InDetail

IBM InfoSphere Data Replication & Federation

An InDetail Paper by Bloor Research
Author : Philip Howard
Publish date : August 2011

Most data federation or data virtualisation vendors do not offer replication or change data capture. Conversely, most suppliers of replication and change data capture do not offer federation. IBM is an exception to the rule.

Philip Howard

Executive summary

Replication is a multi-faceted capability. On the one hand it supports disaster recovery, failover and load balancing while on the other it enables real-time reporting and analytics. Indeed, there is an overlap between these two functions as it is increasingly common to use replicated back-up (disaster recovery) systems for reporting purposes, to relieve pressure on the primary transactional system.

With respect to this support for query capability, replication is a natural complement to data federation, which is also a technology that supports query processes, only this time without actually moving the data. Sometimes one technique will be appropriate and sometimes the other but they are naturally complementary.

Fast facts

IBM's data integration products provide three methods for delivering data: ETL (extract, transform and load), which provides bulk (batch) data loading functionality; data federation, which supports virtual data delivery; and data replication, which allows incremental updates.

The use of ETL processes has been well-established for the best part of twenty years and it is well understood. As a result we will not be discussing this aspect of IBM's portfolio. However, both replication and federation are becoming increasingly popular and relevant, and in this paper we will discuss why that is the case as well as IBM's offerings in this area.

Key findings

In the opinion of Bloor Research the following represent the key facts of which prospective users should be aware with respect to IBM replication and federation:

IBM CDC (changed data capture) and IBM Replication Server are currently separate products. IBM plans to merge the two into a single product, using the CDC nomenclature to refer to a specific IBM technology rather than a product or a general term. Initially this will be a packaging issue but the company plans to integrate them at a technological level also. This integration will be phased to provide deeper and more comprehensive integration.

- IBM CDC is usually targeted and optimised for heterogeneous environments both at the source and target level, while IBM Replication Server is more highly optimised for DB2. A third product, called IBM Classic Event Publisher and Replication Server supports non-DB2 mainframe replication and, where appropriate, acts as a front-end to InfoSphere Replication Server.

- While CDC is usually thought of as a technology for delivering incremental updates to a data warehouse, InfoSphere CDC can also be used to deliver data to Information Server (for data integration and data quality purposes) and to InfoSphere MDM Server for master data management.

- IBM Replication Server offers queue-based replication and is typically referred to as Q Replication. That is, leveraging WebSphere MQ for delivery of data and fast recoverability after a system outage. In addition, InfoSphere Replication Server also supports SQL-based replications, which is especially suitable for supporting data warehousing and similar environments.

- In addition to supporting many sources and targets, IBM's replication technologies also enable a comprehensive set of replication topologies. Indeed, it is difficult to think of any topology that it does not support.

- IBM supports many data sources and targets natively but not yet Netezza, which is a newly acquired IBM company. You can link replication to Netezza via a customized "User Exit" solution using JDBC (which is also the case for many other newer database vendors) and/or the NZ_Load utility but this is not yet productised. Federation Server can access Netezza data through generic ODBC and JDBC interfaces, but does not yet provide a specialised interface to Netezza. IBM will be introducing full native support by both replication and federation products in due course.

- IBM has probably a longer history in data federation than any other vendor, with Data Joiner (as it then was) being first introduced in the last century. It has taken a similar approach with respect to federation as replication in the sense that there is both a Federation Server product and a Classic Federation Server product, where the latter focuses on non-DB2 z Series systems including IMS, VSAM, Sequential/QSAM, ADABAS, IDMS and Datacom.

Executive summary

The bottom line

Most data federation or data virtualisation vendors do not offer replication or change data capture. Conversely, most suppliers of replication and change data capture do not offer federation. IBM is an exception to the rule. Further, most database suppliers that offer replication and specialise in supporting their own databases as a target, do not have the sort of extensive third-party capabilities that IBM can offer.

There is no doubt that federation and replication are frequently complementary. IBM's vision of an integrated suite of products that combine replication, change data capture and federation into a single environment thus represents a step change from what is currently available in the market.

Discussion

Before discussing IBM's products it will make sense to consider what replication and federation provide and why you might want to use either one.

Replication

According to Information Management Magazine (www.information-management.com/glossary/d.html) data replication is "*the process of copying a portion of a database from one environment to another and keeping the subsequent copies of the data in sync with the original source. Changes made to the original source are propagated to the copies of the data in other environments.*" This copying process is typically done in real-time or as close to real-time as possible: in other words, as each update is made to the original data source so that change is propagated to the replicated source. This is why IBM refers to it as incremental delivery.

Traditionally, there have been two different terminologies in play with respect to replication. Replication itself has historically been used to refer to environments where data was replicated in order to support disaster recovery and similar environments while change data capture (CDC) has been used to refer to environments where updates were being propagated to data warehouses for business intelligence purposes. There are differences between these two requirements: for example, you need to ensure transactional integrity when supporting disaster recovery and you need to confirm that updates have been properly received. However, as we shall see, traditional replication is increasingly being used to support business intelligence and other sorts of environments that might previously have been more closely associated with CDC. As a result, the distinction between traditional replication and CDC is becoming blurred and IBM has decided to drop the term CDC completely, using replication throughout.

Having thus clarified what replication is, we need to examine where it is used. There are a variety of applicable use cases:

1. Data warehousing and business intelligence. For loading (near) real-time data you can use replication, either in conjunction with ETL functions or on its own, to incrementally update your data warehouse. This is important for supporting operational BI, call centre operations and (if it is implemented in the data warehouse) master data management.

   In addition, replication can be used to support real-time reporting and query capability against ERP (and similar) environments such as SAP R/3. A further, related use case is when you need to integrate CEP (complex event processing) environments with operational data. The traditional methods of doing this are using polling or database triggers, but the former has latency issues and the latter is complex to program; using replication is an attractive alternative.

2. Synchronisation. There are a number of environments where it is important to synchronise information in operational environments. For example, when booking flights online it is common for users to explore various options before deciding on an airline and flight times. It is usual that low-cost commodity servers are used to support the former and when the user wants to move onto booking then the relevant data needs to be replicated to whatever mission-critical system is used to support actual transactions as opposed to those simply running queries. Another business example is synchronising Point of Sale data with central pricing/inventory systems. For this scenario, the data replication solution must support high volumes of transactions across a many-to-one topology.

3. High/continuous availability. As we have mentioned the traditional role of replication has been to support disaster recovery, both in active-passive and active-active environments, and it is also used to support failover, as well as load balancing, in highly available and continuously available environments. [The difference between continuous and high availability is that the former systems remain available through both planned and unplanned downtime while high availability systems only cater for unplanned downtime.]

4. Data migration. This covers application consolidations, application migration, database migration, migration to a master data management hub and various other scenarios. Traditionally, data migrations were completed with a "big bang" cutover at the end of the project where the old system was turned off at the beginning of a long weekend and then new one turned on at the beginning of the next working week. The problem with this approach is that if it doesn't work (and all too often it doesn't)

Discussion

then this can cause significant problems including, but not limited to, lost revenue, damaged reputation, loss of customer trust, and so on. Over the last few years an alternative approach, known as zero-downtime migrations, has emerged, which guarantees exactly what its name promises: no downtime. At the heart of this approach is replication and it is a requirement for this sort of environment, thanks to its ability to support bidirectional synchronisation with collision detection and recovery, so that both the old and new applications can be updated and/or utilised.

Master data management. Replication may be used to support master data management (MDM) in a variety of guises: to update the MDM hub, whether it is in a data warehouse or in a database of its own; to broadcast updated MDM data to participating applications; and to support high availability and back-up solutions for MDM.

Federation

Information Management Magazine describes data federation as "a method of linking data from two or more physically different locations and making the access/linkage appear transparent, as if the data was co-located." In practice it does four things:

1. It virtualises your data: it makes all relevant data sources, including databases, Excel and flat files, and any data source that can be reached via web services, appear as if they were in one place so that you can query that data (or report against it) as such.

2. It abstracts your data: that is to say, it presents the data to interrogating applications in a consistent fashion regardless of any native structure and syntax that may be in use in the underlying data sources.

3. It federates your data: it allows you to pull data together, from diverse, heterogeneous sources (which may contain either operational or historical data or both) and present that in a holistic manner, while maintaining appropriate security measures.

4. It presents your data in a consistent format to the front-end application (typically, but not always, a BI tool) either through relational views (via SQL) or by means of web services.

Data federation is becoming increasingly popular. Large enterprises typically have many data sources that it is impracticable to consolidate, either because of the costs involved or because you want to combine information from (say) operational sources, content management systems and/or data warehouses. In data warehousing environments it is becoming increasingly common to use data federation to link multiple data marts and warehouses, especially as many enterprises have concluded that the idea of having a single EDW (enterprise data warehouse) is an impractical dream. It is also frequently deployed to link data warehouses with operational sources for real-time reporting and query purposes. Some users of federation like the abstraction layer that it provides, and its ease of use, to the extent that they even use it for single-source queries.

Federation is useful as an interim measure to link relevant data sources after a merger or acquisition: consolidation is a long and costly exercise but the use of data federation can provide a level of integration both rapidly and at relatively low cost. Finally, you can also use federation to support MDM implementations if you implement a registry style MDM solution as opposed to a hub; it can be deployed to support SOA (service oriented architecture) implementations, providing a single, virtualised data access layer; it can be used in conjunction with archiving and test data management; and you can use federation to deliver information to web portals, e-commerce sites and so on.

One other notable feature of federation, in so far as queries are concerned, is the increased developer productivity that it provides. This is partly because there is less hand coding required (estimates suggest a reduction of between 40% and 65% when accessing multiple data sources) and, more particularly, because SQL queries do not have to be decomposed to act across the various databases involved. This is a complex task and requires experienced SQL programmers, thus a subsidiary effect of the automation introduced by using data federation is that it reduces the skill requirement needed.

IBM replication and federation

IBM has a long history in both the data federation and replication markets. The company's product, InfoSphere Federation Server, has gone through a number of nomenclature changes

Discussion

over the years (most notable as Information Integrator) but, when it was first introduced it was known as Data Joiner. This first appeared in the late 1990s, which was well before any other company introduced such a product. At that time (and since) it worked in conjunction with Data Propagator, which was the company's replication offering.

In 2007 IBM acquired DataMirror, another vendor of data replication technology since the 1990s and also a leading supplier of CDC technology. In practice, until now, what were previously DataMirror's products have been marketed by IBM as separate solutions from those that have been developed in-house. However, with the growing interest in replication and federation technologies, as discussed above, IBM has decided to merge its product lines both between the IBM and DataMirror offerings and across its mainframe and distributed systems, where there are also some product differences.

In the remainder of this paper we will discuss the replication and federation offerings that IBM currently has and, briefly, we will describe how the company plans to bring these together into a single, holistic environment.

IBM Replication & Federation

IBM has three products and four options for replication. The products are InfoSphere Replication Server, InfoSphere Classic Replication Server and the InfoSphere Change Data Capture (CDC) family. The InfoSphere Classic Replication Server provides System z data sourcing for InfoSphere Replication Server. That is, all System z sources with the exception of DB2. Both replication server products offer queue (that is WebSphere MQ) based replication but the InfoSphere Replication Server also offers SQL-based replication. IBM's intention is to merge its CDC products (which it acquired with the acquisition of DataMirror) with InfoSphere Replication Server in the distributed space. Initially the two products will be merged in terms of packaging and subsequently in terms of technology.

At present InfoSphere CDC (previously DataMirror Transformation Server) is focused on heterogeneous environments and InfoSphere Replication Server on DB2. The company's intention is to merge these two product suites, initially through packaging but ultimately at a technology level, so that all replication, heterogeneous or otherwise, will be available within a single product (suite). One example of this merging of technology is the recently released Classic CDC product, which introduced a new low impact IMS capture capability for use with the CDC framework (which covers the user interface, transport and apply agents). This same (new) IMS capture agent is also used within the recently released InfoSphere IMS → IMS High Availability solution. This demonstrates IBM's strategy for providing unified agents which can be used in either homogeneous or heterogeneous environments, as required. If we consider the combined offering then the support for heterogeneity is impressive, as illustrated in Table 1, which shows all the environments supported by IBM replication products.

Sources	Targets	Message Queues	Operating systems	Hardware
DB2 z/OS	DB2 z/OS	MQ Series	z/OS	IBM System z
DB2 LUW	DB2 LUW	JMS	AIX	IBM System p
DB2 i	DB2 i	TIBCO	IBM i OS	IBM i Series
IMS	IMS	WebMethods	Red Hat, SUSE Linux for System z	Intel/AMD
VSAM	VSAM*	BEA	Red Hat, SUSE Linux	HP PA-Risc
Informix	Informix		HP-UX	HP Itanium
solidDB	Information Server		Solaris	Sun SPARC
Oracle	Cognos Real Time monitoring		MS Windows	
MS SQL Server	Oracle			
Sybase ASE	Teradata			
Adabas	MS SQL Server			
IDMS	Sybase ASE			
	Netezza**			
	MySQL, Sybase IQ, Greenplum, PostgreSQL etc***			

* Only VSAM to VSAM;
** currently customised via JDBC: planned as both source and target;
*** via JDBC or ODBC, depending on the target.

Table 1: Environments supported by IBM replication products

The most notable thing about this table is the number of targets that are supported. It is common for competitive products from database vendors to focus on their own database products as a target and not offer the degree of heterogeneity supported by IBM.

IBM Replication & Federation

InfoSphere Replication Server

As mentioned, Replication Server supports both message queue based replication and SQL-based replication. This is a significant differentiator: most products do not offer such a choice. Both approaches support partitioned database environments for DB2 running on distributed platforms, automatically merging the changes captured from transaction logs (which is the preferred mechanism, as opposed to using database triggers). In addition, both methods support the replication of just what you need to replicate, using data filtering mechanisms (both horizontal and vertical in the case of SQL-based replication) and the SQL-based method also supports table-at-a-time replication.

The big advantage of queue-based replication is that the use of message queues allows the apply program to receive transactions without having to connect to the source database or subsystem. If either of the replication programs is stopped, messages remain on queues to be processed whenever the program is ready. Because the messages are persistent, the source and target remain synchronised even in the event of a system or device failure. It is also notable that the apply engine is parallelised in order to ensure minimal latency. Further, conflict detection and resolution features allow backup systems to also be used for productive work, supporting active-active environments, thus maximising resource utilisation and enabling application workload distribution across multiple servers. While queue-based replication is perhaps most widely used to support disaster recovery and failover environments, the apply engine also includes transformation capabilities that make it suitable for use in various operational scenarios. Finally, the topologies supported by queue-based replication are very comprehensive. Figure 1 illustrates this fact.

SQL-based replication is most typically used for large-scale data distribution and populating data warehouses and data marts, as it has extensive transformational capabilities, the table-at-a-time replication previously mentioned, and flexible scheduling capabilities that allow replication on a pre-determined, continuous or event-driven basis.

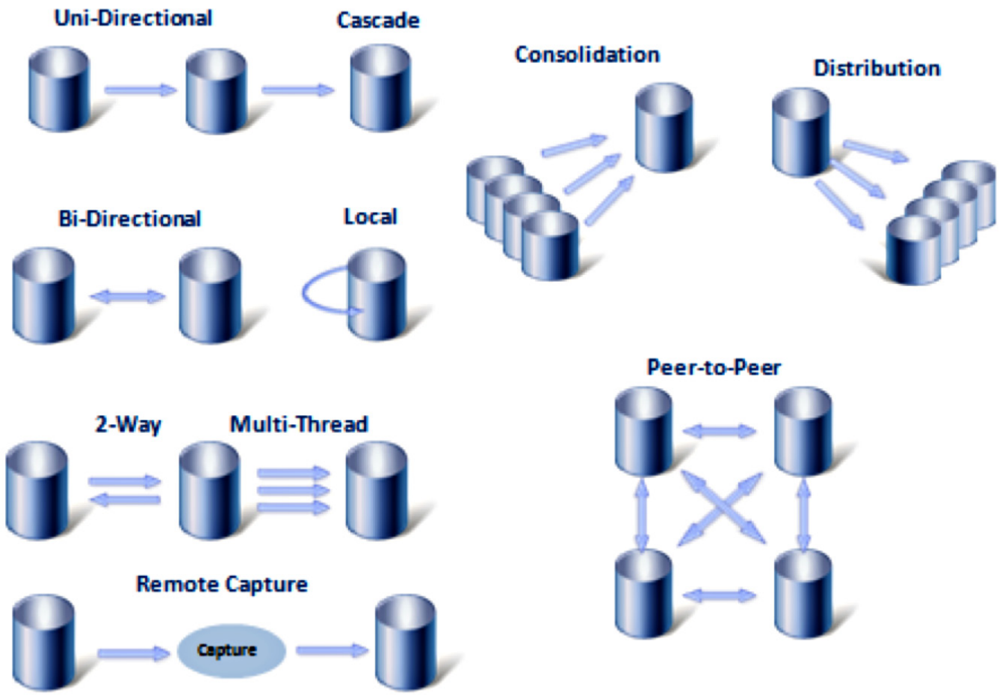
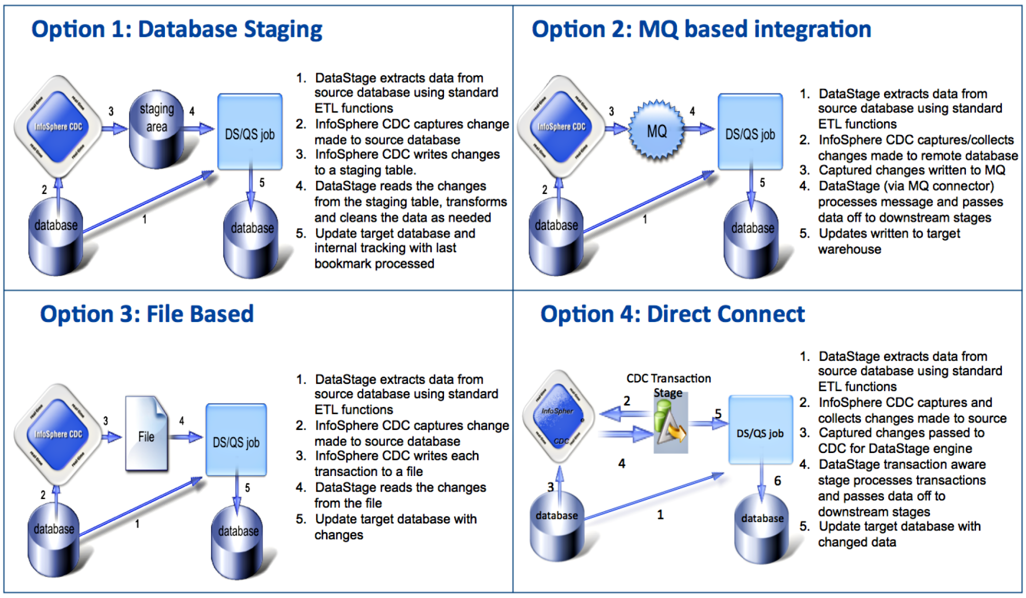


Figure 1: IBM InfoSphere Replication Server topologies

IBM Replication & Federation

InfoSphere CDC

Change data capture works by capturing data from database logs (or similar mechanisms in the case of mainframe environments that do not produce logs as such) and then propagating these changes to the target system in real-time. It is commonly used to trickle feed data into a data warehouse. This has a two-fold advantage: first, it means that the data is available in the warehouse in (near) real-time. Secondly, a further benefit is that the use of CDC reduces or eliminates the need for batch loading of the warehouse. This can be very important where the warehouse is mission-critical (which is increasingly the case) and you cannot afford to close it down for batch updating. This would occur if, for example you are supporting call centres across time zones from your warehouse. Even where the warehouse is not needed 24 x 7, the use of CDC can alleviate pressure on batch windows. As a result, CDC needs to be seen as complementary to ETL processes and Figure 2 illustrates and describes how InfoSphere can integrate with the relevant IBM product: IBM DataStage, as well as the complementary QualityStage, both of which form a part of IBM Information Server. Going beyond what is described in Figure 2, there are additional facilities for integrating CDC and Information Server, notably that IBM has utilities for metadata exchange so that InfoSphere's Metadata Workbench can show end-to-end data lineage (from source to target to BI tools) that spans jobs, leveraging the various CDC → DataStage integrations or even stand-alone CDC subscriptions. In addition, CDC can also be used to deliver data directly to InfoSphere MDM Server.



Option 1: Database Staging

1. DataStage extracts data from source database using standard ETL functions
2. InfoSphere CDC captures change made to source database
3. InfoSphere CDC writes changes to a staging table.
4. DataStage reads the changes from the staging table, transforms and cleans the data as needed
5. Update target database and internal tracking with last bookmark processed

Option 2: MQ based integration

1. DataStage extracts data from source database using standard ETL functions
2. InfoSphere CDC captures/collects changes made to remote database
3. Captured changes written to MQ
4. DataStage (via MQ connector) processes message and passes data off to downstream stages
5. Updates written to target warehouse

Option 3: File Based

1. DataStage extracts data from source database using standard ETL functions
2. InfoSphere CDC captures change made to source database
3. InfoSphere CDC writes each transaction to a file
4. DataStage reads the changes from the file
5. Update target database with changes

Option 4: Direct Connect

1. DataStage extracts data from source database using standard ETL functions
2. InfoSphere CDC captures and collects changes made to source
3. Captured changes passed to CDC for DataStage engine
4. DataStage transaction aware stage processes transactions and passes data off to downstream stages
5. Update target database with changed data

Figure 2: Options for integrating InfoSphere CDC with DataStage/QualityStage

IBM Replication & Federation

InfoSphere Federation Server

InfoSphere Federation Server is a separate product from Replication Server but it is complementary in the sense that sometimes you want to move data from one place to another to support query processing and sometimes you don't or can't. There is the same distinction between classic and non-classic sources as with replication but the supported sources are different, as illustrated in Table 2. However, the Classic Federation Server can also be a data source to the Federation Server on UNIX or Windows, thus allowing federation of Classic sources such as IMS and VSAM with distributed sources such as DB2 LUW and Oracle.

Sources	XML-based	Other	Mainframe (Classic)
DB2 z/OS	XML	WebSphere MQ	IMS
DB2 LUW	Web services	OLE DB	Adabas
DB2 i		Excel	CA-Datacom
Informix		Flat files	CA-IDMS
Oracle		Custom-built Java	VSAM
MS SQL Server		Custom-built C++	Sequential
Sybase ASE			
Teradata			
ODBC			
JDBC			

Table 2: Environments supported by InfoSphere Federation Server

The Federation Server for UNIX and Windows is built from IBM's DB2 technology. This means that, unlike some other products, IBM's data federation benefits from features found in a full relational database server. For example, Federation Server inherits DB2's ability to provide Oracle SQL compatibility. Perhaps more importantly, Federation Server runtime performance benefits from DB2's optimiser.

Perhaps the key issue for all data federation products is runtime performance. A major feature of Federation Server is the strength of its optimisation features, which are built around DB2's cost-based optimiser. These facilities have been leveraged within the product, which knows about the data source wrappers you have defined, collected database statistics (which are automatically refreshed) and so on. Optionally, it can also use information input by the administrator. A number of other features of the optimiser are worth mentioning. One is that the product aims to optimise SQL according to whichever data source it is addressing within the federated environment, though this, of course, is limited to data sources for which there are native connectors. Another notable facility is a "functional compensation" function. Basically, this is the idea that if a non-DB2 data source does not support a given SQL capability then Federation Server can compensate for this lack so that an application will see that data source as if it had that capability.

Yet another significant feature that is provided within Federation Server is support for data caching for applications that query data frequently but modify data only occasionally. What you actually cache is what IBM calls Materialized Query Tables (MQTs, otherwise known as summary tables or materialised views). This cache combines federation and replication. Data is replicated to MQTs, one for each source table. Applications issue all SQL to these MQTs and these queries are satisfied using data

IBM Replication & Federation

from the cache. However, inserts, updates and deletes are redirected to the original source transparently through federation. This allows easy offload of query work to a secondary system without requiring the complexity of bidirectional replication and the management of conflict resolution that comes with it.

Apart from performance, notable features of the federation capabilities offered by IBM include search as well as query capabilities; the ability to publish the results of a query to a message queue; to compose, transform and validate XML documents; and to publish the results of queries either as SQL answer sets or as XML documents. In the case of the former, SQL expressions can then be used to transform the data for data exchange or other purposes, while XML-based query results may be automatically validated against DTDs or XML schemas and can be transformed using XSL. In either case, an appropriate Web service (say, for currency conversion) can be used for transformation. Or the query can be submitted via a Web service request. There are also built-in facilities to allow results to be published to a WebSphere MQ message queue.

It is also important to appreciate that Federation Server isn't necessarily about queries per se. You may, for example, require data from other data sources as a part of a transactional application running locally. The product supports the creation of applications that can insert, update or delete rows from federated data sources, though this will usually be best left to the application API since this will adhere to business rules that may not be implemented within the federated environment.

Summary

Replication and federation extend and complement a variety of information management techniques, including disaster recovery and high availability, data integration, master data management, data warehousing, operational and business intelligence, archival, test data management and so on and so forth. Important factors are the range of data sources and targets that can be supported, performance, scalability and the need to minimise impact on source systems. IBM ticks all of these boxes.

Further Information

Further information about this subject is available from
http://www.BloorResearch.com/update/2101

Bloor Research overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter "noise" and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent over two decades distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

About the author

Philip Howard

Research Director - Data

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up what is now P3ST (Wordsmiths) Ltd in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director. His practice area encompasses anything to do with data and content and he has five further analysts working with him in this area. While maintaining an overview of the whole space Philip himself specialises in databases, data management, data integration, data quality, data federation, master data management, data governance and data warehousing. He also has an interest in event stream/complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to www.IT-Director.com and www.IT-Analysis.com and was previously the editor of both "Application Development News" and "Operating System News" on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and published a number of reports published by companies such as CMI and The Financial Times.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master) and walking the dog.

Copyright & disclaimer

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.

2nd Floor,
145–157 St John Street
LONDON,
EC1V 4PY, United Kingdom

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com