NOVEMBER 2005

Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise

By Colin White, BI Research

TDWI REPORT SERIES

tdwi
THE DATA WAREHOUSING INSTITUTE
A DECADE OF EXCELLENCE

A 101communications Publication

Research Sponsors

Business Objects

Collaborative Consulting

DataFlux

DataMirror Corporation

IBM Information Integration Solutions

Informatica Corporation

SAP America

Sunopsis

Syncsort Incorporated

TDWI REPORT SERIES

NOVEMBER 2005

Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise

By Colin White, BI Research

Table of Contents

About the Author

Colin White is the founder of BI Research. He is well known for his in-depth knowledge of leading-edge business intelligence and business integration technologies, and how they can be used to build a smart and agile business. With over 35 years of IT experience, he has consulted for dozens of companies throughout the world and is a frequent speaker at leading IT events. Colin has written numerous articles on business intelligence and enterprise business integration, and publishes an expert channel and a blog on the Business Intelligence Network. Prior to becoming an independent consultant in 1984, he worked for IBM as an IMS and DB2 specialist, and for Amdahl as a database systems architect.

About TDWI

The Data Warehousing Institute (TDWI), a division of 101communications LLC, is the premier provider of in-depth, high-quality education and training in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the strategies, techniques, and tools required to successfully design, build, and maintain data warehouses. It also fosters the advancement of data warehousing research and contributes to knowledge transfer and the professional development of its Members. TDWI sponsors and promotes a worldwide Membership program, annual educational conferences, regional educational seminars, onsite courses, solution provider partnerships, awards programs for best practices and leadership, resourceful publications, an in-depth research program, and a comprehensive Web site.

About the TDWI Report Series

This series is designed to educate technical and business professionals about new business intelligence technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies, supplemented by surveys of business intelligence professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize an emerging technology discipline or a new approach to solving business intelligence problems. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical business intelligence issues. Please contact Wayne Eckerson (weckerson@tdwi.org) to suggest a topic that meets these requirements.

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who responded to our requests for phone interviews. Second, our report sponsors who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: Jennifer Agee, Bill Grimmer, Denelle Hanlon, Deirdre Hoffman, and Marie McFarland.

For more information about this report or its sponsors, and to view the archived report Webinar, please visit: www.tdwi.org/direport.

Executive Summary

This report is a sequel to TDWI's 2003 report *Evaluating ETL and Data Integration Platforms*. The objective of the present report is to look at how data integration techniques, technologies, applications, and products have evolved since the 2003 report was published. The focus this time is not only on the role of data integration in data warehousing projects, but also on developing an enterprisewide data integration strategy.

Data Integration Is Not Getting Easier

The Challenges of Data Integration. Integrating disparate data has always been a difficult task, and given the data explosion occurring in most organizations, this task is not getting any easier. Over 69 percent of respondents to our survey rated data integration issues as either a *very high* or *high* inhibitor to implementing new applications (see Figure 1). The three main data integration issues (see Figure 2) listed by respondents were data quality and security, lack of a business case and inadequate funding, and a poor data integration infrastructure.

Data Integration: A Barrier to New Application Development

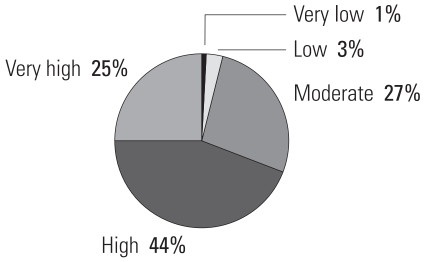


Figure 1. Data integration issues are a major barrier to implementing new applications. Based on 672 respondents.

Top Data Integration Issues

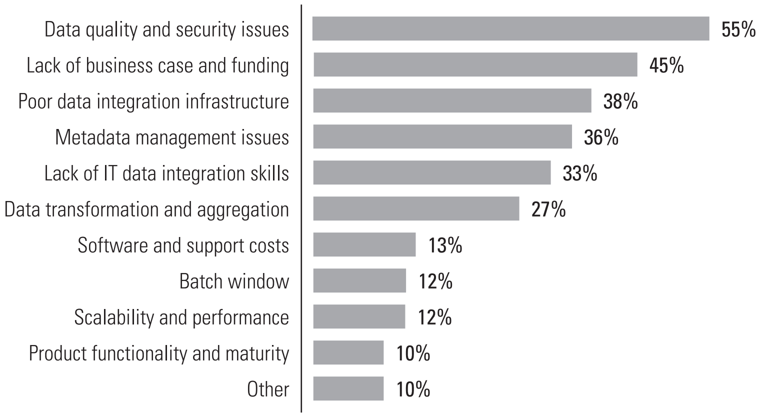


Figure 2: The top inhibitors to the success of data integration projects. Respondents were asked to select up to three. Based on 672 respondents.

Budgets Are Increasing for Data Integration Projects

Funding Data Integration Projects. Although inadequate funding was the second biggest inhibitor to the success of data integration projects, the survey shows that companies are increasing their budgets for data integration projects. Some 36 percent of organizations today devote a *very high* or *high* level of staffing and budget to data integration. The survey shows this number increasing to 55 percent within 18 months (see Figure 3).

Staffing and Budget for Data Integration

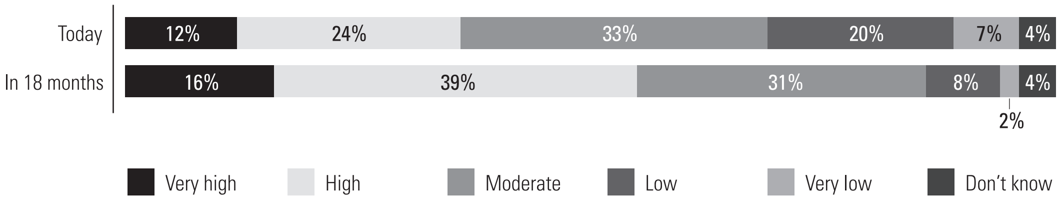


Figure 3. Staffing and budget in the very high and high categories will increase from 36 to 55 percent over the next 18 months. Based on 672 respondents.

Summary of Study Findings: Our survey looked at data integration approaches across a wide range of different companies. The results and follow-up interviews demonstrate that these companies fall into two main groups:

- *Large organizations* that are moving toward an enterprisewide data integration architecture. These companies typically have a multitude of data stores and large amounts of legacy data. They focus on buying an integrated product set and are interested in leading-edge data integration technologies. These organizations also buy high-performance best-of-breed products that work in conjunction with mainline data integration products to handle the integration of large amounts of data. They are also more likely to have a data integration competency center.
- *Medium-sized companies* that are focused on data integration solely from a business intelligence viewpoint and who evaluate products from the perspective of how well they will integrate with the organization's BI tools and applications. These companies often have less legacy data, and are less interested in leading-edge approaches such as right-time data and Web services.

Data Integration Problems Are a Barrier to Success

Many of the ideas and concepts presented in this report apply equally to all companies, regardless of size. The main message is that data integration problems are becoming a barrier to business success and a company must have an enterprisewide data integration strategy if it is to overcome this barrier.

Research Methodology

Report Scope. The report is geared to technical executives responsible for approving, funding, or managing data integration projects, as well as program managers, project managers, and architects responsible for implementing data integration solutions.

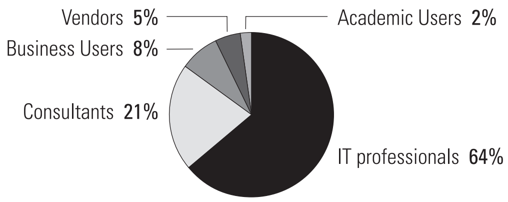
Methodology. The research for this report is based on a survey that TDWI conducted in July 2005, as well as interviews with experts in the field, including end-user organizations, consultants, industry analysts, and report sponsors.

Survey Methodology. TDWI contacted IT professionals in its database and 101communications' database. (TDWI is a business unit of 101communications.) In total, 672 people responded to the survey. Of these respondents, 37 identified themselves as being from the vendor community and 13 as being from the academic community. These latter two groups have been included in the results because in most cases these respondents completed the survey with respect to the use of data integration in their own organizations. Multi-choice questions and rounding account for totals that do not equal 100 percent.

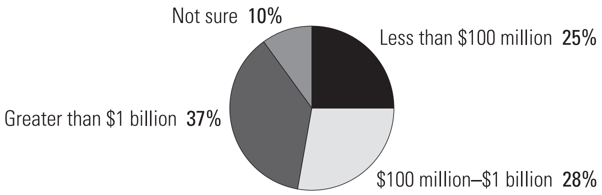
Survey Demographics. A majority of the survey respondents (64 percent) are IT professionals. The remainder are consultants/systems integrators (21 percent), business users (8 percent), vendors (5 percent) or academic users (2 percent.) Respondents to the survey were split fairly evenly among companies with revenues less than $100 million (25 percent), those with revenues between $100 million and $1 billion (28 percent), and those with revenue in excess of $1 billion (37 percent). About a third (35%) of the companies can be considered small and midsize businesses (less than 1,000 employees). The majority of respondents (63 percent) are based in North America and work in a range of industries, but the largest percentage are in consulting, financial services, software, insurance, manufacturing, retail, and healthcare. Consultants and systems integrators were asked to fill out the survey with their most recent client in mind.

Demographics

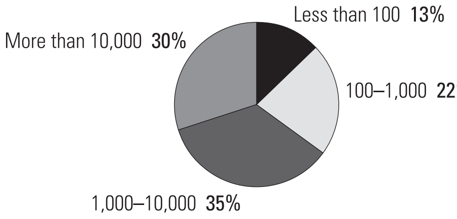
Position



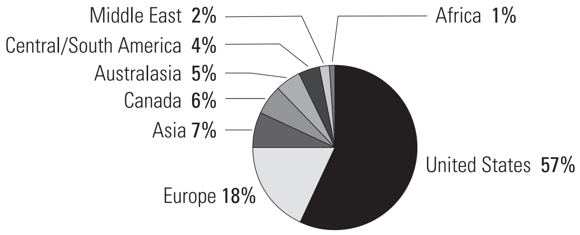
Company Size by Revenue



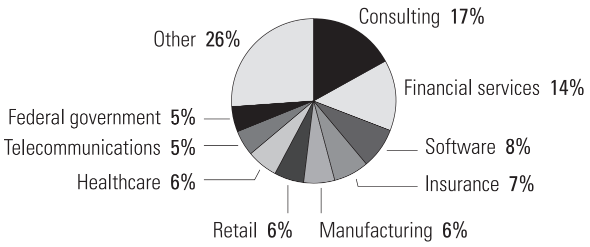
Company Size by Number of Employees



Geography



Industry



The Role of Data Integration in the Enterprise

There Are Four Levels of Enterprise Business Integration

Broadly speaking, enterprise business integration can occur at four different levels in an IT system: *data, application, business process,* and *user interaction* (see Figure 4). Many technologies and tools fit neatly into one of these categories; but, as we will see, there is a trend in the industry toward IT applications supporting multiple integration levels, and it is therefore important to design an integration architecture that can incorporate all four levels of enterprise business integration.

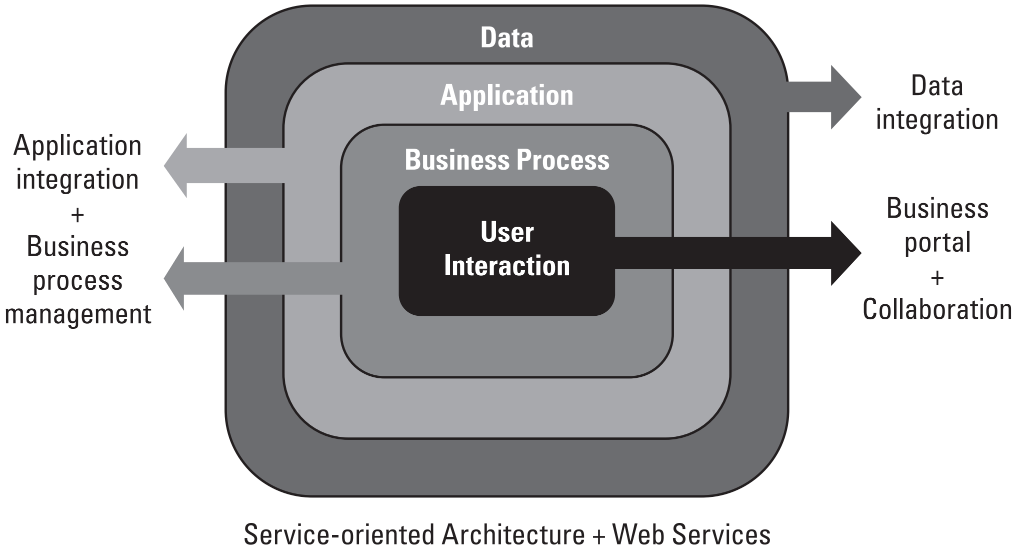

Data
Application
Business Process
User Interaction
Data integration
Business portal + Collaboration
Application integration + Business process management
Service-oriented Architecture + Web Services

Figure 4. The four levels of enterprise business integration (courtesy of BI Research).

Data Integration provides a unified view of the *business data* that is scattered throughout an organization. This unified view can be built using a variety of different techniques and technologies. It may be a physical view of data that has been captured from multiple disparate data sources and *consolidated* into an integrated data store like a data warehouse or operational data store, or it may be a virtual *federated* view of disparate data that is assembled dynamically at data access time. A third option is to provide a view of data that has been integrated by *propagating* data from one database to another—like merging customer data from a CRM database into an ERP database, for example. The approach will depend on the business requirements of the data integration project.

Application Integration provides a unified view of *business applications* that reside within or outside an organization. This unified view is achieved by managing and coordinating the flow of events (transactions, messages, or data) between applications. Application integration, like data integration, offers a variety of different implementation techniques and technologies depending on the requirements of a project.

Business Process Integration provides a unified view of an organization's *business processes*.[1] Business process design tools enable developers to analyze, model, and simulate business processes and their underlying activities. Business process management tools then implement and manage these processes using underlying application integration technologies. The key benefit of business process integration is that the design aspects of business process analysis and design are insulated from physical business process management and application implementation considerations. Increasingly, vendors are supporting both business process integration and application integration with a single product set. Unfortunately, most data warehousing and business intelligence vendors have yet to realize that data integration products should also view business data from a business process perspective.

User Interaction Integration provides users with a single personalized and secure interface to the business content (business processes, applications, and data) they need to do their jobs. This interface also allows users to collaborate and share data with each other. An enterprise portal is an example of a product that supports user interaction integration. A key issue with integration at the user interaction level is that, although the user is given a single unified view of multiple disparate systems, this view will highlight the lack of business process, application, and data integration between those systems; i.e., users will still need to navigate between different applications and data stores.

The four levels of enterprise business integration do not operate in isolation from each other. In a fully integrated business environment, interaction often occurs between the different integration levels. In the data warehousing environment, some data integration tools work with application integration software to capture events from an application workflow, and transform and load the event data into an operational data store (ODS) or data warehouse. The results of analyzing this integrated data are often presented to users through business dashboards that operate under the control of an enterprise portal that implements user interaction integration.

Data Integration Cannot Be Considered In Isolation

It is important for both IT staff and vendors to realize that data integration cannot be considered in isolation. Instead, a data integration strategy and infrastructure must take into account the application, business process, and user interaction integration strategies of the organization. One industry direction here is to build an integrated business environment around a service-oriented architecture (SOA). In an SOA environment business process, application, and data activities and operations are broken down into individual services that can interact with each other. Often an SOA is implemented using Web services because this technology is generally vendor platform independent and easier to implement than earlier SOA approaches.

[1] A business process consists of one or more *activities* that may be supported by business transaction, business intelligence, or business collaboration processing.

Characteristics of Data Integration

Data integration involves a framework of applications, techniques, technologies, and products for providing a unified and consistent view of enterprise business data (see Figure 5).

- *Applications* are custom-built and vendor-developed solutions that utilize one or more data integration products.
- *Products* are off-the-shelf commercial solutions that support one or more data integration technologies.
- *Technologies* implement one or more data integration techniques.
- *Techniques* are technology-independent approaches for doing data integration.

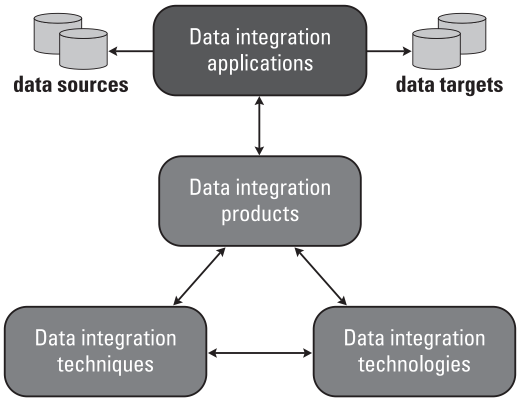

Data integration applications
data sources
data targets
Data integration products
Data integration techniques
Data integration technologies

Figure 5. Components of a data integration solution.

Our discussion of data integration will first review the techniques and technologies used in data integration projects, and then look at how data integration applications and products implement those techniques and technologies.

Data Integration Techniques

Consolidation, Federation, and Propogation Are Techniques for Intergrating Data

There are three main techniques used for integrating data: consolidation, federation, and propagation. These are illustrated in Figure 6.

Data Consolidation captures data from multiple source systems and integrates it into a single persistent data store. This data store may be used for reporting and analysis as in data warehousing, or it can act as a source of data for downstream applications as in an operational data store.

Data Consolidation Introduces Latency

With data consolidation, there is usually a delay, or *latency*, between the time updates occur in source systems and the time those updates appear in the target store. Depending on business needs, this latency may be a few seconds, several hours, or many days. The term *near real time* is often used to describe target data that has a low latency of a few seconds, minutes, or hours. Data with zero latency is known as *real-time* data, but this is difficult to achieve using data consolidation.

Target data stores that contain high-latency data (more than one day, for example) are built using *batch* data integration applications that *pull* data from the source systems at scheduled intervals.

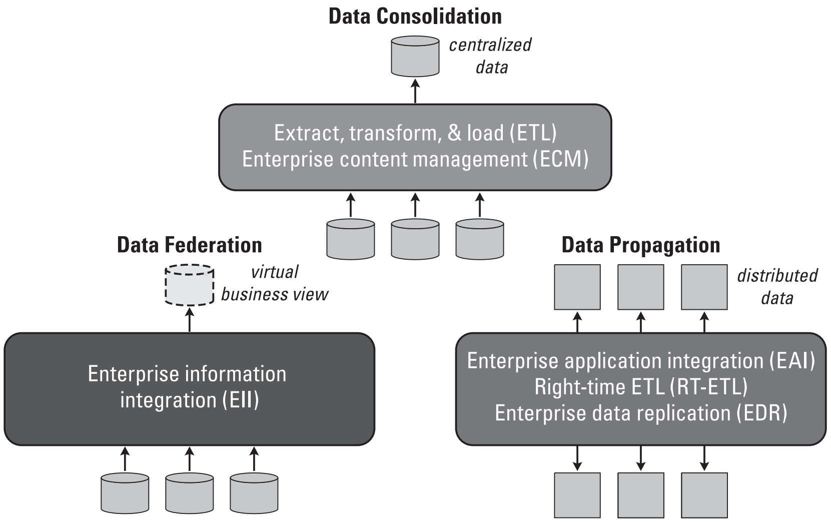

Data Consolidation
centralized data
Extract, transform, & load (ETL)
Enterprise content management (ECM)
Data Federation
virtual business view
Enterprise information integration (EII)
Data Propagation
distributed data
Enterprise application integration (EAI)
Right-time ETL (RT-ETL)
Enterprise data replication (EDR)

Figure 6. Data integration techniques: consolidation, federation, and propagation (courtesy of BI Research).

This pull approach employs data queries that take periodic snapshots of the source data. Although these queries will retrieve the current version of the data, they will not reflect changes that have occurred since the last snapshot—a source record could have been updated several times during the intervening period.

Low-latency target data stores are updated by *online* data integration applications that continuously capture and *push* data changes to the target store from source systems. This push approach requires the data consolidation application to identify the data changes to be captured for data consolidation. Some form of changed data capture (CDC) technique (see discussion on page 12) is usually used to do this. In this case, the capture task will extract all of the data changes that take place in the source data.

Push Mode Is Event-Driven. Pull Mode is on Demand

Pull and push consolidation modes can be used together—an online push application can, for example, accumulate data changes in a staging area, which is queried at scheduled intervals by a batch pull application. It is important to realize that push mode is *event-driven* and pull mode is *on demand* (see Figure 7).

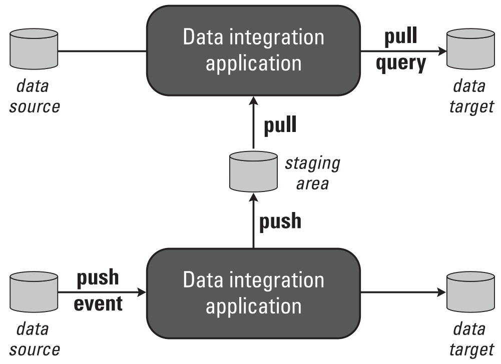


Figure 7. Push and pull modes of data consolidation.

Business applications that process the consolidated data store can query, report on, and analyze the data in the store. They cannot usually update the consolidated data because of problems with synchronizing those updates with source systems. Some data integration products, however, do offer a write capability by providing facilities to handle the data conflicts that may occur between the updated data in the consolidated data store and source systems.

Some applications update the consolidated data store and route the changes back to source systems. An example would be a target data store that is used to build a weekly pricing model. The model could be optimized and updated during the week and then reloaded back into the source system at the beginning of the following week. Another example where the consolidated data is updated is with integrated planning, budgeting, and forecasting tools, which coordinate planning data updates with source systems.

Consolidation Allows for Transformation of Large Data Volumes between Source and Target

The advantage of data consolidation is that it allows large volumes of data to be transformed (restructured, reconciled, cleansed, and/or aggregated) as it flows from source systems to the target data store. The disadvantages are the computing resources required to support the data consolidation process and the amount of disk space required to support the target data store.

Data consolidation is the main approach used by data warehousing applications to build and maintain an operational data store and an enterprise data warehouse. Data consolidation can also be used to build a dependent data mart, but in this case the consolidation process uses a single data source (i.e., an enterprise data warehouse). In a data warehousing environment, ETL (extract, transform, and load) technology is one of the more common technologies used to support data consolidation. Another data consolidation technology is ECM (enterprise content management). Most ECM solutions focus on consolidating and managing unstructured data such as documents, reports, and Web pages.

EII Supports a Federated Approach to Data Integration

Data Federation provides a single virtual view of one or more source data files. When a business application issues a query against this virtual view, a data federation engine retrieves data from the appropriate source data stores, integrates it to match the virtual view and query definition, and sends the results to the requesting business application. By definition, data federation always *pulls* data from source systems on an on-demand basis. Any required data transformation is done as the data is retrieved from the source data files. Enterprise information integration (EII) is an example of a technology that supports a federated approach to data integration.

One of the key elements of a federated system is the metadata used by the data federation engine to access the source data. In some cases, this metadata may consist solely of a virtual view definition that is mapped to the source files. In more advanced solutions, the metadata may also contain detailed information about the amount of data that exists in the source systems and what access paths can be used to access it. This more extensive information can help the federated solution optimize access to the source systems.

Some federated solutions may provide additional business metadata that documents semantic relationships between data elements in the source systems. An example here is customer data. The metadata may contain a common customer identifier that is mapped to the various customer keys in the source systems.

The main advantages of a federated approach are that it provides access to current data and removes the need to consolidate source data into another data store. Data federation, however, is not well suited for retrieving and reconciling large amounts of data, or for applications where there are significant data quality problems in the source data. Another consideration is the potential performance impact and overhead of accessing multiple data sources at run time.

Data Federation May Be Used When the Cost of Data Consolidation Outweighs the Business Benefits It Provides

Data federation may be used when the cost of data consolidation outweighs the business benefits it provides. Operational query and reporting is an example where this may be the case. Data federation can be of benefit when data security policies and license restrictions prevent source data being copied. Syndicated data usually falls into this latter category. It can also be used as a short-term data integration solution following a company merger or acquisition.

Organizations Should Use Data Integration Products That Support Both Data Consolidation and Federation

The source data investigation and profiling required for data federation is similar to that needed with data consolidation. Organizations should therefore use data integration products that support both data consolidation and federation, or at least products that can share the metadata used for consolidation and federation.

Data Propagation applications copy data from one location to another. These applications usually operate online and push data to the target location; i.e., they are event-driven. Updates to a source system may be propagated asynchronously or synchronously to the target system. Synchronous propagation requires that updates to both source and target systems occur in the same physical transaction. Regardless of the type of synchronization used, propagation guarantees the delivery of the data to the target. This guarantee is a key distinguishing feature of data propagation. Most synchronous data propagation technologies support a two-way exchange of data between a data source and a data target. Enterprise application integration (EAI) and enterprise data replication (EDR) are examples of technologies that support data propagation.

Data Propagation Can Be Used for Real-time or Near-real-time Movement of Data

The big advantage of data propagation is that it can be used for the real-time or near-real-time movement of data. Other benefits include guaranteed data delivery and two-way data propagation. The availability of many of these facilities will vary by product. Data propagation can also be used for workload balancing, backup and recovery, and disaster recovery.

Data propagation implementations vary considerably in both performance and data restructuring and cleansing capabilities. Some enterprise data replication products can support high volume data movement and restructuring, whereas EAI products are often limited in their bulk data movement and data restructuring capabilities. Part of the reason for these differences is that enterprise data replication has a data-centric architecture, whereas EAI is message- or transaction-centric.

A Hybrid Approach. The techniques used by data integration applications will depend on both business and technology requirements. It is quite common for a data integration application to use a *hybrid* approach that involves several data integration techniques. A good example here is customer data integration (CDI) where the objective is to provide a harmonized view of customer information.

A simple approach to CDI is to build a consolidated customer data store that contains customer data captured from source systems. The latency of the information in the consolidated store will depend on whether data is consolidated online or in batches, and how often updates are applied to the store.

Another approach to CDI is data federation, where virtual business views of the customer data in source systems are defined. These views are used by business applications to access current customer information in the source systems. The federated approach may also employ a metadata reference file to connect related customer information based on a common key.

A hybrid data consolidation and data federation approach may also be appropriate. Common customer data (name, address, etc.) could be consolidated in a single store, but customer data that is unique to a specific source application (customer orders, for example) could be federated. This hybrid approach can be extended further using data propagation. If a customer updates his or her name and address during a Web store transaction, this change could be sent to the consolidated data store and then propagated to other source systems such as a retail store customer database.

Changed Data Capture

Rebuilding the Target Data Store Is Impractical for All but Small Data Stores

Data consolidation and data propagation both create and maintain copies of source data. The challenge with both these approaches is how to handle the data changes that occur in source systems. One approach is to rebuild the target data store on a regular basis to keep it close to the currency of the source data, but this is impractical except for small data stores. Some form of changed data capture (CDC) capability is required to handle this issue.

If the source data contains a time-stamp showing when the data was last modified, it could be used to locate the data that has changed since the CDC application ran last. Unless a new record or version of the data is created each time it is modified, the CDC application will only see the most recent change to an individual record, not all changes since the last time the application ran.

If the source data is not time-stamped then source business applications could be modified either to create a time-stamp or to maintain a separate data file or message queue of the data changes. Packaged application vendors such as SAP often provide CDC facilities at the application level to handle this. In a service-oriented architecture, a Web service could be called to record the changes. Tapping in an EAI workflow is another option.

A common approach in relational DBMS applications is to add database update triggers that take a copy of the modified data. Another source for finding data changes is the DBMS recovery log. Enterprise data replication (EDR) solutions often support CDC using DBMS triggers and/or recovery logs. Triggers have more impact on the performance of source applications because the trigger and source data update processing is usually done in the same physical transaction. Processing of the recovery log, on the other hand, has less impact because it happens asynchronously from the data update processing.

Time-Stamping and Versioning Aid CDC

Time-stamping and versioning are quite common in unstructured data applications, which eases the CDC task. When a document is created or modified, the document metadata is usually updated to reflect the date and time of the activity. Many unstructured data systems also create a new version of a document each time it is modified.

There are many different ways of implementing CDC. If this capability is important to you, it is essential that you evaluate data integration solutions carefully to see if they support CDC, and to assess the performance impact of the CDC approach on source systems.

Data Quality Considerations

We saw at the beginning of this report that data quality issues are the leading inhibitor to successful data integration projects. It is also a key factor in choosing the data integration techniques, technologies, and products to be used in a project. Data quality by itself could be the subject of a completely separate study, but it is worth looking briefly at how data quality affects data integration.

Two aspects of data quality need to be considered in a data integration project. The first is the analysis of the source data stores to ascertain their contents and data quality. Data profiling tools are of value here. These tools are provided by both data integration and third-party vendors.

The second area of data quality for consideration is the cleansing of poor quality data. This is often done by inserting a data transformation process in the data integration workflow. Data transformation includes data restructuring, cleansing, reconciliation, and aggregation. It is during the data cleansing and reconciliation tasks where data quality issues are addressed most often. These tasks may be handled solely by the data integration tool itself, or by a third-party product.

Participants in our study had much to say about data quality in data integration projects. Comments included:

"The Biggest Issue for Business Users Today Is Data Quality"

- "When selecting a data integration approach, you need to focus on the needs of business users and the problems they are trying to solve," says John Williams, director of process and technology services at Collaborative Consulting. "The biggest issue for business users today is data quality, and this issue is a major stumbling block in data integration projects. This problem is caused largely by companies not treating data as a corporate asset, and not putting the required resources and management support in place to help solve this problem."
- "Data quality is a big driver for deploying data integration products," says Sachin Chawla, vice president of data integration at Business Objects. "This is a big issue for companies struggling to satisfy compliance legislation."
- "One of the key business drivers for data integration in our customer base is improving data quality," says Tony Fisher, president and general manager of DataFlux. This data quality requirement is strongly related to business application drivers like regulatory compliance in areas like HIPAA and Sarbanes-Oxley. An increasing number of companies are setting up data integration and data quality compliance centers. About 10 to 15 percent of our customers have done this."
- "Data quality is important in data integration projects, and its management should be handled by all data integration products," says Yves de Montcheuil, director of product management at Sunopsis. "Companies should not be forced into having to justify a separate data quality tool."

"Data Quality Management Should Be Handled by All Data Integration Products"

These comments highlight the issues surrounding data quality. They also emphasize that data quality is more than just a technology issue. It requires senior management to treat data as a corporate asset and to realize that the value of this asset depends on its quality.

Data Integration Technologies

A wide range of technologies are available for implementing the data integration techniques outlined above. This section reviews three of the main ones: extract, transform, and load (ETL); enterprise information integration (EII); and enterprise application integration (EAI). It also briefly reviews enterprise data replication (EDR) and ECM (enterprise content management). Master data management (MDM) and customer data integration (CDI), which are really data integration applications, are also discussed because they are often thought of as data integration technologies.

Extract, Transform, and Load

As the name implies, ETL technology extracts data from source systems, transforms it to satisfy business requirements, and loads the results into a target destination. Sources and targets are usually databases and files, but they can also be other types of data stores such as a message queue. ETL supports a consolidation approach to data integration.

Data can be extracted in schedule-driven pull mode or event-driven push mode. Both modes can take advantage of changed data capture. Pull mode operation supports data consolidation and is typically done in batch. Push mode operation is done online by propagating data changes to the target data store.

Data transformation may involve data record restructuring and reconciliation, data content cleansing and/or data content aggregation. Data loading may cause a complete refresh of a target data store or may be done by updating the target destination. Interfaces used here include *de facto* standards like ODBC, JBDC, JMS, for example, or native database and application interfaces.

ETL Products Have Improved and Evolved Recently

Early ETL solutions involved running batch jobs at scheduled intervals to capture data from flat files and relational databases and consolidate it into a data warehouse database managed by a relational DBMS. Over recent years, commercial ETL vendors have made a wide range of improvements and extensions to their products. Examples here include:

- Additional sources—legacy data, application packages, XML files, Web logs, EAI sources, Web services, unstructured data
- Additional targets—EAI targets, Web services
- Improved data transformation—user defined exits, data profiling and data quality management, support for standard programming languages, DBMS engine exploitation, Web services
- Better administration—job scheduling and tracking, metadata management, error recovery
- Better performance—parallel processing, load balancing, caching, support for native DBMS application and data load interfaces
- Improved usability—better visual development interfaces
- Enhanced security—support for external security packages and extranets
- Support for a data federation approach to data integration

These enhancements extend the use of ETL products beyond consolidating data for data warehousing to include a wide range of other enterprise data integration projects.

More information about ETL technology can be found in the TDWI research report *Evaluating ETL and Data Integration Platforms*. Another TDWI report, *Developing a BI Strategy for CRM/ERP Data*, is a valuable reference for organizations wishing to use ETL tools to capture and integrate data from packaged CRM and ERP applications.

The Batch ETL Market Has Flattened Out

In our survey, 57 percent of respondents rated their batch ETL usage as *high* (see Figure 8). Adding a *medium* rating to the result increases the figure to 81 percent. The survey also asked what the likely usage of batch ETL will be in two years. The result was 58 percent for *high* usage, and 82 percent for *high* and *medium*. As expected, these figures demonstrate that the batch ETL market has flattened out because most organizations use it.

ETL Use in Organizations

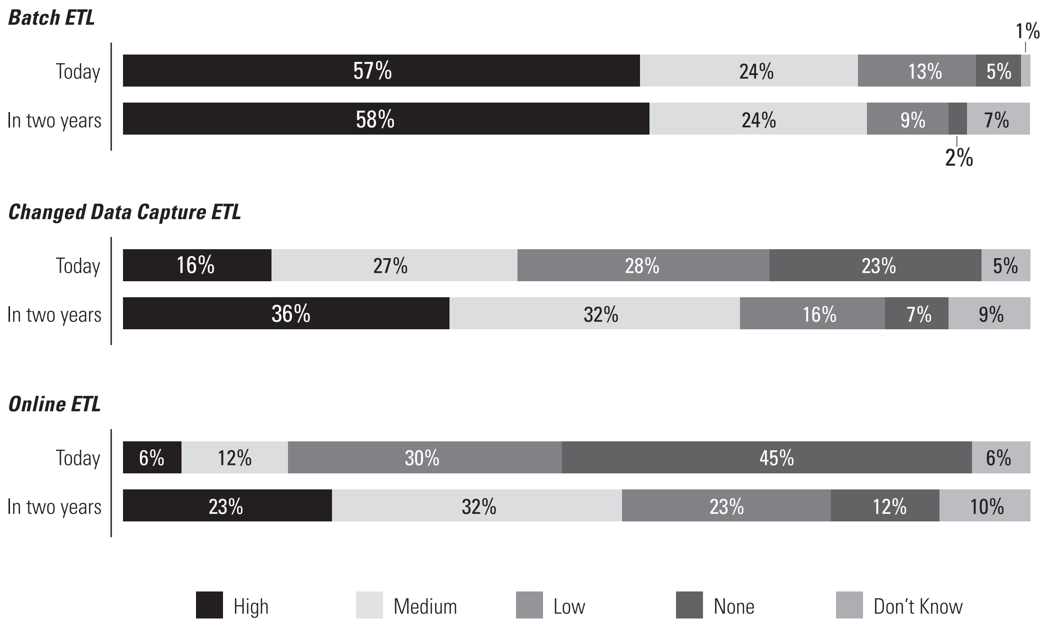


Figure 8. Batch ETL use is flat, but changed data capture and online ETL use will grow over the next two years. Based on 672 respondents.

CDC and Online ETL Operations Are Growing

The picture changes when looking at the growth figures for changed data capture (CDC) and online ETL operations. Our survey shows 16 percent of respondents rated their usage of CDC in ETL today as *high*. This number grows to 36 percent in two years. The equivalent figures for online ETL (called real-time or trickle-feed ETL in the survey) were 6 percent and 23 percent respectively. These growth trends are due primarily to shrinking batch windows and the increasing need for low-latency data. It is interesting to note that combining the *high* and *medium* usage figures for the two-year projection of online ETL gives a result of 55 percent. This clearly shows the industry is moving from batch to online ETL usage.

Enterprise Information Integration

EII provides a virtual business view of dispersed data. This view can be used for demand-driven query access to operational business transaction data, a data warehouse, and/or unstructured information. EII supports a data federation approach to data integration.

EII Enables Applications to See Dispersed Data As Though It Resided in a Single Database

The objective of EII is to enable applications to see dispersed data as though it resided in a single database. EII shields applications from the complexities of retrieving data from multiple locations, where the data may differ in semantics and formats, and may employ different data interfaces.

In its basic form, EII access to dispersed data involves breaking down a query issued against a virtual view into subcomponents, and sending each subcomponent for processing to the location where the required data resides. The EII product then combines the retrieved data and sends the final result to the application that issued the query. More advanced EII solutions contain sophisticated performance facilities that tune this process for optimal performance.

EII products have evolved from two different technology backgrounds—relational DBMS and XML. The trend of the industry, however, is toward products supporting both SQL (ODBC and JDBC) and XML (XQuery and XPath) data interfaces. Almost all EII products are based on Java.

A DDBMS Provides Transparent Access to Distributed Data

Products vary considerably in their features. Query optimization and performance are key areas where products differ. EII products that originate from a DBMS background often take advantage of the research done in developing distributed database management systems (DDBMS). The objective of a DDBMS is to provide transparent, full read and write access to distributed data. One of the main issues in the DDBMS field concerns the performance impact of distributed processing on mission-critical applications. This is especially the case when supporting full write access to the distributed data. Another problem is the complexity of handling multiple heterogeneous data sources.

To overcome the problems identified in DDBMS research, most EII products provide *read-only* access to heterogeneous data. Some products provide limited update capabilities, however. Another important performance option is the ability of the EII product to cache results and allow administrators to define rules that determine when the data in the cache is valid or needs to be refreshed.

Distinguishing features to look for when evaluating EII products include the data sources and targets supported (including Web services and unstructured data), transformation capabilities, metadata management, source data update capabilities, authentication and security options, performance, and caching.

In our survey, 5 percent of respondents rated their EII use as *high* (see Figure 9). Adding a *medium* rating to the result increases the figure to 19 percent. These figures grow to 22 percent and 52 percent respectively in two years, indicating considerable interest in exploiting EII technology in the future.

EII Use in Organizations

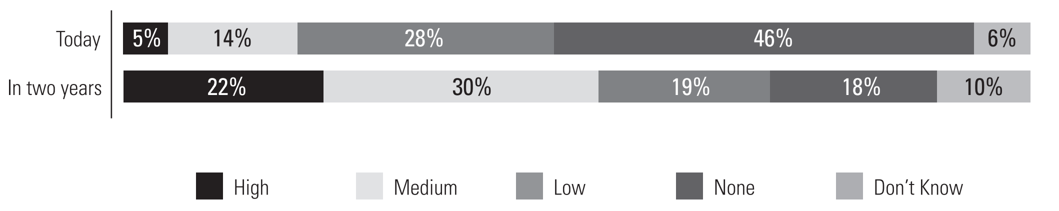


Figure 9. EII use is low at present but its usage is likely to grow rapidly. Based on 672 respondents.

Our study participants were asked for their opinions about where to use EII technology and where they saw EII technology heading. One large financial institution said they used EII for operational reporting, for accessing low usage data, and as a stop gap following an acquisition. Three specific EII applications were identified:

- An operational reporting application in a PeopleSoft environment.
- A compliance system that needed to access and analyze a variety of different data stores using an ODBC interface.
- A custom-built J2EE reporting dashboard that accesses a variety of data sources, including IBM DB2, Informix, Microsoft SQL Server, and Sybase. Several of these sources were used

because of a recent corporate acquisition. The dashboard is a temporary solution pending the creation of a data warehouse that will consolidate this data.

"EII is gaining traction for certain types of applications," says Ivan Chang, vice president of product marketing at Informatica. "There is a lot of consolidation going on in industries like pharmaceuticals and financial services, and EII can provide a temporary data integration solution while the data from the acquired or merged organization is brought into the data warehouse. Some companies also like the abstraction layer provided by EII because of its flexibility and simplicity. This abstraction layer can be used as a data source for an ETL product."

"There is a lot of interest about EII in the marketplace," says Philip Russom, senior manager of research and services at TDWI. "The two main uses of EII are reporting applications that need frequent access to current data, and dashboard applications that need to access a variety of different data stores. Ultimately in the future, I think EII will be assimilated into database products and BI reporting systems."

EII versus ETL

EII Data Federation Cannot Replace the Traditional ETL Data Consolidation Approach

It is important to emphasize that EII data federation cannot replace the traditional ETL data consolidation approach used for data warehousing. A fully federated data warehouse is not recommended because of performance and data consistency issues. EII should be used instead to extend and enhance a data warehousing environment to address specific business needs.

A Fully Federated Data Warehouse Is Not Recommended Because of Performance and Data Consistency Issues

EII is a powerful technology for solving certain types of data access problems, but it is essential to understand the trade-off of using federated data. One issue is that federated queries may need access to an operational business transaction system. Complex EII query processing against such a system can affect the performance of the operational applications running on that system. An EII approach can reduce this impact by sending less complex and more specific queries to the operational system.

Another potential problem with EII is how to transform data coming from multiple source systems. This is a similar problem that must be addressed when designing the ETL processes for building a data warehouse. The same detailed profiling and analysis of the data sources and their relationships to the targets is required. Sometimes, it will become clear that a data relationship is too complex, or the source data quality too poor, to allow federated access. EII does not in any way reduce the need for detailed modeling and analysis. It may in fact require more rigor in the design process, because of the real-time nature of data transformation in an EII environment.

In Some Circumstances, EII May Be an Appropriate Alternative to Using ETL

The following are circumstances when EII may be an appropriate alternative to using ETL:

- *On-demand access to volatile data.* Capturing and consolidating rapidly changing data from operational transaction systems can be costly, and some data latency will always occur in the consolidation process. EII can be used to directly access live operational source data. The performance and security aspects of accessing the live data, however, must be considered carefully. A federated query can also be used to access both live data and historical data warehouse information in the same request.
- *Direct write access to the source data.* Updating a consolidated copy of the source data is generally not advisable because data integrity issues between the original source data and the copy can occur. Not all EII products support write access, however.

- *It is difficult to consolidate the original source data.* When users require access to widely heterogeneous data and content, it may be difficult to bring all the structured and unstructured data together in a single consolidated copy.
- *Federated queries cost less than data consolidation.* The cost and performance impact of using federated queries should be compared with the network, storage, and maintenance costs of using ETL to consolidate data in a single store. There may be a case for a federated approach when source data volumes are too large to justify consolidation, or when only a small percentage of the consolidated data is ever used.
- *It is forbidden to make copies of the source data.* Creating a copy of sensitive source data or data that is controlled by another organization may be impractical for security, privacy, or licensing reasons.

Consider EII When User Needs Are Not Known in Advance

- *User needs are not known in advance.* Giving users immediate and self-service access to data is an obvious argument in favor of EII. Caution is required here, however, because of the potential for users to create queries that give poor response times and negatively impact both source system and network performance. In addition, because of semantic inconsistencies across data stores within organizations, there is a risk that such queries could return incorrect answers.

The arguments in favor of ETL are the opposite of those for EII:

- *Read-only access to reasonably stable data is required.* Creating regular copies of a data source isolates business users from the ongoing changes to source data, and enables data to be fixed at a certain moment in time to enable detailed analyses.

Use ETL When Users Need Historical or Trend Data

- *Users need historical or trend data.* This data is seldom available in operational data sources, but it can be built up over time through the ETL data consolidation process.
- *Data access performance and availability are key requirements.* Users want fast access to local data for complex query processing and analysis.
- *User needs are repeatable and can be predicted in advance.* When queries are well defined, repeated, and require access to only a known subset of the source data, it makes sense to create a copy of the data in a consolidated data store for access and use. This is particularly true when certain users need a view of the data that differs substantially from the way the data is stored in the source systems.

Use ETL for Complex Data Transformation

- *Data transformation is complex.* It is inadvisable for performance reasons to do complex data transformation online as a part of an EII federated data query. This transformation complexity may be caused not only by the need to restructure and aggregate data, but also by the need to cleanse and reconcile it to improve data quality.

Look for Vendors That Support Both EII and ETL

Both ETL and EII have a role to play in data warehousing and data integration, and organizations will need to implement both technologies. Rather than buying two separate products for ETL and EII, companies should look for vendors that support both technologies in a single integrated product set with shared metadata.

ETL vendors are beginning to offer an EII capability, which may be provided by the ETL product itself, or by using the services of a third-party product. Some ETL products use EII services behind the scenes to access heterogeneous data.

There are circumstances when the EII component within the offering must be deployed by itself on a separate system. An enterprise portal or dashboard application that employs EII to access a variety of data stores is an example of such a situation. In this case, the deployment of a complete data integration product set on the portal platform is not required and may be cost prohibitive.

Enterprise Application Integration

EAI Is Usually Employed for Real-time Operational Business Transaction Processing

EAI integrates application systems by allowing them to communicate and exchange business transactions, messages, and data with each other using standard interfaces. It enables applications to access data transparently without knowing its location or format. EAI is usually employed for real-time operational business transaction processing. It supports a data propagation approach to data integration.

The direction of the EAI industry is toward the use of an enterprise service bus (ESB) that supports the interconnection of legacy and packaged applications, and also Web services that form part of a service-oriented architecture (SOA).

From a data integration perspective, EAI can be used to transport data between applications and to route real-time event data to other data integration applications like an ETL process. Access to application sources and targets is done via Web services, Microsoft .NET interfaces, Java-related capabilities such as JMS, legacy application interfaces and adapters, etc.

EAI is designed to propagate small amounts of data from one application to another. This propagation can be synchronous or asynchronous, but is nearly always done within the scope of a single business transaction. In the case of asynchronous propagation, the business transaction may be broken down into multiple physical transactions. An example would be a travel request that is broken down in separate but coordinated airline, hotel, and car reservations.

Data transformation and metadata capabilities in an EAI system are focused toward simple transaction and message structures, and they cannot usually support the complex data structures handled by ETL products. In this regard, EAI does not compete with ETL.

Organizations See the Need to Merge the Benefits of EAI and ETL

In our survey, 9 percent of respondents rated their EAI usage as *high* (see Figure 10). Adding a *medium* rating increases the figure to 29 percent. These figures grow to 26 percent and 58 percent respectively in two years. It is important to point out that the question relates to the use of EAI for data integration, as opposed to the use of EAI in the organization overall. The two-year EAI projection of 58 percent is consistent with the 55 percent growth figure for online ETL use mentioned earlier. This suggests that organizations see the need to merge the event-driven benefits of EAI with the transformation and consolidation power of ETL.

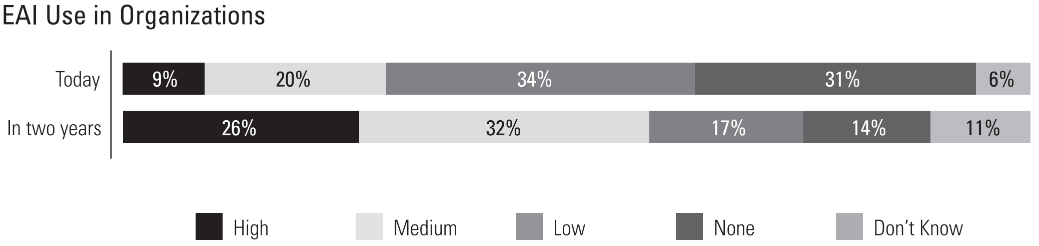


Figure 10. EAI growth is consistent with the growth in online ETL use shown in Figure 8. This suggests the two technologies will be used together. Based on 672 respondents.

EAI versus ETL

EAI and ETL Are Not Competing Technologies

Although some vendors would have you believe otherwise, EAI and ETL are not competing technologies. There are many situations where they can be used in conjunction with each other—EAI can act as an input source for ETL, and ETL can act as service to EAI.

One of the main objectives of EAI is to provide transparent access to the wide range of applications that exist in an organization. An EAI-to-ETL interface could therefore be used to give an ETL product access to this application data. This interconnection could be built using a Web service or a message queue. Such an interface eliminates the need for ETL vendors to develop point-to-point adapters for these application data sources. Also, given that EAI is focused on real-time processing, the EAI-to-ETL interface can also act as a real-time event source for ETL applications that require low-latency data. The interface can also be used as a data target by an ETL application.

Although several ETL and EAI vendors have announced marketing and technology relationships, the interfaces they provide are often still in their infancy. Potential users need to evaluate carefully the functionality and performance of these interfaces. It is expected, however, that the quality of these interfaces will steadily improve. At present, instead of using a dynamic EAI-to-ETL interface, many organizations are using EAI products to create data files, which are then input to ETL applications.

In the reverse direction, EAI applications can use ETL as a service. Several ETL vendors already allow developers to define ETL tasks as Web services. These ETL Web services can be invoked by EAI applications. This not only adds additional transformation power to the EAI environment, but also supports code and metadata reuse.

Enterprise Data Replication

Several other data integration technologies are worth mentioning. Data replication, for example, supports both the data propagation and CDC approaches to data integration.

EDR Is Used Extensively in Data Integration Projects, Often Packaged into Other Solutions

Although EDR is not as visible as ETL, EII, and EAI, it is nevertheless used extensively in data integration projects. One of the reasons for this lack of visibility is that EDR often is packaged into other solutions. All the major relational DBMS vendors, for example, provide data replication capabilities. Also, companies offering CDC solutions often employ data replication. EDR is used not only for data integration, but also for backup and recovery, and data mirroring and workload balancing scenarios.

The survey results show that 44 percent of respondents rate as *high* or *medium* their usage of EDR today (see Figure 11). This usage is projected to grow by 6 percent to 50 percent in two years.

Data Replication Use in Organizations

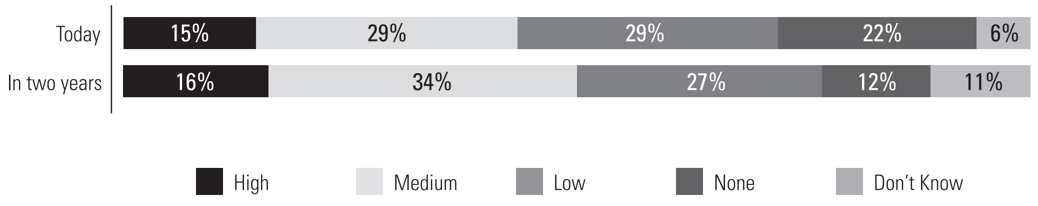


Figure 11. Total high and medium use ratings of EDR today is 44 percent. Based on 672 respondents.

EDR tools vary in their capabilities. Replication tools often employ database triggers and/or recovery logs to capture source data changes and propagate them to one or more remote databases. As discussed earlier, using recovery logs has less impact on source applications. In most cases propagation occurs asynchronously from the source transactions that produce the updates. Some EDR products, however, support two-way data synchronous propagation between multiple databases. Several also allow data to be transformed as it flows between databases.

One of the more significant differences between EDR and EAI is that data replication is designed for the transfer of data between databases, whereas EAI is designed for the movement of messages and transactions between applications. EDR typically involves considerably more data than EAI.

Data replication is sometimes used in conjunction with an ETL tool. For example, one company we interviewed was using EDR to continuously capture and transfer high volume data changes to a staging area. These changes were then extracted at regular intervals from the staging area by a batch ETL tool that consolidated them in a data warehouse. The data replication product was used because the ETL tool could not handle the volume of data changes involved.

Integrating Unstructured Data

Several EII Vendors Provide Federated Access to Unstructured Data

Most of the data integration technologies discussed so far focus on structured data. This is changing, however. Several EII vendors now provide federated access to unstructured data sources, particularly text-based documents. ETL vendors are also working on the processing of unstructured data.

"We are starting to see companies wanting to incorporate unstructured and semi-structured data into data integration workflows," says Ivan Chang, vice president of product marketing at Informatica. "One example here is the financial and healthcare industries who want to handle semi-structured information that is defined in industry-standard formats like SWIFT, HL7, and HIPAA."

Applications that employ ETL and EII to process unstructured data often want to integrate or relate the results to structured information. An example would be a marketing application that retrieves product sales analytics and related product information about advertising and market surveys.

ECM Products Add a Content Management Layer on Top of a Shared Data Store

Another technology that handles the integration of unstructured data is enterprise content management (ECM), which is focused on the consolidation of documents, Web information, and rich media. ECM products concentrate on the sharing and management of large quantities of unstructured data for a wide user population. These products add a content management layer on top of a shared data store. This layer provides metadata management, versioning, templates, and workflow.

An ECM content store can act as a data source for an EII or ETL application. The key here is not simply to provide access to unstructured data, but also to access the metadata that describes the structure, contents, and business meaning of that data. This is analogous to the issues associated with accessing and integrating packaged application data where the metadata is again important to understanding the business meaning of the data. In both cases, it is important to evaluate not only what data and application sources are supported, but also the level of integration with the source data and metadata.

Integrating Unstructured Data Is Becoming More Important

In our survey, 15 percent of respondents rated their use of ECM as *high* or *medium* for data integration (see Figure 12). This figure will grow to 43 percent within two years, which demonstrates the growing importance of integrating unstructured data.

Enterprise Content Management Use in Organizations

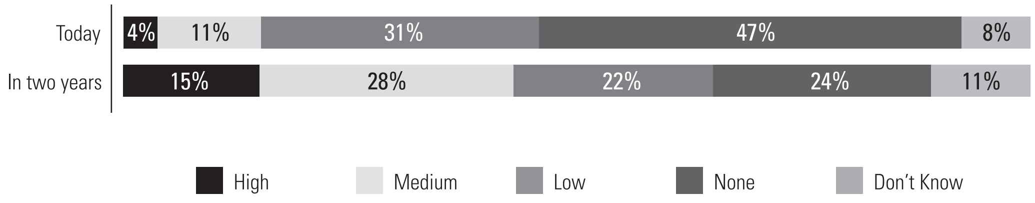


Figure 12. ECM use (total of high and medium ratings) will grow from 15 to 43 percent within two years. Based on 672 respondents.

To determine what types of data were involved in data integration projects and to gauge how many organizations were integrating unstructured data, the survey asked respondents to list the types of source data involved in their integration projects. The results are shown in Figure 13.

Types of Source Data Involved in Data Integration Projects

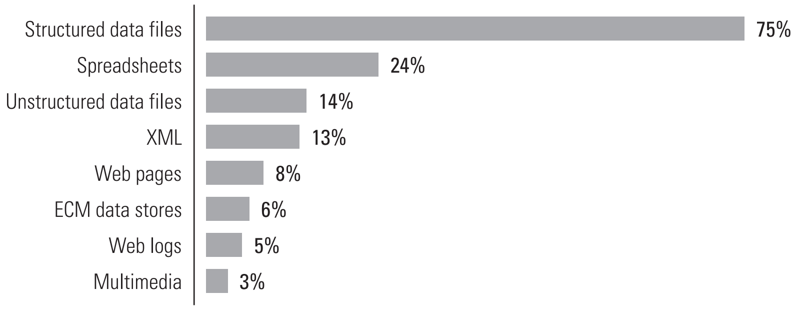


Figure 13. Structured data has the highest usage (75 percent). Other high-use data sources include spreadsheets (24 percent), unstructured data (14 percent), and XML (13 percent). Based on 672 respondents.

Many Organizations Still Source Data from Spreadsheets

As expected, our research shows the most common type of data used in integration projects is structured data (75 percent of respondents). Results for unstructured data and semi-structured data were 14 percent and 13 percent respectively. These latter figures are similar to those for ECM usage in data integration. The most surprising result was the use of spreadsheet data in integration projects. Some 24 percent of organizations were sourcing data from spreadsheets.

For structured data, our research also looked into the types of database systems that serve as data sources in integration projects. The results are shown in Figure 14. The top two DBMS data sources were relational databases (79 percent) and mainframe/legacy databases (41 percent). The latter figure clearly shows the amount of legacy data that still exists in companies. The result of 32 percent for application package databases shows the increasing influence of packaged applications and the growing need to integrate data from these applications with other enterprise data.

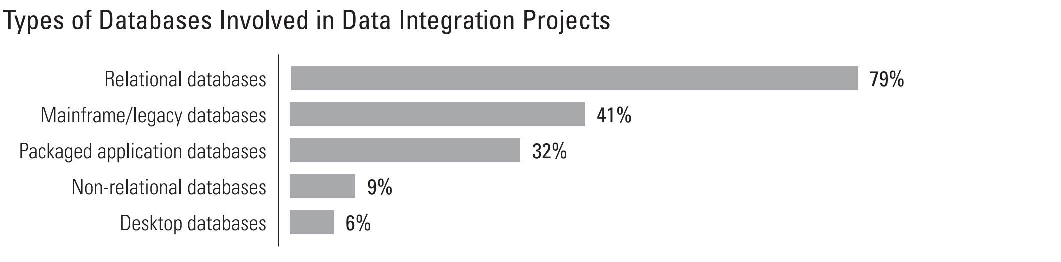


Figure 14. High-use database sources are relational (79 percent), legacy/mainframe (41 percent), and packaged application (32 percent). Based on 672 respondents who could select more than one answer.

Master Data Management and Customer Data Integration

MDM does the job of providing and maintaining a consistent view of an organization's reference data, which may be scattered across a range of application systems. The type of data involved in this process varies by industry and organization, but examples include customers, parts, employees, and finances. Most MDM applications at present concentrate on handling customer data because this aids the sales and marketing process, and can thus help improve revenues. A new buzzword for customer MDM solutions is customer data integration, or CDI.

Our research shows that 20 percent of organizations use MDM in data integration projects to a *high* or *medium* degree today. This figure is projected to grow to 50 percent within two years (see Figure 15).

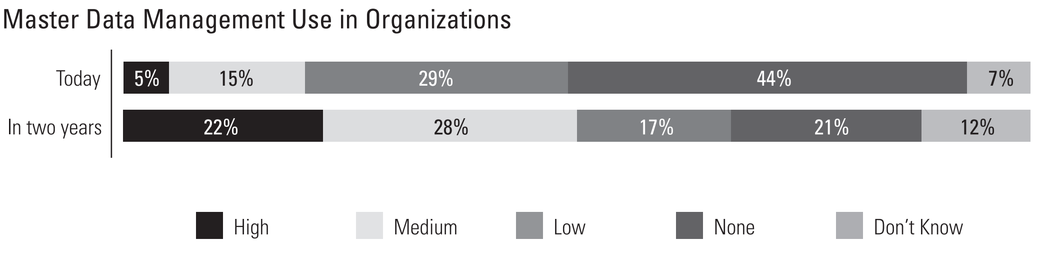


Figure 15. MDM use (total of high and medium ratings) will grow from 20 to 50 percent within two years. Based on 672 respondents.

MDM and CDI Are Business Applications, Not Technologies

MDM and CDI are often discussed as technologies, but in reality they are business applications. The objective of both MDM and CDI is to provide a consistent view of dispersed data. This view is created using data integration techniques and technologies, and may be used by business transaction applications and/or analytic applications. The actual techniques and technologies used will depend on application requirements, such as data latency and the need to update or just read the integrated data. What MDM and CDI add to data integration are the business semantics about the reference data as it relates to the business domain and industry involved. The value of the MDM or CDI solution therefore arises not only from the technology platform provided, but also from the power of the business semantic layer. MDM and CDI data stores can act as data sources for data warehousing applications.

In our survey we also investigated the use of MDM and CDI business applications. MDM applications are deployed in 30 percent of organizations and CDI applications in 50 percent of organizations. The MDM figure is consistent with the 20 percent *high* and *medium* use total shown in Figure 15 for MDM usage from a *technology* perspective.

Several people interviewed in the study said they saw increasing demand for MDM applications.

"We're Seeing a Lot of Demand from Customers to Help Complete Their MDM Projects"

"While MDM is still in its infancy, we're seeing a lot of demand from customers to help complete their MDM projects," says Andrew Manby, business strategist in the IBM Information Integration Solutions Group. "In order to be successful, every MDM project should employ data integration and its related best practices. While there are tremendous synergies between the data lifecycle in data warehousing and MDM, they are distinct disciplines. The MDM lifecycle comprises five phases:

1. Data assessment
2. Data harmonization
3. Loading the MDM systems (product information management, customer data integration, for example)
4. Creating operational processes that deliver data integrity
5. Putting in place data governance for ongoing assessment and evaluation

MDM represents a fantastic opportunity for data warehouse practitioners to broaden their skill sets and add even more value to their businesses."

"There are two aspects to master data management," says Lothar Schubert, director of solutions marketing at SAP Labs. "The first is the metadata that defines what a business entity such as a customer means to the business. The second aspect is the technologies used to maintain the consistency and integrity of the master data. For example, the master data may be used only for analysis by a BI application. In this case a batch asynchronous process may be sufficient to maintain the master data. If you are trying to maintain the integrity of master data in operational systems, however, then a synchronous approach to data harmonization is often the preferred approach."

Defining the business meaning of data in MDM applications is complex and requires a thorough understanding of how the data is used throughout the organization.

"One of our biggest needs for master data at present is enterprisewide customer information," says Brian Hickie, vice president of business intelligence at McKesson Pharmaceutical. "The pharmaceutical unit, for which I am responsible from a business intelligence perspective, has its own data warehouse based on SAP BW. Other business units not on SAP BW use other data warehousing technologies or have no data warehouse at all. We have recently started to extract, load, and integrate customer data across the enterprise as a part of a central McKesson-wide customer information store. One of the biggest challenges we face in integrating this enterprisewide customer data is designing and implementing customer hierarchies. From a business perspective, a customer has different roles depending on the business unit with which it is transacting. The need to see our customers in these different roles poses significant challenges in viewing and understanding the customers holistically across the enterprise."

Data Integration Application Design and Usage

Designing data integration applications requires applying the techniques and technologies outlined earlier to satisfy business needs and solve business issues. This requires a detailed understanding of the type of data integration application required, the business problems the application must address, and the characteristics of the data that the application needs.

Understanding Data Flow Is Important

To understand the different types of data integration applications that may exist in an organization, it is useful to examine how data flows in an IT system (see Figure 16). There are three main types of processing to consider: business transaction (BTx), business intelligence (BI), and business collaboration.

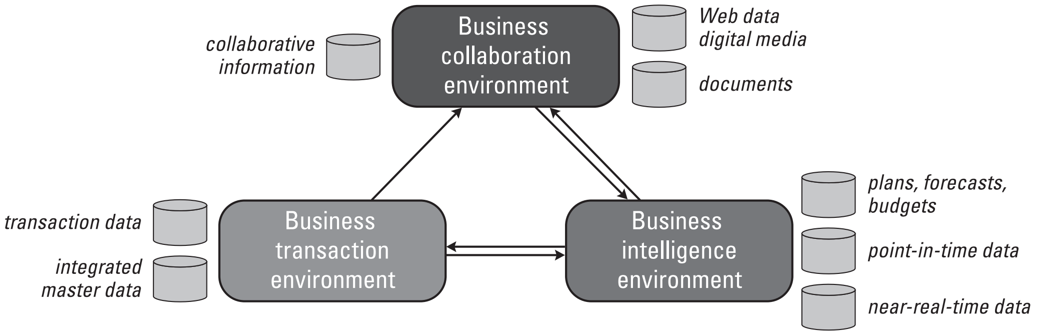

collaborative information
Business collaboration environment
Web data
digital media
documents
transaction data
integrated master data
Business transaction environment
Business intelligence environment
plans, forecasts, budgets
point-in-time data
near-real-time data

Figure 16. Data flow in an IT system (courtesy of BI Research).

BTx processing applications are responsible for running day-to-day business operations and they store *detailed* transaction data about the operations in data repositories that are managed by a variety of different files and databases. In most companies, these data repositories are scattered throughout the organization, and there is often a need to integrate them for operational reporting, creating consistent master data, and so forth.

BI applications report on and analyze business operations and produce information to help business users understand, improve, and optimize business operations. This information may be produced by working directly against *real-time* BTx data, but it more commonly comes from processing the data stored in an operational data store or data warehouse. A data warehousing environment enables data from various BTx data sources to be captured and consolidated into a data repository. The data warehouse repository is managed by a database system that supports data languages such as SQL and XQuery for data access and manipulation.

Data Latency Varies with Application Requirements

Data in a data warehouse repository reflects different moments in time. Strategic and tactical BI applications employ *summarized* and *historical* data, whereas operational BI applications require *detailed* data whose currency may only be a few seconds or minutes behind the live BTx system from which the data has been captured. The time delay, or latency, between when data is updated by a BTx application and when it is added to the data warehouse repository will vary depending on BI application requirements and on the cost of implementing near-real-time data integration.

Often it is better to use the term *right time* rather than real time or near real time, when discussing data latency. This term emphasizes that the latency in a data integration scenario should be based on the latency required by applications to satisfy business needs.

"We see a significant amount of interest in right-time data integration," says Gary O'Connell, director of product marketing in the IBM Information Integration Solutions Group. "One industry where there is high demand is in retail. These companies want to be able to monitor sales and promotion information through a trading day. They want to be able to look at demand and inventory levels, and balance their replenishment processes appropriately. We are not talking about real time, but about updating information several times throughout the day. The problem for these companies is determining best practices for using and deploying right-time data integration."

"We see two main uses of right-time data integration emerging," says SAP's Lothar Schubert. "The first is for consolidation of data from worldwide systems when customers are dealing with multiple time zones and tight batch windows. The second is with specific applications that need to merge and marry current data with historical data. Sales and customer relationship management are examples of applications in this second category. A call center application could, for example, provide a CSR with analytical insights based on information collected at the point of contact with a customer and on historical information about that customer. It can then help the CSR to make special offers based on the geographical or buying pattern demographics of the customer."

In our survey, 45 percent of respondents indicated they were using some form of right-time data integration into a data warehouse or operational data store. The latency of the data varied as shown in Figure 17. Of the 301 respondents answering the question, 45 percent had latencies shorter than an hour, and 20 percent had latencies between one and 12 hours.

Latency of Data in Right-time Data Integration Projects

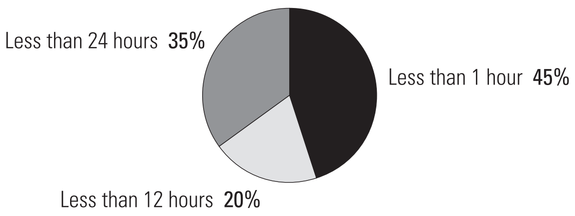


Figure 17. Forty-five percent of right-time data warehouses have a data latency of less than one hour. Based on 301 respondents.

Right-time Data Integration Captures Data Changes and Events Occurring in Source Systems and Loads Them into the Target Data Store

Right-time data integration involves capturing data changes and events occurring in source systems and loading them into the target data store. As discussed earlier, there are a variety of ways of doing this, and our survey respondents were asked to identify their method. The results from 297 respondents illustrated in Figure 18 show that the most popular approaches (usage rated as either *high* or *medium*) were right-time/online ETL (67 percent), changed data capture (51 percent), and data from message queues (42 percent). These methods are of course not mutually exclusive. A third of respondents were using Web services and events from an EAI server.

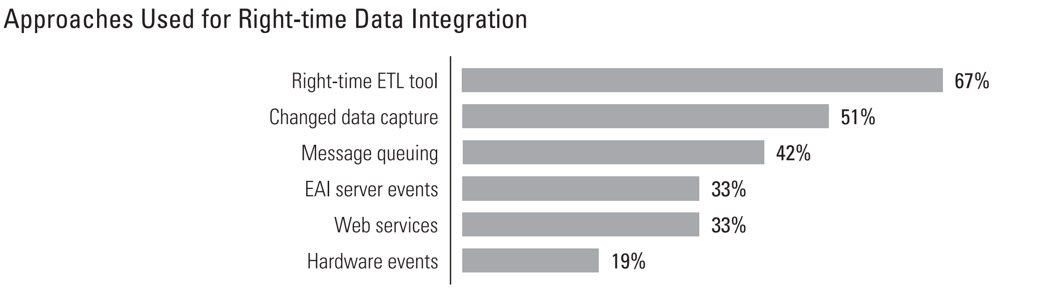


Figure 18. Top three approaches were right-time ETL, changed data capture, and message queuing. Based on 297 respondents who could select more than one answer.

Immature Integration Architecture Inhibits Right-time Data Integration

In our research, the biggest inhibitors to right-time data integration were immature data integration architecture, inability to understand the benefits of right-time data integration, and the performance impact on operational systems. Implementation costs and a limited number of right-time data projects were also key inhibitors. The results are shown in Figure 19.

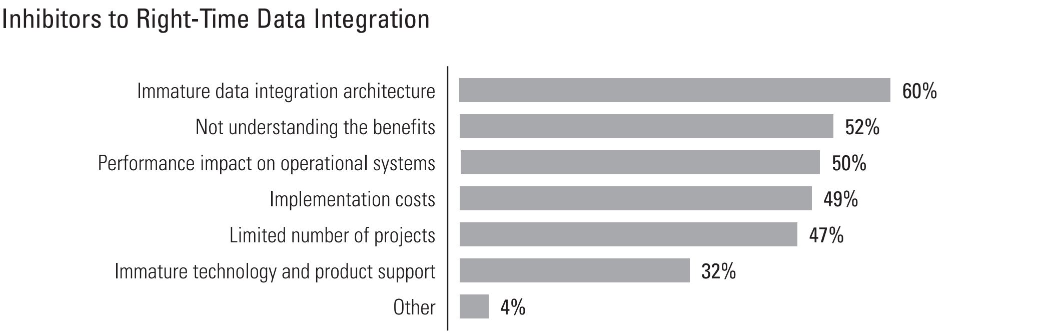


Figure 19. Top three inhibitors were immature data integration architecture, understanding the business benefits, and the performance impact on operational systems. Based on 300 respondents who could select more than one answer.

The TDWI Research Report *Building the Real-Time Enterprise* discusses the building of right-time data stores in more detail.

Putting Analytics into a Business Context Improves Business Decision Making and Action Taking

Performance management. A recent trend in BI is toward performance management applications that put BI analytics into a business context; i.e., they relate analytical results to business plans, forecasts, and budgets. Putting analytics into a business context improves business decision making and action taking because the results become *actionable*. If you know today's sales figures are 10 percent below target, then you can decide how to fix this problem and take the appropriate action. The planning, budgeting, and forecasting data that the business analytics are compared against also need to be integrated into a data planning repository, which may be a component of a data warehouse.

A Content Repository Is Usually Managed by a Database System

Putting analytics into a business context creates *business information*. This information may be embedded in documents, reports, spreadsheets, presentations, Web pages, audio, video, and e-mail. This unstructured information may be consolidated and managed in a content repository, which supports business semantics (i.e., business metadata) like author, date produced, etc. A content repository also provides facilities like versioning, workflow, templates, and search tools. Like a data repository, a content repository is usually managed by a database system.

When business users receive information from a BI system, they use their expertise or *knowledge* to make decisions and take actions. The decision- and action-taking process may involve interacting with other users, which is supported by *business collaboration applications*. This user-driven approach to decision making and action taking is manual or un-programmed. If the knowledge of the business user can be captured as a set of best business practices in the form of a set of *business rules*, then the decision- and action-taking process can be programmed and automated.

Survey respondents were asked to identify the main types of business applications being supported by data integration projects. The results are shown in Figure 20. The top three applications were historical data reporting (80 percent), ad hoc query (73 percent), and operational data reporting (65 percent). These three were followed by strategic/tactical analysis and performance management (62 percent), and operational analysis and performance management (57 percent).

These results show much greater use of data integration in business intelligence and data warehousing projects than in business transaction processing projects. This number is probably affected by the makeup of the audience who responded to the survey and by the fact that enterprise data integration often begins in the business intelligence area.

"Among our customers, we still see a strong demand for data integration solutions that support a data warehousing environment," says Philip On, senior product marketing manager at Business Objects. "For example, many of our customers, particularly those who have Crystal Reports, still do not have a solid data foundation for their BI processing, and they need a data warehouse to solve this problem. This type of customer is looking for a data integration solution that works well with their BI product set."

The results also show the healthy use of operational business intelligence, which often involves integrating business intelligence processing with business transaction processing. This trend has political implications for the way IT departments are organized and funded for providing enterprise integration solutions. This topic will be discussed in more detail later in this report.

"For us the move toward an enterprisewide view of data integration is not new," says Jennifer St. Louis, product manager at DataMirror. "Data warehousing used to be the first point of contact into a company for data integration, but this is not the case anymore. We are finding that companies are looking for data integration products that handle not only data warehousing requirements, but also needs in other project areas like e-business. Companies want to leverage their investments in data integration software."

"Companies Want to Leverage Their Investments in Data Integration Software"

Types of Applications Supported by Data Integration Projects

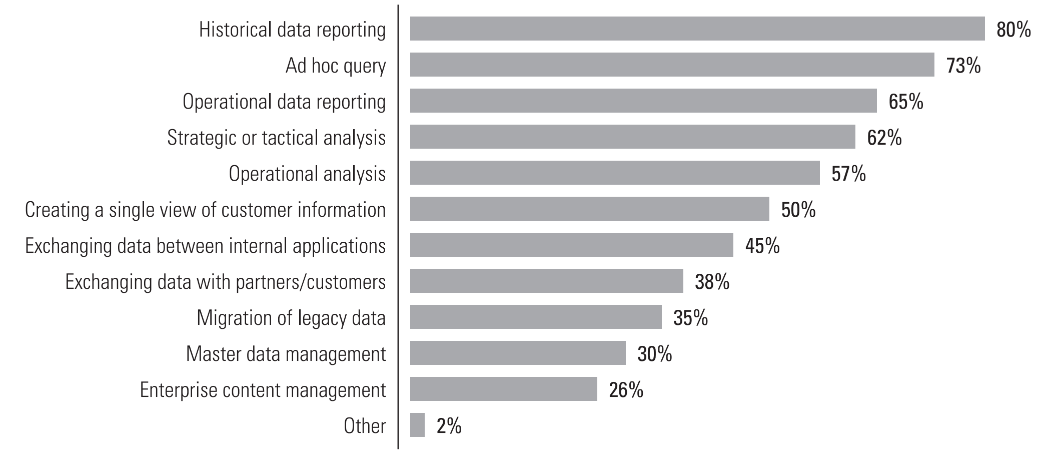


Figure 20. Greatest application use of data integration is for reporting and ad hoc query. Based on 671 respondents who could select more than one answer.

Choosing the Right Data Integration Products

It can be seen from the discussion so far that there are several application characteristics or variables that affect the choice of techniques and technologies (and thus products) for doing data integration. These variables are shown here.

Data Integration Application Variables

- Source data type
  - Structured
  - Semi-structured (e.g., XML)
  - Unstructured
  - Packaged application
  - EAI
  - Web service
  - Metadata
- Source data organization
  - Homogeneous or heterogeneous
  - Centralized or distributed (integrated data and metadata)
  - Federated (integrated metadata) or dispersed (no integrated metadata)
- Source data transformation requirements
  - Data restructuring
  - Data cleansing
  - Data reconciliation
  - Data aggregation
- Target data currency (latency) and access
  - Real time
  - Near real time
  - Point in time
  - Read-only or read-write
- Data integration technique and mode
  - consolidation, federation, propagation, changed data capture
  - event push or on-demand pull
  - synchronous or asynchronous
- Data integration technology
  - ETL, EII, EAI, EDR, ECM
- Data scale
  - Number of data sources
  - Data store size
  - Data store volatility

Over time, as the number of integration applications increases, most of the data integration variables outlined in the sidebar will be experienced in IT projects. It is rarely the case, however, that an organization designs a data integration architecture that supports all of these variables up front. In reality, the data integration architecture is often built piecemeal using best-of-breed products that are added to the architecture to satisfy the needs of each new application.

Data Integration Architecture Is Often Built Piecemeal

Best-of-Breed Products versus Integrated Solutions

As the use of enterprise and data integration matures in an organization, interest usually grows in building a more coherent data integration framework and in reducing the number of products involved in data integration. Product selection then often shifts toward a data integration suite of products from a single vendor, and filling any gaps with best-of-breed products.

Our research points to increasing interest in vendors that provide a complete data integration solution. This factor came third in importance, after total cost of ownership and product features, in response to our question about the main criteria used in choosing a data integration vendor (see Figure 21).

Main Criteria for Choosing a Data Integration Vendor

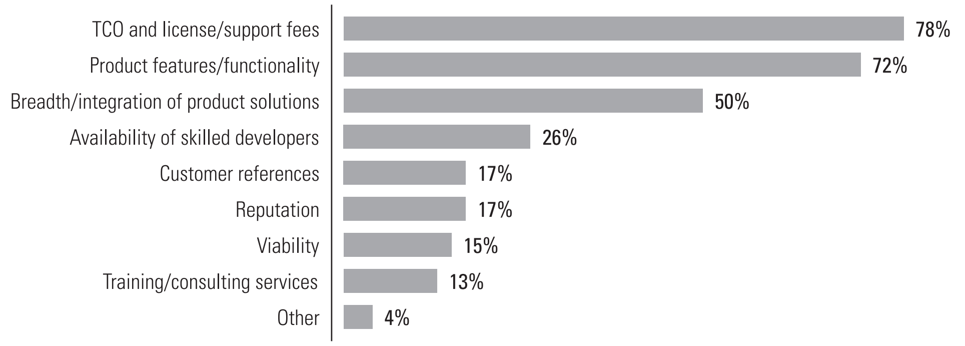


Figure 21. The most important criteria are total cost of ownership, product features, and breadth of the vendor's solution. Our 672 respondents were asked to select up to three criteria.

Custom-Built Solutions

Previous TDWI reports have discussed at length the topic of whether to build a data integration solution or buy an off-the-shelf tool from a vendor. In our 2003 ETL study, 55 percent of respondents were using some form of custom-built solution in their data warehouse. In the current study, we asked companies whether they were still using custom-built approaches for data integration—58 percent said they made either *high* or *medium* use of in-house-developed features. This shows there has been no decline in the use of custom-built solutions since the original study. Significantly, this figure means these companies are either using custom-built solutions exclusively or are using these solutions in conjunction with vendor data integration products.

The Use of Custom-Built Solutions Has Not Declined

When asked to predict the use of custom-built features in two years, some 48 percent of respondents indicated their usage would still be *high* or *medium*. The *high* rating declined from 25 percent to 15 percent during this period.

Product Selection Criteria

The results in Figure 21 show that the second most significant criterion in choosing a vendor was the features offered by the product set. Our survey gave respondents a list of features and asked them to rank the features *very important, important,* or *not important*. The four top features ranked as very important (see Figure 22) were performance (70 percent of respondents), transformation power (68 percent), security (56 percent), and data sources and targets supported (53 percent).

Most Important Data Integration Features

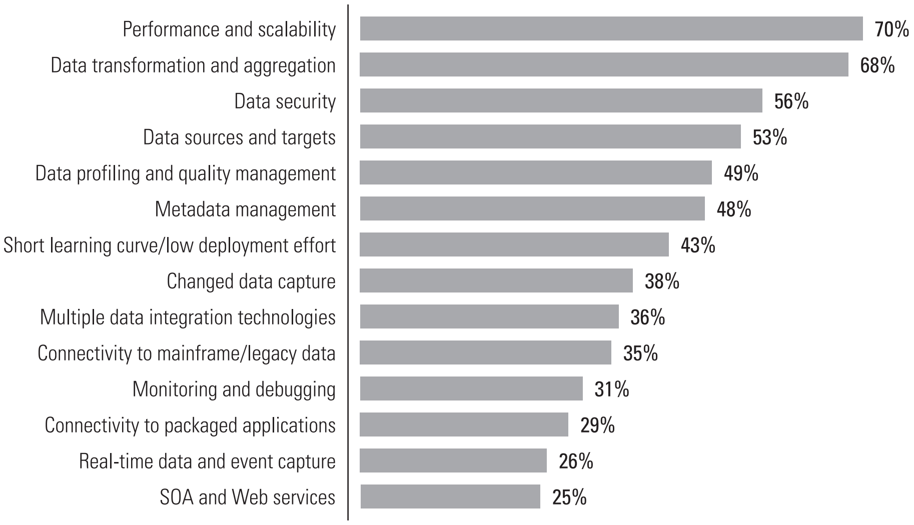


Figure 22. The four most important features are performance, data transformation, security, and data sources/ targets supported. Our 672 respondents could select more than one answer.

In was not surprising that performance was top of the requirements list. Given the data growth in most organizations, performance is becoming a critical issue for many.

Performance Is Becoming More Important As Data Volume Grows

"The data volumes involved in data integration projects have grown dramatically in the 30 years Syncsort has been in business," says Andrew Coleman, director of software engineering at Syncsort. "The big data users have evolved from the banking industry to industries like retail and telecommunications. If you consider how many times a week you use a bank ATM or buy products from a retail store, and then compare this to the number of phone calls you make in the same period, you get an idea of how data volumes have increased. As the industry moves toward new technologies like RFID, the data volumes will get even bigger. Our customers are looking for solutions that enable them to efficiently transfer and transform data ranging in volume from megabytes to hundreds of gigabytes. In many cases these solutions are used in conjunction with data integration ETL tools from other vendors."

Although the requirement for Web services (25 percent of respondents had this need) was lower in the list than expected, the result is consistent with the requirement for real-time data (26 percent of respondents). These two requirements are related. The need for high performance may also be an issue here since Web services is still an immature technology from a performance standpoint.

"Web services and XML are important industry directions, and our products support them, but these technologies are not yet ready for high-performance and high-volume data integration," says Yves de Montcheuil of Sunopsis.

Developing a Data Integration Strategy

To determine where companies are with their data integration strategy, we asked survey respondents about their organization's approach to data integration (see Figure 23). Some 18 percent of respondents have built a common enterprisewide data integration architecture. As expected, a large number (41 percent) use separate data integration architectures for business transaction processing and data warehousing. In 38 percent of companies, each group or project develops its own data integration approach.

Strategies for Data Integration

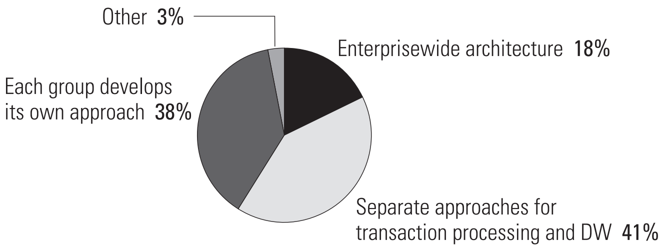


Figure 23. Eighteen percent of organizations have a common enterprisewide data integration architecture. Based on 672 respondents.

82% Do Not Have an Enterprise Approach to Data Integration

The lack of an enterprise approach to data integration in 82 percent of organizations demonstrates why data integration is becoming an inhibitor to new application development. To solve this problem all organizations should have a long-term objective to create a flexible enterprise data integration architecture that provides the techniques, technologies, and products to support new data integration applications (see Figure 24). The architecture should evolve over time as new application requirements are uncovered, and as new data integration technologies and products are introduced. This architecture is especially important for organizations that have a complex heterogeneous data environment involving large volumes of data.

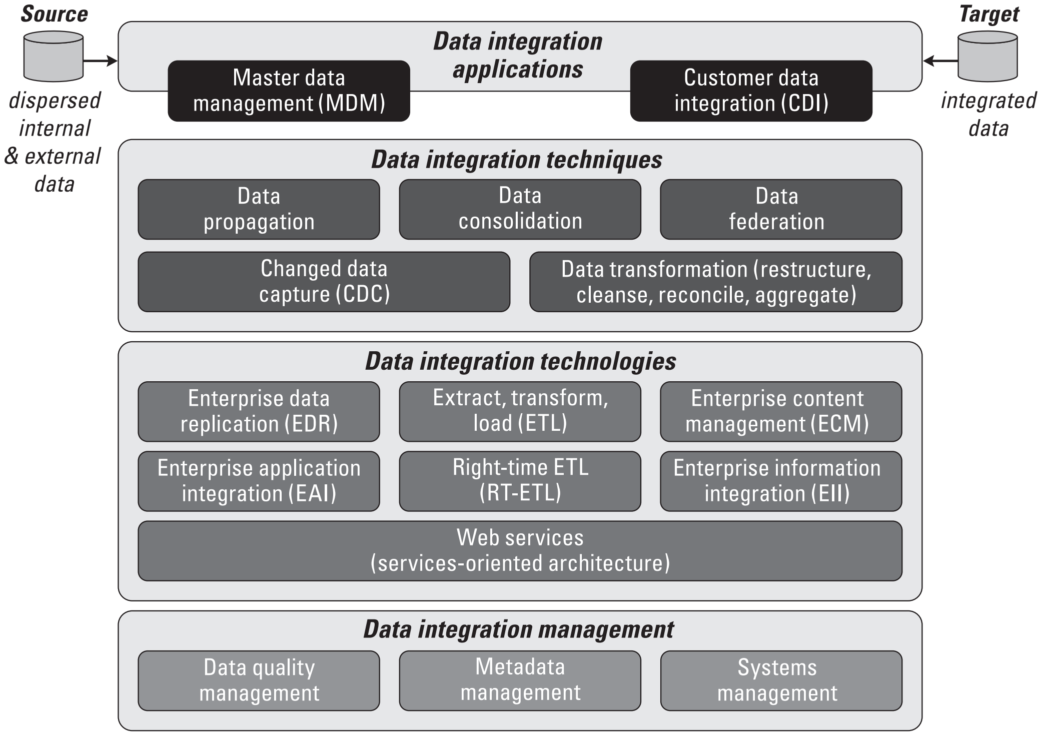

Source
dispersed internal & external data
Data integration applications
Master data management (MDM)
Customer data integration (CDI)
Target
integrated data
Data integration techniques
Data propagation
Data consolidation
Data federation
Changed data capture (CDC)
Data transformation (restructure, cleanse, reconcile, aggregate)
Data integration technologies
Enterprise data replication (EDR)
Extract, transform, load (ETL)
Enterprise content management (ECM)
Enterprise application integration (EAI)
Right-time ETL (RT-ETL)
Enterprise information integration (EII)
Web services (services-oriented architecture)
Data integration management
Data quality management
Metadata management
Systems management

Figure 24. Enterprise data integration architecture (courtesy of BI Research).

Data Integration Strategy Involoves More Than Building an Architecture

There is more to creating a data integration strategy, however, than just building an enterprise data integration architecture. There is also the need to share skills across data integration projects and to capture best business practices.

As business transaction, business intelligence, and business collaboration processing become more intertwined, there will be the need to solve political problems and also possibly to reorganize the IT organization to bring together the various factions involved not only in data integration, but also other enterprise integration disciplines.

Competency Centers. Many companies are developing their enterprise data integration strategies using the services of a data integration competency center. The center's objective is not only to design and support an enterprisewide integration architecture and to provide a shared pool of data integration skills and resources, but also to bring together all of the organization's business integration disciplines into a single group.

Our survey showed that 11 percent of respondents have a data integration competency center, and 28 percent plan to have one. Follow-on interviews indicated that while most companies had created a competency center to provide integration services and support to line-of-business development groups, many also created the center to eliminate the political battles and turf wars over which IT department should deploy new integration applications. Often, these centers not only cover data integration, but also the complete field of enterprise integration and business intelligence.

One company, for example, merged its data warehousing competency group and application integration competency group to form an enterprisewide integration competency center. The company saw a lot of interaction between the two groups, which often provided joint solutions. All of the data warehousing, BI, and EAI analyst and developer expertise is in the combined center, which acts as an integration services sub-contractor to project teams.

Another organization we interviewed, which was using a combination of EAI, ETL, and EII integration technologies, merged its separate integration expertise groups into a common integration services group. This group, however, did not subcontract its services—the costs of the group were instead a component of the overall IT budget.

"Data Integration Has Moved Beyond Data Warehousing"

"Data integration has moved beyond data warehousing," says Informatica's Ivan Chang. "Many of our clients have similar data integration requirements across many projects. These projects involve both data warehousing and non-data warehousing applications. To share best practices and implement a common data integration strategy, companies are creating data integration competency centers. Many of our customers have deployed some form of integration competency center. Several of these have merged data integration expertise into their EAI competency centers. This is especially the case when near-real-time data integration is required."

"Most Global 2000 Companies Have Already Implemented a Data Integration Competency Center"

"Most global 2000 companies have already implemented a data integration competency center of some kind," says TDWI's Philip Russom. "It is important to realize though that this competency center is responsible not only for providing most data integration expertise and development, but also for related activities like data quality, data profiling, and metadata management."

Summary

This study looked at data integration approaches across a wide range of different companies and applications. The study results show that these companies fall into two main groups:

- *Large organizations* that are moving toward building an enterprisewide data integration architecture. These companies typically have a multitude of data stores and large amounts of legacy data. They focus on buying an integrated product set and are interested in leading-edge data integration technologies. These organizations also buy high-performance best-of-breed products that work in conjunction with mainline data integration products to handle the integration of large amounts of data. They are also more likely to have a data integration competency center.
- *Medium-sized companies* that are focused on data integration solely from a business intelligence viewpoint and who evaluate products from the perspective of how well they will integrate with the organization's BI tools and applications. These companies often have less legacy data, and are less interested in leading-edge approaches such as right-time data and Web services.

Where Does Your Company Fit?

In evaluating and applying the contents of this report, it is important to understand which of the two categories your company fits into, and thus how sophisticated a data integration environment your company needs. Nonetheless, many of the ideas and concepts presented in this report apply equally to all companies, regardless of size. The main message of this report is that data integration problems are becoming a barrier to business success and your company must have an enterprisewide data integration strategy if it is to overcome this barrier.

Research Sponsors

Business Objects
3030 Orchard Parkway
San Jose, CA 95134
408.953.6000
Fax: 408.953.6001
www.businessobjects.com

Business Objects is the world's leading BI software company. Business Objects helps organizations gain better insight into their business, improve decision making, and optimize enterprise performance. The company's business intelligence platform, BusinessObjectsTM XI, offers the BI industry's most advanced and complete platform for reporting, query and analysis, performance management, and data integration. BusinessObjects XI includes Crystal Reports®, the industry standard for enterprise reporting. Business Objects has also built the industry's strongest and most diverse partner community, with more than 3,000 partners worldwide. In addition, the company offers consulting and education services to help customers effectively deploy their business intelligence projects.

DataMirror Corporation
3100 Steeles Avenue East, Suite 1100
Markham, ON,
L3R 8T3, CANADA
905.415.0310 or 800.362.5955
Fax: 905.415.0340
info@datamirror.com
www.datamirror.com

DataMirror (Nasdaq: DMCX; TSX: DMC), a leading provider of real-time data integration, protection, and Java database solutions, improves the integrity and reliability of information across all of the systems that create and store data. DataMirror's flexible and affordable integration solutions allow customers to easily and continuously detect, translate, and communicate all information changes throughout the enterprise. DataMirror helps customers make better decisions by providing access to the continuous, accurate information they need to take timely action and move forward faster.

More than 2,000 companies have gained tangible competitive advantage from DataMirror software. DataMirror is headquartered in Markham, Canada, and has offices around the globe. For more information, visit www.datamirror.com.

Collaborative Consulting
877.376.9900
info@collaborative.ws
www.collaborative.ws

Collaborative Consulting is a leading professional services organization dedicated to helping clients optimize their business and technology capabilities. We combine a powerful blend of business knowledge and market-leading technology expertise with an effective partnership approach, allowing us to understand even the most complex business problems. By first assessing an organization's existing technology abilities and business processes with exceptional accuracy, we determine the best solutions and execute flawlessly. Aligning business and technology initiatives enables our clients to achieve superior, cost-effective business solutions. Founded in 1999, the organization serves clients from offices across the U.S., with headquarters in Woburn, MA. Collaborative's Web site is www.collaborative.ws.

IBM Information Integration Solutions
50 Washington Street
Westboro, MA 01581
508.336.3669
www.ibm.com

IBM WebSphere Information Integration
The IBM® WebSphere® Information Integration platform integrates and transforms any data and content to deliver information clients can trust for their critical business initiatives. It provides breakthrough productivity, flexibility, and performance, so clients and their customers and partners have the right information for running and growing their businesses. It helps clients understand, cleanse, and enhance information, while governing its quality to ultimately provide authoritative information. Integrated across the extended enterprise and delivered when the client needs it, this consistent, timely, and complete information can enrich business processes, enable key contextual insights, and inspire confident business decision making.

DataFlux
4001 Weston Parkway, Suite 300
Cary, NC 27513
877.846.3589
info@dataflux.com
www.dataflux.com

DataFlux enables organizations to analyze, improve, and control their data through an integrated technology platform. Through its enterprise data quality integration solutions, companies can build a solid information foundation that delivers a unified view of customer, product or supplier data. A wholly owned subsidiary of SAS (www.sas.com), DataFlux helps customers enhance the effectiveness of their data-driven initiatives, including customer data integration (CDI), enterprise resource planning (ERP), legacy data migration, and compliance. To learn more about DataFlux, visit www.dataflux.com.

Turning integration into insight.™

Informatica Corporation
2100 Seaport Boulevard
Redwood City, CA 94063
650.385.5000
Fax: 650.385.5500
www.informatica.com

Informatica Corporation is a leading provider of data integration software. Using Informatica products, companies can access, integrate, visualize, and audit their enterprise information assets to help improve business performance, increase customer profitability, streamline supply chain operations, and proactively manage regulatory compliance. More than 2,200 companies worldwide rely on Informatica for their end-to-end enterprise data integration needs. For more information, call 650.385.5000 (800.970.1179 in the U.S.), or visit www.informatica.com/.

SAP America
3899 West Chester Pike
Newtown Square, PA 19073
610.661.4600 or 800.SAP.USA
Fax: 610.661.4024

SAP is the world's leading provider of business software solutions. Today, more than 26,150 customers in over 120 countries run more than 88,700 installations of SAP® software—from distinct solutions addressing the needs of small and midsize businesses to enterprise-scale suite solutions for global organizations. Powered by the SAP NetWeaver™ platform to drive innovation and enable business change, mySAP™ Business Suite solutions help enterprises worldwide improve customer relationships, enhance partner collaboration, and create efficiencies across supply chains and business operations. SAP industry solutions support the unique business processes of more than 25 industry segments, including high-tech, retail, public-sector, and financial services. With subsidiaries in more than 50 countries, SAP is listed on several exchanges, including the Frankfurt stock exchange and NYSE under the symbol "SAP."

Sunopsis
6 Lincoln Knoll Lane, Suite 100
Burlington, MA 01803
782.238.1770
info-us@sunopsis.com
www.sunopsis.com

Sunopsis is the leading provider of Simply Faster Integration software, optimized for data warehousing and business intelligence projects. Sunopsis products leverage a business rules approach to data integration and use a unique E-LT (Extract-Load & Transform) approach to their execution. This leverages the existing database engines and executes the ETL processes where they can run the most efficiently, achieving 10 times better performance than traditional ETL tools at a lower cost. Quick to deploy, Sunopsis is the most cost-effective solution on the market today. Visit our Web site to find out why leading companies throughout the world have chosen Sunopsis.

Syncsort Incorporated
50 Tice Boulevard
Woodcliff Lake, NJ 07677
201.930.9700
Fax: 201.930.8290
www.syncsort.com/company/
contact/home.htm

Syncsort Incorporated is a leading developer of high-performance, data management, data warehousing, and data protection software for mainframe, UNIX, and Windows environments. For more than 35 years, Syncsort has built a reputation for superior product performance and technical support. An independent market research firm has named Syncsort one of the top Data Warehouse 100 Vendors for seven years in a row. In addition, Syncsort was recently chosen as the leading provider of data acquisition and integration products. Syncsort's products are used to protect data in distributed environments, speed data warehouse processing, improve database loads, and speed query processing.

THE DATA WAREHOUSING INSTITUTE
A DECADE OF EXCELLENCE

5200 Southcenter Blvd.
Suite 250
Seattle, WA 98188
Phone: 206.246.5059
Fax: 206.246.5952
E-mail: info@tdwi.org
www.tdwi.org