up-to-date by taking a fresh backup any time that the cluster configuration is changed. To restore the CCD, shut down the cluster software on all machines in the cluster, move the `conf` directory to `conf.old`, and ″untar″ the backup copy. The cluster can then be started with the new CCD.

DB2 Considerations

The following topics are covered in this section:

- “Applications Connecting to an HA Instance”
- “Disk Layout for EE and EEE Instances” on page 302
- “Home Directory Layout for EE and EEE Instances” on page 303
- “Logical Hosts and DB2 UDB EEE” on page 304
- “DB2 Installation Location and Options” on page 305
- “Database and Database Manager Configuration Parameters” on page 305
- “Crash Recovery” on page 306
- “High Availability through Data Replication” on page 306.

Applications Connecting to an HA Instance

Applications that rely on a highly available DB2 instance must be able to reconnect in the event of a failover. Since the host name and IP address of a logical host remain the same, there is no need to connect to a different host name or to recatalog the database.

Consider a cluster with two machines and one DB2 Universal Database Enterprise Edition (EE) instance. The EE instance will normally reside on one of the machines in the cluster. Clients of the HA instance will connect to the logical IP address (or host name) of the logical host associated with the HA instance.

According to an HA client, there are two types of failover. One type occurs if the machine that is hosting the HA instance crashes. The other type occurs when the HA instance is given an opportunity to shut down gracefully.

If a machine crashes and takes down the HA instance, both existing connections and new connections to the database will hang. The connections hang because there are no machines on the network with the IP address that the clients were using for the database. If the database is shut down gracefully, a `db2stop force` breaks existing connections to the database, and an error message is returned.

During the failover, the logical IP address associated with the database is offline, either because the SC2.2 software took it offline, or because the machine that was hosting the logical host crashed. At this point, any new connections to the database will hang for a short period of time.

The logical IP address associated with the database is eventually brought up on another machine before DB2 is started. At this stage, a connection to the database will not hang, but will receive a communication error, because DB2 has not yet been started on the system. DB2 clients that were still connected to the database will also begin receiving communication errors. Although the clients still believe they are connected, the machine that has started hosting the logical IP address has no knowledge of any existing connections. The connections are simply reset, and the DB2 client receives a communication error. After a short time, DB2 will be started on the machine, and a successful connection to the database can be made. At this point, the database may be inconsistent, and clients may have to wait for it to recover.

When designing an application for an HA environment, it is not necessary to write special code for the stages where the database connections hang. The connections only hang for a short period of time while the Sun Cluster software moves the logical IP address. Any data service running on Sun Cluster will experience the same hanging connections during this stage. No matter how the database comes down, the clients will receive an error message, and must try to reconnect until successful. From the client’s perspective, it is as if the HA instance went down, and was brought back up on the same machine. In a controlled failover, it appears to the client that it was forced off, and that it can later reconnect to the database on the same machine. In an uncontrolled failover, it appears to the client that the database server crashed, and was soon brought back up on the same machine.

Disk Layout for EE and EEE Instances

DB2 expects the disk devices or file systems it requires to appear the same on each machine in the cluster. To ensure that this happens, the required disks or file systems should be configured in such a way that they follow the logical host associated with the HA instance, and will have the same path names on each machine in the cluster.

Both DMS and SMS table spaces are supported in an HA environment. Device containers for DMS table spaces must use raw devices created by the volume manager, which are either mirrored, or in a RAID configuration. Regular disk devices, such as `/dev/rdsk/c20t0d0s0` should not be used because:

- It increases the possibility that the device could be written to from more than one machine at the same time.
- The controller number may be different on another machine.

If DB2 is failed over in this situation, the disk devices it requires will not look the same as they did on the other machine, and it will not start. File containers for DMS table spaces, and containers for SMS table spaces, must reside on mounted file systems. The file systems for a logical host are mounted automatically when they are included in the `vfstab` file for the logical host.

The `vfstab` file for a logical host is in the path:

```
/etc/opt/SUNWcluster/conf/hanfs/vfstab.<logical_host>
```

where *logical_host* is the name of the logical host that is associated with the `vfstab` file.

Each logical host has its own `vfstab` file, which contains file systems that are to be mounted after the disk groups for the logical host have been transferred to the current machine, but before the HA services are started. The Sun Cluster software will try to mount any file system that is properly defined after running **fsck** (file system check), to ensure the health of the file system. If **fsck** fails, the file system will not be mounted, and an error message is logged.

Note: If a process has an open file, or its current working directory is under a mount point, the mount will fail. To prevent this, ensure that no processes are left under the mount points contained in the logical host's `vfstab` file.

Any convention can be used for the file system layout of an EEE instance when using SMS table spaces. Following is the convention used by the hadb2_setup utility:

```
scadmin@crackle(190)# pwd
/export/ha_home/db2eee/db2eee
scadmin@crackle(191)# ls -l
total 18
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0000 -> /log0/disks/db2eee/NODE0000
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0001 -> /log0/disks/db2eee/NODE0001
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0002 -> /log0/disks/db2eee/NODE0002
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0003 -> /log0/disks/db2eee/NODE0003
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0004 -> /log0/disks/db2eee/NODE0004
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0005 -> /log1/disks/db2eee/NODE0005
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0006 -> /log1/disks/db2eee/NODE0006
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0007 -> /log1/disks/db2eee/NODE0007
lrwxrwxrwx 1 root build 28 Aug 12 19:08 NODE0008 -> /log1/disks/db2eee/NODE0008
scadmin@crackle(192)#
```

The instance owner is `db2eee`, and the default database directory for the `db2eee` instance is `/export/ha_home/db2eee`. Logical host `log0` is hosting database partitions 0, 1, 2, and 3, while logical host `log1` is hosting database partitions 4, 5, 6, 7, and 8.

For each database partition, there is a corresponding `NODE`*`xxxx`* directory. The node directories for the database partitions point to a directory under the associated logical host file system.

When choosing a path convention, ensure that:

1. The disks for the file system are in a disk group of the logical host responsible for the database partitions that need them.
2. The file systems that hold containers are mounted through the logical host’s `vfstab` file.

Home Directory Layout for EE and EEE Instances

For an EE instance, the home directory should be a file system that is defined in the `vfstab` file for a logical host. This directory will be available before DB2 is started, and is transferred with DB2 to wherever the logical host is moved in the cluster. Each machine has its own copy of the `vfstab` file, and care should be taken to ensure that it has the same contents on each machine. Following is an example of the home directory for an EE instance:

```
/log0/home/db2ee
```

where `/log0` is the logical host file system for the logical host `log0`, and `db2ee` is the name of the DB2 instance. This home directory path should be placed in the `/etc/passwd` file on each machine in the cluster that could host the ″db2ee″ instance.

For an EEE instance, there are two ways to set up the home directory. For a hot standby configuration, the home directory can be set up in the same way as for an EE instance. For a mutual takeover configuration, HA-NFS must be used for the home directory, and must be configured properly *before* setting up the EEE instance.

One of the machines in the cluster must export the file system for the EEE instance, using the `dfstab` file for a chosen logical host. The `dfstab` file contains file systems that should be exported through NFS when a machine is hosting a logical host. Each machine has its own copy of the `dfstab` file, and care should be taken to ensure that it has the same contents on each machine.

Information for the HA-NFS file system is placed in the `hadb2tab` file (through the `hadb2_setup` program). When an HA agent reads the information for the instance, it automatically mounts the HA-NFS file system for the instance (see "The hadb2tab File" on page 307).

The mount point for the HA-NFS file system is typically `/export/ha_home`. On each machine in the cluster, this would be NFS mounted from the logical host that is exporting the HA-NFS directory. The EEE instance owner's home directory is placed under this directory and is called:

```
/export/ha_home/<instance>
```

where *instance* is the name of the instance owner.

One could have a home directory for an instance on each machine, to avoid having to mount or unmount it. Doing this requires extra administrative overhead to ensure that the home directories remain identical on each machine. Failure to do so can prevent DB2 from starting properly, or cause it to start with a different configuration. This is *not* a supported configuration.

Logical Hosts and DB2 UDB EEE

A logical host is usually chosen to host one or more database partitions, as well as export the HA-NFS file system. For example, if there are four database partitions and two machines in the cluster, there should be one logical host for each machine (Figure 61 on page 305). One logical host could host two database partitions, and export the HA-NFS file system, while the other logical host could host the remaining two database partitions.

By default, a DB2 UDB EEE instance allocates enough resources to successfully add up to two database partitions to a machine that already has one or more live database partitions for that instance. For example, if there are four database partitions for a single instance on a cluster, this will only be a concern if there is one database partition per logical host, or one logical host is hosting three database partitions. In either case, it is possible to have three database partitions fail over to a machine that is already hosting a database partition for the same instance.

The DB2_NUM_FAILOVER_NODES registry variable can be used to increase the amount of resource reserved for database partitions that are failed over.

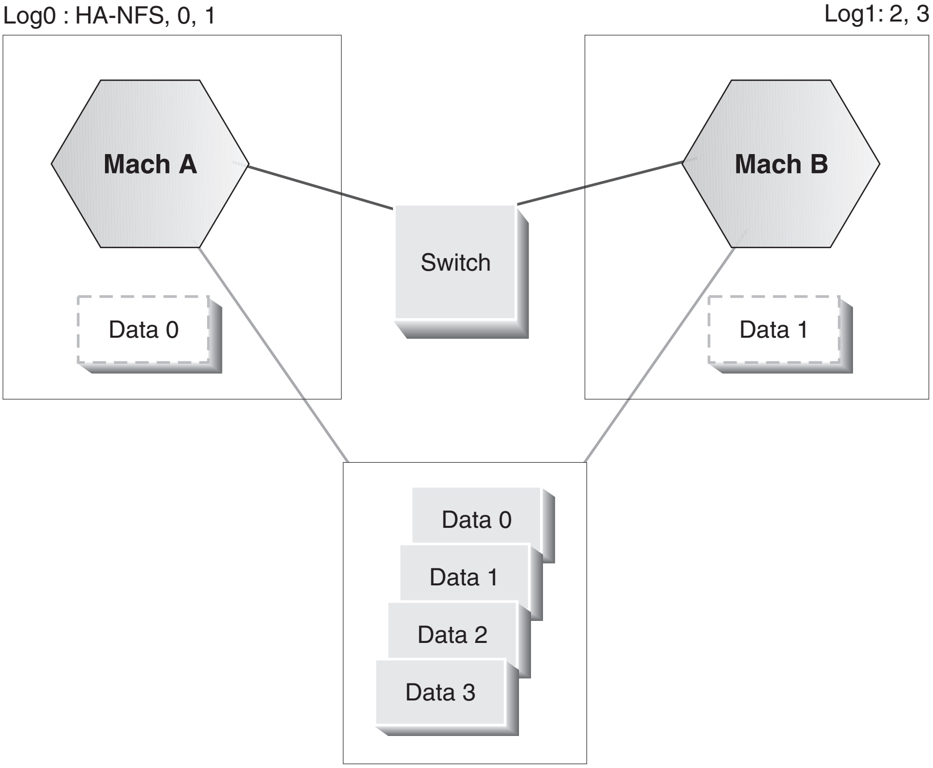


Figure 61. One Logical Host For Each Machine

DB2 Installation Location and Options

The file system on which DB2 is installed should be mirrored, or at least be in a RAID configuration. If DB2 is installed on regular disks, disk failure is more likely; the resulting failover is considered preventable, and decreases the stability of the cluster.

DB2 cannot be installed on disks in a disk group for a logical host, because the HA agent always needs to have access to the DB2 libraries. If the HA agents do not have access to the DB2 libraries, they will fail. DB2 must be installed normally on each machine in the cluster.

Database and Database Manager Configuration Parameters

The database manager configuration parameters can be changed after a failover, and before DB2 is started, by using the `pre_db2start` script (see “User Scripts” on page 309). This executable script is run (if it exists) under the `sqllib/ha` directory of the instance owner’s home directory. As the name suggests, it is run just before **db2start**. The same arguments that are passed to the control methods are passed to the `pre_db2start` script, unless the instance is an EEE instance. For an EEE instance, the `pre_db2start` script is also passed the node number for the **db2start** command.

Crash Recovery

Crash recovery in an HA environment is the same as it would in a regular environment. Even if the HA instance is brought up on a different machine from the one on which it crashed, the files and disk devices for the instance will look the same, and the actions needed to recover the database will not be different. For more information about crash recovery and other forms of database recovery, see "Recovering a Database" in the *Administration Guide: Implementation*.

Although a database can be restarted manually (or through one of the user scripts), it is recommended that the *autorestart* database configuration parameter be set to `ON`, especially for an EEE instance. This will minimize the amount of time that the database is in an inconsistent state.

High Availability through Data Replication

Data availability can also be enhanced through replication. By replicating data between two servers, a form of high availability is achieved. If one of the servers goes down, the other server should be able to take over and continue to provide the data service.

However, because the replication is done asynchronously, some changes may not have been propagated to the other server when that server goes down.

The DB2 High Availability Agent

The DB2 high availability agent acts like a mediator between DB2 and SC2.x. It provides a way for the Sun Cluster 2.2 software to control DB2 in a clustered environment, without having intimate knowledge of DB2. There is one agent for both EE and EEE instances. The agent supports both administrative instances and database instances.

Registering the hadb2 Service

To work with SC2.2, the DB2 HA agent must be registered. Registering a data service tells SC2.2 which control methods are available, and in which directory they reside. A special script called `hadb2_reg`, which is shipped with the HA agent, can register the `hadb2` service for both EE and EEE instances. The `hadb2_reg` script needs to be run only once for the entire cluster.

Although there is only one set of control methods for the DB2 HA agent, the way they are registered depends on whether or not an EEE instance will be used in a mutual takeover configuration. For an EE instance or EEE instance in a hot standby configuration, HA-NFS is not used; therefore, the "-d nfs" switch, which tells the SC2.2 software that the hadb2 service is dependent on HA-NFS, is not needed.

The actual command that hadb2_reg uses to register the DB2 V7.1 control methods for an EEE instance is:

```
hareg -r hadb2 -b /opt/IBMdb2/V7.1/ha -m
START=hadb2_start,START_NET=hadb2_startnet,STOP_NET=hadb2_stopnet,
FM_START=hadb2_fmstart,FM_STOP=hadb2_fmstop
-t START_NET=$TIMEOUT,STOP_NET=$TIMEOUT -d nfs
```

The -b switch tells SC2.x to look in the opt/IBMdb2/V7.1/ha directory for all of the control methods. The -m switch defines the actual control methods for the hadb2 service. The -t switch defines the timeout for the START_NET and STOP_NET control methods. For a detailed description of each control method, refer to the Sun Cluster documentation.

The hadb2_unreg script can be used to unregister the hadb2 service and, like hadb2_reg, needs to be run only once for the cluster.

The hadb2tab File

The hadb2tab file is the main configuration file for the DB2 HA agent. Each control method consults this file to find out which instances are highly available. The hadb2tab file is located under the /var/db2/v71/ directory for DB2 UDB Version 7.1. The file supports multiple instances, and each non-commented line represents a different HA instance. Following is an example of an hadb2tab file:

```
<scadmin@thrash(203)# cat hadb2tab
EEE DATA db2eee jolt ON /export/ha_home /log0/home #Added by DB2 HA software
EE ADMIN db2ee log1 ON  -               -          #Added by DB2 HA software
```

The first field indicates to the DB2 HA agent whether the instance is an EE instance, or an EEE instance. The second field indicates whether the instance is a data instance, or an administrative instance. The third field contains the user name of the HA instance. The fourth field is the logical host or the HA-NFS host for the instance, depending on whether it is an EE or an EEE instance. The fifth field indicates whether fault monitoring for the instance is turned on or off. The last two fields are the local mount point, and the remote HA-NFS directory, respectively. These fields should be set to - (hyphen) if they are not used, and should only be used with an EEE mutual takeover configuration. Comments are allowed in the hadb2tab file if the information on the line before a "#" marker is either of zero length, or a valid definition of an instance.

Control Methods

Control methods for SC2.2 agents can be a set of scripts or programs. The agent for DB2 on Solaris is a set of programs that includes the following methods:

START_NET
hadb2_startnet, used to start DB2

STOP_NET
hadb2_stopnet, used to stop DB2

FM_START
hadb2_fmstart, used to start the fault monitor for DB2

FM_STOP
hadb2_fmstop, used to stop the fault monitor for DB2

For more information about these control methods, refer to the Sun Cluster documentation.

For EE instances, the logical host that is associated with the instance is defined right in the `hadb2tab` file. For EEE instances, however, the control method must also look in:

```
~<instance>/sqllib/ha/hadb2-eee.cfg
```

where `~<instance>` is the home directory of the instance owner. This file contains one line for each database partition, and is used to associate database partitions with logical hosts. An example of a valid `hadb2-eee.cfg` file is:

```
crackle % cat hadb2-eee.cfg
NODE:log0 0
NODE:log0 1
NODE:log1 2
NODE:log1 3
```

The instance or database partitions follow the corresponding logical host around the cluster. The logical host can move to any machine in the cluster that is supported by the underlying hardware and SC2.2. If the configuration is properly set up, DB2 will support any topology that is supported by the SC2.2 software.

After reading all of the information for an instance, the control method knows which logical hosts are associated with the instance. After parsing the command line arguments, the control method also knows which logical hosts are hosted, and which are not hosted by the current machine.

The following table shows the actions that are taken, depending on which control method is being run, and whether the logical hosts associated with the database partition or instance are hosted on the current machine.

Control Method	Associated logical host(s) are hosted	Associated logical host(s) are not hosted
START_NET	Start DB2 instance or database partitions	No action
STOP_NET	No action	Stop DB2 instance or database partitions
FM_START	Start fault monitor for instance	No action
FM_STOP	No action	Stop fault monitoring for instance

The control methods that perform start actions are only concerned with the logical hosts that are currently being hosted, and the control methods that perform stop actions are only concerned with the logical hosts that are not currently being hosted.

The control methods also need to mount the HA-NFS directory in a special way if HA-NFS is being used. If the local mount point and directory for HA-NFS are not defined as `-` (hyphen), the control method runs a `statvfs(2)` on the local mount point. If the file system type for the local mount point is not `nfs`, the agent attempts to mount the file system using information from the `hadb2tab` line. If the mount point and the directory for HA-NFS are defined as `-` (hyphen), the `vfstab` file of the corresponding logical host is required to mount the file system containing the home directory of the instance. The local mount point and the remote directory for HA-NFS should only be defined as `-` (hyphen) for EE and EEE hot standby configurations.

User Scripts

These scripts are run from the control methods to add additional functionality; they are passed the same command line arguments as the control methods are passed, and are written by the system administrator or the database administrator.

If a program must be run from within a script that is not run in the background, consider backgrounding the program with `nohup(1)`. The `nohup` program protects the executed program from the SIGHUP (or hangup) signal. Without `nohup`, a program that is run in the background from a script may die as a result of a SIGHUP signal when the script is finished.

The control methods run the following scripts:

- `/var/db2/v61/failover`
- `~<instance>/sqllib/ha/pre_db2start`
- `~<instance>/sqllib/ha/post_db2start`

- ˜<instance>%s/sqllib/ha/post_failover
- ˜<instance>/sqllib/ha/pre_db2stop
- ˜<instance>/sqllib/ha/fm_warning

where ˜*instance* is the home directory of the HA instance.

With the exception of the fm_warning script, each user script is run with the same arguments as the control method that invoked it. When using EEE instances, the database partition number is also passed (as the last argument) to the user script.

The /var/db2/v71/failover script is invoked at the beginning of the START_NET method, and runs in the background. Such a script can be used, for example, to e-mail support staff in the event of a failover. Following is an example of a failover script:

```
#!/bin/ksh

# E-mail or page support staff to notify them that a failover has occurred.

echo "Failover occurred on machine 'hostname':Running $0!"
   |/bin/mail admin@sphere.torolab.ibm.com
```

To e-mail successfully from a script, sendmail(1m) must be properly configured on the system.

As its name suggests, the pre_db2start script is run just before **db2start** is invoked. This script can be used for such tasks as changing database manager configuration parameters. It is given a maximum of 20 seconds to complete. For EEE instances, this script is run before **db2start** is invoked on each database partition. This script is applicable only to data instances, not to administrative instances.

Similarly, the post_db2start script is run just *after* **db2start** is invoked. This script can be used for such tasks as restarting databases. It is run in the background to ensure that its execution time does not interfere with other instances. This script is applicable only to data instances, not to administrative instances.

The post_failover script under the instance owner’s home directory, is run after processing the instance. This script can be used to notify client applications that DB2 is now functional, to activate databases, or to send administrators a status file. It is run in the background to prevent its execution time from delaying actions against the other HA instances. Following is an example of a post-failover script:

```
#!/bin/ksh
#

# Send the status file to the administrato-r.
mail admin@sphere.torolab.ibm.com </tmp/HA.info.db2eee
```

Both the START_NET and the STOP_NET method of the DB2 HA agent create a status file after processing each instance. The name of the status file is:

```
/tmp/HA.info.<instance>
```

where *instance* is the user name of the instance owner. The status file contains the start and stop report for the instance, as well as the time it took to run the control method. Following is an example of a status file:

```
scadmin@crackle(173)# cat /tmp/HA.info.db2eee
-----  Elapsed Time: 00:00:18          -----
-----  Elapsed Time: 00:00:00  (HA-NFS) -----

NODE     ACTION     RESULT     TRIES     RC
----     ------     ------     -----     --
   4     stop       success        3    1064
   5     stop       success        1    1064
   6     stop       success        2    1064
   7     stop       success        2    1064
   8     stop       success        1    1064
---------------------------------------------
```

The `pre_db2stop` script is run just before **db2stop** is invoked. This script can be used to notify client applications that DB2 is about to stop. It is given a maximum of 20 seconds to complete. This script is applicable only to data instances, not to administrative instances.

The fault monitor will also run a user script when DB2 is restarted because of an unexpected shutdown. This script is called:

```
~<instance>/sqllib/ha/fm_warning
```

The `fm_warning` script can be used to notify the system administrator that DB2 was restarted by the fault monitor. The system administrator should try to find out why DB2 shut down unexpectedly, and take appropriate actions to prevent this from happening again. The `fm_warning` script is run in the background.

Other Considerations

If an HA data service is turned off, only the stop methods are run during a failover or cluster reconfiguration; the other methods are run only if the HA data service is properly registered and turned on.

Ensure that each machine in the cluster has enough resources to run all of the data services for which it may be responsible. Resources such as CPU load, memory, swap and kernel parameters must be considered before the cluster goes into production. For example, if a machine in the cluster may need to run two DB2 instances, the kernel parameter requirements for that machine will be the sum of what is needed for each instance.

Fault Monitor

If fault monitoring is turned on, the fault monitor will be started during a cluster reconfiguration or failover. If DB2 is not started by the START_NET script, the fault monitor itself will start DB2. The fault monitor can detect if DB2 did not start, or if it shut down for unknown reasons. Because of this, it is important not to shut down DB2 manually when the fault monitor is turned on. The fault monitor will see this as an unexpected shutdown, and restart DB2. If this happens too many times, it will fail over the appropriate logical host.

When fault monitoring is enabled for an instance, the correct way to start or stop the instance manually is to first turn off fault monitoring or the `hadb2` service. Both of these actions can be initiated through the **hadb2_setup** command using the `-f` and `-s` switches (see “The hadb2_setup Command” on page 317).

Note: Do not use more than one instance for the same logical host. If more than one instance is associated with a logical host, a healthy instance may be failed over along with an unhealthy one.

EEE Considerations

When deciding which database partitions to associate with a logical host, it is important to consider how they will fail over. Consider a two-machine cluster that is to be used with four database partitions between the two machines, as shown in Figure 62 on page 313.

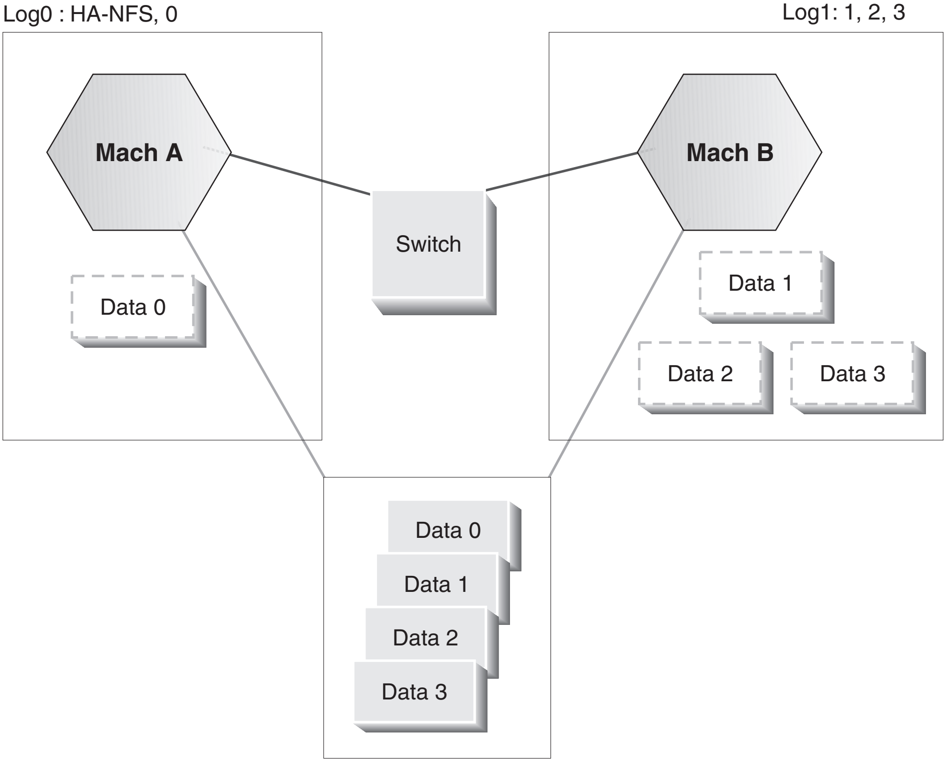


Figure 62. Two-machine Cluster with Four Database Partitions

You could associate one logical host with each database partition, and one for HA-NFS. In this case, there could be a problem if all of the logical hosts are being hosted by one system. If that system fails, all of the logical hosts must be moved off the system at the same time. Unfortunately, the Sun Cluster software does not move the logical hosts in any predictable order, and it is possible for a logical host that has a database partition associated with it to move before the logical host with HA-NFS. It is usually a good idea to group database partitions together, according to what would be hosted on a single system. This means that two database partitions that are normally hosted on one machine should be associated with a single logical host.

The `db2nodes.cfg` file used by an EEE instance is updated to indicate the machine on which the database partitions are residing. For example, if all of the database partitions are on a machine called ″crackle″, the `db2nodes.cfg` file resembles the following:

```
scadmin@crackle(193)# cat db2nodes.cfg
0 crackle 0 204.152.65.33
1 crackle 1 204.152.65.33
2 crackle 2 204.152.65.33
3 crackle 3 204.152.65.33
```

```
4 crackle 4 204.152.65.33
5 crackle 5 204.152.65.33
6 crackle 6 204.152.65.33
7 crackle 7 204.152.65.33
8 crackle 8 204.152.65.33
```

If some of these database partitions are moved to a machine called ″thrash″, the `db2nodes.cfg` file is updated as follows:

```
scadmin@crackle(193)# cat db2nodes.cfg
0 crackle 0 204.152.65.33
1 crackle 1 204.152.65.33
2 crackle 2 204.152.65.33
3 crackle 3 204.152.65.33
4 thrash 0 204.152.65.34
5 thrash 1 204.152.65.34
6 thrash 2 204.152.65.34
7 thrash 3 204.152.65.34
8 thrash 4 204.152.65.34
```

Notice that both the host name and the switch name are changed to reflect the machine name ″thrash″, and that the port numbers are also different.

The HA.config File

If it exists, the `/etc/HA.config` file can contain a number of configuration options, including the following:

```
scadmin@thrash(204)# cat /etc/HA.config
SYSLOG_FACILITY=LOG_LOCAL3
SYSLOG_LPRIORITY=LOG_INFO
SYSLOG_EPRIORITY=LOG_ERR
USE_INTERCONNECT=auto
SWITCH_NAME=204.152.65.18
DEBUG_LEVEL=2
FAILS_PER_HOUR=2
FAILS_PER_DAY=4
FAILS_PER_WEEK=10
FM_FAIL_SEV=soft
DB2START_TIMEOUT=60
DB2STOP_TIMEOUT=500
SCRIPT_USER=bin
```

Note: If the `HA.config` file does not exist, default values are used.

The SYSLOG_FACILITY variable sets the SYSLOG facility for logging both messages and errors. The SYSLOG_LPRIORITY and SYSLOG_EPRIORITY variables set the SYSLOG priority for logging informational messages and error messages, respectively.

Some changes may be needed to enable the SYSLOG daemon to log information from the DB2 HA agent. For example, one of the following two lines added to the /etc/syslog.conf file will tell the SYSLOG daemon to write information to a log file.

```
*.notice                                      /var/adm/SC.x
local3.info                                   /var/adm/SC.LOG_LOCAL3
```

A Sun Cluster usually has a high speed interconnect. To use the high speed interconnect with DB2, set USE_INTERCONNECT to auto or to override. The auto setting (the default) uses the Sun internal logical network interface. This interface will be transferred to another physical interface if the initial interface fails. If USE_INTERCONNECT is set to override, the switch name is taken from the SWITCH_NAME variable. Another option is to set USE_INTERCONNECT to no, which specifies that high speed interconnect is not to be used.

DEBUG_LEVEL specifies how much information is to be logged during a failover. It is a number between 0 and 10, where 10 is the highest debug level. The information is logged at the specified SYSLOG priority and facility. If any problems are encountered, set the debug level to the maximum level, configure SYSLOG to log the output from the HA agents, and send the SYSLOG output to IBM service.

Three of the variables help the DB2 fault monitor decide when to fail over a logical host: FAILS_PER_HOUR, FAILS_PER_DAY, and FAILS_PER_WEEK. Every HA environment is different; you must decide how many DB2 failures are acceptable. After each ″acceptable″ failure, DB2 is restarted on the same machine. When one of these three failure thresholds is exceeded, the logical host associated with the instance or database partition is failed over.

The FM_FAIL_SEV variable specifies whether the failover is ″soft″ or ″hard″. For more information, refer to the Sun Cluster documentation on hactl(1m).

The DB2START_TIMEOUT and DB2STOP_TIMEOUT variables specify the maximum number of seconds that **db2start** and **db2stop** are allowed to run. After the specified interval has passed, the HA agent considers the operation to have failed, and try to restart the instance.

There are some user scripts that are not associated with any particular instance. Normally, these scripts are run as root; this can be overridden by the SCRIPT_USER variable, which can be set to specify the user ID that can run these scripts.

How Control Methods Run DB2 Commands

The DB2 HA agent uses the **su** command to run commands as the instance owner. The actual command would look something like:

```
su - <instance> -c "db2stop"
```

where *instance* is the user name of the instance.

It is important to ensure that the `.profile` file of the instance owner is **su**-″friendly″. If it is not, the **su** command may not work properly. Invoke the **su** command manually, or from a script, to verify that the command can run successfully.

Setup

Before you read this section, be sure that you are familiar with the SC2.2 software. This section assumes that you know how to set up SC2.2 and HA-NFS, and that you know how to use your volume manager. Along with the other required patches for DB2, the following patches are required for the HA agent:

```
Solaris 2.6:
   105210-17 (or later)
   105786-05 (or later)
```

Note: There are no required patches for Solaris 7 (Solaris 2.7).

Common Installation Steps

1. Install SC2.2 on all machines in the cluster. During installation, SC2.2 will ask which agents to install. Since DB2 is not shipped with SC2.2, it is not in the list of agents. The agent for DB2 will be installed with DB2 and registered through the **hadb2_reg** command.
2. Configure the logical hosts with disk groups and logical IP addresses.

Setup on DB2 UDB Enterprise Edition

1. Create the home directory for the instance under the logical host file system of a logical host.
2. Install DB2 on all machines in the cluster.
3. Create the instance on the machine in the cluster that currently has the home directory for the instance.
4. Add the user for the instance to the other machines in the cluster, ensuring that the numeric user ID is the same.
5. Register the `hadb2` service using the **hadb2_reg** command.
6. Run the **hadb2_setup** command to set up HA for the instance.

Setup on DB2 UDB Enterprise - Extended Edition

1. Create the home directory for the HA instance owner:
   a. For hot standby, create the home directory for the instance under the logical host file system of a logical host.

   b. For mutual takeover, configure HA-NFS, and export the home directory from one of the logical hosts. On one of the machines, mount the HA-NFS directory under the chosen mount point.
2. Install DB2 on all machines in the cluster.
3. Create the instance on the machine that has the HA-NFS file system mounted.
4. Add the user for the instance to the other machines in the cluster, ensuring that the numeric user ID is the same.
5. Register the `hadb2` service using the **hadb2_reg** command.
6. Run the **hadb2_setup** command to set up HA for the instance.

Note: Using NIS to define the information for the HA instance is not recommended, because NIS can introduce a single point of failure.

The hadb2_setup Command

The **hadb2_setup** command is the central point of the programs that come with the DB2 HA agent. It can be used to set up an instance, to modify it, or to delete it. It can also be used to turn the hadb2_setup service on and off. With this command, there is no need to manually edit the `hadb2tab` file.

Note: The **hadb2_setup** command performs actions only on the machine on which it runs. Changes made to one machine should also be made to the other machines in the cluster.

The following arguments are supported:

```
To add an EE instance:
---------------------
   hadb2_setup -a -i <instance> -f [on|off] -h <logical_host> -p [DATA|ADMIN] -t EE

   For example:
      hadb2_setup -a -i db2ee -f off -h log1 -p DATA -t EE



To add an EEE instance:
----------------------
   hadb2_setup -a -i <instance> -f [on|off] -h <nfs_host> -l <mount_point> \
      -r <ha-nfs_dir> -p [DATA|ADMIN] -t EEE -n "<node_info>"

   For example:
      hadb2_setup -a -i db2eee -f off -h ha-sun1 -l /export/ha_home \
         -r /log0/home -p DATA -t EEE -n "log0[0,10,20],log1[30,40,50]"



To delete an instance:
---------------------
   hadb2_setup -d -i <instance>
```

```
To modify an instance:
---------------------
   hadb2_setup -m -i <instance> [-f [on|off] | -l <mount_point> | \
      -h <host> | -p [DATA|ADMIN] | -r <ha-nfs_dir> | -t [EE|EEE] ]



Other options:
-------------
   -s   <on|off>            Bring hadb2 up or down (for all HA instances)
   -y                       Assume yes for safety checks
```

To turn the hadb2 service on or off, specify the -s switch. This is equivalent to using the **hareg** command with the -n and -y switches, and specifying the hadb2 service. For more information about the **hareg(1m)** command, refer to the Sun Cluster documentation.

The fault monitor for the instance can be turned off using the -f switch. This has the effect of stopping the fault monitor for the instance on the local machine, as well as modifying the hadb2tab file to reflect the fact that fault monitoring is turned off.

For EE instances, turning off fault monitoring on all machines is recommended in case the instance fails over. For EEE instances, fault monitoring must be turned off on all machines that are hosting database partitions for the instance before it is shut down manually.

To delete an instance, use the -d switch. This only removes the instance from the hadb2tab file, and does not remove or modify any other files or directories. Since the hadb2tab file is the main configuration file for the HA-DB2 agent, removing an instance from this file makes the control methods unaware of its existence.

To modify an instance, use the -m switch. This only changes information in the hadb2tab file, and does not remove or modify any other files or directories. The -m switch can be used with any switch that pertains to information in the hadb2tab file. The db2nodes.cfg file and the hadb2-eee.cfg file must be changed manually after the initial setup, because the **hadb2_setup** command does not support modifying these files.

Adding an instance is somewhat more involved.

For EE instances, the following arguments are required:

```
   hadb2_setup -a -i <instance> -f <fm> -h <logical_host> -t <EEE_or_EE>
      -p <purpose>
```

where *instance* is the name of the instance to be added, *fm* specifies whether fault monitoring is initially turned on or off, *logical_host* is the associated logical host, *EEE_or_EE* is set to `EE`, and *purpose* can be either `DATA` or `ADMIN`.

For EEE instances, the following arguments are required:

```
hadb2_setup -a -i <instance> -f <fm> -h <nfs_host> -t <EEE_or_EE> -p
   <purpose> -l <mount_point> -r <HA-NFS_directory> -n <node_info>
```

where *instance* is the name of the instance to be added, *fm* specifies whether fault monitoring is initially turned on or off, *nfs_host* is the host name for the logical host that is exporting the HA-NFS file system, *EEE_or_EE* is set to `EEE`, *purpose* can be either `DATA` or `ADMIN`, *mount_point* is the local mount point for the HA-NFS directory, *HA-NFS_directory* is the HA-NFS directory, and *node_info* is the information that associates database partitions with a logical host. For example:

```
hadb2_setup -a -i db2eee -f on -h jolt -l /export/ha_home -p DATA -t EEE -r
   /log1/home -n "log0[0,1],log1[2,3]"
```

When adding an EEE instance, the node information must be enclosed by quotation marks. In this example, the instance ″db2eee″ will be associated with two logical hosts, ″log0″ and ″log1″. Database partitions ″0″ and ″1″ of the ″db2eee″ instance will be associated with the logical host ″log0″, and database partitions ″2″ and ″3″ will be associated with logical host ″log1″.

Use the **hadb2_setup** command to add an instance to all machines in the cluster. The instance can then be started by forcing a cluster reconfiguration, or by turning `hadb2` service off and then on. This can be done, either through the **hareg** command, or with the `-s` switch of the **hadb2_setup** command. If the instance does not start, see “Troubleshooting” on page 323.

When the **hadb2_setup** command adds an EEE instance, the following actions are performed transparently:

- Checking the specified information. This includes ensuring that the user exists on the system, and that HA-NFS is running.
- Creating a `db2nodes.cfg` file.
- Creating an `hadb2-eee.cfg` file.
- Creating a `.rhosts` file for the EEE instance.
- Creating symbolic links from the default database path to the associated logical hosts data directories.
- Adding a line to the `hadb2tab` file.

To prevent configuration errors, and to ensure that the HA instance will be able to start after the **hadb2_setup** command runs, the command performs a significant amount of testing before a new instance is added.

The `db2nodes.cfg` file is created and seeded with information corresponding to the current cluster status. For example, if logical host ″log0″ is being hosted by the machine ″crackle″, the entries for the database partitions associated with ″log0″ will contain the machine name ″crackle″ and the high speed interconnect for ″crackle″:

```
scadmin@crackle(193)# cat db2nodes.cfg
0 crackle 0 204.152.65.33
1 crackle 1 204.152.65.33
2 thrash 0 204.152.65.34
3 thrash 1 204.152.65.34
```

The `hadb2-eee.cfg` file is created only on the basis of the node information that is specified on the command. There is one line per database partition:

```
sphere % cat hadb2-eee.cfg
NODE:log0 0
NODE:log0 1
NODE:log1 2
NODE:log1 3
```

The `.rhost` file is required for DB2 UDB EEE, and should contain all host names (or IP addresses) for each machine in the cluster. For example:

```
crackle db2eee
204.152.65.1 db2eee
204.152.65.17 db2eee
thrash db2eee
204.152.65.2 db2eee
204.152.65.18 db2eee
crackle db2eee
jolt db2eee
bump db2eee
thrash.torolab.ibm.com db2eee
crackle.torolab.ibm.com db2eee
```

In accordance with a file system layout for SMS tables spaces, the **hadb2_setup** command sets up a number of directories and symbolic links. These include:

- A directory called ″data″ under the logical host file system for each logical host.
- A node directory (under this ″data″ directory) for each database partition associated with the logical host.
- Symbolic links in the default database path, located under `~<instance>`, where *~instance* is the home directory of the instance. There is one symbolic link for each database partition that points to the corresponding node directory. For more information, see “Disk Layout for EE and EEE Instances” on page 302.

Failover Time

Failover time is measured from when data is first unavailable to when it is available again. A number of events that occur during a failover can contribute significantly to the failover time:

- Disk deporting and importing.

  Deporting and importing disks usually does not take a very long time compared to other events, although it does contribute to the overall down time. The more disks that need to be moved from one machine to another during a failover, the longer the process takes. If there are defective disks, the process can take even longer.

- Fsck of the file systems that are mounted for a logical host.

  Before the file systems of the logical host can be mounted, they must pass an `fsck` to ensure the health of the file system. The larger the file system, the longer this process takes. By using a journalled file system, this time can be drastically reduced. Since journalled file systems are normally used in an HA environment, the `fsck` time is usually not an issue.

- User scripts called from the HA agent.

  The HA agent will call user scripts if they exist and are executable. Some of these scripts are run synchronously, and can add to the time it takes to bring up the HA instances. Ensure that they run as quickly as possible; consider running any external programs called by these scripts in the background.

- HA-NFS.

  For a single EEE instance in a mutual takeover configuration, HA-NFS must be used for the home directory of the instance owner. HA-NFS adds to failover time because of the grace period for *lockd* (defined in the HA agent for HA-NFS), which is 90 seconds when running HA-NFS. This affects failover times, because any process that locks a file on the HA-NFS file system after a failover must wait until the grace period is over. The HA agent for DB2 is the first process to lock a file under the instance owner's home directory after a failover, and it records the time it takes to obtain the first lock. This time is displayed in the status report after a failover.

- Starting DB2.

  Starting DB2 contributes only a small amount to the failover time. For an EE instance, it contributes about 5-15 seconds on average. For an EEE instance, it contributes about 10 seconds, plus about 5 seconds per database partition that is being failed over. If three database partitions are being failed over, for example, the failover time contributed by starting these three database partitions will be approximately 25 seconds. This does *not* include crash recovery for the databases of the instance.

- Database crash recovery.

Crash recovery often contributes to the majority of down time associated with a failover. How long it takes to recover a database depends on a number of factors, including:

- Client workload. Only changes to the database are logged in the transaction logs. If the client workload is mostly read-only operations, relatively few transactions must be applied to the database during crash recovery.
- Disk and machine speed. The speed of the disks and the machine that is hosting the HA instance also contributes to the time it takes to recover the database. The faster the system, the shorter the crash recovery time.
- Value of the *softmax* database configuration parameter. The value of *softmax* is the percentage of the log file size at which a soft checkpoint is to be taken, and a log control file is to be written. The log control file is used during crash recovery to determine which log records are truly necessary to restore the database to a consistent state. Reducing this value will cause the database manager to trigger the page cleaners more often, and take more frequent soft checkpoints; although performance is reduced, database recovery is faster.
- Whether the instance is EE or EEE. If the instance is an EEE instance, the database restart operations will be done in parallel. Each database partition is responsible for restarting its own portion of the databases. If there are 50 GB of data for a database, an instance with four database partitions will be able to recover the database roughly four times faster than an EE instance can.

Troubleshooting

The following table identifies problems that you might encounter, their probable causes, and actions that you can take to solve them.

Table 29. Troubleshooting High Availability on Sun Cluster 2.2

Symptom	Possible cause	Action
Cannot mount logical host file system	The logical host file system is normally mounted and unmounted during the failover of a logical host. During failover, there should be no active processes or open files under the logical host file system. In rare cases, processes that cannot be killed have their current working directory under the logical host file system. To find out if a process is under the mount point, use `fuser(1m)`, or a GNU utility called `lsof`. Error messages are produced when the logical host file system cannot be mounted.[a]	Reboot the system, or move the logical host file system to another name and recreate it. Doing this allows the frozen process to stay under the directory (since it can't be killed), and allows the mount to take place.[b]
The db2start or db2stop time-out does not work	A SIGALRM signal may not break out of a blocking system call. Instead, the system call will restart as if the SA_RESTART flag were set with `sigaction()`. This causes time-outs for the DB2 HA agents to be ignored, and the agent method will hang instead of recovering from a hung **db2start** or **db2stop** command.	Apply the required patch, 105210-17 (or later), for Solaris 2.6.
Logging into an instance hangs	Although there are numerous reasons why this can happen, the most common reasons include NFS problems and the `/usr/sbin/quota` program.	Check the NFS mounts to ensure that they are healthy, and look for quota processes owned by the instance owner. At the discretion of the system administrator, changing the quota program to a symbolic link to `/bin/true` may solve the problem. This is not a recommended solution, but it may work.
I just set up an EEE instance, but it does not start	The **hadb2_setup** command does not add ports to the `/etc/services` file; it is expected that the administrator will add them manually. An error message is returned.[c]	Ensure that you have appropriate ports named in the `/etc/services` file.

Table 29. Troubleshooting High Availability on Sun Cluster 2.2 (continued)

Symptom	Possible cause	Action
START_NET method cannot start DB2		Turn off fault monitoring to ensure that the instance does not get failed over. Log in as the instance owner, and try to start DB2 manually. 1. Ensure that the hadb2tab configuration file has the correct instance type specified. For example, having a db2nodes.cfg file for an EE administrative instance will cause problems, and the HA agent methods will not be able to recover from this. 2. Ensure that the .rhosts file exists, and has valid entries in it. 3. Ensure that the HA-NFS file system is shared with root permissions for all machines in the cluster. 4. Check the kernel parameters, and ensure that they are correct. 5. Ensure that the /etc/services file contains entries for the instance.
The instance only works on one machine	• The numeric *uid* for the instance may not be the same on each machine in the cluster. • The kernel parameters may not be valid on each machine in the cluster. • The hadb2tab file may not be the same on each machine in the cluster. • Other configuration files, such as the logical host vfstab file, may not be the same on each machine in the cluster.	If none of these causes appears to apply, try logging in as the instance owner, and start DB2 manually. For EE instances, this should work if the logical host that is hosting the instance is being hosted by the current machine. For EEE instances, this should work from any machine in the cluster that can host the database partitions.

Table 29. Troubleshooting High Availability on Sun Cluster 2.2 (continued)

Symptom	Possible cause	Action
su - <instance> -c ″db2start″ does not work	• The `.profile` for the instance may not be **su**-″friendly″. • There is a known problem with the Bourne shell (`/bin/sh`), in which the **su** command works manually, but not through the HA agent.	• As root, try running this command manually, and ensure that it works before trying again through the HA agent. • Switch to the Korn shell (`/bin/ksh`), if necessary.
My EEE instance cannot start, but the home directory is mounted	The HA-NFS directory may not have been exported with ″root″ permissions to the machines in the cluster. Both DB2 and the HA agents require this to run properly.	To test this, try to create a file (as root) under the instance owner's home directory.
Trying to access the EEE instance directory returns a ″Stale NFS file handle″ error	There may still be processes under the instance owner's home directory.	Unmount the instance owner's home directory, and allow the HA agent to remount it. The HA agent will remount it if the `hadb2` service is turned off and on again (see a description of the `-s` switch on the **hadb2_setup** command in "The hadb2_setup Command" on page 317).

Table 29. Troubleshooting High Availability on Sun Cluster 2.2 (continued)

Symptom	Possible cause	Action
Control methods do not run successfully through SC2.2	The hadb2 service may not be registered with the Sun Cluster software, or it may not be turned on.	If the control methods appear to run normally from the command line, check the SYSLOG files for error messages that may help to explain the problem. Ensure that the hadb2 service is registered with the Sun Cluster software, and that it is turned on. Running the methods manually is useful for debugging a problem.[d] The methods should be run as root and given the appropriate command line arguments. If the list of logical hosts is nil, the argument should be given as "". The double quotation marks without a blank space separator denotes a blank argument. For example: `hadb2_startnet log0,log1 "" 600` The first argument, log0,log1, tells the hadb2_startnet method that logical hosts log0 and log1 are being hosted by the current machine. The second argument is nil, which tells the hadb2_startnet method that there are no other logical hosts being hosted on other machines in the cluster (all of them are on the current machine). The third argument tells the method that SC2.2 will time out after 600 seconds.
User scripts do not run	The user scripts can only be run if they exist in the appropriate directories and are executable.	Check file ownership and attributes. If a script still fails to run, contact IBM service. Forward a directory listing of the script that does not run, and SYSLOG output for a failover or a cluster reconfiguration that should have run the script.

Table 29. Troubleshooting High Availability on Sun Cluster 2.2 (continued)

Symptom	Possible cause	Action
Information is not being logged to the file specified in /etc/syslog.conf		Use `touch(1)` to create the file that is specified in the `/etc/syslog.conf` file, and then restart the SYSLOG daemon.

[a] Error messages that are produced when the logical host file system cannot be mounted may look something like the following:

```
Aug 17 11:14:01 rash ID[SUNWcluster.loghost.1170]: importing data1
Aug 17 11:14:06 rash ID[SUNWcluster.scnfs.3040]: mount -F ufs -o ""
     /dev/vx/dsk/data1/data1-stat /log1 failed.
Aug 17 11:14:07 rash ID[SUNWcluster.ccd.ccdd.5304]: error freeze cmd =
     /opt/SUNWcluster/bin/loghost_sync
CCDSYNC_POST_ADDU LOGHOST_CM:log1:rash /etc/opt/SUNWcluster/conf/ccd.database
     2 "0 1" 1 error code = 1
```

[b] For example:

```
scadmin@rash(218)# ps -fe | egrep db2
db2ee 1984 1 0 0:01 <defunct>

Solution:

   scadmin@rash(229)# cd /
   scadmin@rash(230)# mv /log1 /log1.bkp
   scadmin@rash(231)# mkdir /log1
```

[c] The error message may look something like the following:

```
SQL6030N START or STOP DATABASE MANAGER failed. Reason code "13".
```

[d] For example, if the `hadb2_startnet` method cannot find `libdb2.so.1`, but it runs normally through the Sun Cluster software, no errors will be reported. Running the method manually results in the following:

```
scadmin@crackle(213)# hadb2_startnet '''log0,log1' 600
ld.so.1: hadb2_startnet: fatal: libdb2.so.1: open failed:
   No such file or directory
Killed
```

Part 5. Appendixes

Appendix A. Using the DB2 Library

The DB2 Universal Database library consists of online help, books (PDF and HTML), and sample programs in HTML format. This section describes the information that is provided, and how you can access it.

To access product information online, you can use the Information Center. For more information, see “Accessing Information with the Information Center” on page 345. You can view task information, DB2 books, troubleshooting information, sample programs, and DB2 information on the Web.

DB2 PDF Files and Printed Books

DB2 Information

The following table divides the DB2 books into four categories:

DB2 Guide and Reference Information
These books contain the common DB2 information for all platforms.

DB2 Installation and Configuration Information
These books are for DB2 on a specific platform. For example, there are separate *Quick Beginnings* books for DB2 on OS/2, Windows, and UNIX-based platforms.

Cross-platform sample programs in HTML
These samples are the HTML version of the sample programs that are installed with the Application Development Client. The samples are for informational purposes and do not replace the actual programs.

Release notes
These files contain late-breaking information that could not be included in the DB2 books.

The installation manuals, release notes, and tutorials are viewable in HTML directly from the product CD-ROM. Most books are available in HTML on the product CD-ROM for viewing and in Adobe Acrobat (PDF) format on the DB2 publications CD-ROM for viewing and printing. You can also order a printed copy from IBM; see “Ordering the Printed Books” on page 341. The following table lists books that can be ordered.

On OS/2 and Windows platforms, you can install the HTML files under the `sqllib\doc\html` directory. DB2 information is translated into different languages; however, all the information is not translated into every language. Whenever information is not available in a specific language, the English information is provided

On UNIX platforms, you can install multiple language versions of the HTML files under the `doc/%L/html` directories, where *%L* represents the locale. For more information, refer to the appropriate *Quick Beginnings* book.

You can obtain DB2 books and access information in a variety of ways:

- "Viewing Information Online" on page 344
- "Searching Information Online" on page 348
- "Ordering the Printed Books" on page 341
- "Printing the PDF Books" on page 340

Table 30. DB2 Information

Name	Description	Form Number PDF File Name	HTML Directory
	DB2 Guide and Reference Information		
Administration Guide	*Administration Guide: Planning* provides an overview of database concepts, information about design issues (such as logical and physical database design), and a discussion of high availability.	SC09-2946 db2d1x70	db2d0
	Administration Guide: Implementation provides information on implementation issues such as implementing your design, accessing databases, auditing, backup and recovery.	SC09-2944 db2d2x70	
	Administration Guide: Performance provides information on database environment and application performance evaluation and tuning.	SC09-2945 db2d3x70	
	You can order the three volumes of the *Administration Guide* in the English language in North America using the form number SBOF-8934.		
Administrative API Reference	Describes the DB2 application programming interfaces (APIs) and data structures that you can use to manage your databases. This book also explains how to call APIs from your applications.	SC09-2947 db2b0x70	db2b0

Table 30. DB2 Information (continued)

Name	Description	Form Number PDF File Name	HTML Directory
Application Building Guide	Provides environment setup information and step-by-step instructions about how to compile, link, and run DB2 applications on Windows, OS/2, and UNIX-based platforms.	SC09-2948 db2axx70	db2ax
APPC, CPI-C, and SNA Sense Codes	Provides general information about APPC, CPI-C, and SNA sense codes that you may encounter when using DB2 Universal Database products. Available in HTML format only.	No form number db2apx70	db2ap
Application Development Guide	Explains how to develop applications that access DB2 databases using embedded SQL or Java (JDBC and SQLJ). Discussion topics include writing stored procedures, writing user-defined functions, creating user-defined types, using triggers, and developing applications in partitioned environments or with federated systems.	SC09-2949 db2a0x70	db2a0
CLI Guide and Reference	Explains how to develop applications that access DB2 databases using the DB2 Call Level Interface, a callable SQL interface that is compatible with the Microsoft ODBC specification.	SC09-2950 db2l0x70	db2l0
Command Reference	Explains how to use the Command Line Processor and describes the DB2 commands that you can use to manage your database.	SC09-2951 db2n0x70	db2n0
Connectivity Supplement	Provides setup and reference information on how to use DB2 for AS/400, DB2 for OS/390, DB2 for MVS, or DB2 for VM as DRDA application requesters with DB2 Universal Database servers. This book also details how to use DRDA application servers with DB2 Connect application requesters. Available in HTML and PDF only.	No form number db2h1x70	db2h1

Table 30. DB2 Information (continued)

Name	**Description**	**Form Number** **PDF File Name**	**HTML Directory**
Data Movement Utilities Guide and Reference	Explains how to use DB2 utilities, such as import, export, load, AutoLoader, and DPROP, that facilitate the movement of data.	SC09-2955 db2dmx70	db2dm
Data Warehouse Center Administration Guide	Provides information on how to build and maintain a data warehouse using the Data Warehouse Center.	SC26-9993 db2ddx70	db2dd
Data Warehouse Center Application Integration Guide	Provides information to help programmers integrate applications with the Data Warehouse Center and with the Information Catalog Manager.	SC26-9994 db2adx70	db2ad
DB2 Connect User's Guide	Provides concepts, programming, and general usage information for the DB2 Connect products.	SC09-2954 db2c0x70	db2c0
DB2 Query Patroller Administration Guide	Provides an operational overview of the DB2 Query Patroller system, specific operational and administrative information, and task information for the administrative graphical user interface utilities.	SC09-2958 db2dwx70	db2dw
DB2 Query Patroller User's Guide	Describes how to use the tools and functions of the DB2 Query Patroller.	SC09-2960 db2wwx70	db2ww
Glossary	Provides definitions for terms used in DB2 and its components. Available in HTML format and in the *SQL Reference*.	No form number db2t0x70	db2t0
Image, Audio, and Video Extenders Administration and Programming	Provides general information about DB2 extenders, and information on the administration and configuration of the image, audio, and video (IAV) extenders and on programming using the IAV extenders. It includes reference information, diagnostic information (with messages), and samples.	SC26-9929 dmbu7x70	dmbu7
Information Catalog Manager Administration Guide	Provides guidance on managing information catalogs.	SC26-9995 db2dix70	db2di

Table 30. DB2 Information (continued)

Name	Description	Form Number PDF File Name	HTML Directory
Information Catalog Manager Programming Guide and Reference	Provides definitions for the architected interfaces for the Information Catalog Manager.	SC26-9997 db2bix70	db2bi
Information Catalog Manager User's Guide	Provides information on using the Information Catalog Manager user interface.	SC26-9996 db2aix70	db2ai
Installation and Configuration Supplement	Guides you through the planning, installation, and setup of platform-specific DB2 clients. This supplement also contains information on binding, setting up client and server communications, DB2 GUI tools, DRDA AS, distributed installation, the configuration of distributed requests, and accessing heterogeneous data sources.	GC09-2957 db2iyx70	db2iy
Message Reference	Lists messages and codes issued by DB2, the Information Catalog Manager, and the Data Warehouse Center, and describes the actions you should take. You can order both volumes of the Message Reference in the English language in North America with the form number SBOF-8932.	Volume 1 GC09-2978 db2m1x70 Volume 2 GC09-2979 db2m2x70	db2m0
OLAP Integration Server Administration Guide	Explains how to use the Administration Manager component of the OLAP Integration Server.	SC27-0787 db2dpx70	n/a
OLAP Integration Server Metaoutline User's Guide	Explains how to create and populate OLAP metaoutlines using the standard OLAP Metaoutline interface (not by using the Metaoutline Assistant).	SC27-0784 db2upx70	n/a
OLAP Integration Server Model User's Guide	Explains how to create OLAP models using the standard OLAP Model Interface (not by using the Model Assistant).	SC27-0783 db2lpx70	n/a
OLAP Setup and User's Guide	Provides configuration and setup information for the OLAP Starter Kit.	SC27-0702 db2ipx70	db2ip
OLAP Spreadsheet Add-in User's Guide for Excel	Describes how to use the Excel spreadsheet program to analyze OLAP data.	SC27-0786 db2epx70	db2ep

Table 30. DB2 Information (continued)

Name	Description	Form Number PDF File Name	HTML Directory
OLAP Spreadsheet Add-in User's Guide for Lotus 1-2-3	Describes how to use the Lotus 1-2-3 spreadsheet program to analyze OLAP data.	SC27-0785 db2tpx70	db2tp
Replication Guide and Reference	Provides planning, configuration, administration, and usage information for the IBM Replication tools supplied with DB2.	SC26-9920 db2e0x70	db2e0
Spatial Extender User's Guide and Reference	Provides information about installing, configuring, administering, programming, and troubleshooting the Spatial Extender. Also provides significant descriptions of spatial data concepts and provides reference information (messages and SQL) specific to the Spatial Extender.	SC27-0701 db2sbx70	db2sb
SQL Getting Started	Introduces SQL concepts and provides examples for many constructs and tasks.	SC09-2973 db2y0x70	db2y0
SQL Reference, Volume 1 and Volume 2	Describes SQL syntax, semantics, and the rules of the language. This book also includes information about release-to-release incompatibilities, product limits, and catalog views. You can order both volumes of the *SQL Reference* in the English language in North America with the form number SBOF-8933.	Volume 1 SC09-2974 db2s1x70 Volume 2 SC09-2975 db2s2x70	db2s0
System Monitor Guide and Reference	Describes how to collect different kinds of information about databases and the database manager. This book explains how to use the information to understand database activity, improve performance, and determine the cause of problems.	SC09-2956 db2f0x70	db2f0
Text Extender Administration and Programming	Provides general information about DB2 extenders and information on the administration and configuring of the text extender and on programming using the text extenders. It includes reference information, diagnostic information (with messages) and samples.	SC26-9930 desu9x70	desu9

Table 30. DB2 Information (continued)

Name	Description	Form Number PDF File Name	HTML Directory
Troubleshooting Guide	Helps you determine the source of errors, recover from problems, and use diagnostic tools in consultation with DB2 Customer Service.	GC09-2850 db2p0x70	db2p0
What's New	Describes the new features, functions, and enhancements in DB2 Universal Database, Version 7.	SC09-2976 db2q0x70	db2q0
	DB2 Installation and Configuration Information		
DB2 Connect Enterprise Edition for OS/2 and Windows Quick Beginnings	Provides planning, migration, installation, and configuration information for DB2 Connect Enterprise Edition on the OS/2 and Windows 32-bit operating systems. This book also contains installation and setup information for many supported clients.	GC09-2953 db2c6x70	db2c6
DB2 Connect Enterprise Edition for UNIX Quick Beginnings	Provides planning, migration, installation, configuration, and task information for DB2 Connect Enterprise Edition on UNIX-based platforms. This book also contains installation and setup information for many supported clients.	GC09-2952 db2cyx70	db2cy
DB2 Connect Personal Edition Quick Beginnings	Provides planning, migration, installation, configuration, and task information for DB2 Connect Personal Edition on the OS/2 and Windows 32-bit operating systems. This book also contains installation and setup information for all supported clients.	GC09-2967 db2c1x70	db2c1
DB2 Connect Personal Edition Quick Beginnings for Linux	Provides planning, installation, migration, and configuration information for DB2 Connect Personal Edition on all supported Linux distributions.	GC09-2962 db2c4x70	db2c4
DB2 Data Links Manager Quick Beginnings	Provides planning, installation, configuration, and task information for DB2 Data Links Manager for AIX and Windows 32-bit operating systems.	GC09-2966 db2z6x70	db2z6

Table 30. DB2 Information (continued)

Name	Description	Form Number PDF File Name	HTML Directory
DB2 Enterprise - Extended Edition for UNIX Quick Beginnings	Provides planning, installation, and configuration information for DB2 Enterprise - Extended Edition on UNIX-based platforms. This book also contains installation and setup information for many supported clients.	GC09-2964 db2v3x70	db2v3
DB2 Enterprise - Extended Edition for Windows Quick Beginnings	Provides planning, installation, and configuration information for DB2 Enterprise - Extended Edition for Windows 32-bit operating systems. This book also contains installation and setup information for many supported clients.	GC09-2963 db2v6x70	db2v6
DB2 for OS/2 Quick Beginnings	Provides planning, installation, migration, and configuration information for DB2 Universal Database on the OS/2 operating system. This book also contains installation and setup information for many supported clients.	GC09-2968 db2i2x70	db2i2
DB2 for UNIX Quick Beginnings	Provides planning, installation, migration, and configuration information for DB2 Universal Database on UNIX-based platforms. This book also contains installation and setup information for many supported clients.	GC09-2970 db2ixx70	db2ix
DB2 for Windows Quick Beginnings	Provides planning, installation, migration, and configuration information for DB2 Universal Database on Windows 32-bit operating systems. This book also contains installation and setup information for many supported clients.	GC09-2971 db2i6x70	db2i6
DB2 Personal Edition Quick Beginnings	Provides planning, installation, migration, and configuration information for DB2 Universal Database Personal Edition on the OS/2 and Windows 32-bit operating systems.	GC09-2969 db2i1x70	db2i1
DB2 Personal Edition Quick Beginnings for Linux	Provides planning, installation, migration, and configuration information for DB2 Universal Database Personal Edition on all supported Linux distributions.	GC09-2972 db2i4x70	db2i4

Table 30. DB2 Information (continued)

Name	Description	Form Number PDF File Name	HTML Directory
DB2 Query Patroller Installation Guide	Provides installation information about DB2 Query Patroller.	GC09-2959 db2iwx70	db2iw
DB2 Warehouse Manager Installation Guide	Provides installation information for warehouse agents, warehouse transformers, and the Information Catalog Manager.	GC26-9998 db2idx70	db2id
	Cross-Platform Sample Programs in HTML		
Sample programs in HTML	Provides the sample programs in HTML format for the programming languages on all platforms supported by DB2. The sample programs are provided for informational purposes only. Not all samples are available in all programming languages. The HTML samples are only available when the DB2 Application Development Client is installed. For more information on the programs, refer to the *Application Building Guide*.	No form number	db2hs
	Release Notes		
DB2 Connect Release Notes	Provides late-breaking information that could not be included in the DB2 Connect books.	See note #2.	db2cr
DB2 Installation Notes	Provides late-breaking installation-specific information that could not be included in the DB2 books.	Available on product CD-ROM only.	
DB2 Release Notes	Provides late-breaking information about all DB2 products and features that could not be included in the DB2 books.	See note #2.	db2ir

Notes:

1. The character *x* in the sixth position of the file name indicates the language version of a book. For example, the file name `db2d0e70` identifies the English version of the *Administration Guide* and the file name `db2d0f70` identifies the French version of the same book. The following letters are used in the sixth position of the file name to indicate the language version:

Language	**Identifier**
Brazilian Portuguese	b

Bulgarian	u
Czech	x
Danish	d
Dutch	q
English	e
Finnish	y
French	f
German	g
Greek	a
Hungarian	h
Italian	i
Japanese	j
Korean	k
Norwegian	n
Polish	p
Portuguese	v
Russian	r
Simp. Chinese	c
Slovenian	l
Spanish	z
Swedish	s
Trad. Chinese	t
Turkish	m

2. Late breaking information that could not be included in the DB2 books is available in the Release Notes in HTML format and as an ASCII file. The HTML version is available from the Information Center and on the product CD-ROMs. To view the ASCII file:
   - On UNIX-based platforms, see the `Release.Notes` file. This file is located in the `DB2DIR/Readme/%L` directory, where *`%L`* represents the locale name and `DB2DIR` represents:
     - `/usr/lpp/db2_07_01` on AIX
     - `/opt/IBMdb2/V7.1` on HP-UX, PTX, Solaris, and Silicon Graphics IRIX
     - `/usr/IBMdb2/V7.1` on Linux.
   - On other platforms, see the `RELEASE.TXT` file. This file is located in the directory where the product is installed. On OS/2 platforms, you can also double-click the **IBM DB2** folder and then double-click the **Release Notes** icon.

Printing the PDF Books

If you prefer to have printed copies of the books, you can print the PDF files found on the DB2 publications CD-ROM. Using the Adobe Acrobat Reader, you can print either the entire book or a specific range of pages. For the file name of each book in the library, see Table 30 on page 332.

You can obtain the latest version of the Adobe Acrobat Reader from the Adobe Web site at http://www.adobe.com.

The PDF files are included on the DB2 publications CD-ROM with a file extension of PDF. To access the PDF files:

1. Insert the DB2 publications CD-ROM. On UNIX-based platforms, mount the DB2 publications CD-ROM. Refer to your *Quick Beginnings* book for the mounting procedures.
2. Start the Acrobat Reader.
3. Open the desired PDF file from one of the following locations:
   - On OS/2 and Windows platforms:

     `x:\doc\language` directory, where *x* represents the CD-ROM drive and *language* represent the two-character country code that represents your language (for example, EN for English).
   - On UNIX-based platforms:

     `/cdrom/doc/%L` directory on the CD-ROM, where */cdrom* represents the mount point of the CD-ROM and *%L* represents the name of the desired locale.

You can also copy the PDF files from the CD-ROM to a local or network drive and read them from there.

Ordering the Printed Books

You can order the printed DB2 books either individually or as a set (in North America only) by using a sold bill of forms (SBOF) number. To order books, contact your IBM authorized dealer or marketing representative, or phone `1-800-879-2755` in the United States or `1-800-IBM-4YOU` in Canada. You can also order the books from the Publications Web page at http://www.elink.ibmlink.ibm.com/pbl/pbl.

Two sets of books are available. SBOF-8935 provides reference and usage information for the DB2 Warehouse Manager. SBOF-8931 provides reference and usage information for all other DB2 Universal Database products and features. The contents of each SBOF are listed in the following table:

Table 31. Ordering the printed books

SBOF Number	Books Included
SBOF-8931	• Administration Guide: Planning • Administration Guide: Implementation • Administration Guide: Performance • Administrative API Reference • Application Building Guide • Application Development Guide • CLI Guide and Reference • Command Reference • Data Movement Utilities Guide and Reference • Data Warehouse Center Administration Guide • Data Warehouse Center Application Integration Guide • DB2 Connect User’s Guide • Installation and Configuration Supplement • Image, Audio, and Video Extenders Administration and Programming • Message Reference, Volumes 1 and 2 • OLAP Integration Server Administration Guide • OLAP Integration Server Metaoutline User’s Guide • OLAP Integration Server Model User’s Guide • OLAP Integration Server User’s Guide • OLAP Setup and User’s Guide • OLAP Spreadsheet Add-in User’s Guide for Excel • OLAP Spreadsheet Add-in User’s Guide for Lotus 1-2-3 • Replication Guide and Reference • Spatial Extender Administration and Programming Guide • SQL Getting Started • SQL Reference, Volumes 1 and 2 • System Monitor Guide and Reference • Text Extender Administration and Programming • Troubleshooting Guide • What’s New
SBOF-8935	• Information Catalog Manager Administration Guide • Information Catalog Manager User’s Guide • Information Catalog Manager Programming Guide and Reference • Query Patroller Administration Guide • Query Patroller User’s Guide

DB2 Online Documentation

Accessing Online Help

Online help is available with all DB2 components. The following table describes the various types of help.

Type of Help	Contents	How to Access...
Command Help	Explains the syntax of commands in the command line processor.	From the command line processor in interactive mode, enter: `? command` where *command* represents a keyword or the entire command. For example, `? catalog` displays help for all the CATALOG commands, while `? catalog database` displays help for the CATALOG DATABASE command.
Client Configuration Assistant Help *Command Center Help* *Control Center Help* *Data Warehouse Center Help* *Event Analyzer Help* *Information Catalog Manager Help* *Satellite Administration Center Help* *Script Center Help*	Explains the tasks you can perform in a window or notebook. The help includes overview and prerequisite information you need to know, and it describes how to use the window or notebook controls.	From a window or notebook, click the **Help** push button or press the **F1** key.

Message Help	Describes the cause of a message and any action you should take.	From the command line processor in interactive mode, enter:
		`? XXXnnnnn`
		where *XXXnnnnn* represents a valid message identifier.
		For example, `? SQL30081` displays help about the SQL30081 message.
		To view message help one screen at a time, enter:
		`? XXXnnnnn | more`
		To save message help in a file, enter:
		`? XXXnnnnn > filename.ext`
		where *filename.ext* represents the file where you want to save the message help.

Type of Help	Contents	How to Access...
SQL Help	Explains the syntax of SQL statements.	From the command line processor in interactive mode, enter: `help` *`statement`* where *statement* represents an SQL statement. For example, `help SELECT` displays help about the SELECT statement. **Note:** SQL help is not available on UNIX-based platforms.
SQLSTATE Help	Explains SQL states and class codes.	From the command line processor in interactive mode, enter: `?` *`sqlstate`* `or ?` *`class code`* where *sqlstate* represents a valid five-digit SQL state and *class code* represents the first two digits of the SQL state. For example, `? 08003` displays help for the 08003 SQL state, while `? 08` displays help for the 08 class code.

Viewing Information Online

The books included with this product are in Hypertext Markup Language (HTML) softcopy format. Softcopy format enables you to search or browse the information and provides hypertext links to related information. It also makes it easier to share the library across your site.

You can view the online books or sample programs with any browser that conforms to HTML Version 3.2 specifications.

To view online books or sample programs:

- If you are running DB2 administration tools, use the Information Center.
- From a browser, click **File —>Open Page**. The page you open contains descriptions of and links to DB2 information:
  - On UNIX-based platforms, open the following page:

    ```
    INSTHOME/sqllib/doc/%L/html/index.htm
    ```

    where *%L* represents the locale name.
  - On other platforms, open the following page:

    ```
    sqllib\doc\html\index.htm
    ```

    The path is located on the drive where DB2 is installed.

If you have not installed the Information Center, you can open the page by double-clicking the **DB2 Information** icon. Depending on the system you are using, the icon is in the main product folder or the Windows Start menu.

Installing the Netscape Browser

If you do not already have a Web browser installed, you can install Netscape from the Netscape CD-ROM found in the product boxes. For detailed instructions on how to install it, perform the following:

1. Insert the Netscape CD-ROM.
2. On UNIX-based platforms only, mount the CD-ROM. Refer to your *Quick Beginnings* book for the mounting procedures.
3. For installation instructions, refer to the `CDNAV`*`nn`*`.txt` file, where *nn* represents your two character language identifier. The file is located at the root directory of the CD-ROM.

Accessing Information with the Information Center

The Information Center provides quick access to DB2 product information. The Information Center is available on all platforms on which the DB2 administration tools are available.

You can open the Information Center by double-clicking the Information Center icon. Depending on the system you are using, the icon is in the Information folder in the main product folder or the Windows **Start** menu.

You can also access the Information Center by using the toolbar and the **Help** menu on the DB2 Windows platform.

The Information Center provides six types of information. Click the appropriate tab to look at the topics provided for that type.

Tasks
Key tasks you can perform using DB2.

Reference
DB2 reference information, such as keywords, commands, and APIs.

Books
DB2 books.

Troubleshooting
Categories of error messages and their recovery actions.

Sample Programs
Sample programs that come with the DB2 Application Development Client. If you did not install the DB2 Application Development Client, this tab is not displayed.

Web
DB2 information on the World Wide Web. To access this information, you must have a connection to the Web from your system.

When you select an item in one of the lists, the Information Center launches a viewer to display the information. The viewer might be the system help viewer, an editor, or a Web browser, depending on the kind of information you select.

The Information Center provides a find feature, so you can look for a specific topic without browsing the lists.

For a full text search, follow the hypertext link in the Information Center to the **Search DB2 Online Information** search form.

The HTML search server is usually started automatically. If a search in the HTML information does not work, you may have to start the search server using one of the following methods:

On Windows

Click **Start** and select **Programs —> IBM DB2 —> Information —> Start HTML Search Server**.

On OS/2

Double-click the **DB2 for OS/2** folder, and then double-click the **Start HTML Search Server** icon.

Refer to the release notes if you experience any other problems when searching the HTML information.

Note: The Search function is not available in the Linux, PTX, and Silicon Graphics IRIX environments.

Using DB2 Wizards

Wizards help you complete specific administration tasks by taking you through each task one step at a time. Wizards are available through the Control Center and the Client Configuration Assistant. The following table lists the wizards and describes their purpose.

Note: The Create Database, Create Index, Configure Multisite Update, and Performance Configuration wizards are available for the partitioned database environment.

Wizard	Helps You to...	How to Access...
Add Database	Catalog a database on a client workstation.	From the Client Configuration Assistant, click **Add**.
Backup Database	Determine, create, and schedule a backup plan.	From the Control Center, right-click the database you want to back up and select **Backup —> Database Using Wizard**.

Wizard	Helps You to...	How to Access...
Configure Multisite Update	Configure a multisite update, a distributed transaction, or a two-phase commit.	From the Control Center, right-click the **Databases** folder and select **Multisite Update**.
Create Database	Create a database, and perform some basic configuration tasks.	From the Control Center, right-click the **Databases** folder and select **Create —> Database Using Wizard**.
Create Table	Select basic data types, and create a primary key for the table.	From the Control Center, right-click the **Tables** icon and select **Create —> Table Using Wizard**.
Create Table Space	Create a new table space.	From the Control Center, right-click the **Table Spaces** icon and select **Create —> Table Space Using Wizard**.
Create Index	Advise which indexes to create and drop for all your queries.	From the Control Center, right-click the **Index** icon and select **Create —> Index Using Wizard**.
Performance Configuration	Tune the performance of a database by updating configuration parameters to match your business requirements.	From the Control Center, right-click the database you want to tune and select **Configure Performance Using Wizard**. For the partitioned database environment, from the Database Partitions view, right-click the first database partition you want to tune and select **Configure Performance Using Wizard**.
Restore Database	Recover a database after a failure. It helps you understand which backup to use, and which logs to replay.	From the Control Center, right-click the database you want to restore and select **Restore —> Database Using Wizard**.

Setting Up a Document Server

By default, the DB2 information is installed on your local system. This means that each person who needs access to the DB2 information must install the same files. To have the DB2 information stored in a single location, perform the following steps:

1. Copy all files and subdirectories from `\sqllib\doc\html` on your local system to a Web server. Each book has its own subdirectory that contains all the necessary HTML and GIF files that make up the book. Ensure that the directory structure remains the same.

2. Configure the Web server to look for the files in the new location. For information, refer to the NetQuestion Appendix in the *Installation and Configuration Supplement*.
3. If you are using the Java version of the Information Center, you can specify a base URL for all HTML files. You should use the URL for the list of books.
4. When you are able to view the book files, you can bookmark commonly viewed topics. You will probably want to bookmark the following pages:
   - List of books
   - Tables of contents of frequently used books
   - Frequently referenced articles, such as the ALTER TABLE topic
   - The Search form

For information about how you can serve the DB2 Universal Database online documentation files from a central machine, refer to the NetQuestion Appendix in the *Installation and Configuration Supplement*.

Searching Information Online

To find information in the HTML files, use one of the following methods:

- Click **Search** in the top frame. Use the search form to find a specific topic. This function is not available in the Linux, PTX, or Silicon Graphics IRIX environments.
- Click **Index** in the top frame. Use the index to find a specific topic in the book.
- Display the table of contents or index of the help or the HTML book, and then use the find function of the Web browser to find a specific topic in the book.
- Use the bookmark function of the Web browser to quickly return to a specific topic.
- Use the search function of the Information Center to find specific topics. See “Accessing Information with the Information Center” on page 345 for details.

Appendix B. Naming Rules

Use the naming rules shown below when you provide names for the following databases and database objects:

- Database Names
- Database and Database Alias Names
- User IDs and Passwords
- Schema Names
- Group and User Names
- Object Names.

Do not use IBM SQL or ISO/ANSI SQL92 reserved words to name tables, views, columns, indexes, or authorization IDs. A list of these words is included in the *SQL Reference*.

Refer to the *Quick Beginnings* manuals for naming rules about authorization IDs (including user names and group names) and workstations, and for additional platform restrictions.

Database Names

Every time a new database is created, the database manager creates a separate directory to store the control files and data files for that database.

The naming scheme for these directories is `SQL00001` through `SQL`*`nnnnn`*, where `SQL00001` contains control files associated with the first database created, `SQL00002` contains control files for the second database created, and so on.

These directories are maintained automatically. To avoid potential directory naming problems, do not create your own directories using the same naming schema as used by the database manager, and do not manipulate directories that have already been created by the database manager.

Database and Database Alias Names

Database names are the identifying names you or your users provide as part of the CREATE DATABASE command or API. These names must be unique within the location in which they are cataloged. For example, on UNIX based implementations of DB2, this location is a directory path, while on OS/2 implementations, it is a drive letter.

Database alias names are local synonyms given to local or remote databases. These names must be unique within the System Database Directory, in which all aliases are stored for the individual instance of the database manager. When a new database is created, the alias defaults to the database name. As a result, you cannot create a database using a name that exists as a database alias, even if there is no database with that name.

When naming a database or a database alias, the name you specify:

- Can contain 1 to 8 characters
- Must begin with one of the following:
  - A through Z (converts lowercase letters to uppercase)
  - @, #, or $
- Can include:
  - A through Z (converts lowercase letters to uppercase)
  - 0 through 9
  - @, #, $, and _ (underscore)

Note: To avoid potential problems, do not use the special characters @, #, and $ in a database name if you intend to use the database in a communications environment. Also, because these characters are not common to all keyboards, do not use them if you plan to use the database in another country. Finally, on Windows NT systems, ensure that no instance name is the same as a service name.

User IDs and Passwords

When creating a user ID or password, the name you create:

- Cannot be any of the following:
  - USERS, ADMINS, GUESTS, PUBLIC, LOCAL, or any SQL reserved word listed in the *SQL Reference*.
- Cannot begin with:
  - SQL, SYS, or IBM
- Can include:
  - A through Z

    Note: Some operating systems allow case-sensitive user IDs and passwords. You should check your operating system documentation to see if this is the case.
  - 0 through 9
  - @, #, or $
- User IDs cannot exceed 30 characters.

Note: You may be required to perform password maintenance tasks. Since such tasks are required at the server, and many users are not able or comfortable working with the server environment, carrying these tasks can pose a significant challenge. DB2 UDB provides a way to update and verify passwords without having to be at the server. For example, DB2 for OS/390 Version 5 supports this method of changing a user's password. If an error message SQL1404N "Password expired" is received, use the CONNECT statement to change the password as follows:

```
CONNECT TO <database> USER <userid> USING <password>
   NEW <new_password> VERIFY <new_password>
```

The "Password change" dialog of the DB2 Client Configuration Assistant (CCA) can also be used to change the password. For more information about these methods of changing the password, refer to the *SQL Reference* and the CCA online help.

Schema Names

The following schema names are reserved words and must not be used:

- SYSCAT
- SYSFUN
- SYSIBM
- SYSSTAT.

In general, you should avoid schema names that begin with SYS to avoid potential migration problems in the future. The database manager will not allow you to create triggers, user-defined types or user-defined functions using a schema name beginning with SYS.

It is also recommended that you not use SESSION as a schema name. Declared temporary tables must be qualified by SESSION. It is therefore possible to have an application declare a temporary table with a name identical to that of a persistent table, in which case the application logic can become overly complicated. Avoid the use of the schema SESSION, except when dealing with declared temporary tables.

Group and User Names

On UNIX based systems, groups and users can have the same name. For the GRANT statement, you must specify whether you are referring to a group or to a user. For the REVOKE statement, specifying user or group depends on whether or not there are multiple rows in the authorization catalog tables for the GRANTEE with different values of GRANTEETYPE.

On OS/2, groups and users cannot have the same name.

On Windows NT, Local Group names, Global Group names, and User IDs cannot have the same name.

Group names cannot exceed 8 characters.

Object Names

Database objects include the following:

- Schemas
- Tables
- Views
- Columns
- Indexes
- User-defined functions (UDFs)
- User-defined types (UDTs)
- Triggers
- Aliases
- Table spaces
- Stored procedures
- Methods
- Nodegroups
- Buffer pools
- Event monitors.

When naming database objects, the name you specify:

- Can contain 1 to 18 characters (bytes)

  Note: There are exceptions:
  - Schemas and columns allow 1 to 30 characters
  - Tables, views, correlation names, and alias names allow 1 to 128 characters.
- Must begin with one of the following:
  - A through Z (converts lowercase letters to uppercase)
  - A valid accented letter (such as ö)
  - A multibyte character, except multibyte spaces (for multibyte environments)
- Can include:
  - A through Z (converts lowercase letters to uppercase)

  - A valid accented letter (such as ö)
  - 0 through 9
  - @, #, $, and _ (underscore)
  - Multibyte characters, except multibyte spaces (for multibyte environments)

Keywords can be used. If the keyword is used in a context where it could also be interpreted as an SQL keyword, it must be specified as a delimited identifier. Refer to the *SQL Reference* for information on delimited identifiers.

For maximum portability, use the IBM SQL and ISO/ANSI SQL92 reserved words. For a list of these words, refer to the *SQL Reference*.

Notes:

1. Using delimited identifiers, it is possible to create an object that violates these naming rules; however, subsequent use could lead to error situations. To avoid potential problems with the use and operation of your database, **do not** violate the above rules.

   For example, if you create a column with a + or a − sign included in the name, and you subsequently use that column in an index, you will experience problems when you attempt to reorganize the table.

Federated Database Object Names

Federated database objects include:

- Index specifications
- Nicknames
- Servers
- Wrappers
- Function mappings
- Type mappings
- User mappings.

Limits apply when naming federated database objects. A complete list of object names and associated identifier limits and requirements is located in the *SQL Reference*. In summary, object names:

- Have limits. Nicknames, mapping, index specification, server, and wrapper names cannot exceed 128 bytes.
- Must begin with one of the following:
  - A through Z (names without quotation marks are converted to uppercase)
  - A valid accented letter (such as ö)

  - A multibyte character, except multibyte spaces (for multibyte environments)
- Must follow internal naming conventions. Non-leading characters can include:
  - A through Z
  - A valid accented letter (such as ö)
  - 0 through 9
  - @, #, $, and _ (underscore)
  - Multibyte characters, except multibyte spaces (for multibyte environments)

Keywords can be used. If the keyword is used in a context where it could also be interpreted as an SQL keyword, it must be specified as a delimited identifier. Refer to the *SQL Reference* for information on delimited identifiers.

For maximum portability, use the IBM SQL and ISO/ANSI SQL92 reserved words. For a list of these words, refer to the *SQL Reference*.

Options (server, nickname) and option settings are limited to 255 bytes.

How Case-Sensitive Values Are Preserved in a Federated System

With distributed requests, you sometimes need to specify identifiers and passwords that are case-sensitive at the data source. To ensure that the case is correct when they are passed to the data source, follow these guidelines:

- Specify them in the required case, and enclose them in double quotation marks.
- If you are specifying a user ID, set the fold_id server option to ″n″ (“No, don’t change case”) for the data source. If you are specifying a password, set the fold_pw server option to ″n″ for the data source.

  There is an alternative for user IDs and passwords. If a data source requires a user ID to be in lowercase, you can specify it in any case and set the fold_id server option to ″l″ (“Send this ID to the data source in lowercase”). If the data source requires the ID to be in uppercase, you can specify it in any case and set fold_id to ″u″ (“Send this ID to the data source in uppercase”). In the same way, if a data source requires a password to be in lowercase or uppercase, you can meet this requirement by setting the fold_pw server option to ″l″ or ″u″.

  For more information about server options, see ″Using Server Options to Help Define Data Sources and Facilitate Authentication Processing″ in the *Administration Guide: Implementation* .
- If you enclose a case-sensitive identifier or a password in double quotation marks at an operating system command prompt, you must ensure that the system parses the double quotation marks correctly. To do this:

- On a UNIX based operating system, enclose the statement in single quotation marks.
- On the Windows NT operating system, precede each quotation mark with a backward slash.

For example, many delimited identifiers in DB2 family data sources are case-sensitive. Suppose you want to create a nickname, NICK1, for a DB2 for CS view, "my_schema"."wkly_sal", that resides in a data source called NORBASE.

At the command prompt for a UNIX based system, you would type:

```
db2 'create nickname nick1 for norbase."my_schema"."wkly_sal"'
```

At a Windows NT command prompt, you would type:

```
db2 create nickname nick1 for norbase.\"my_schema\".\"wkly_sal\"
```

If you enter the statement from the DB2 interactive mode command prompt, or if you specify it in an application program, you do not need the single quotation marks or the slashes. For example, at the DB2 command prompt on either a UNIX based system or Windows NT, you would type:

```
create nickname nick1 for norbase."my_schema"."wkly_sal"
```

Appendix C. Planning Database Migration

This section provides you with an overview of the migration process. Note that DB2 UDB Version 6 databases do not need to be migrated to Version 7. Detailed information about migrating your DB2 UDB Version 5.*x* databases can be found in the *Quick Beginnings* manual for your operating system.

When you migrate your database:

- The following database entities are migrated:
  - Database configuration file
  - Database system catalog tables
  - Database directories
  - Database log file header
- System catalog tables are changed as follows:
  - New columns are added.
  - New tables are created.
  - A set of catalog views is migrated, and a set of new catalog views is created in the SYSCAT schema.
  - A set of updatable catalog views is created in the SYSSTAT schema.
  - A set of general purpose scalar functions is kept, and a set of new general purpose scalar functions is created in the SYSFUN schema. Only the SYSFUN.DIFFERENCE scalar function is dropped and recreated during database migration.
- A database history file and its shadow are created in the database directory. This file contains a summary of backup information that can be used if a database must be restored, and it is updated whenever specific operations are performed on the database. A summary of backup information is also kept for backup and restore operations on a table space.

Migration Considerations

To successfully migrate a database created with a previous version of the database manager, you must consider the following:

- "Migration Restrictions" on page 358
- "Security and Authorization" on page 358
- "Storage Requirements" on page 358
- "Release-to-Release Incompatibilities" on page 358

Migration Restrictions

There are certain pre-conditions or restrictions that you should be aware of before attempting to migrate your database to Version 7:

- Migration is only supported from V5.x or V6. Migration from DB2 V1.2 Parallel Edition is not supported. Earlier versions of DB2 (Database Manager) must be migrated to V5.x or V6 before being migrated to V7.
- Issuing the migration command from a V7 client to migrate a database to a V7 server is supported; however, issuing the migration command from an older DB2 client to migrate a database to a V7 server is not supported.
- Migration between platforms is not supported.
- User objects within your database cannot have V7 reserved schema names as object qualifiers. These reserved schema names include: SYSCAT, SYSSTAT, and SYSFUN.
- User-defined distinct types using the names BIGINT, REAL, DATALINK, or REFERENCE must be renamed before migrating the database.
- You cannot migrate a database that is in one of the following states:
  - Backup pending
  - Roll-forward pending
  - One or more table spaces not in a normal state
  - Transaction inconsistent
- Restoration of down-level (V5.x or V6) database backups is supported, but the rolling forward of down-level logs is not supported.

Security and Authorization

You need SYSADM authority to migrate your database.

Storage Requirements

Space is required for both the old and the new catalogs during the migration. The amount of disk space required will vary, depending on the complexity of the database, as well as the number and size of the database objects. These objects include all tables and views. You should make available at least two times the amount of disk space that the database catalog currently occupies. If there is not enough disk space, migration fails.

If your SYSCAT table space is an SMS type of table space, you should also consider updating the database configuration parameters that are associated with the log files. You should increase the values of *logfilsiz*, *logprimary*, and *logsecond* to prevent the space for these log files from running out (SQL1704N with reason code 3). If this happens, increase the log space parameters, and re-issue the MIGRATE DATABASE command.

Release-to-Release Incompatibilities

Consider the impact of incompatibilities between the two versions of the product when planning to migrate a database.

To take advantage of Version 7 enhancements, you should tune your database and database manager configuration after migrating your databases. To facilitate this, you can record and compare configuration parameter values from before and after migration. (For a description of the GET DATABASE CONFIGURATION command and the GET DATABASE MANAGER CONFIGURATION command, refer to the *Command Reference*.)

Migrating a Database

Following are the steps you must take to migrate your database. The database manager must be started before migration can begin.

PRE-MIGRATION:

Note: The pre-migration steps must be done on a previous release (that is, on your current release before migrating to, or installing, the new release).

1. Verify that there are no unresolved issues that pertain to “Migration Restrictions” on page 358.
2. Disconnect all applications and end users from each database being migrated (use the LIST APPLICATIONS command, or the FORCE APPLICATIONS command, as necessary).
3. Use the DB2CKMIG pre-migration utility to determine if the database can be migrated (for detailed information about using this utility, see the *Quick Beginnings* book for your platform). Note that on Windows NT or OS/2, you are prompted to run this tool during installation, but on UNIX based systems, this tool is invoked automatically during instance migration.
4. Back up your database.

   Migration is not a recoverable process. If you back up your database before the Version 6 reserved schema names are changed, you will not be able to restore the database using DB2 UDB Version 7. To restore the database, you will have to use your previous version of the database manager.

   Attention! If you do not have a backup of your database, and the migration fails, you will have no way of restoring your database using DB2 UDB Version 7, or your previous version of the database manager.

   You should also be aware that any database transactions done between the time the backup was taken and the time that the upgrade to Version 7 is completed are not recoverable. That is, if at some time following the completion of the installation and migration to Version 7, the database needs to be restored (to a Version 7 level), the logs written before Version 7 installation cannot be used in roll-forward recovery.

MIGRATION:

5. Migrate the database using one of the following:

- The MIGRATE DATABASE command
- The RESTORE DATABASE command, when restoring a full backup of the database
- The sqlemgdb - Migrate Database API.

On OS/2: The DB2CIDMG migration utility, which works in a Configuration/Installation/Distribution (CID) architecture environment, is only available on DB2 for OS/2. It permits remote unattended installation and configuration on LAN-based workstations. You must have NetView DM/2 on your LAN to use CID migration.

On UNIX based systems: The *Quick Beginnings* book for your platform describes what to do if you do not want to migrate all databases in a given instance.

POST-MIGRATION:

6. Optionally, use the DB2UIDDL utility to facilitate the management of a staged migration of unique indexes on your own schedule. (DB2 Version 5 databases that were created in Version 5 do not require this tool to take advantage of deferred uniqueness checking, because all unique indexes created in Version 5 have these semantics already. However, for databases that were previously migrated to Version 5, these semantics are not automatic, unless you use the DB2UIDDL utility to change the unique indexes.) This utility generates CREATE UNIQUE INDEX statements for unique indexes on user tables, and writes them to a file. Running this file as a DB2 CLP command file results in the unique index being converted to Version 7 semantics. For detailed information about using this utility, refer to one of the *Quick Beginnings* books.
7. Optionally, issue the RUNSTATS command against tables that are particularly critical to the performance of SQL queries. Old statistics are retained in the migrated database, and are not updated unless you invoke the RUNSTATS command.
8. Optionally, use the DB2RBIND utility to revalidate all packages, or allow package revalidation to occur implicitly when a package is first used.
9. Optionally, migrate Explain tables if you are planning to use them in Version 7. For more information, see "SQL Explain Facility" in the *Administration Guide: Performance*.
10. Tune your database and database manager configuration parameters to take advantage of Version 7 enhancements.

Appendix D. Incompatibilities Between Releases

This section identifies the incompatibilities that exist between DB2 Universal Database and previous releases of DB2.

An *incompatibility* is a part of DB2 Universal Database that works differently than it did in a previous release of DB2. If used in an existing application, it will produce an unexpected result, necessitate a change to the application, or reduce performance. In this context, ″application″ refers to:

- Application program code
- Third-party utilities
- Interactive SQL queries
- Command or API invocation.

Incompatibilities introduced with DB2 Universal Database Version 6 and Version 7 are described. They are grouped according to the following categories:

- System Catalog Views
- Application Programming
- SQL
- Database Security and Tuning
- Utilities and Tools
- Connectivity and Coexistence
- Configuration Parameters.

Each incompatibility section includes a description of the incompatibility, the symptom or effect of the incompatibility, and possible resolutions. There is also an indicator at the beginning of each incompatibility description that identifies the operating system to which the incompatibility applies:

WIN Microsoft Windows platforms supported by DB2

UNIX UNIX based platforms supported by DB2

OS/2 OS/2

Note: As of DB2 Universal Database Version 6, Version 1.x and Version 2.x clients, including the clients packaged with DB2 Parallel Edition Version 1.2 servers, are no longer supported.

DB2 Universal Database Planned Incompatibilities

This section describes future incompatibilities that users of DB2 Universal Database should keep in mind when coding new applications, or when modifying existing applications. This will facilitate migration to future versions of DB2 UDB.

Read-only Views in a Future Version of DB2 Universal Database

WIN	UNIX	OS/2

Change
The system catalog views will be read-only views. The SYSSTAT views will continue to be updatable.

Symptom
UPDATE statements that used to work against columns in the SYSCAT views will fail.

Explanation
Tools or applications are coded to change values in the catalog by updating the column as defined in the SYSCAT view.

Resolution
Change the tool or application to change the catalog by updating the column as defined in the SYSSTAT view.

PK_COLNAMES and FK_COLNAMES in a Future Version of DB2 Universal Database

WIN	UNIX	OS/2

Change
The SYSCAT.REFERENCES columns PK_COLNAMES and FK_COLNAMES will no longer be available.

Symptom
Column does not exist and an error is returned.

Explanation
Tools or applications are coded to use the obsolete PK_COLNAMES and FK_COLNAMES columns.

Resolution
Change the tool or application to use the SYSCAT.KEYCOLUSE view instead.

COLNAMES No Longer Available in a Future Version of DB2 Universal Database

WIN	UNIX	OS/2

Change
The SYSCAT.INDEXES column COLNAMES will no longer be available.

Symptom
Column does not exist and an error is returned.

Explanation
Tools or applications are coded to use the obsolete COLNAMES column.

Resolution
Change the tool or application to use the SYSCAT.INDEXCOLUSE view instead.

DB2 Universal Database Version 7 Incompatibilities

This section identifies incompatibilities introduced in DB2 Universal Database Version 7.

Application Programming

Query Patroller Universal Client

WIN	UNIX	OS/2

Change: This new version of the client application enabler (CAE) will only work with Query Patroller Server Version 7, because there are new stored procedures. CAE is the application interface to DB2 through which all applications must eventually pass to access the database.

Symptom: If this CAE is run against a back-level server, message SQL29001 is returned.

Object Transform Functions and Structured Types

WIN	UNIX	OS/2

Change: There is a minor and remotely possible incompatibility between a pre-Version 7 client and a Version 7 server that relates to changes that have been made to the SQLDA. As described in the *Application Development Guide*, byte 8 of the second SQLVAR can now take on the value X'12' (in addition to the values X'00' and X'01'). Applications that do not anticipate the new value may be affected by this extension.

Resolution: Because there may be other extensions to this field in future releases, developers are advised to only test for explicitly defined values.

Versions of Class and Jar Files Used by the JVM

WIN	UNIX	OS/2

Change: Previously, once a Java stored procedure or user-defined function (UDF) was started, the Java Virtual Machine (JVM) locked all files given in the CLASSPATH (including those in `sqllib/function`). The JVM used these files until the database manager was stopped. Depending on the environment in which you run a stored procedure or UDF (that is, depending on the value of the *keepdari* database manager configuration parameter, and whether or not the stored procedure is fenced), refreshing classes will let you replace class and jar files without stopping the database manager. This is different from the previous behavior.

Changed Functionality of Install, Replace, and Remove Jar Commands

WIN	UNIX	OS/2

Change: Previously, installation of a jar caused the flushing of all DARI (Database Application Remote Interface) processes. This way, a new stored procedure class was guaranteed to be picked up on the next call. Currently, no jar commands flush DARI processes. To ensure that classes from newly installed or replaced jars are picked up, you must explicitly issue the SQLEJ.REFRESH_CLASSES command.

Another incompatibility introduced by not flushing DARI processes is the fact that for fenced stored procedures, with the value of the *keepdari* database manager configuration parameter set to "YES", clients may get different versions of the jar files. Consider the following scenario:

1. User A replaces a jar and does not refresh classes.
2. User A then calls a stored procedure from the jar. Assuming that this call uses the same DARI process, User A will get an old version of the jar file.
3. User B calls the same stored procedure. This call uses a new DARI, which means that the newly created class loader will pick up the new version of the jar file.

In other words, if classes are not refreshed after jar operations, a stored procedure from different versions of jars may be called, depending on which

DARI processes are used. This differs from the previous behavior, which ensured (by flushing DARI processes) that new classes were always used.

32-bit Application Incompatibility

	UNIX	

Change: 32-bit executables (DB2 applications) will not run against the new 64-bit database engine.

Symptom: The application fails to link. When you attempt to link 32-bit objects against the 64-bit DB2 application library, an operating system linker error message is displayed.

Resolution: The application must be recompiled as a 64-bit executable, and relinked against the new 64-bit DB2 libraries.

Changing the Length Field of the Scratchpad

WIN	UNIX	OS/2

Change: Any user-defined function (UDF) that changes the length field of the scratchpad passed to the UDF will now receive SQLCODE -450.

Symptom: A UDF that changes the length field of the scratchpad fails. The invoking statement receives SQLCODE -450, with the schema and the specific name of the function filled in.

Resolution: Rewrite the UDF body to not change the length field of the scratchpad.

SQL

Applications that Use Regular Tables Qualified by the Schema SESSION

WIN	UNIX	OS/2

Change: The schema SESSION is the only schema allowed for temporary tables, and is now used by DB2 to indicate that a SESSION-qualified table may refer to a temporary table. However, SESSION is not a keyword reserved for temporary tables, and can be used as a schema for regular base tables. An application, therefore, may find a SESSION.T1 real table and a SESSION.T1 declared temporary table existing simultaneously. If, when a package is being bound, a static statement that includes a table reference qualified (explicitly or implicitly) by ″SESSION″ is encountered, neither a section nor dependencies for this statement are stored in the catalogs. Instead, this section will need to be incrementally bound at run time. This will place a copy of the section in the dynamic SQL cache, where the cached copy will be private only to the unique instance of the application. If, at run time, a declared temporary table matching the table name exists, the declared temporary table is used, even if a permanent base table of the same name exists.

Symptom: In Version 6 (and earlier), any package with static statements involving tables qualified by SESSION would always refer to a permanent base table. When binding the package, a section, as well as relevant dependency records for that statement, would be saved in the catalogs. In Version 7, these statements are not bound at bind time, and could resolve to a declared temporary table of the same name at run time. Thus, the following situations can arise:

- Migrating from Version 5. If such a package existed in Version 5, it will be bound again in Version 6, and the static statements will now be incrementally bound. This could affect performance, because these incrementally bound sections behave like cached dynamic SQL, except that the cached dynamic section cannot be shared among other applications (even different instances of the same application executable).
- Migrating from Version 6 to Version 7. If such a package existed in Version 6, it will not necessarily be bound again in Version 7. Instead, the statements will still execute as regular static SQL, using the section that was saved in the catalog at original bind time. However, if this package is rebound (either implicitly or explicitly), the statements in the package with SESSION-qualified table references will no longer be stored, and will require incremental binding. This could degrade performance.

To summarize, any packages bound in Version 7 with static statements referring to SESSION-qualified tables will no longer perform like static SQL, because they require incremental binding. If, in fact, the application process issues a DECLARE GLOBAL TEMPORARY TABLE statement for a table that has the same name as an existing SESSION-qualified table, view, or alias, references to those objects will always be taken to refer to the declared temporary table.

Resolution: If possible, change the schema names of permanent tables so that they are not ″SESSION″. Otherwise, there is no recourse but to be aware of the performance implications, and the possible conflict with declared temporary tables that may occur.

The following query can be used to identify tables, views, and aliases that may be affected if an application uses temporary tables:

```
select tabschema, tabname from SYSCAT.TABLES where tabschema = 'SESSION'
```

The following query can be used to identify Version 7 bound packages that have static sections stored in the catalogs, and whose behavior might change if the package is rebound (only relevant when moving from Version 6 to Version 7):

```
select pkgschema, pkgname, bschema, bname from syscat.packagedep
   where bschema = 'SESSION' and btype in ('T', 'V', 'I')
```

Utilities and Tools

Data Links File Manager and File System Filter on Solaris

	UNIX	

Change: Data Links File Manager and File System Filter are not supported on Solaris OS 2.5.1.

db2set on AIX and Solaris

	UNIX	

Change: The command ″db2set -ul (user level)″ and its related functions are not ported to AIX or Solaris.

Connectivity and Coexistence

32-bit Client Incompatibility

WIN	UNIX	OS/2

Change: 32-bit clients cannot attach to instances or connect to databases on 64-bit servers.

Symptom: If both the client and the server are running Version 7 code, SQL1434N is returned; otherwise, the attachment or connection fails with SQLCODE -30081.

Resolution: Use 64-bit clients.

DB2 Universal Database Version 6 Incompatibilities

This section identifies incompatibilities introduced in DB2 Universal Database Version 6.

System Catalog Views

System Catalog Views in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: In the system catalog views, new codes have been introduced: ″U″ for typed tables, and ″W″ for typed views.

Symptom: Queries that search for tables and views in the system catalogs, using the type code ″T″ for tables and ″V″ for views, will no longer find typed tables and views.

Explanation: Several system catalogs, including the system catalog views named TABLES, PACKAGEDEP, TRIGDEP, and VIEWDEP, have a column named TYPE or BTYPE containing a one-letter type code. In Version 5.2, the type code ″T″ was used for all tables, and ″V″ was used for all views. In Version 6, untyped tables will continue to have a type code of ″T″ and typed tables will have a new type code of ″U″. Similarly, untyped views will continue to have a type code of ″V″ and typed views will have a new type code of ″W″. Also, a new kind of table called a hierarchy table, not directly created by users but used by the system to implement table hierarchies, will appear in the system catalog tables with a type code of ″H″.

Resolution: Change the tool or application to recognize the codes for typed tables and views. If the tool or application needs a logical view of tables, then type codes ″T″, ″U″, ″V″, and ″W″ should be used. If the tool or application needs a physical view of tables, including hierarchy tables, then type codes ″T″ and ″H″ should be used.

Primary and Foreign Key Column Names in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: Data type change to two SYSCAT.REFERENCES columns, PK_COLNAMES and FK_COLNAMES, from VARCHAR(320) to VARCHAR(640).

Symptom: Primary key or foreign key column names are truncated, are not correct, or are missing.

Explanation: When column names greater than 18 bytes in length are used in a primary key or a foreign key, the format under which the list of column names are stored in these two columns cannot remain the same. The 20-byte blank delimited column names following the column whose length (*n*) is greater than 18 will be shifted *n*-18 bytes to the right. As well, if the list of column names exceeds 640 bytes, the column will contain the empty string.

Resolution: The SYSCAT.KEYCOLUSE view contains the list of columns that make up a primary, foreign, as well as a unique key, and should be used instead of the columns in SYSCAT.REFERENCES. Alternatively, users can restrict the length of column names to 18 bytes, or restrict the total length of the list of columns to 640 bytes.

SYSCAT.VIEWS Column TEXT in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: View text in the SYSCAT.VIEWS column TEXT will no longer be split across multiple rows. The data type is changed from VARCHAR(3600) to CLOB(64K).

Symptom: The complete view text is not given by the tool or the application.

Explanation: Tools or applications that were coded to expect no more than 3600 (or perhaps 3900) bytes returned from the TEXT column at one time are not handling the increased size of this field. The mechanism for retrieving multiple rows and reconstructing the view text using the SEQNO field is no longer necessary. The SEQNO value will always be 1.

Resolution: Change the tool or application to be able to handle values from the TEXT column that are greater than 3600 bytes. Alternatively, the view TEXT could be rewritten to fit within 3600 bytes.

SYSCAT.STATEMENTS Column TEXT in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: Statement text in the SYSCAT.STATEMENTS column TEXT will no longer be split across multiple rows. The data type is changed from VARCHAR(3600) to CLOB(64K).

Symptom: The complete statement text is not given by the tool or the application.

Explanation: Tools or applications that were coded to expect no more than 3600 (or perhaps 3900) bytes returned from the TEXT column at one time are not handling the increased size of this field. The mechanism for retrieving multiple rows and reconstructing the statement text using the SEQNO field is no longer necessary. The SEQNO value will always be 1.

Resolution: Change the tool or application to be able to handle values from the TEXT column that are greater than 3600 bytes. Alternatively, the statement TEXT could be rewritten to fit within 3600 bytes.

SYSCAT.INDEXES Column COLNAMES in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: The SYSCAT.INDEXES column COLNAMES data type is changed from VARCHAR(320) to VARCHAR(640).

Symptom: Column names are missing from an index.

Explanation: Tools or applications coded to retrieve data from a column with data type VARCHAR(320) cannot handle the increased size of this field.

Resolution: The SYSCAT.INDEXCOLUSE view contains the list of columns that make up an index, and should be used instead of the COLNAMES column. Alternatively, remove a column from the index, or reduce the size of the column name, so that the list of column names (with the leading + or −) will fit within 320 bytes.

SYSCAT.CHECKS Column TEXT in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: CHECKS Column TEXT data type is changed from CLOB(32K) to CLOB(64K).

Symptom: Check constraint clause is incomplete.

Explanation: Tools or applications coded to retrieve data from a column with data type CLOB(32K) cannot handle the increased size of this field.

Resolution: Change the tool or application to be able to handle values from the TEXT column that are longer than 32 KB. Alternatively, rewrite the check constraint clause to use fewer characters, so that it will fit within 32 KB.

Column Data Type to BIGINT in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: Several system catalog view columns have had their data type changed from INTEGER to BIGINT.

Symptom: Values are much smaller (or larger) than expected, especially statistical information.

Explanation: Tools or applications coded to retrieve data from a column with data type INTEGER cannot handle the increased size of this field.

Resolution: Change the tool or application to be able to handle values that are greater than the maximum, or less than the minimum value that can be stored in an INTEGER field. Alternatively, change the underlying structure or SQL code that causes the value to fall outside of what can be represented in an INTEGER field.

Column Mismatch in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: New columns are not inserted at the end of views in the SYSCAT view definition.

Symptom: Re-preprocessing fails with several column mismatches or column data type mismatches.

Explanation: New columns are introduced to the system catalog views and placed in a position that is useful in an ad hoc query environment; specifically, shorter columns are placed before very long columns, and the REMARKS column is always last.

Resolution: Explicitly name the columns in the select list instead of coding ″SELECT *″.

SYSCAT.COLUMNS and SYSCAT.ATTRIBUTES in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: SYSCAT.COLUMNS and SYSCAT.ATTRIBUTES now contain entries for inherited columns and attributes.

Symptom: Queries against SYSCAT.COLUMNS to retrieve the columns of a typed table or view, and queries against SYSCAT.ATTRIBUTES to retrieve the attributes of a structured type, may return more rows in Version 6 than in Version 5.2 if the subject of the query is a subtable, subview, or subtype.

Explanation: In Version 5.2, for a given table, view, or structured type, the COLUMNS and ATTRIBUTES catalogs contained entries only for columns and attributes that were introduced by that table, view, or type. Columns and attributes that were inherited from supertables or supertypes were not represented in the catalogs. However, in Version 6, the COLUMNS and ATTRIBUTES catalogs will contain entries for inherited columns and attributes.

Resolution: Change the tool or application to recognize the new entries in the COLUMNS and ATTRIBUTES catalogs.

OBJCAT Views No Longer Supported in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: The recursive catalog views in the OBJCAT schema of Version 5.2 are no longer part of the shipped DB2 Universal Database product.

Symptom: Queries written against the OBJCAT catalog views will no longer run successfully.

Resolution: Most of the information formerly in the OBJCAT views has been incorporated into the regular SYSCAT catalog views. In most cases, you can obtain the information from the system catalog views. If you migrate from Version 5.2, and the OBJCAT catalog views exist, they should be dropped. This can be done by running the CLP script called `objcatdp.db2`, found under the `misc` subdirectory of the `sqllib` directory.

You can also create your own set of OBJCAT views that are equivalent to the catalog views supported in Version 5.2.

In version 5.2, ″Appendix E″ of the *SQL Reference* warned users that the OBJCAT catalog views were temporary, and would not be supported in future releases.

Dependency Codes Changed in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: In the system catalog views, the hierarchic dependencies formerly denoted by code ″H″ are now denoted by code ″O″.

Symptom: Queries that search for hierarchic dependencies by code ″H″ in the catalog views will no longer work correctly.

Explanation: Several system catalogs, including the system catalog views named PACKAGEDEP, TRIGDEP, and VIEWDEP, have a column named

BTYPE. In Version 5.2, the OBJCAT views denoted hierarchic dependencies by code ″H″. In Version 6, these dependencies are denoted by code ″O″.

Resolution: Revise these queries to search for code ″O″.

SYSIBM Base Catalog Tables in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: Following are changes to the SYSIBM base catalog tables, which you may still be using instead of the SYSCAT views:

- Deleted fields (but still in the SYSCAT views):
  - SYSSTMT.SEQNO
  - SYSVIEWS.SEQNO
- Renamed catalog table: SYSTRIGDEP changed to SYSDEPENDENCIES. As well, the columns BCREATOR and DCREATOR were renamed to BSCHEMA and DSCHEMA, respectively. The view SYSCAT.TRIGDEP did not change.
- Deleted fields (were never in the SYSCAT views):
  - SYSATTRIBUTES.DEFAULT_VALUE
  - SYSATTRIBUTES.NULLS
  - SYSCOLUMNS.SERVERTYPE
  - SYSDATATYPES.REFREP_TYPENAME
  - SYSDATATYPES.REFREP_TYPESCHEMA
  - SYSDATATYPES.REFREP_LENGTH
  - SYSDATATYPES.REFREP_SCALE
  - SYSDATATYPES.REFREP_CODEPAGE
  - SYSINDEXES.TEXT

    (Was in the view, but reserved for future use only.)
  - SYSPLANDEP.PUBLICPRIV
  - SYSSECTION.SEQNO
  - SYSTABAUTH.UPDATE_BY_COLS
  - SYSTABAUTH.REF_BY_COLS
  - SYSTABLES.MINPDLENGTH
  - SYSTABLESPACES.READONLY
  - SYSTABLESPACES.REMOVABLEMEDIA
- Data type changes:
  - SYSSECTION.SECTION, from VARCHAR(3600) to CLOB(10M)
  - SYSPLANDEP.COLUSAGE, from VARCHAR(3000) FOR BIT DATA to BLOB(5K)

Application Programming

VARCHAR Data Type in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: Maximum possible size of VARCHAR (VARGRAPHIC) data type has increased from 4000 characters (2000 double byte characters) to 32672 characters (16336 double byte characters) in Version 6.

Symptom: An application that uses fixed length buffers of 4000 bytes for a VARCHAR (VARGRAPHIC) data type has the potential for buffer overwrite or truncation, if it fetches a VARCHAR field that is longer than 4000 bytes into a buffer that is too small. The CLI function - `SQLGetTypeInfo()` now returns the size of VARCHAR as 32672. CLI applications that use this value in table DDLs may get errors because table spaces of sufficient page size are not available. For more information about table space page size, see “User Table Data” on page 117.

Resolution: When coding the application, it is recommended that you first describe the columns of the result set (using the DESCRIBE statement), and then use buffers whose size is based on the length returned from the DESCRIBE statement.

Java Programming Positioned UPDATE and DELETE in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: When programming Java in Version 6, positioned UPDATE and DELETE statements use the default authorization identifier of the person that bound the cursor package. This is different from Version 5.2, in which the authorization identifier of the person running the package was used.

Symptom: The package containing the positioned UPDATE and DELETE statements may not run, because the authorization identifier of the person who bound the package does not have sufficient authority.

Resolution: The authorization identifier of the person who binds the package must be granted sufficient authority to run the positioned UPDATE and DELETE statements in the package. Grant the correct privileges and then rebind the package.

Syntax Change in FOR UPDATE Clause in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: In Version 5.2, the FOR UPDATE clause in a SELECT statement can be used in an SQLJ program to identify the columns that can be updated in subsequent positioned UPDATE statements. The syntax has changed for Version 6.

Symptom: You will receive the error message SQJ0204E if a SELECT statement contains a FOR UPDATE clause.

Resolution: Remove the FOR UPDATE clause from the SELECT statement. Specify an updatable iterator through the iterator declaration clause. For example:

```
#sql public iterator DelByName implements sqlj.runtime.ForUpdate(String EmpNo)
   with updateColumns = (salary);
```

If you want to explicitly identify what columns are updateable, specify them through the `updateColumns` keyword, used in conjunction with the WITH clause.

For more information about positioned iterator declarations, refer to the *Application Development Guide*.

Character Name Sizes in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: DB2 Universal Database Version 6 supports 128-byte table, view, and alias names, and 30-byte column names. Previous support was for 18-byte names for each of these entities.

USER and CURRENT SCHEMA special registers were CHAR(8), and are now VARCHAR(128). The CURRENT EXPLAIN MODE special register was CHAR(8), and is now VARCHAR(254). The output for TYPE_SCHEMA and TABLE_SCHEMA built-in functions was CHAR(8), and is now VARCHAR(128).

Symptom: If applications that were developed before Version 6 are run against a Version 6 database that does not use the longer limits, application behavior should not change at all. However, running these applications against a Version 6 database that *does* use longer names could result in certain side effects, depending on how these applications were coded.

Following are some examples:

- Consider an existing application that FETCHes a table or column name (typically from a catalog view) into a host variable that was defined to be 18 bytes long. Since 18 bytes was the limit on the size of the table or column name until Version 6, this application may not bother to check the `sqlwarn1` bit of the SQLCA. It will assume (incorrectly) that truncation will never occur.
- Consider an application that FETCHes a table or column name (typically from a catalog view) into an SQLDA, where the size of the `sqldata` field was allocated on the basis of the `sqllen` field from a DESCRIBE of the SELECT. This will result in the correct (untruncated) result being returned to the application, even though the size of the table or column names may have increased. If other application logic operates on the assumption that column names are limited to 18 bytes, longer names that are returned may be handled in an unexpected way; for example, the display of longer column names may be truncated at 18 bytes.
- Since the SQLCA token field (`sqlerrmc`) is limited to 70 bytes, existing applications that attempt to insert a row into a table may be affected. In response to error SQL0204N, such applications determine the name of the table from the SQLCA `sqlerrmc` field, and then perform some operations based on that object name. With earlier versions of DB2, the table or schema identifier limit guaranteed that the entire table name would be included in the SQLCA. This is not the case in Version 6.
- An application using a back-level API will only get the first 18 bytes of a table name.
- Existing CLI and ODBC applications that use the schema functions (such as SQLTables(), or SQLColumns(), and others) will be affected when connecting to a server with support for greater than 18-byte names. Although there will be truncation warnings, the application may not check for this warning, and may proceed with a truncated name.

Resolution: The best way to resolve problems of this type is to recode the application to handle longer table and column names. Otherwise, ensure that these applications are not run against Version 6 databases that use greater than 18-byte names.

PC/IXF Format Changes in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: DB2 Universal Database Version 6 supports 128-byte table, view, and alias names, and 30-byte column names. Previous support was for 18-byte names for each of these entities.

Symptom: A DB2 Universal Database Version 5 client cannot import a PC/IXF file that was exported by a DB2 Universal Database Version 6 client (error SQL3059N). A PC/IXF file (exported from a DB2 Universal Database Version 6 client) cannot be loaded into a DB2 Universal Database Version 5 database (error SQL3059N).

Resolution: Use compatible versions of DB2 Universal Database when importing or loading PC/IXF data.

SQLNAME in a Non-doubled SQLVAR in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: DB2 Universal Database Version 6 supports 30-byte column names. The former support was for 18-byte names. In Version 5, the documented behavior was that ″0xFF″ is placed in the 30th byte of an SQLNAME field for a non-doubled SQLVAR; for system-generated names and for user-specified column names specified in an ″AS″ clause, ″0x00″ is also placed in the 30th byte.

In Version 6, ″0xFF″ is returned in the 30th byte only if the name is system-generated.

Symptom: Any applications that rely on the 30th byte of the SQLNAME field to determine whether it is a user-specified column name or a system-generated name may receive unexpected logic checks if the user-specifed column name is 30 characters long. This should be a rare occurrence.

Resolution: These applications should be modified to only check for ″0xFF″ in the 30th byte of the SQLNAME field if the length of that field is less than 30. In this case, the name is user-generated.

Obsolete DB2 CLI/ODBC Configuration Keywords in DB2 Universal Database Version 6

WIN		

Change: When migrating to a new version of DB2 UDB, you can change the behavior of the DB2 CLI/ODBC driver by specifying a set of optional keywords in the `db2cli.ini` file.

In Version 6, the TRANSLATEDLL and TRANSLATEOPTION keywords became obsolete.

Symptom: These keywords will be ignored if they still exist. You may notice behavioral changes based on the removal of these settings.

Resolution: You will need to review the new list of valid parameters to decide what the appropriate keywords and settings are for your environment. For information about these keywords, refer to the *CLI Guide and Reference*.

Event Monitor Output Stream Format in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: Event monitor output streams have no version control. As a result, adding support for greater than 18-byte table names requires moving to an output stream format.

Symptom: Applications that parse the event monitor output streams will no longer work properly.

Resolution: There are two options:
- Update the application to use the new data stream.
- Set the registry variable

  ```
  DB2OLDEVMON=evmonname1,evmonname2,...
  ```

  where *evmonname* is the name of the event monitor that you want written in the old data format. Note that any new fields in the event monitor will not be accessible under the old data format.

SQL

DATALINK Columns in DB2 Universal Database Version 6

	UNIX	

Change: DATALINK values inserted under DB2 Universal Database Version 6 will require an extra four bytes of space in the column value descriptor.

Symptom: When DATALINK columns created in Version 5.2 are updated, an additional four bytes are required on the data page to store the new column value. As a result, there may not be enough space in the data page to complete the update, and it may have to be moved to a new page. This could cause the update to run out of space.

Resolution: You will need to add more space on your system to allow for updates.

SYSFUN String Function Signatures in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: A number of string functions in the SYSFUN schema now have improved versions defined in the SYSIBM schema (built-in functions). The function names are LCASE, LTRIM, RTRIM, and UCASE.

Symptom: When preparing statements or creating views, the returned data type from any of these functions may be different in Version 6. This occurs because the built-in functions (under the SYSIBM schema) are usually resolved before functions in the SYSFUN schema are resolved.

Resolution: No action is required. The built-in function is usually preferred over the function in the SYSFUN schema. The previous version behavior can be restored by switching the SQL path (so that SYSFUN precedes SYSIBM), but performance will be degraded. The previous version function can also be invoked by qualifying the function name with the schema name SYSFUN.

Migrated packages, views, summary tables, triggers, and constraints that reference these functions continue to use the version from the SYSFUN schema, unless explicit action is taken, such as explicitly binding the package or recreating the view, summary table, trigger, or constraint.

SYSTABLE Column Change With New Integrity State in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: The ″U″ states in the CONST_CHECKED column of SYSCAT.TABLES changes differently when a SET INTEGRITY ... OFF statement is run.

Symptom: Prior to Version 6, any ″U″ state in the CONST_CHECKED column changed to an ″N″ state when a SET INTEGRITY ... OFF statement was run. The ″U″ state now changes to a ″W″ state.

Resolution: No action is required. The new ″W″ state in the CONST_CHECKED column is used to indicate that the constraints type was previously checked by the user, and that some data in the table may need to be checked for integrity.

The ″N″ state does not clarify whether there exists any old data that has not yet been verified by the database manager. On a subsequent SET INTEGRITY ... IMMEDIATE CHECKED INCREMENTAL statement, the database manager must return an error, because data integrity cannot be guaranteed if only new changes have been checked. On the other hand, the ″W″ state can be changed back to the ″U″ state (if the INCREMENTAL option is specified) to indicate that the user is still responsible for the integrity of data in the table. If the INCREMENTAL option is not specified, the database manager will choose full processing, change the ″W″ state to a ″Y″ state, and assume responsibility for maintaining data integrity.

Database Security and Tuning

Creating Databases Using Clients in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: The method used by clients to create a database.

Symptom: Using a back-level client to create a database will result in errors.

Resolution: When using a client to create a database, ensure that the client and the server are running the same level of DB2 code.

SELECT Privilege Required on Hierarchy in DB2 Universal Database Version 6

WIN 32-bit	UNIX	OS/2

Change: Specification of the ONLY keyword (for a table) now requires that the user have SELECT privilege on all subtables of the specified typed table. Similarly, specification of the ONLY keyword (for a view) now requires that the user have SELECT privilege on all subviews of the specified typed table. Previous versions of DB2 only required SELECT privilege on the specified table or view.

Symptom: There are two possible symptoms:

- An authorization error (SQLCODE -551, SQLSTATE 42501) occurs when rebinding a package containing an SQL statement that specifies the ONLY keyword in a FROM clause, if the authorization ID under which the package was bound lacks the SELECT privilege on the subtables of the named typed table (or view).
- If the definition of a view or trigger contains the ONLY keyword in a FROM clause, the view or trigger will continue to work normally. However, the definition of the view or trigger can no longer be used to create a new view or trigger, unless the creator holds the SELECT privilege on all of the subtables of the named table (or view).

Resolution: The authorization ID that needs to rebind a package, or to create a new view or trigger, should be granted SELECT privilege on all subtables (and subviews) of the table (or view) specified following the ONLY keyword.

Obsolete Profile Registry and Environment Variables in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: The following profile registry or environment variables are obsolete:

- DB2_VECTOR

Resolution: These variables are no longer needed.

Utilities and Tools

Current Explain Mode in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: The type of the ″CURRENT EXPLAIN MODE″ special register has changed from CHAR(8) to VARCHAR(254).

Symptom: If the application assumes that the type is still CHAR(8), the value may be truncated from 254 to 8 bytes.

Resolution: Redefine the type of all host variables that read the special register, from CHAR(8) to VARCHAR(254).

This change is required to accommodate two new values for the ″CURRENT EXPLAIN MODE″ special register. These new values are ″EVALUATE INDEXES″ and ″RECOMMEND INDEXES″.

The USING and SORT BUFFER Parameters in DB2 Universal Database Version 6

WIN	UNIX	OS/2

Change: As of Version 6, the USING and SORT BUFFER parameters of the LOAD command are no longer supported. These parameters are ignored.

Symptom: A warning message is returned, stating that the USING and SORT BUFFER parameters are no longer supported, and will be ignored by the load utility.

Resolution: Ignore the warning message. For additional information, refer to the *Data Movement Utilities Guide and Reference*.

Connectivity and Coexistence

Replace RUMBA with PCOMM in DB2 Universal Database Version 6

WIN		

Change: In Version 6, RUMBA is replaced by PCOMM on Windows NT, Windows 98, and Windows 95 (but not on Windows 3.1).

Symptom: None.

Resolution: None.

Configuration Parameters

Obsolete Database Configuration Parameters

WIN	UNIX	OS/2

Change: The following database configuration parameters are obsolete:

- DL_NUM_BACKUP (replaced by NUM_DB_BACKUP database configuration parameter)

Resolution: Remove all references to these parameters from your applications.

Appendix E. National Language Support (NLS)

This section contains information about the national language support (NLS) provided by DB2, including information about countries, languages, and code pages (code sets) supported, and how to configure and use DB2 NLS features in your databases and applications.

Country Code and Code Page Support

Table 32 on page 384 shows the languages and code sets supported by the database servers, and how these values are mapped to country code and code page values that are used by the database manager.

The following is an explanation of each column in the table:

- **Code Page** shows the IBM-defined code page as mapped from the operating system code set.
- **Group** shows whether a code page is single-byte (″S″) or multi-byte (″D″). The ″-n″ is a number used to create a letter-number combination. Matching combinations show where connection and conversion is allowed by DB2. For example, all ″S-1″ groups can work together.
- **Code Set** shows the code set associated with the supported language. The code set is mapped to the DB2 code page.
- **Tr.** shows the two letter territory identifier.
- **Country Code** shows the country code that is used by the database manager internally to provide country-specific support.
- **Locale** shows the locale values supported by the database manager.
- **OS** shows the operating system that supports the languages and code sets.
- **Country Name** shows the name of the country or countries.

Table 32. Supported Languages and Code Sets

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
437	S-1	IBM-437	AL	355	-	OS2	Albania
850	S-1	IBM-850	AL	355	-	OS2	Albania
819	S-1	ISO8859-1	AL	355	sq_AL	AIX	Albania
850	S-1	IBM-850	AL	355	Sq_AL	AIX	Albania
819	S-1	iso88591	AL	355	-	HP	Albania
1051	S-1	roman8	AL	355	-	HP	Albania
819	S-1	ISO8859-1	AL	355	-	Sun	Albania
1252	S-1	1252	AL	355	-	WIN	Albania
1275	S-1	1275	AL	355	-	Mac	Albania
37	S-1	IBM-37	AL	355	-	HOST	Albania
1140	S-1	IBM-1140	AL	355	-	HOST	Albania
864	S-6	IBM-864	AA	785	-	OS2	Arabic Countries
1046	S-6	IBM-1046	AA	785	Ar_AA	AIX	Arabic Countries
1089	S-6	ISO8859-6	AA	785	ar_AA	AIX	Arabic Countries
1089	S-6	iso88596	AA	785	ar_SA.iso88596	HP	Arabic Countries
1256	S-6	1256	AA	785	-	WIN	Arabic Countries
420	S-6	IBM-420	AA	785	-	HOST	Arabic Countries
437	S-1	IBM-437	AU	61	-	OS2	Australia
850	S-1	IBM-850	AU	61	-	OS2	Australia
819	S-1	ISO8859-1	AU	61	en_AU	AIX	Australia
850	S-1	IBM-850	AU	61	En_AU	AIX	Australia
819	S-1	iso88591	AU	61	-	HP	Australia
1051	S-1	roman8	AU	61	-	HP	Australia
819	S-1	ISO8859-1	AU	61	en_AU	Sun	Australia
819	S-1	ISO8859-1	AU	61	en_AU	SCO	Australia
1252	S-1	1252	AU	61	-	WIN	Australia
1275	S-1	1275	AU	61	-	Mac	Australia
37	S-1	IBM-37	AU	61	-	HOST	Australia
1140	S-1	IBM-1140	AU	61	-	HOST	Australia
437	S-1	IBM-437	AT	43	-	OS2	Austria
850	S-1	IBM-850	AT	43	-	OS2	Austria
819	S-1	ISO8859-1	AT	43	ge_AT	AIX	Austria
850	S-1	IBM-850	AT	43	Ge_AT	AIX	Austria
819	S-1	iso88591	AT	43	-	HP	Austria
1051	S-1	roman8	AT	43	-	HP	Austria
819	S-1	ISO8859-1	AT	43	de_AT	SCO	Austria
819	S-1	ISO-8859-1	AT	43	de_AT	Linux	Austria
819	S-1	ISO8859-1	AT	43	de_AT	Sun	Austria
1252	S-1	1252	AT	43	-	WIN	Austria
1275	S-1	1275	AT	43	-	Mac	Austria
37	S-1	IBM-37	AT	43	-	HOST	Austria
1140	S-1	IBM-1140	AT	43	-	HOST	Austria

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
915	S-5	ISO8859-5	BY	375	-	OS2	Belarus
915	S-5	ISO8859-5	BY	375	be_BY	AIX	Belarus
1131	S-5	IBM-1131	BY	375	-	OS2	Belarus
1251	S-5	1251	BY	375	-	WIN	Belarus
1283	S-5	1283	BY	375	-	Mac	Belarus
1025	S-5	IBM-1025	BY	375	-	HOST	Belarus
437	S-1	IBM-437	BE	32	-	OS2	Belgium
850	S-1	IBM-850	BE	32	-	OS2	Belgium
819	S-1	ISO8859-1	BE	32	nl_BE	AIX	Belgium
850	S-1	IBM-850	BE	32	Nl_BE	AIX	Belgium
819	S-1	iso88591	BE	32	-	HP	Belgium
819	S-1	ISO8859-1	BE	32	fr_BE	SCO	Belgium
819	S-1	ISO8859-1	BE	32	nl_BE	SCO	Belgium
819	S-1	ISO-8859-1	BE	32	nl_BE	Linux	Belgium
819	S-1	ISO8859-1	BE	32	nl_BE	Sun	Belgium
1252	S-1	1252	BE	32	-	WIN	Belgium
1275	S-1	1275	BE	32	-	Mac	Belgium
500	S-1	IBM-500	BE	32	-	HOST	Belgium
1148	S-1	IBM-1148	BE	32	-	HOST	Belgium
855	S-5	IBM-855	BG	359	-	OS2	Bulgaria
915	S-5	ISO8859-5	BG	359	-	OS2	Bulgaria
915	S-5	ISO8859-5	BG	359	bg_BG	AIX	Bulgaria
915	S-5	iso88595	BG	359	bg_BG.iso88595	HP	Bulgaria
1251	S-5	1251	BG	359	-	WIN	Bulgaria
1283	S-5	1283	BG	359	-	Mac	Bulgaria
1025	S-5	IBM-1025	BG	359	-	HOST	Bulgaria
850	S-1	IBM-850	BR	55	-	OS2	Brazil
850	S-1	IBM-850	BR	55	-	AIX	Brazil
819	S-1	ISO8859-1	BR	55	pt_BR	AIX	Brazil
819	S-1	ISO8859-1	BR	55	-	HP	Brazil
819	S-1	ISO8859-1	BR	55	pt_BR	SCO	Brazil
819	S-1	ISO8859-1	BR	55	pt_BR	Sun	Brazil
819	S-1	ISO-8859-1	BR	55	pt_BR	Linux	Brazil
1252	S-1	1252	BR	55	-	WIN	Brazil
37	S-1	IBM-37	BR	55	-	HOST	Brazil
1140	S-1	IBM-1140	BR	55	-	HOST	Brazil

Table 32. Supported Languages and Code Sets (continued)

Code Page	Group	Code-Set	Tr.	Country Code	Locale	OS	Country Name
850	S-1	IBM-850	CA	1	-	OS2	Canada
850	S-1	IBM-850	CA	1	En_CA	AIX	Canada
819	S-1	ISO8859-1	CA	1	en_CA	AIX	Canada
819	S-1	iso88591	CA	1	fr_CA.iso88591	HP	Canada
1051	S-1	roman8	CA	1	fr_CA.roman8	HP	Canada
819	S-1	ISO8859-1	CA	1	en_CA	SCO	Canada
819	S-1	ISO8859-1	CA	1	fr_CA	SCO	Canada
819	S-1	ISO8859-1	CA	1	en_CA	Sun	Canada
819	S-1	ISO8859-1	CA	1	en_CA	Sun	Canada
819	S-1	ISO-8859-1	CA	1	en_CA	Linux	Canada
1252	S-1	1252	CA	1	-	WIN	Canada
1275	S-1	1275	CA	1	-	Mac	Canada
37	S-1	IBM-37	CA	1	-	HOST	Canada
1140	S-1	IBM-1140	CA	1	-	HOST	Canada
863	S-1	IBM-863	CA	2	-	OS2	Canada (French)
1381	D-4	IBM-1381	CN	86	-	OS2	China (PRC)
1386	D-4	GBK	CN	86	-	OS2	China (PRC)
1383	D-4	IBM-eucCN	CN	86	zh_CN	AIX	China (PRC)
1386	D-4	GBK	CN	86	Zh_CN.GBK	AIX	China (PRC)
1383	D-4	hp15CN	CN	86	zh_CN.hp15CN	HP	China (PRC)
1383	D-4	eucCN	CN	86	zh_CN	SCO	China (PRC)
1383	D-4	eucCN	CN	86	zh_CN.eucCN	SCO	China (PRC)
1383	D-4	gb2312	CN	86	zh	Sun	China (PRC)
1381	D-4	IBM-1381	CN	86	-	WIN	China (PRC)
1386	D-4	GBK	CN	86	-	WIN	China (PRC)
935	D-4	IBM-935	CN	86	-	HOST	China (PRC)
1388	D-4	IBM-1388	CN	86	-	HOST	China (PRC)
852	S-2	IBM-852	HR	385	-	OS2	Croatia
912	S-2	ISO8859-2	HR	385	hr_HR	AIX	Croatia
912	S-2	iso88592	HR	385	hr_HR.iso88592	HP	Croatia
912	S-2	ISO8859-2	HR	385	hr_HR.ISO8859-2	SCO	Croatia
912	S-2	ISO-8859-2	HR	385	hr_HR	Linux	Croatia
1250	S-2	1250	HR	385	-	WIN	Croatia
1282	S-2	1282	HR	385	-	Mac	Croatia
870	S-2	IBM-870	HR	385	-	HOST	Croatia
852	S-2	IBM-852	CZ	421	-	OS2	Czech Republic
912	S-2	ISO8859-2	CZ	421	cs_CZ	AIX	Czech Republic
912	S-2	iso88592	CZ	421	cs_CZ.iso88592	HP	Czech Republic
912	S-2	ISO8859-2	CZ	421	cs_CZ.ISO8859-2	SCO	Czech Republic
912	S-2	ISO-8859-2	CZ	421	cs_CZ	Linux	Czech Republic
1250	S-2	1250	CZ	421	-	WIN	Czech Republic
1282	S-2	1282	CZ	421	-	Mac	Czech Republic
870	S-2	IBM-870	CZ	421	-	HOST	Czech Republic

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
850	S-1	IBM-850	DK	45	-	OS2	Denmark
819	S-1	ISO8859-1	DK	45	da_DK	AIX	Denmark
850	S-1	IBM-850	DK	45	Da_DK	AIX	Denmark
819	S-1	iso88591	DK	45	da_DK.iso88591	HP	Denmark
1051	S-1	roman8	DK	45	da_DK.roman8	HP	Denmark
819	S-1	ISO8859-1	DK	45	da	SCO	Denmark
819	S-1	ISO8859-1	DK	45	da_DA	SCO	Denmark
819	S-1	ISO8859-1	DK	45	da_DK	SCO	Denmark
819	S-1	ISO8859-1	DK	45	da	Sun	Denmark
819	S-1	ISO8859-1	DK	45	da	Sun	Denmark
819	S-1	ISO-8859-1	DK	45	da_DK	Linux	Denmark
1252	S-1	1252	DK	45	-	WIN	Denmark
1275	S-1	1275	DK	45	-	Mac	Denmark
277	S-1	IBM-277	DK	45	-	HOST	Denmark
1142	S-1	IBM-1142	DK	45	-	HOST	Denamrk
922	S-10	IBM-922	EE	372	-	OS2	Estonia
922	S-10	IBM-922	EE	372	Et_EE	AIX	Estonia
922	S-10	IBM-922	EE	372	-	WIN	Estonia
1122	S-10	IBM-1122	EE	372	-	HOST	Estonia
437	S-1	IBM-437	FI	358	-	OS2	Finland
850	S-1	IBM-850	FI	358	-	OS2	Finland
819	S-1	ISO8859-1	FI	358	fi_FI	AIX	Finland
850	S-1	IBM-850	FI	358	Fi_FI	AIX	Finland
819	S-1	iso88591	FI	358	fi_FI.iso88591	HP	Finland
819	S-1	ISO8859-1	FI	358	fi	SCO	Finland
819	S-1	ISO8859-1	FI	358	fi_FI	SCO	Finland
819	S-1	ISO8859-1	FI	358	sv_FI	SCO	Finland
819	S-1	ISO8859-1	FI	358	-	Sun	Finland
819	S-1	ISO-8859-1	FI	358	fi_FI	Linux	Finland
1051	S-1	roman8	FI	358	-	HP	Finland
1252	S-1	1252	FI	358	-	WIN	Finland
1275	S-1	1275	FI	358	-	Mac	Finland
278	S-1	IBM-278	FI	358	-	HOST	Finland
1143	S-1	IBM-1143	FI	358	-	HOST	Finland
855	S-5	IBM-855	MK	389	-	OS2	FYR Macedonia
915	S-5	ISO8859-5	MK	389	-	OS2	FYR Macedonia
915	S-5	ISO8859-5	MK	389	mk_MK	AIX	FYR Macedonia
915	S-5	iso88595	MK	389	-	HP	FYR Macedonia
1251	S-5	1251	MK	389	-	WIN	FYR Macedonia
1283	S-5	1283	MK	389	-	Mac	FYR Macedonia
1025	S-5	IBM-1025	MK	389	-	HOST	FYR Macedonia

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
437	S-1	IBM-437	FR	33	-	OS2	France
850	S-1	IBM-850	FR	33	-	OS2	France
819	S-1	ISO8859-1	FR	33	fr_FR	AIX	France
850	S-1	IBM-850	FR	33	Fr_FR	AIX	France
819	S-1	iso88591	FR	33	fr_FR.iso88591	HP	France
1051	S-1	roman8	FR	33	fr_FR.roman8	HP	France
819	S-1	ISO8859-1	FR	33	fr	Sun	France
819	S-1	ISO8859-1	FR	33	fr	SCO	France
819	S-1	ISO8859-1	FR	33	fr_FR	SCO	France
819	S-1	ISO-8859-1	FR	33	fr_FR	Linux	France
1252	S-1	1252	FR	33	-	WIN	France
1275	S-1	1275	FR	33	-	Mac	France
297	S-1	IBM-297	FR	33	-	HOST	France
1147	S-1	IBM-1147	FR	33	-	HOST	France
437	S-1	IBM-437	DE	49	-	OS2	Germany
850	S-1	IBM-850	DE	49	-	OS2	Germany
819	S-1	ISO8859-1	DE	49	de_DE	AIX	Germany
850	S-1	IBM-850	DE	49	De_DE	AIX	Germany
819	S-1	iso88591	DE	49	de_DE.iso88591	HP	Germany
1051	S-1	roman8	DE	49	de_DE.roman8	HP	Germany
819	S-1	ISO8859-1	DE	49	de	SCO	Germany
819	S-1	ISO8859-1	DE	49	de_DE	SCO	Germany
819	S-1	ISO8859-1	DE	49	de	Sun	Germany
819	S-1	ISO-8859-1	DE	49	de_DE	Linux	Germany
1252	S-1	1252	DE	49	-	WIN	Germany
1275	S-1	1275	DE	49	-	Mac	Germany
273	S-1	IBM-273	DE	49	-	HOST	Germany
1141	S-1	IBM-1141	DE	49	-	HOST	Germany
819	S-1	ISO8859-1	DE	49	De_DE.88591	SINIX	Germany
819	S-1	ISO8859-1	DE	49	De_DE.6937	SINIX	Germany
813	S-7	ISO8859-7	GR	30	-	OS2	Greece
869	S-7	IBM-869	GR	30	-	OS2	Greece
813	S-7	ISO8859-7	GR	30	el_GR	AIX	Greece
813	S-7	iso88597	GR	30	el_GR.iso88597	HP	Greece
813	S-7	ISO8859-7	GR	30	el_GR.ISO8859-7	SCO	Greece
813	S-7	ISO-8859-7	GR	30	gr_GR	Linux	Greece
737	S-7	737	GR	30	-	WIN	Greece
1253	S-7	1253	GR	30	-	WIN	Greece
1280	S-7	1280	GR	30	-	Mac	Greece
423	S-7	IBM-423	GR	30	-	HOST	Greece
875	S-7	IBM-875	GR	30	-	HOST	Greece

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
852	S-2	IBM-852	HU	36	-	OS2	Hungary
912	S-2	ISO8859-2	HU	36	hu_HU	AIX	Hungary
912	S-2	iso88592	HU	36	hu_HU.iso88592	HP	Hungary
912	S-2	ISO8859-2	HU	36	hu_HU.ISO8859-2	SCO	Hungary
912	S-2	ISO-8859-2	HU	36	hu_HU	Linux	Hungary
1250	S-2	1250	HU	36	-	WIN	Hungary
1282	S-2	1282	HU	36	-	Mac	Hungary
870	S-2	IBM-870	HU	36	-	HOST	Hungary
850	S-1	IBM-850	IS	354	-	OS2	Iceland
819	S-1	ISO8859-1	IS	354	is_IS	AIX	Iceland
850	S-1	IBM-850	IS	354	Is_IS	AIX	Iceland
819	S-1	iso88591	IS	354	is_IS.iso88591	HP	Iceland
1051	S-1	roman8	IS	354	is_IS.roman8	HP	Iceland
819	S-1	ISO8859-1	IS	354	is	SCO	Iceland
819	S-1	ISO8859-1	IS	354	is_IS	SCO	Iceland
819	S-1	ISO8859-1	IS	354	-	Sun	Iceland
819	S-1	ISO-8859-1	IS	354	is_IS	Linux	Iceland
1252	S-1	1252	IS	354	-	WIN	Iceland
1275	S-1	1275	IS	354	-	Mac	Iceland
871	S-1	IBM-871	IS	354	-	HOST	Iceland
1149	S-1	IBM-1149	IS	354	-	HOST	Iceland
437	S-1	IBM-437	IE	353	-	OS2	Ireland
850	S-1	IBM-850	IE	353	-	OS2	Ireland
819	S-1	ISO8859-1	IE	353	en_IE	AIX	Ireland
850	S-1	IBM-850	IE	353	En_IE	AIX	Ireland
819	S-1	iso88591	IE	353	-	HP	Ireland
1051	S-1	roman8	IE	353	-	HP	Ireland
819	S-1	ISO8859-1	IE	353	en_IE	Sun	Ireland
819	S-1	ISO8859-1	IE	353	en_IE.ISO8859-1	SCO	Ireland
819	S-1	ISO-8859-1	IE	353	en_IE	Linux	Ireland
1252	S-1	1252	IE	353	-	WIN	Ireland
1275	S-1	1275	IE	353	-	Mac	Ireland
285	S-1	IBM-285	IE	353	-	HOST	Ireland
1146	S-1	IBM-1146	IE	353	-	HOST	Ireland
806	S-12	IBM-806	IN	91	hi_IN	-	India
1137	S-12	IBM-1137	IN	91	-	HOST	India
862	S-8	IBM-862	IL	972	-	OS2	Israel
916	S-8	ISO8859-8	IL	972	iw_IL	AIX	Israel
916	S-8	ISO-8859-8	IL	972	iw_IL	Linux	Israel
1255	S-8	1255	IL	972	-	WIN	Israel
424	S-8	IBM-424	IL	972	-	HOST	Israel

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
437	S-1	IBM-437	IT	39	-	OS2	Italy
850	S-1	IBM-850	IT	39	-	OS2	Italy
819	S-1	ISO8859-1	IT	39	it_IT	AIX	Italy
850	S-1	IBM-850	IT	39	It_IT	AIX	Italy
819	S-1	iso88591	IT	39	it_IT.iso88591	HP	Italy
1051	S-1	roman8	IT	39	it_IT.roman8	HP	Italy
819	S-1	ISO8859-1	IT	39	it	SCO	Italy
819	S-1	ISO8859-1	IT	39	it_IT	SCO	Italy
819	S-1	ISO8859-1	IT	39	it	Sun	Italy
819	S-1	ISO-8859-1	IT	39	it_IT	Linux	Italy
1252	S-1	1252	IT	39	-	WIN	Italy
1275	S-1	1275	IT	39	-	Mac	Italy
280	S-1	IBM-280	IT	39	-	HOST	Italy
1144	S-1	IBM-1144	IT	39	-	HOST	Italy
932	D-1	IBM-932	JP	81	-	OS2	Japan
942	D-1	IBM-942	JP	81	-	OS2	Japan
943	D-1	IBM-943	JP	81	-	OS2	Japan
954	D-1	IBM-eucJP	JP	81	ja_JP	AIX	Japan
932	D-1	IBM-932	JP	81	Ja_JP	AIX	Japan
954	D-1	eucJP	JP	81	ja_JP.eucJP	HP	Japan
5039	D-1	SJIS	JP	81	ja_JP.SJIS	HP	Japan
954	D-1	eucJP	JP	81	ja	SCO	Japan
954	D-1	eucJP	JP	81	ja_JP	SCO	Japan
954	D-1	eucJP	JP	81	ja_JP.EUC	SCO	Japan
954	D-1	eucJP	JP	81	ja_JP.eucJP	SCO	Japan
954	D-1	eucJP	JP	81	ja	Sun	Japan
954	D-1	EUC-JP	JP	81	ja_JP	Linux	Japan
943	D-1	IBM-943	JP	81	-	WIN	Japan
930	D-1	IBM-930	JP	81	-	HOST	Japan
939	D-1	IBM-939	JP	81	-	HOST	Japan
5026	D-1	IBM-5026	JP	81	-	HOST	Japan
5035	D-1	IBM-5035	JP	81	-	HOST	Japan
1390	D-1		JP	81	-	HOST	Japan
1399	D-1		JP	81	-	HOST	Japan
949	D-3	IBM-949	KR	82	-	OS2	Korea, South
970	D-3	IBM-eucKR	KR	82	ko_KR	AIX	Korea, South
970	D-3	eucKR	KR	82	ko_KR.eucKR	HP	Korea, South
970	D-3	eucKR	KR	82	ko_KR.eucKR	SGI	Korea, South
970	D-3	5601	KR	82	ko	Sun	Korea, South
1363	D-3	1363	KR	82	-	WIN	Korea, South
933	D-3	IBM-933	KR	82	-	HOST	Korea, South
1364	D-3	IBM-1364	KR	82	-	HOST	Korea, South

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
437	S-1	IBM-437	Lat	3	-	OS2	Latin America
850	S-1	IBM-850	Lat	3	-	OS2	Latin America
819	S-1	ISO8859-1	Lat	3	-	AIX	Latin America
850	S-1	IBM-850	Lat	3	-	AIX	Latin America
819	S-1	iso88591	Lat	3	-	HP	Latin America
819	S-1	ISO8859-1	Lat	3	-	Sun	Latin America
819	S-1	ISO-8859-1	Lat	3	-	Linux	Latin America
1051	S-1	roman8	Lat	3	-	HP	Latin America
1252	S-1	1252	Lat	3	-	WIN	Latin America
1275	S-1	1275	Lat	3	-	Mac	Latin America
284	S-1	IBM-284	Lat	3	-	HOST	Latin America
1145	S-1	IBM-1145	Lat	3	-	HOST	Latin America
921	S-10	IBM-921	LV	371	-	OS2	Latvia
921	S-10	IBM-921	LV	371	Lv_LV	AIX	Latvia
921	S-10	IBM-921	LV	371	-	WIN	Latvia
1112	S-10	IBM-1112	LV	371	-	HOST	Latvia
921	S-10	IBM-921	LT	370	-	OS2	Lithuania
921	S-10	IBM-921	LT	370	Lt_LT	AIX	Lithuania
921	S-10	IBM-921	LV	370	-	WIN	Lithuania
1112	S-10	IBM-1112	LV	370	-	HOST	Lithuania
437	S-1	IBM-437	NL	31	-	OS2	Netherlands
850	S-1	IBM-850	NL	31	-	OS2	Netherlands
819	S-1	ISO8859-1	NL	31	nl_NL	AIX	Netherlands
850	S-1	IBM-850	NL	31	Nl_NL	AIX	Netherlands
819	S-1	iso88591	NL	31	nl_NL.iso88591	HP	Netherlands
1051	S-1	roman8	NL	31	nl_NL.roman8	HP	Netherlands
819	S-1	ISO8859-1	NL	31	nl	SCO	Netherlands
819	S-1	ISO8859-1	NL	31	nl_NL	SCO	Netherlands
819	S-1	ISO8859-1	NL	31	nl	Sun	Netherlands
819	S-1	ISO-8859-1	NL	31	nl_NL	Linux	Netherlands
1252	S-1	1252	NL	31	-	WIN	Netherlands
1275	S-1	1275	NL	31	-	Mac	Netherlands
37	S-1	IBM-37	NL	31	-	HOST	Netherlands
1140	S-1	IBM-1140	NL	31	-	HOST	Netherlands
850	S-1	IBM-850	NZ	64	-	OS2	New Zealand
850	S-1	IBM-850	NZ	64	En_NZ	AIX	New Zealand
819	S-1	ISO8859-1	NZ	64	en_NZ	AIX	New Zealand
819	S-1	ISO8859-1	NZ	64	-	HP	New Zealand
819	S-1	ISO8859-1	NZ	64	en_NZ	SCO	New Zealand
819	S-1	ISO8859-1	NZ	64	en_NZ	Sun	New Zealand
1252	S-1	1252	NZ	64	-	WIN	New Zealand
37	S-1	IBM-37	NZ	64	-	HOST	New Zealand
1140	S-1	IBM-1140	NZ	64	-	HOST	New Zealand

Table 32. Supported Languages and Code Sets (continued)

Code Page	Group	Code-Set	Tr.	Country Code	Locale	OS	Country Name
850	S-1	IBM-850	NO	47	-	OS2	Norway
819	S-1	ISO8859-1	NO	47	no_NO	AIX	Norway
850	S-1	IBM-850	NO	47	No_NO	AIX	Norway
819	S-1	iso88591	NO	47	no_NO.iso88591	HP	Norway
1051	S-1	roman8	NO	47	no_NO.roman8	HP	Norway
819	S-1	ISO8859-1	NO	47	no	SCO	Norway
819	S-1	ISO8859-1	NO	47	no_NO	SCO	Norway
819	S-1	ISO8859-1	NO	47	no	Sun	Norway
819	S-1	ISO-8859-1	NO	47	no_NO	Linux	Norway
1252	S-1	1252	NO	47	-	WIN	Norway
1275	S-1	1275	NO	47	-	Mac	Norway
277	S-1	IBM-277	NO	47	-	HOST	Norway
1142	S-1	IBM-1142	NO	47	-	HOST	Norway
852	S-2	IBM-852	PL	48	-	OS2	Poland
912	S-2	ISO8859-2	PL	48	pl_PL	AIX	Poland
912	S-2	iso88592	PL	48	pl_PL.iso88592	HP	Poland
912	S-2	ISO8859-2	PL	48	pl_PL.ISO8859-2	SCO	Poland
912	S-2	ISO-8859-2	PL	48	pl_PL	Linux	Poland
1250	S-2	1250	PL	48	-	WIN	Poland
1282	S-2	1282	PL	48	-	Mac	Poland
870	S-2	IBM-870	PL	48	-	HOST	Poland
860	S-1	IBM-860	PT	351	-	OS2	Portugal
850	S-1	IBM-850	PT	351	-	OS2	Portugal
819	S-1	ISO8859-1	PT	351	pt_PT	AIX	Portugal
850	S-1	IBM-850	PT	351	Pt_PT	AIX	Portugal
819	S-1	iso88591	PT	351	pt_PT.iso88591	HP	Portugal
1051	S-1	roman8	PT	351	pt_PT.roman8	HP	Portugal
819	S-1	ISO8859-1	PT	351	pt	SCO	Portugal
819	S-1	ISO8859-1	PT	351	pt_PT	SCO	Portugal
819	S-1	ISO8859-1	PT	351	pt	Sun	Portugal
819	S-1	ISO-8859-1	PT	351	pt_PT	Linux	Portugal
1252	S-1	1252	PT	351	-	WIN	Portugal
1275	S-1	1275	PT	351	-	Mac	Portugal
37	S-1	IBM-37	PT	351	-	HOST	Portugal
1140	S-1	IBM-1140	PT	351	-	HOST	Portugal
852	S-2	IBM-852	RO	40	-	OS2	Romania
912	S-2	ISO8859-2	RO	40	ro_RO	AIX	Romania
912	S-2	iso88592	RO	40	ro_RO.iso88592	HP	Romania
912	S-2	ISO8859-2	RO	40	ro_RO.ISO8859-2	SCO	Romania
912	S-2	ISO-8859-2	RO	40	ro_RO	Linux	Romania
1250	S-2	1250	RO	40	-	WIN	Romania
1282	S-2	1282	RO	40	-	Mac	Romania
870	S-2	IBM-870	RO	40	-	HOST	Romania

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
866	S-5	IBM-866	RU	7	-	OS2	Russia
915	S-5	ISO8859-5	RU	7	-	OS2	Russia
915	S-5	ISO8859-5	RU	7	ru_RU	AIX	Russia
915	S-5	iso88595	RU	7	ru_RU.iso88595	HP	Russia
915	S-5	ISO8859-5	RU	7	ru_RU.ISO8859-5	SCO	Russia
915	S-5	ISO-8859-5	RU	7	ru_RU	Linux	Russia
1251	S-5	1251	RU	7	-	WIN	Russia
1283	S-5	1283	RU	7	-	Mac	Russia
1025	S-5	IBM-1025	RU	7	-	HOST	Russia
855	S-5	IBM-855	SP	381	-	OS2	Serbia/Montenegro
915	S-5	ISO8859-5	SP	381	-	OS2	Serbia/Montenegro
915	S-5	ISO8859-5	SP	381	sr_SP	AIX	Serbia/Montenegro
915	S-5	iso88595	SP	381	-	HP	Serbia/Montenegro
1251	S-5	1251	SP	381	-	WIN	Serbia/Montenegro
1283	S-5	1283	SP	381	-	Mac	Serbia/Montenegro
1025	S-5	IBM-1025	SP	381	-	HOST	Serbia/Montenegro
852	S-2	IBM-852	SK	422	-	OS2	Slovakia
912	S-2	ISO8859-2	SK	422	sk_SK	AIX	Slovakia
912	S-2	iso88592	SK	422	sk_SK.iso88592	HP	Slovakia
912	S-2	ISO8859-2	SK	422	sk_SK.ISO8859-2	SCO	Slovakia
1250	S-2	1250	SK	422	-	WIN	Slovakia
1282	S-2	1282	SK	422	-	Mac	Slovakia
870	S-2	IBM-870	SK	422	-	HOST	Slovakia
852	S-2	IBM-852	SI	386	-	OS2	Slovenia
912	S-2	ISO8859-2	SI	386	sl_SI	AIX	Slovenia
912	S-2	iso88592	SI	386	sl_SI.iso88592	HP	Slovenia
912	S-2	ISO8859-2	SI	386	sl_SI.ISO8859-2	SCO	Slovenia
912	S-2	ISO-8859-2	SI	386	sl_SI	Linux	Slovenia
1250	S-2	1250	SI	386	-	WIN	Slovenia
1282	S-2	1282	SI	386	-	Mac	Slovenia
870	S-2	IBM-870	SI	386	-	HOST	Slovenia
437	S-1	IBM-437	ZA	27	-	OS2	South Africa
850	S-1	IBM-850	ZA	27	-	OS2	South Africa
819	S-1	ISO8859-1	ZA	27	en_ZA	AIX	South Africa
850	S-1	IBM-850	ZA	27	En_ZA	AIX	South Africa
819	S-1	iso88591	ZA	27	-	HP	South Africa
1051	S-1	roman8	ZA	27	-	HP	South Africa
819	S-1	ISO8859-1	ZA	27	-	Sun	South Africa
819	S-1	ISO8859-1	ZA	27	en_ZA.ISO8859-1	SCO	South Africa
1252	S-1	1252	ZA	27	-	WIN	South Africa
1275	S-1	1275	ZA	27	-	Mac	South Africa
285	S-1	IBM-285	ZA	27	-	HOST	South Africa
1146	S-1	IBM-1146	ZA	27	-	HOST	South Africa

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
437	S-1	IBM-437	ES	34	-	OS2	Spain
850	S-1	IBM-850	ES	34	-	OS2	Spain
819	S-1	ISO8859-1	ES	34	es_ES	AIX	Spain
850	S-1	IBM-850	ES	34	Es_ES	AIX	Spain
819	S-1	iso88591	ES	34	es_ES.iso88591	HP	Spain
1051	S-1	roman8	ES	34	es_ES.roman8	HP	Spain
819	S-1	ISO8859-1	ES	34	es	Sun	Spain
819	S-1	ISO8859-1	ES	34	es	SCO	Spain
819	S-1	ISO8859-1	ES	34	es_ES	SCO	Spain
819	S-1	ISO-8859-1	ES	34	es_ES	Linux	Spain
1252	S-1	1252	ES	34	-	WIN	Spain
1275	S-1	1275	ES	34	-	Mac	Spain
284	S-1	IBM-284	ES	34	-	HOST	Spain
1145	S-1	IBM-1145	ES	34	-	HOST	Spain
437	S-1	IBM-437	SE	46	-	OS2	Sweden
850	S-1	IBM-850	SE	46	-	OS2	Sweden
819	S-1	ISO8859-1	SE	46	sv_SE	AIX	Sweden
850	S-1	IBM-850	SE	46	Sv_SE	AIX	Sweden
819	S-1	iso88591	SE	46	sv_SE.iso88591	HP	Sweden
1051	S-1	roman8	SE	46	sv_SE.roman8	HP	Sweden
819	S-1	ISO8859-1	SE	46	sv	SCO	Sweden
819	S-1	ISO8859-1	SE	46	sv_SE	SCO	Sweden
819	S-1	ISO8859-1	SE	46	sv	Sun	Sweden
819	S-1	ISO-8859-1	SE	46	sv_SE	Linux	Sweden
1252	S-1	1252	SE	46	-	WIN	Sweden
1275	S-1	1275	SE	46	-	Mac	Sweden
278	S-1	IBM-278	SE	46	-	HOST	Sweden
1143	S-1	IBM-1143	SE	46	-	HOST	Sweden
437	S-1	IBM-437	CH	41	-	OS2	Switzerland
850	S-1	IBM-850	CH	41	-	OS2	Switzerland
819	S-1	ISO8859-1	CH	41	de_CH	AIX	Switzerland
850	S-1	IBM-850	CH	41	De_CH	AIX	Switzerland
819	S-1	iso88591	CH	41	-	HP	Switzerland
1051	S-1	roman8	CH	41	-	HP	Switzerland
819	S-1	ISO8859-1	CH	41	de_CH	SCO	Switzerland
819	S-1	ISO8859-1	CH	41	fr_CH	SCO	Switzerland
819	S-1	ISO8859-1	CH	41	it_CH	SCO	Switzerland
819	S-1	ISO8859-1	CH	41	de_CH	Sun	Switzerland
819	S-1	ISO-8859-1	CH	41	de_CH	Linux	Switzerland
1252	S-1	1252	CH	41	-	WIN	Switzerland
1275	S-1	1275	CH	41	-	Mac	Switzerland
500	S-1	IBM-500	CH	41	-	HOST	Switzerland
1148	S-1	IBM-1148	CH	41	-	HOST	Switzerland

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
938	D-2	IBM-938	TW	88	-	OS2	Taiwan
948	D-2	IBM-948	TW	88	-	OS2	Taiwan
950	D-2	big5	TW	88	-	OS2	Taiwan
950	D-2	big5	TW	88	Zh_TW	AIX	Taiwan
964	D-2	IBM-eucTW	TW	88	zh_TW	AIX	Taiwan
950	D-2	big5	TW	88	zh_TW.big5	HP	Taiwan
964	D-2	eucTW	TW	88	zh_TW.eucTW	HP	Taiwan
950	D-2	big5	TW	88	big5	Sun	Taiwan
964	D-2	cns11643	TW	88	zh_TW	Sun	Taiwan
950	D-2	big5	TW	88	-	WIN	Taiwan
937	D-2	IBM-937	TW	88	-	HOST	Taiwan
874	S-20	TIS620-1	TH	66	-	OS2	Thailand
874	S-20	TIS620-1	TH	66	Th_TH	AIX	Thailand
874	S-20	tis620	TH	66	th_TH.tis620	HP	Thailand
874	S-20	TIS620-1	TH	66	-	WIN	Thailand
838	S-20	IBM-838	TH	66	-	HOST	Thailand
857	S-9	IBM-857	TR	90	-	OS2	Turkey
920	S-9	ISO8859-9	TR	90	tr_TR	AIX	Turkey
920	S-9	iso88599	TR	90	tr_TR.iso88599	HP	Turkey
920	S-9	ISO8859-9	TR	90	tr_TR.ISO8859-9	SCO	Turkey
920	S-9	ISO-8859-9	TR	90	tr_TR	Linux	Turkey
1254	S-9	1254	TR	90	-	WIN	Turkey
1281	S-9	1281	TR	90	-	Mac	Turkey
1026	S-9	IBM-1026	TR	90	-	HOST	Turkey
437	S-1	IBM-437	GB	44	-	OS2	U.K.
850	S-1	IBM-850	GB	44	-	OS2	U.K.
819	S-1	ISO8859-1	GB	44	en_GB	AIX	U.K.
850	S-1	IBM-850	GB	44	En_GB	AIX	U.K.
819	S-1	iso88591	GB	44	en_GB.iso88591	HP	U.K.
1051	S-1	roman8	GB	44	en_GB.roman8	HP	U.K.
819	S-1	ISO8859-1	GB	44	en_UK	Sun	U.K.
819	S-1	ISO8859-1	GB	44	en_GB	SCO	U.K.
819	S-1	ISO8859-1	GB	44	en	SCO	U.K.
819	S-1	ISO-8859-1	GB	44	en_GB	Linux	U.K.
1252	S-1	1252	GB	44	-	WIN	U.K.
1275	S-1	1275	GB	44	-	Mac	U.K.
285	S-1	IBM-285	GB	44	-	HOST	U.K.
1146	S-1	IBM-1146	GB	44	-	HOST	U.K.
819	S-1	88591	GB	44	En_GB.88591	SINIX	U.K.
819	S-1	ISO8859-1	GB	44	En_GB.6937	SINIX	U.K.
1125	S-5	IBM-1125	UA	380	-	OS2	Ukraine
1124	S-5	IBM-1124	UA	380	uk_UA	AIX	Ukraine
1251	S-5	1251	UA	380	-	WIN	Ukraine
1123	S-5	IBM-1123	UA	380	-	HOST	Ukraine

Table 32. Supported Languages and Code Sets (continued)

Code				Country			
Page	Group	Code-Set	Tr.	Code	Locale	OS	Country Name
437	S-1	IBM-437	US	1	-	OS2	USA
850	S-1	IBM-850	US	1	-	OS2	USA
819	S-1	ISO8859-1	US	1	en_US	AIX	USA
850	S-1	IBM-850	US	1	En_US	AIX	USA
819	S-1	iso88591	US	1	en_US.iso88591	HP	USA
1051	S-1	roman8	US	1	en_US.roman8	HP	USA
819	S-1	ISO8859-1	US	1	en_US	Sun	USA
819	S-1	ISO8859-1	US	1	en_US	SGI	USA
819	S-1	ISO8859-1	US	1	en_US	SCO	USA
819	S-1	ISO-8859-1	US	1	en_US	Linux	USA
1252	S-1	1252	US	1	-	WIN	USA
1275	S-1	1275	US	1	-	Mac	USA
37	S-1	IBM-37	US	1	-	HOST	USA
1140	S-1	IBM-1140	US	1	-	HOST	USA
1163	S-11	IBM-1163	VN	84	-	OS2	Vietnam
1163	S-11	IBM-1163	VN	84	vi_VN	AIX	Vietnam
1258	S-11	1258	VN	84	-	WIN	Vietnam
1164	S-11	IBM-1164	VN	84	-	HOST	Vietnam

```
The following map to Arabic Countries (AA):
-------------------------------------------
    /* Arabic (Saudi Arabia) */
    /* Arabic (Iraq) */
    /* Arabic (Egypt) */
    /* Arabic (Libya) */
    /* Arabic (Algeria) */
    /* Arabic (Morocco) */
    /* Arabic (Tunisia) */
    /* Arabic (Oman) */
    /* Arabic (Yemen) */
    /* Arabic (Syria) */
    /* Arabic (Jordan) */
    /* Arabic (Lebanon) */
    /* Arabic (Kuwait) */
    /* Arabic (United Arab Emirates) */
    /* Arabic (Bahrain) */
    /* Arabic (Qatar) */
```

```
The following map to English (US):
----------------------------------
    /* English (Jamaica) */
    /* English (Carribean) */
```

Table 32. Supported Languages and Code Sets (continued)

```
Code                         Country
Page   Group  Code-Set  Tr.  Code Locale           OS        Country Name
----   -----  --------  --    ---  -----          ----      --------------
```

```
The following map to Latin America (Lat):
----------------------------------------
    /* Spanish (Mexican) */
    /* Spanish (Guatemala) */
    /* Spanish (Costa Rica) */
    /* Spanish (Panama) */
    /* Spanish (Dominican Republic) */
    /* Spanish (Venezuela) */
    /* Spanish (Colombia) */
    /* Spanish (Peru) */
    /* Spanish (Argentina) */
    /* Spanish (Ecuador) */
    /* Spanish (Chile) */
    /* Spanish (Uruguay) */
    /* Spanish (Paraguay) */
    /* Spanish (Bolivia) */
```

Note: The Solaris code page 950 does not support the following characters in IBM 950:

Code Range	Description	Sun Big-5	IBM Big-5
C6A1-C8FE	Symbols	Reserved area	Symbols
F9D6-F9FE	ETen extension	Reserved area	ETen extension
F286-F9A0	IBM selected chars	Reserved area	IBM selected

Note: Euro-symbol support is provided with this version of DB2 UDB. Microsoft Windows ANSI code pages are modified according to the latest definition from Microsoft to include the Euro-symbol in position 0x80. This position was previously undefined. In addition, the definition of code page 850 has changed to replace the character Dotless i (found at position 0xD5) with the Euro-symbol. DB2 UDB uses the new definitions of these code pages as the default to provide Euro-symbol support. This implementation is the appropriate default for current DB2 UDB customers who require Euro-symbol support, and should not impact other customers. However, if you want to continue to use the previous definition of these code pages, you can copy the following files:

- 12520850.cnv
- 08501252.cnv

- IBM00850.ucs
- IBM01252.ucs

from this directory

```
sqllib/conv/alt/
```

to this directory

```
sqllib/conv/
```

after installation is complete. You may want to back up the existing `IBM01252.usc` and `IBM00850.ucs` files before copying the non-Euro versions over them. After copying the files, you will not have Euro currency symbol support from DB2 UDB.

Deriving Code Page Values

The *application code page* is derived from the active environment when the database connection is made. If the DB2CODEPAGE registry variable is set, its value is taken as the application code page. However, it is not necessary to set the DB2CODEPAGE registry variable, because DB2 will determine the appropriate code page value from the operating system. Setting the DB2CODEPAGE registry variable to incorrect values may cause unpredictable results.

The *database code page* is derived from the value specified (explicitly or by default) at the time the database is created. For example, the following defines how the *active environment* is determined in different operating environments:

UNIX
On UNIX based operating systems, the active environment is determined from the locale setting, which includes information about language, territory and code set.

OS/2
On OS/2, primary and secondary code pages are specified in the `CONFIG.SYS` file. You can use the **chcp** command to display and dynamically change code pages within a given session.

Macintosh
For the Macintosh operating system, if the DB2CODEPAGE registry variable is not set, the Macintosh code page is derived from the Regional version code from the installed script.

Windows
For the Windows operating system, if the DB2CODEPAGE registry variable is not set, the Windows code page is derived from the

country ID, as specified in the iCountry value in the [intl] section of the Windows WIN.INI file.

Windows 32-bit operating systems
For all Windows 32-bit operating systems, if the DB2CODEPAGE registry variable is not set, the code page is derived from the ANSI code page setting in the Registry.

For a complete list of environment mappings for code page values, see Table 32 on page 384.

Character Sets

The database manager does not, in general, restrict the character set available to an application. For a detailed explanation of multi-byte character sets (MBCS) supported by DB2, refer to the *Application Development Guide*.

Character Set for Identifiers

The basic character set that can be used in database names consists of the single-byte uppercase and lowercase Latin letters (A...Z, a...z), the Arabic numerals (0...9) and the underscore character (_). This list is augmented with three special characters (#, @, and $) to provide compatibility with host database products. However, these special characters should be used with care in an NLS environment, because they are not included in the NLS host (EBCDIC) invariant character set.

When naming database objects (such as tables and views), program labels, host variables, cursors, and elements from the extended character set (for example, letters with diacritical marks) can also be used. Precisely which characters are available depends on the code page in use. If you are using the database in a multiple code page environment, you must ensure that all code pages support any elements from the extended character set that you plan to use. For information about delimited identifiers that have characters outside of the extended character set, but which can be used in SQL statements, refer to the *SQL Reference*.

Extended Character Set Definition for DBCS Identifiers

In DBCS environments, the extended character set consists of all the characters in the basic character set, plus the following:

- All double-byte characters in each DBCS code page, except the double-byte space, are valid letters.
- The double-byte space is a special character.
- The single-byte characters available in each mixed code page are assigned to various categories as follows:

Category
  Valid Code Points within each Mixed Code Page

Digits x30-39

Letters
  x23-24, x40-5A, x61-7A, xA6-DF (A6-DF for code pages 932 and 942 only)

Special Characters
  All other valid single-byte character code points

Coding SQL Statements

The coding of SQL statements is not language dependent. SQL keywords can be typed in uppercase, lowercase, or mixed case. The names of database objects and host variables, as well as program labels in an SQL statement cannot contain characters that are outside of the extended character set as described above.

Bidirectional CCSID Support

The following BiDi attributes are required for correct handling of bidirectional data on different platforms:

```
- Text type (LOGICAL or VISUAL)
- Shaping (SHAPED or UNSHAPED)
- Orientation (RIGHT-TO-LEFT or LEFT-TO-RIGHT)
- Numeral shape (ARABIC or HINDI)
- Symmetric swapping (YES or NO)
```

Because default values on different platforms are not the same, problems can occur when DB2 data is moved from one platform to another. For example, the Windows operating system uses LOGICAL UNSHAPED data, while OS/390 usually uses SHAPED VISUAL data. Therefore, without support for bidirectional attributes, data sent from DB2 Universal Database for OS/390 to DB2 UDB on Windows 32-bit operating systems may display incorrectly.

Bidirectional-specific CCSIDs

DB2 supports bidirectional data attributes through special bidirectional Coded Character Set Identifiers (CCSIDs). The following bidirectional CCSIDs have been defined and are implemented with DB2 UDB:

CCSID	CCSID	Code	String
(dec)	(hex)	Page	Type
00420	x'01A4'	420	4
00424	x'01A8'	424	4
08612	x'21A4'	420	5
08616	x'21A8'	424	10
00856	x'0358'	856	5
00862	x'035E'	862	4
00864	x'0360'	864	5

```
00916   x'0394'   916      5
01046   x'0416'  1046      5
01089   x'0441'  1089      5
01255   x'04E7'  1255      5
01256   x'04E8'  1256      5

62208   x'F300'   856      4
62209   x'F301'   862     10
62210   x'F302'   916      4
62211   x'F303'   424      5
62213   x'F305'   862      5
62215   x'F307'  1255      4
62218   x'F30A'   864      4
62220   x'F30C'   856      6
62221   x'F30D'   862      6
62222   x'F30E'   916      6
62223   x'F30F'  1255      6
62224   x'F310'   420      6
62225   x'F311'   864      6
62226   x'F312'  1046      6
62227   x'F313'  1089      6
62228   x'F314'  1256      6
62229   x'F315'   424      8
62230   x'F316'   856      8
62231   x'F317'   862      8
62232   x'F318'   916      8
62233   x'F319'   420      8
62234   x'F31A'   420      9
62235   x'F31B'   424      6
62236   x'F31C'   856     10
62237   x'F31D'  1255      8
62238   x'F31E'   916     10
62239   x'F31F'  1255     10
62240   x'F320'   424     11
62241   x'F321'   856     11
62242   x'F322'   862     11
62243   x'F323'   916     11
62244   x'F324'  1255     11

62245   x'F325'   424     10
62246   x'F326'  1046      8
62247   x'F327'  1046      9
62248   x'F328'  1046      4
62249   x'F329'  1046     12
62250   x'F32A'   420     12
```

where CDRA string types are defined as:

```
 String | Text  | Numerical  | Orientation | Shaping   | Symmetrical
  Type  | Type  |   Shape    |             |           |  Swapping
--------+-------+------------+-------------+-----------+-------------
   4     Visual   Passthru        LTR       Shaped         OFF
   5     Implicit  Arabic         LTR       Unshaped       ON
   6     Implicit  Arabic         RTL       Unshaped       ON
   7(*)  Visual    Arabic    Contextual(*)  Unshaped-Lig   OFF
```

8	Visual	Arabic	RTL	Shaped	OFF
9	Visual	Passthru	RTL	Shaped	ON
10	Implicit	Passthru	Contextual-L	Unshaped	ON
11	Implicit	Passthru	Contextual-R	Unshaped	ON
12	Implicit	Arabic	RTL	Shaped	ON

Note: (*) Field orientation is left-to-right (LTR) when the first alphabetic character is a Latin character, and right-to-left (RTL) when it is a bidirectional (RTL) character. Characters are unshaped, but LamAlef ligatures are kept, and are not broken into constituents.

DB2 Universal Database Implementation of Bidirectional Support

Bidirectional layout transformations are implemented in DB2 Universal Database using the new CCSID definitions. For the new BiDi-specific CCSIDs, layout transformations are performed instead of, or in addition to, code page conversions. To use this support, the DB2BIDI registry variable must be set to YES. By default, this variable is not set. It is used by the server for all conversions, and can only be set when the server is started. Setting DB2BIDI to YES may have some performance impact because of additional checking and layout transformations.

To specify a particular bidirectional CCSID in a non-DRDA environment, select the CCSID (from the above table) that matches the characteristics of your client, and set DB2CODEPAGE to that value. If you already have a connection to the database, you must issue a TERMINATE command, and then reconnect to allow the new setting for DB2CODEPAGE to take effect. If you select a CCSID that is not appropriate for the code page or string type of your client platform, you may get unexpected results. If you select an incompatible CCSID (for example, the Hebrew CCSID for connection to an Arabic database), or if DB2BIDI has not been set for the server, you will receive an error message when you try to connect.

For DRDA environments, if the HOST EBCDIC platform also supports these bidirectional CCSIDs, you only need to set the DB2CODEPAGE value. However, if the HOST platform does not support these CCSIDs, you must also specify a CCSID override for the HOST database server to which you are connecting. This is necessary because, in a DRDA environment, code page conversions and layout transformations are performed by the receiver of data. However, if the HOST server does not support these bidirectional CCSIDs, it does not perform layout transformation on the data that it receives from DB2 UDB. If you use a CCSID override, the DB2 UDB client performs layout transformation on the outbound data as well. For information about setting a CCSID override, refer to the *DB2 Connect User's Guide*.

CCSID override is not supported for cases where the HOST EBCDIC platform is the client, and DB2 UDB is the server.

DB2 Connect Implementation of Bidirectional Support

When data is exchanged between DB2 Connect and a database on the server, it is usually the receiver that performs conversion on the incoming data. The same convention would normally apply to bidirectional layout transformations, and is in addition to the usual code page conversion. DB2 Connect has the optional ability to perform bidirectional layout transformation on data it is about to send to the server database, in addition to data received from the server database.

In order for DB2 Connect to perform bidirectional layout transformation on outgoing data for a server database, the bidirectional CCSID of the server database must be overridden. This is accomplished through the use of the BIDI parameter in the PARMS field of the DCS database directory entry for the server database.

Note: If you want DB2 Connect to perform layout transformation on the data it is about to send to the DB2 host database, even though you do not have to override its CCSID, you must still add the BIDI parameter to the PARMS field of the DCS database directory. In this case, the CCSID that you should provide is the default DB2 host database CCSID.

The BIDI parameter is to be specified as the ninth parameter in the PARMS field, along with the bidirectional CCSID with which you want to override the default server database bidirectional CCSID:

```
",,,,,,,,BIDI=xyz"
```

where *xyz* is the CCSID override.

Note: The registry variable DB2BIDI must be set to YES for the BIDI parameter to take effect.

A list of the bidirectional CCSIDs that are supported, along with their string types, can be found in “Bidirectional-specific CCSIDs” on page 400.

The use of this feature is best described with an example.

Suppose you have a Hebrew DB2 client running CCSID 62213 (bidirectional string type 5), and you want to access a DB2 host database running CCSID 00424 (bidirectional string type 4). However, you know that the data contained in the DB2 host database is based on CCSID 08616 (bidirectional string type 6).

There are two problems here: The first is that the DB2 host database does not know the difference in the bidirectional string types with CCSIDs 00424 and 08616. The second problem is that the DB2 host database does not recognize the DB2 client CCSID (62213). It only supports CCSID 00862, which is based on the same code page as CCSID 62213.

You will need to ensure that data sent to the DB2 host database is in bidirectional string type 6 format to begin with, and also let DB2 Connect know that it has to perform bidirectional transformation on data it receives from the DB2 host database. You will need to use following catalog command for the DB2 host database:

```
db2 catalog dcs database nydb1 as telaviv parms ",,,,,,,,BIDI=08616"
```

This command tells DB2 Connect to override the DB2 host database CCSID of 00424 with 08616. This override includes the following processing:

1. DB2 Connect connects to the DB2 host database using CCSID 00862.
2. DB2 Connect performs bidirectional layout transformation on the data it is about to *send* to the DB2 host database. The transformation is from CCSID 62213 (bidirectional string type 5) to CCSID 62221 (bidirectional string type 6).
3. DB2 Connect performs bidirectional layout transformation on data it *receives* from the DB2 host database. This transformation is from CCSID 08616 (bidirectional string type 6) to CCSID 62213 (bidirectional string type 5).

Note: In some cases, use of a bidirectional CCSID may cause the SQL query itself to be modified in such a way that it is not recognized by the DB2 server. Specifically, you should avoid using IMPLICIT CONTEXTUAL and IMPLICIT RIGHT-TO-LEFT CCSIDs when a different string type can be used. CONTEXTUAL CCSIDs can produce unpredictable results if the SQL query contains quoted strings. Avoid using quoted strings in SQL statements; use host variables whenever possible.

If a specific bidirectional CCSID is causing problems that cannot be rectified by following these recommendations, set DB2BIDI to NO.

Collating Sequences

The database manager compares character data using a *collating sequence*. This is an ordering for a set of characters that determines whether a particular character sorts higher, lower, or the same as another. For example, a collating sequence can be used to indicate that lowercase and uppercase versions of a particular character are to be sorted equally.

The collating sequence is specified at database creation time, and cannot be modified later.

The database manager allows databases to be created with custom collating sequences, using the application programming interface (API). For information about implementing a custom collating sequence table, refer to the *Application Development Guide*.

Note: Character string data defined with the FOR BIT DATA attribute, and BLOB data, is sorted using the binary sort sequence.

General Concerns

Once a collating sequence is defined, all future character comparisons for that database will be performed with that collating sequence. Except for character data defined as FOR BIT DATA or BLOB data, the collating sequence will be used for all SQL comparisons and ORDER BY clauses, and also in setting up indexes and statistics. For more information about how the database collating sequence is used, see "String Comparisons" in the *SQL Reference*.

Potential problems can occur in the following cases:

- An application merges sorted data from a database with application data that was sorted using a different collating sequence.
- An application merges sorted data from one database with sorted data from another, but the databases have different collating sequences.
- An application makes assumptions about sorted data that are not true for the relevant collating sequence. For example, numbers collating lower than alphabetics may or may not be true for a particular collating sequence.

A final point to remember is that the results of any sort based on a direct comparison of character code points will only match query results that are ordered using an identity collating sequence.

Federated Database Concerns

Your choice of database collating sequence can affect federated system performance. If a data source uses the same collating sequence as the DB2 federated database, DB2 can push down order-dependent processing involving character data to the data source. If a data source collating sequence does not match the DB2 collating sequence, data is retrieved, and all order-dependent processing on character data is done locally (this can reduce performance).

To determine whether a data source and DB2 have the same collating sequence, consider the following:

- National language support.

  The collating sequence is related to the language supported on a server. Compare DB2 NLS information to data source NLS information.
- Data source characteristics.

Some data sources are created using case-insensitive collating sequences, which can yield results that are different from DB2 in order-dependent operations.

- Customization.

  Some data sources provide multiple options for collating sequences, or allow the collating sequence to be customized.

Choose the collating sequence for a DB2 federated database based on the mix of data sources that will be accessed from that database. For example:

- If a DB2 database will access mostly Oracle databases with the same code page (NLS) as DB2, specify the identity sequence at database creation time (Oracle databases use an equivalent collating sequence).
- If a DB2 database will access only DB2 UDB databases, ensure that you match collating sequence values.

For information about setting up an MVS collating sequence, refer to the *Administrative API Reference* for examples under the description of the **sqlecrea** - Create Database API. These examples contain collation tables for the EBCIDIC 500, 37, and 5026/5035 code pages.

After you set the collating sequence for the DB2 database, ensure that you set the *collating_sequence* server option for each data source server. This option specifies whether the collating sequence of a given data source server matches the collating sequence of the DB2 database.

Set the collating_sequence option to ″Y″ if the collating sequences match. This setting allows the DB2 optimizer to consider order-dependent processing at a data source, which can improve performance. However, if the data source collating sequence is not the same as the DB2 database collating sequence, you may receive incorrect results. For example, if your plan uses merge joins, the DB2 optimizer will push down ordering operations to the data sources as much as possible. If the data source collating sequence is not the same, the join result set may not be correct.

Set the collating_sequence option to ″N″ if the collating sequences do not match. Use this value when data source collating sequences differ from DB2, or when the data source collating operations might be case insensitive. For example, in a case-insensitive data source with an English code page, TOLLESON, ToLLeSoN, and tolleson would all be considered equal. Set the collating_sequence option to ″N″ if you are not sure whether the collating sequence at the data source is identical to the DB2 collating sequence.

Datetime Values

The datetime data types are described below. Although datetime values can be used in certain arithmetic and string operations, and are compatible with certain strings, they are neither strings nor numbers.

Date

A *date* is a three-part value (year, month, and day). The range of the year part is 0001 to 9999. The range of the month part is 1 to 12. The range of the day part is 1 to *x*, where *x* depends on the month.

The internal representation of a date is a string of 4 bytes. Each byte consists of 2 packed decimal digits. The first 2 bytes represent the year, the third byte the month, and the last byte the day.

The length of a DATE column, as described in the SQLDA, is 10 bytes, which is the appropriate length for a character string representation of the value.

Time

A *time* is a three-part value (hour, minute, and second) designating a time of day under a 24-hour clock. The range of the hour part is 0 to 24. The range of the other parts is 0 to 59. If the hour is 24, the minute and second specifications are zero.

The internal representation of a time is a string of 3 bytes. Each byte is 2 packed decimal digits. The first byte represents the hour, the second byte the minute, and the last byte the second.

The length of a TIME column, as described in the SQLDA, is 8 bytes, which is the appropriate length for a character string representation of the value.

Time Stamp

A *time stamp* is a seven-part value (year, month, day, hour, minute, second, and microsecond) designating a date and a time of day as defined above, except that the time includes the specification of microseconds.

The internal representation of a time stamp is a string of 10 bytes. Each byte is 2 packed decimal digits. The first 4 bytes represent the date, the next 3 bytes the time, and the last 3 bytes the microseconds.

The length of a TIMESTAMP column, as described in the SQLDA, is 26 bytes, which is the appropriate length for a character string representation of the value.

String Representations of Datetime Values

Values whose data types are DATE, TIME, or TIMESTAMP are represented in an internal form that is transparent to the SQL user. Dates, times, and time stamps can also, however, be represented by character strings, and these representations directly concern the SQL user, because there are no constants or variables whose data types are DATE, TIME, or TIMESTAMP. Thus, to be retrieved, a datetime value must be assigned to a character string variable. The character string representation is normally the default format of datetime values associated with the country code of the client, unless overridden by specification of the ″F″ format option when the program is precompiled or bound to the database. For a list of the string formats for the various country codes, see Table 35 on page 410.

When a valid string representation of a datetime value is used in an operation with an internal datetime value, the string representation is converted to the internal form of the date, time, or time stamp before the operation is performed. Valid string representations of datetime values are defined in the following sections.

Date Strings

A string representation of a date is a character string that starts with a digit and has a length of at least 8 characters. Trailing blanks may be included; leading zeros may be omitted from the month part and the day part of the date.

Valid string formats for dates are listed in Table 33. Each format is identified by name, and includes an associated abbreviation and an example of its use.

Table 33. Formats for String Representations of Dates

Format Name	Abbreviation	Date Format	Example
International Standards Organization	ISO	yyyy-mm-dd	1991-10-27
IBM USA standard	USA	mm/dd/yyyy	10/27/1991
IBM European standard	EUR	dd.mm.yyyy	27.10.1991
Japanese Industrial Standard Christian era	JIS	yyyy-mm-dd	1991-10-27
Site-defined (Local)	LOC	Depends on database country code	—

Time Strings

A string representation of a time is a character string that starts with a digit and has a length of at least 4 characters. Trailing blanks may be included; a leading zero may be omitted from the hour part of the time, and seconds may be omitted entirely. If you choose to omit seconds, an implicit specification of 0 seconds is assumed. Thus, `13.30` is equivalent to `13.30.00`.

Valid string formats for times are listed in Table 34 on page 409. Each format is identified by name, and includes an associated abbreviation and an example of its use.

Table 34. Formats for String Representations of Times

Format Name	Abbreviation	Time Format	Example
International Standards Organization	ISO	hh.mm.ss	13.30.05
IBM USA standard	USA	hh:mm AM or PM	1:30 PM
IBM European standard	EUR	hh.mm.ss	13.30.05
Japanese Industrial Standard Christian Era	JIS	hh:mm:ss	13:30:05
Site-defined (Local)	LOC	Depends on application country code	—

Notes:

1. In ISO, EUR, or JIS time string format, `.ss` (or `:ss`) is optional.
2. In USA time string format, the minutes specification can be omitted, indicating an implicit specification of 00 minutes. Thus, 1 PM is equivalent to 1:00 PM.
3. In USA time string format, the hours specification cannot be greater than 12, and cannot be 0, except in the special case of 00:00 AM. Using the ISO format of the 24-hour clock, the correspondence between the USA format and the 24-hour clock is as follows:
   - 12:01 AM through 12:59 AM corresponds to 00.01.00 through 00.59.00.
   - 01:00 AM through 11:59 AM corresponds to 01.00.00 through 11.59.00.
   - 12:00 PM (noon) through 11:59 PM corresponds to 12.00.00 through 23.59.00.
   - 12:00 AM (midnight) corresponds to 24.00.00, and 00:00 AM (midnight) corresponds to 00.00.00.

Time Stamp Strings

A string representation of a time stamp is a character string that starts with a digit and has a length of at least 16 characters. The complete string representation of a time stamp has the form *yyyy-mm-dd-hh.mm.ss.nnnnnn*. Trailing blanks may be included; leading zeros may be omitted from the month, day, or hour part of the time stamp, and microseconds may be truncated or omitted entirely. If you choose to omit any digit of the microseconds part, an implicit specification of 0 is assumed. Thus, `1991-3-2-8.30.00` is equivalent to `1991-03-02-08.30.00.000000`.

MBCS Considerations

Date and time stamp strings must contain only single-byte characters and digits.

Date and Time Formats: The character string representation of date and time formats is the default format of datetime values associated with the country code of the application. This default format can be overridden by specifying the ″F″ format option when the program is precompiled or bound to the database.

Following is a description of the input and output formats for date and time:

- Input Time Format
  - There is no default input time format
  - All time formats are allowed as input for all country codes.
- Output Time Format
  - The default output time format is equal to the local time format.
- Input Date Format
  - There is no default input date format
  - Where the local format for date conflicts with an ISO, JIS, EUR, or USA date format, the local format is recognized for date input. For example, see the UK entry in Table 35.
- Output Date Format
  - The default output date format is shown in Table 35.

    Note: Table 35 also shows a listing of the string formats for the various country codes.

Table 35. Date and Time Formats by Country Code

Country Code	Local Date Format	Local Time Format	Default Output Date Format	Input Date Formats
355 Albania	yyyy-mm-dd	JIS	LOC	LOC, USA, EUR, ISO
785 Arabic	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
001 Australia (1)	mm-dd-yyyy	JIS	LOC	LOC, USA, EUR, ISO
061 Australia	dd-mm-yyyy	JIS	LOC	LOC, USA, EUR, ISO
032 Belgium	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
055 Brazil	dd.mm.yyyy	JIS	LOC	LOC, EUR, ISO
359 Bulgaria	dd.mm.yyyy	JIS	EUR	LOC, USA, EUR, ISO
001 Canada	mm-dd-yyyy	JIS	USA	LOC, USA, EUR, ISO
002 Canada (French)	dd-mm-yyyy	ISO	ISO	LOC, USA, EUR, ISO

Table 35. Date and Time Formats by Country Code (continued)

Country Code	Local Date Format	Local Time Format	Default Output Date Format	Input Date Formats
385 Croatia	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
042 Czech Republic	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
045 Denmark	dd-mm-yyyy	ISO	ISO	LOC, USA, EUR, ISO
358 Finland	dd/mm/yyyy	ISO	EUR	LOC, EUR, ISO
389 FYR Macedonia	dd.mm.yyyy	JIS	EUR	LOC, USA, EUR, ISO
033 France	dd/mm/yyyy	JIS	EUR	LOC, EUR, ISO
049 Germany	dd/mm/yyyy	ISO	ISO	LOC, EUR, ISO
030 Greece	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
036 Hungary	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
354 Iceland	dd-mm-yyyy	JIS	LOC	LOC, USA, EUR, ISO
091 India	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
972 Israel	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
039 Italy	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
081 Japan	mm/dd/yyyy	JIS	ISO	LOC, USA, EUR, ISO
082 Korea	mm/dd/yyyy	JIS	ISO	LOC, USA, EUR, ISO
001 Latin America (1)	mm-dd-yyyy	JIS	LOC	LOC, USA, EUR, ISO
003 Latin America	dd-mm-yyyy	JIS	LOC	LOC, EUR, ISO
031 Netherlands	dd-mm-yyyy	JIS	LOC	LOC, USA, EUR, ISO
047 Norway	dd/mm/yyyy	ISO	EUR	LOC, EUR, ISO
048 Poland	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
351 Portugal	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
086 People's Republic of China	mm/dd/yyyy	JIS	ISO	LOC, USA, EUR, ISO

Table 35. Date and Time Formats by Country Code (continued)

Country Code	Local Date Format	Local Time Format	Default Output Date Format	Input Date Formats
040 Romania	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
007 Russia	dd/mm/yyyy	ISO	LOC	LOC, EUR, ISO
381 Serbia/ Montenegro	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
042 Slovakia	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
386 Slovenia	yyyy-mm-dd	JIS	ISO	LOC, USA, EUR, ISO
034 Spain	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
046 Sweden	dd/mm/yyyy	ISO	ISO	LOC, EUR, ISO
041 Switzerland	dd/mm/yyyy	ISO	EUR	LOC, EUR, ISO
088 Taiwan	mm-dd-yyyy	JIS	ISO	LOC, USA, EUR, ISO
066 Thailand (2)	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
090 Turkey	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
044 UK	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO
001 USA	mm-dd-yyyy	JIS	USA	LOC, USA, EUR, ISO
084 Vietnam	dd/mm/yyyy	JIS	LOC	LOC, EUR, ISO

Notes:

1. Countries using the default C locale are assigned country code 001.
2. yyyy in Buddhist era is equivalent to Gregorian + 543 years (Thailand only).

Unicode/UCS-2 and UTF-8 Support in DB2 UDB

These two standards are documented here.

Introduction

The Unicode character encoding standard is a fixed-length, character encoding scheme that includes characters from almost all of the living languages of the world. Unicode characters are usually shown as "U+*xxxx*", where *xxxx* is the hexadecimal code of the character.

Each character is 16 bits (2 bytes) wide, regardless of the language. While the resulting 65000 code elements are sufficient for encoding most of the characters of the major languages of the world, the Unicode standard also provides an extension mechanism that allows the encoding of as many as one million more characters. This extension reserves a range of code values (U+D800 to U+D8FF, known as ″surrogates″) for encoding some 32-bit characters as two successive code elements.

The International Standards Organization (ISO) and the International Electrotechnical Commission (IEC) standard 10646 (ISO/IEC 10646) specifies the Universal Multiple-Octet Coded Character Set (UCS) that has a 2-byte version (UCS-2) and a 4-byte version (UCS-4). The 2-byte version of this ISO standard is identical to Unicode without surrogates. ISO 10646 also defines an extension technique for encoding some UCS-4 codes in a UCS-2 encoded string. This extension, called UTF-16, is identical to Unicode with surrogates.

DB2 UDB supports UCS-2; that is, Unicode without surrogates.

Connection of a UTF-8 (code page 1208) client to a non-Unicode database is not supported.

UTF-8

With UCS-2 or Unicode encoding, ASCII and control characters are also two bytes long, and the lead byte is zero. For example, NULL is U+0000, and the uppercase ″A″ is represented by U+0041. This could be a major problem for ASCII-based applications and ASCII file systems, because in a UCS-2 string, extraneous NULLs can appear anywhere in the string. A transformation algorithm, known as UTF-8, can be used to circumvent this problem for programs that rely on ASCII code being invariant.

UTF-8 (UCS Transformation Format 8) is an algorithmic transformation that transforms fixed-length UCS-4 characters into variable-length byte strings. In UTF-8, ASCII characters are represented by their usual single-byte codes, but non-ASCII characters in UCS-2 become two or three bytes long. In other words, UTF-8 transforms UCS-2 characters into a multi-byte code set, for which ASCII is invariant. The number of bytes for each UCS-2 character in UTF-8 format can be determined from the following table:

```
UCS-2 (hex)     UTF-8 (binary)                 Description
------------    --------------------------     ----------------
0000 to 007F    0xxxxxxx                       ASCII
0080 to 07FF    110xxxxx 10xxxxxx              up to U+07FF
0800 to FFFF    1110xxxx 10xxxxxx 10xxxxxx     other UCS-2

NOTE: The range D800 to DFFF is to be excluded from treatment
      by the third row of this table which governs the UCS-4 range
      0000 0800 to 0000 FFFF.
```

In each of the above, a series of *x*’s is the UCS bit representation of the character. For example, U0080 transforms into 11000010 10000000.

UCS-2/UTF-8 Implementation in DB2 UDB

Code Page/CCSID Numbers

Within IBM, the UCS-2 code page has been registered as code page 1200. All code pages are defined with growing character sets; that is, when new characters are added to a code page, the code page number does not change. Code page 1200 always refers to the current version of Unicode/UCS-2, and has been used for UCS-2 support in DB2 UDB.

A specific version of the UCS standard, as defined by Unicode 2.0 and ISO/IEC 10646-1, has also been registered within IBM as CCSID 13488. This CCSID has been used internally by DB2 UDB for storing graphic string data in euc-Japan and euc-Taiwan databases. CCSID 13488 and code page 1200 both refer to UCS-2, and are handled the same way, except for the value of their ″double-byte″ (DBCS) space:

CP/CCSID	Single-byte (SBCS) space	Double-byte (DBCS) space
1200	N/A	U+0020
13488	N/A	U+3000

```
NOTE: In a UCS-2 database, U+3000 has no special meaning.
```

Regarding the conversion tables, since code page 1200 is a superset of CCSID 13488, the same (superset) tables are used for both.

Within IBM, UTF-8 has been registered as CCSID 1208 with growing character set (sometimes also referred to as code page 1208). As new characters are added to the standard, this number (1208) will not change. The number 1208 is used as the multi-byte code page number for DB2's UCS-2/UTF-8 support.

DB2 UDB supports UCS-2 as a new multi-byte code page. The MBCS code page number is 1208, which is the database code page number, and the code page of character string data within the database. The double-byte code page number for UCS-2 is 1200, which is the code page of graphic string data within the database. When a database is created in UCS-2/UTF-8, CHAR, VARCHAR, LONG VARCHAR, and CLOB data are stored in UTF-8, and GRAPHIC, VARGRAPHIC, LONG VARGRAPHIC, and DBCLOB data are stored in UCS-2. We will simply refer to this as a UCS-2 database.

Creating a UCS-2 Database

By default, databases are created in the code page of the application creating them. Therefore, if you create your database from a UTF-8 client (for example, the UNIVERSAL locale of AIX), or if the DB2CODEPAGE registry variable on the client is set to 1208, your database will be created as a UCS-2 database. Alternatively, you can explicitly specify ″UTF-8″ as the CODESET name, and use any valid two letter TERRITORY code supported by DB2 UDB.

For example, to create a UCS-2 database with the territory code for the United States, issue:

```
   DB2 CREATE DATABASE dbname USING CODESET UTF-8 TERRITORY US
```

To create a UCS-2 database using the **sqlecrea** API, you should set the values in *sqledbcountryinfo* accordingly. For example, set SQLDBCODESET to `UTF-8`, and SQLDBLOCALE to any valid territory code (for example, `US`).

The default collating sequence for a UCS-2 database is IDENTITY, which provides UCS-2 code point order. Therefore, by default, all UCS-2/UTF-8 characters are ordered and compared according to their UCS-2 code point sequence.

All culturally-sensitive parameters, such as date or time format, decimal separator, and others, are based on the current territory of the client.

A UCS-2 database allows connection from every single-byte and multi-byte code page supported by DB2 UDB. Code page character conversions between the client's code page and UTF-8 are automatically performed by the database manager. Data in graphic string types is always in UCS-2, and does not go through code page conversions. The command line processor (CLP) environment is an exception. If you select graphic string (UCS-2) data from the CLP, the returned graphic string data is converted (by the CLP) from UCS-2 to the code page of your client environment.

Every client is limited by the character repertoire, the input method, and the fonts supported by its environment, but the UCS-2 database itself accepts and stores all UCS-2 characters. Therefore, every client usually works with a subset of UCS-2 characters, but the database manager allows the entire repertoire of UCS-2 characters.

When characters are converted from a local code page to UTF-8, there may be expansion in the number of bytes. There is no expansion for ASCII characters, but other UCS-2 characters expand by a factor of two or three. The number of bytes of each UCS-2 character in UTF-8 format can be determined from the table in "UTF-8" on page 413.

Data Types

All data types supported by DB2 UDB are also supported in a UCS-2 database. In particular, graphic string data is supported for a UCS-2 database, and is stored in UCS-2/Unicode. Every client, including SBCS clients, can work with graphic string data types in UCS-2/Unicode when connected to a UCS-2 database.

A UCS-2 database is like any MBCS database where character string data is measured in number of bytes. When working with character string data in

UTF-8, one should not assume that each character is one byte. In multi-byte UTF-8 encoding, each ASCII character is one byte, but non-ASCII characters take two or three bytes each. This should be taken into account when defining CHAR fields. Depending on the ratio of ASCII to non-ASCII characters, a CHAR field of size *n* bytes can contain anywhere from *n*/3 to *n* characters.

Using character string UTF-8 encoding versus the graphic string UCS-2 data type also has an impact on the total storage requirements. In a situation where the majority of characters are ASCII, with some non-ASCII characters in between, storing UTF-8 data may be a better alternative, because the storage requirements are closer to one byte per character. On the other hand, in situations where the majority of characters are non-ASCII characters that expand to three-byte UTF-8 sequences (for example ideographic characters), the UCS-2 graphic-string format may be a better alternative, because every UCS-2 character requires exactly two bytes, rather than three bytes, for each corresponding character in the UTF-8 format.

In MBCS environments, SQL scalar functions that operate on character strings, such as LENGTH, SUBSTR, POSSTR, MAX, MIN, and the like, operate on the number of ″bytes″ rather than number of ″characters″. The behavior is the same in a UCS-2 database, but you should take extra care when specifying offsets and lengths for a USC-2 database, because these values are always defined in the context of the database code page. That is, in the case of a UCS-2 database, these offsets should be defined in UTF-8. Since some single-byte characters require more than one byte in UTF-8, SUBSTR indexes that are valid for a single-byte database may not be valid for a UCS-2 database. If you specify incorrect indexes, SQLCODE -191 (SQLSTATE 22504) is returned. For a description of the behavior of these functions, refer to the *SQL Reference*.

SQL CHAR data types are supported (in the C language) by the `char` data type in user programs. SQL GRAPHIC data types are supported by `sqldbchar` in user programs. Note that, for a UCS-2 database, `sqldbchar` data is always in big-endian (high byte first) format. When an application program is connected to a UCS-2 database, character string data is converted between the application code page and UTF-8 by DB2 UDB, but graphic string data is always in UCS-2.

Identifiers

In a UCS-2 database, all identifiers are in multi-byte UTF-8. Therefore, it is possible to use any UCS-2 character in identifiers where the use of a character in the extended character set (for example, an accented character, or a multi-byte character) is allowed by DB2 UDB. For details about which identifiers allow the use of extended characters, see “Appendix B. Naming Rules” on page 349.

Clients can enter any character that is supported by their SBCS or MBCS environment, and all the characters in the identifiers will be converted to UTF-8 by the database manager. Two points must be taken into account when specifying national language characters in identifiers for a UCS-2 database:

- Each non-ASCII character requires two or three bytes. Therefore, an *n*-byte identifier can only hold somewhere between *n*/3 and *n* characters, depending on the ratio of ASCII to non-ASCII characters. If you have only one or two non-ASCII (for example, accented) characters, the limit is closer to *n* characters, while for an identifier that is completely non-ASCII (for example, in Japanese), only *n*/3 characters can be used.
- If identifiers are to be entered from different client environments, they should be defined using the common subset of characters available to those clients. For example, if a UCS-2 database is to be accessed from Latin-1, Arabic, and Japanese environments, all identifiers should realistically be limited to ASCII.

UCS-2 Literals

UCS-2 literals can be specified in two ways:

- As a graphic string constant, using the G'...' or N'....' format described in the ″Graphic String Constants″ section of the ″Language Elements″ chapter in the *SQL Reference*. Any literal specified in this way will be converted by the database manager from the application code page to UCS-2.
- As a UCS-2 hexadecimal string, using the UX'....' or GX'....' format. The constant specified between the quotation marks after UX or GX must be a multiple of 4 hexadecimal digits. Each four-digit group represents one UCS-2 code point.

When using the command line processor (CLP), the first method is easier if the UCS-2 character exists in the local application code page (for example, when entering any code page 850 character from a terminal that is using code page 850). The second method should be used for characters that are outside of the application code page repertoire (for example, when specifying Japanese characters from a terminal that is using code page 850).

Pattern Matching in a UCS-2 Database

Pattern matching is one area where the behavior of existing MBCS databases is slightly different from the behavior of a UCS-2 database.

For MBCS databases in DB2 UDB, the current behavior is as follows: If the match-expression contains MBCS data, the pattern can include both SBCS and MBCS characters. The special characters in the pattern are interpreted as follows:

- An SBCS underscore refers to one SBCS character.
- A DBCS underscore refers to one MBCS character.

- A percent (either SBCS or DBCS) refers to a string of zero or more SBCS or MBCS characters.

If the match-expression contains graphic string DBCS data, the expressions contain only DBCS characters. The special characters in the pattern are interpreted as follows:

- A DBCS underscore refers to one DBCS character.
- A DBCS percent sign refers to a string of zero or more DBCS characters.

In a UCS-2 database, there is really no distinction between ″single-byte″ and ″double-byte″ characters; every UCS-2 character occupies two bytes. Although the UTF-8 format is a ″mixed-byte″ encoding of UCS-2 characters, there is no real distinction between SBCS and MBCS characters in UTF-8. Every character is a UCS-2 character, regardless of the number of its bytes that are in UTF-8 format. When specifying a character string, or a graphic string expression, an underscore refers to one UCS-2 character, and a percent sign refers to a string of zero or more UCS-2 characters.

On the client side, the character string expressions are in the code page of the client, and will be converted to UTF-8 by the database manager. SBCS client code pages do not have a DBCS percent sign or a DBCS underscore, but every supported code page contains a single-byte percent sign (corresponding to U+0025) and a single-byte underscore (corresponding to U+005F). The interpretation of special characters for a UCS-2 database is as follows:

- An SBCS underscore (corresponding to U+0025) refers to one UCS-2 character in a graphic string expression, or to one UTF-8 character in a character string expression.
- An SBCS percent sign (corresponding to U+005F) refers to a string of zero or more UCS-2 characters in a graphic string expression, or to a string of zero or more UTF-8 characters in a character string expression.

DBCS code pages also support a DBCS percent sign (corresponding to U+FF05) and a DBCS underscore (corresponding to U+FF3F). These characters have no special meaning for a UCS-2 database.

For the optional ″escape expression″, which specifies a character to be used to modify the special meaning of the underscore and percent sign characters, only ASCII characters, or characters that expand into a two-byte UTF-8 sequence, are supported. If you specify an escape character that expands to a three-byte UTF-8 value, an error message (error SQL0130N, SQLSTATE 22019) is returned.

Import/Export/Load Considerations

The DEL, ASC, and PC/IXF file formats are supported for a UCS-2 database, as described in this section. The WSF format is not supported.

When exporting from a UCS-2 database to an ASCII delimited (DEL) file, all character data is converted to the application code page. Both character string and graphic string data are converted to the same SBCS or MBCS code page of the client. This is expected behavior for the export of any database, and cannot be changed, because the entire delimited ASCII file can have only one code page. Therefore, if you export to a delimited ASCII file, only those UCS-2 characters that exist in your application code page will be saved. Other characters are replaced with the default substitution character for the application code page. For UTF-8 clients (code page 1208), there is no data loss, because all UCS-2 characters are supported by UTF-8 clients.

When importing from an ASCII file (DEL or ASC) to a UCS-2 database, character string data is converted from the application code page to UTF-8, and graphic string data is converted from the application code page to UCS-2. There is no data loss. If you want to import ASCII data that has been saved under a different code page, you should change the data file code page before issuing the IMPORT command. One way to accomplish this is to set DB2CODEPAGE to the code page of the ASCII data file.

The range of valid ASCII delimiters for SBCS and MBCS clients is identical to what is currently supported by DB2 UDB for those clients. The range of valid delimiters for UTF-8 clients is 0x01 to 0x7F, with the usual restrictions. For a complete list of these restrictions, refer to the "Export/Import/Load Utility File Formats" appendix in the *Data Movement Utilities Guide and Reference*.

When exporting from a UCS-2 database to a PC/IXF file, character string data is converted to the SBCS/MBCS code page of the client. graphic string data is not converted, and is stored in UCS-2 (code page 1200). There is no data loss.

When importing from a PC/IXF file to a UCS-2 database, character string data is assumed to be in the SBCS/MBCS code page stored in the PC/IXF header, and graphic string data is assumed to be in the DBCS code page stored in the PC/IXF header. Character string data is converted by the import utility from the code page specified in the PC/IXF header to the code page of the client, and then from the client code page to UTF-8 (by the INSERT statement). graphic string data is converted by the import utility from the DBCS code page specified in the PC/IXF header directly to UCS-2 (code page 1200).

The load utility places the data directly into the database and, by default, assumes data in ASC or DEL files to be in the code page of the database. Therefore, by default, no code page conversion takes place for ASCII files. When the code page for the data file has been explicitly specified (using the `codepage` modifier), the load utility uses this information to convert from the specified code page to the database code page before loading the data. For

PC/IXF files, the load utility always converts from the code pages specified in the IXF header to the database code page (1208 for CHAR, and 1200 for GRAPHIC).

The code page for DBCLOB files is always 1200 for UCS-2. The code page for CLOB files is the same as the code page for the data files being imported, loaded or exported. For example, when loading or importing data using the PC/IXF format, the CLOB file is assumed to be in the code page specified by the PC/IXF header. If the DBCLOB file is in ASC or DEL format, the load utility assumes that CLOB data is in the code page of the database (unless explicitly specified otherwise using the `codepage` modifier), while the import utility assumes it to be in the code page of the client application.

The `nochecklengths` modifier is always specified for a UCS-2 database, because:

- Any SBCS can be connected to a database for which there is no DBCS code page
- Character strings in UTF-8 format usually have different lengths than those in client code pages.

For more information about the load, import, and export utilities, refer to the *Data Movement Utilities Guide and Reference*.

Incompatibilities

For applications connected to a UCS-2 database, graphic string data is always in UCS-2 (code page 1200). For applications connected to non-UCS-2 databases, the graphic string data is in the DBCS code page of the application, or not allowed if the application code page is SBCS. For example, when a 932 client is connected to a Japanese non-UCS-2 database, the graphic string data is in code page 301. For the 932 client applications connected to a UCS-2 database, the graphic string data is in UCS-2.

Appendix F. Notices

IBM may not offer the products, services, or features discussed in this document in all countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Canada Limited
Office of the Lab Director
1150 Eglinton Ave. East
North York, Ontario
M3C 1H7
CANADA

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this information and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement, or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information may contain examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information may contain sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Each copy or any portion of these sample programs or any derivative work must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_. All rights reserved.

Trademarks

The following terms, which may be denoted by an asterisk(*), are trademarks of International Business Machines Corporation in the United States, other countries, or both.

ACF/VTAM
AISPO
AIX
AIX/6000
AIXwindows
AnyNet
APPN
AS/400
BookManager
CICS
C Set++
C/370
DATABASE 2
DataHub
DataJoiner
DataPropagator
DataRefresher
DB2
DB2 Connect
DB2 Extenders
DB2 OLAP Server
DB2 Universal Database
Distributed Relational Database Architecture
DRDA
eNetwork
Extended Services
FFST
First Failure Support Technology
IBM
IMS
IMS/ESA
LAN DistanceMVS
MVS/ESA
MVS/XA
Net.Data
OS/2
OS/390
OS/400
PowerPC
QBIC
QMF
RACF
RISC System/6000
RS/6000
S/370
SP
SQL/DS
SQL/400
System/370
System/390
SystemView
VisualAge
VM/ESA
VSE/ESA
VTAM
WebExplorer
WIN-OS/2

The following terms are trademarks or registered trademarks of other companies:

Microsoft, Windows, and Windows NT are trademarks or registered trademarks of Microsoft Corporation.

Java or all Java-based trademarks and logos, and Solaris are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Tivoli and NetView are trademarks of Tivoli Systems Inc. in the United States, other countries, or both.

UNIX is a registered trademark in the United States, other countries or both and is licensed exclusively through X/Open Company Limited.

Other company, product, or service names, which may be denoted by a double asterisk(**) may be trademarks or service marks of others.

Index

Contacting IBM

If you have a technical problem, please review and carry out the actions suggested by the *Troubleshooting Guide* before contacting DB2 Customer Support. This guide suggests information that you can gather to help DB2 Customer Support to serve you better.

For information or to order any of the DB2 Universal Database products contact an IBM representative at a local branch office or contact any authorized IBM software remarketer.

If you live in the U.S.A., then you can call one of the following numbers:

- 1-800-237-5511 for customer support
- 1-888-426-4343 to learn about available service options

Product Information

If you live in the U.S.A., then you can call one of the following numbers:

- 1-800-IBM-CALL (1-800-426-2255) or 1-800-3IBM-OS2 (1-800-342-6672) to order products or get general information.
- 1-800-879-2755 to order publications.

http://www.ibm.com/software/data/
The DB2 World Wide Web pages provide current DB2 information about news, product descriptions, education schedules, and more.

http://www.ibm.com/software/data/db2/library/
The DB2 Product and Service Technical Library provides access to frequently asked questions, fixes, books, and up-to-date DB2 technical information.

Note: This information may be in English only.

http://www.elink.ibmlink.ibm.com/pbl/pbl/
The International Publications ordering Web site provides information on how to order books.

http://www.ibm.com/education/certify/
The Professional Certification Program from the IBM Web site provides certification test information for a variety of IBM products, including DB2.

ftp.software.ibm.com

Log on as anonymous. In the directory `/ps/products/db2`, you can find demos, fixes, information, and tools relating to DB2 and many other products.

comp.databases.ibm-db2, bit.listserv.db2-l

These Internet newsgroups are available for users to discuss their experiences with DB2 products.

On Compuserve: GO IBMDB2

Enter this command to access the IBM DB2 Family forums. All DB2 products are supported through these forums.

For information on how to contact IBM outside of the United States, refer to Appendix A of the *IBM Software Support Handbook*. To access this document, go to the following Web page: http://www.ibm.com/support/, and then select the IBM Software Support Handbook link near the bottom of the page.

Note: In some countries, IBM-authorized dealers should contact their dealer support structure instead of the IBM Support Center.

Part Number: CT7XVNA

Printed in the United States of America
on recycled paper containing 10%
recovered post-consumer fiber.

SC09-2946-00

CT7XVNA