Delivering information you can trust
June 2006

IBM Information Management software

Deploying integration services with IBM Information Server

ON DEMAND BUSINESS™

Contents

To compete effectively in today's increasingly crowded marketplace, companies are shifting their focus to reducing costs and increasing sales through business agility, responsiveness and timeliness of information. To succeed, businesses need to coordinate activities among analytical, operational and transactional systems. Historically, this process has been very challenging, hindered by both technological and organizational obstacles.

Much of the recent and well-documented innovation and investment by businesses to reduce latency has been centered on what the analyst firm Gartner calls the "real-time enterprise.[1]" Despite the recent industry buzz, the concept is not new. It has been fundamental to information technology from the time of its inception. What has changed are the underlying technologies and the renewed emphasis on delivering information in a form that enables businesses to react quickly—whether the response is automated or manual. The timing of the delivery is not always true real-time, but rather "right-time"—tailored to the appropriate response requirements of the information.

Traditionally, response to business events has been predicated on the ability to recognize these events within the semantics and the scope of transactional and operational systems. But this approach has a fundamental problem: The source systems are limited in their ability to recognize data or events that might be interesting to the business, particularly when that data transcends the boundaries of a single system. Business process management has enhanced the ability to recognize business events by providing a broader cross-application scope and some degree of shared semantics. Although this new technology can help define and recognize new categories of business events, large categories of information—and thus potential business events—are still not being considered.

Latency is typically defined as the elapsed time between the occurrence of a business event and an appropriate action or response to the event. So the key to reducing latency is tied to improving the ability to recognize and respond to business events. Two related components—both process issues—work together. The first is in the time lapse from when a business event occurs to when information on the event is recognized; the second occurs with the time lapse from when the event is recognized to when action is taken.

Meanwhile, data-centric technologies have focused on providing context to the data that originates from the transactional and operational systems. This context is provided by decoupling the data from the source systems, organizing it together in new ways, reusing the new information in consistent and standard procedures and looking at its changed behavior across time.

The interesting dichotomy is the fact that most companies separate their data- and process-centric efforts, with little or no sharing of knowledge, technologies or approach. The organizational groups that control the transactional and operational systems are often completely separate from those that control the analytical systems, with each side using different tools and technologies.

The data-centric approach focuses almost entirely on creating information in a format that enables the business to clearly recognize important and interesting events, while relegating the response to delayed and mostly manual efforts. The process-centric approach focuses on shortening response times to the limited events it can recognize. Clearly, the optimal scenario would involve harnessing the broader and more business-tailored information from the data-centric approach to allow the process-centric approach to respond to it.

This paper provides a road map for IBM DataStage® customers to leverage their existing software and skill set investments to achieve this goal and gain greater business agility by deploying information as a service to the business in a service-oriented architecture (SOA).

Coordinating analytical, operational and transactional systems to improve real-time recognition and response

Traditionally, analytical data has been created in batch because the data structure differs significantly from transactional and operational systems; it also is optimized for specific types of analysis. Because this structure is complicated, creating it becomes a very complex and processor-intensive transformation task. In addition, the data is usually derived from multiple source systems, requiring more matching and merging to arrive at a unified view. It also is viewed in multiple dimensions, such as across time, to provide additional context to the information, which adds processing requirements. With this amount of processing, the analytical data must be stored separately and the extraction and transformation processes must be completed during batch cycles.

Beyond this data latency issue, another factor is the separation of groups within the organization that control the systems. Those that control transactional and operational systems differ from the groups that control analytical systems. The process side of the organization does not have the access knowledge and technologies for getting to analytical data and the analytical technologies do not support the same standards. Therefore, process-oriented technologies do not have the ability to include analytical data transformation logic as a component of an event process or transaction.

To overcome these issues, analytical data and data creation routines need to be published to the process-centric groups using technologies and standards that fit into their enterprise architectures and do not compromise their real-time performance requirements. By deploying information services using standards such as Enterprise JavaBeans™ (EJBs), Java™ Messaging Services (JMS) or Web services—that is, deploying information as a service—those with access knowledge to the analytic systems can facilitate the reuse of standardized integration procedures by application developers as they create or enhance new and existing business processes.

Leveraging IBM Information Server in a service-oriented architecture

IBM Information Server delivers all of the functions required to integrate, enrich and deliver information you can trust for key business initiatives. With rich functionality, broad connectivity to heterogeneous sources and unified metadata, it provides a strong foundation for the enterprise information architecture. Information Server removes traditional barriers between data-centric and process-centric projects and provides a way for analytical processing to be harnessed within real-time business processes. The net result is the ability to automate response across the broadest scope of recognized business events.

With the IBM Information Services Director product module (formerly RTI) of the Information Server, data-centric processing can occur in real time. This real-time processing is possible because of the unique combination of massively parallel processing (MPP) capabilities of DataStage and the high volume connectivity of the Java Platform Enterprise Edition (J2EE™) foundation in Information Server. By leveraging these capabilities, Information Server is capable of handling the transformation and data quality processing demands of analytical data creation, and the request volume and time sensitivity of real-time processing. Information Services Director publishes complex data transformations and access routines as Web services, EJBs or JMS objects that can be called by Java or Microsoft® .NET application infrastructures, enterprise application integration (EAI) platforms, business process management tools, or most products and technologies used within the transactional and operational arenas.

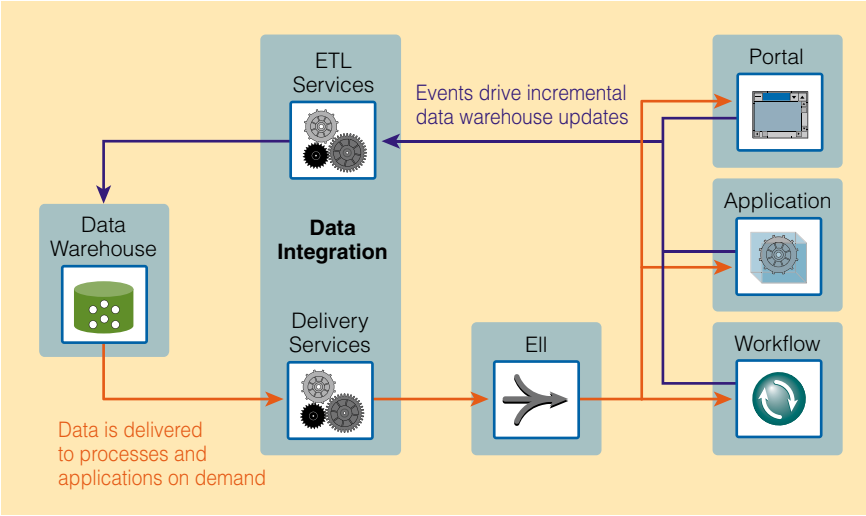
IBM customers are using Information Server capabilities to blur the lines between analytical and transactional/operational data. The result: availability of the best data at all times, to all people or processes. IBM sees four general categories for using Information Server.

On demand data warehousing

The value of analytical data for use in managerial decision making and as operational instruments for business processes and applications has led to increasing pressure to reduce the latency of this data. At the recent Gartner Symposium ITxpo 2006, Gartner stated that "Data warehouses are moving quickly from being passive, offline, after-the-fact reporting structures that are disconnected from business processes. They are gradually becoming deeply embedded in the business process structure, used for real-time decision making and collaboration, and are directly interactive with operational systems."[2]

On demand data warehousing takes advantage of Information Server to help reduce or eliminate data latency within the warehouse and publish data from the data warehouse into services that are easily consumed by applications and processes, as shown in Figure 1. As data is created within source systems, Information Services Director can help push this data into DataStage where it is transformed and populated to operational stores, warehouses and datamarts on an event-driven basis, triggered by applications, EAI or business processes. Companies can then take advantage of the best-in-class transformation and processing capabilities of DataStage without waiting for a batch window. Data experts can publish services for their data that are easily consumed by applications and processes, without requiring application developers to understand the complex schemas and sources associated with the data warehouse.

Figure 1: On demand data warehousing with IBM Information Server



One unique capability of Information Services Director is to leverage existing DataStage job logic: the logic for data transformation can be reused rather than being re-created. In fact, Information Services Director allows jobs to be designed without any knowledge of how they will be accessed. Job developers do not need to know about JMS queues, SOAP message headers or EJB structures. The ability to reuse this logic enables a quick and smooth transition to on demand data warehousing using DataStage, without extensive rework or redesign.

With less inherent data latency across the spectrum, DataStage can provide data integration services across analytical and transactional/operational stores. Analytical and operational data can be brought closer together, allowing automated operational decision-making to leverage richer analytical data. These data integration services can be easily called from any application, portal or development project. They also can be called from enterprise information integration (EII) platforms to provide advanced data matching and transformation services to federated queries. Once these services are in a familiar service-oriented structure, external applications and development teams are more likely to use them to get the best available data and to access it in a standard way.

One example is a leading pharmaceutical company that uses the IBM platform to reduce latency in operational and analytical data stores. By using Web services interfaces within DataStage processes, the company can access various data sources both within and outside the enterprise. Web services calls from its application middleware also trigger DataStage processes. Reduced latency of its operational and analytical systems facilitates better decision making. By allowing its partners access to this information through secure Web services, the company can provide its customers and suppliers with an on demand view of information that is vital to optimizing their supply chain.

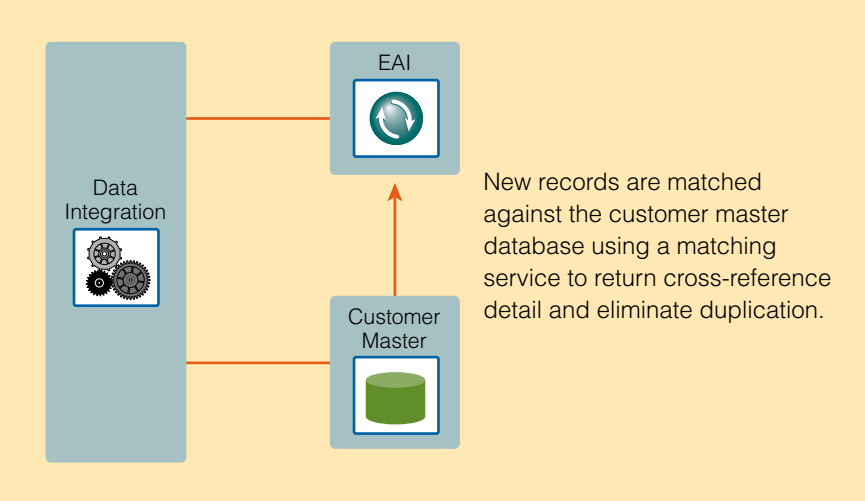
Master data management

When trying to reduce process latency, companies discover early that their most vital data is often stored across many systems, with little or no consistency. This discovery forces development projects to go to extraordinary lengths to reach the correct data sources and rationalize the data into a single de-duplicated semantic representation. In most cases, it also means that applications and users never have a complete picture of the data. A common example is customer marketing and customer service initiatives, where obtaining a single view of the customer remains an elusive goal for most organizations.

Master data management goes a step beyond on demand data warehousing by creating authoritative sources of common reference data that can be used throughout organizational operations. The types of data typically targeted include customer, product and inventory data. Enterprises often choose this common data because it is frequently accessed across many applications and the consistency of this data is very important to the business. Creating these master stores improves the consistency and reliability of information for everyone and allows new development efforts to reuse proven, standard access mechanisms rather than re-creating them.

In this scenario, DataStage is used to create a master database that is kept up to date with Information Services Director as transactions flow through source systems. The master database can house either complete matched records of reference data, or it can simply contain a cross-reference table of identifiers from the various source systems. In the latter case, when a resource requires the reference data, DataStage—or an external technology like EAI or EII—dynamically assembles a complete record from the source systems using these identifiers. See Figure 2.

Figure 2: Master data management with IBM Information Server



To populate the master database, Information Server utilizes both transformation and best-in-class data matching technologies of DataStage and QualityStage (both product modules of Information Server). These matching routines are typically used in batch during the initial data population, and then reused in real-time to match new records as they are entered into source systems. QualityStage is also used to help ensure that data is centrally standardized, resulting in more consistent and accurate data. DataStage can easily involve systems—such as legacy mainframe systems—that are traditionally very difficult to reach, providing a complete organizational view of matched reference data.

The master database provides not only a single place to get reference data, but it also helps reduce the burden on development projects, which are often hampered by the difficulty in understanding and accessing the different data sources. Even when developers know the data location and can get to it, the task of matching and merging records is often very tedious and difficult. DataStage allows data managers—the people who know the data best—to focus on these tasks and easily create reusable information services that development projects can leverage.

Even better, because these services are based on a service-oriented architecture, they abstract the underlying complexities of data sources and integration from the developer. Developers are presented with business-level services that do not require them to be experts in source systems or data. In fact, when developers use these services, they do not need to know anything about Information Server. This insulation of application development tasks from data management tasks best leverages the competencies and skills of each group and leads to more unified and standard systems.

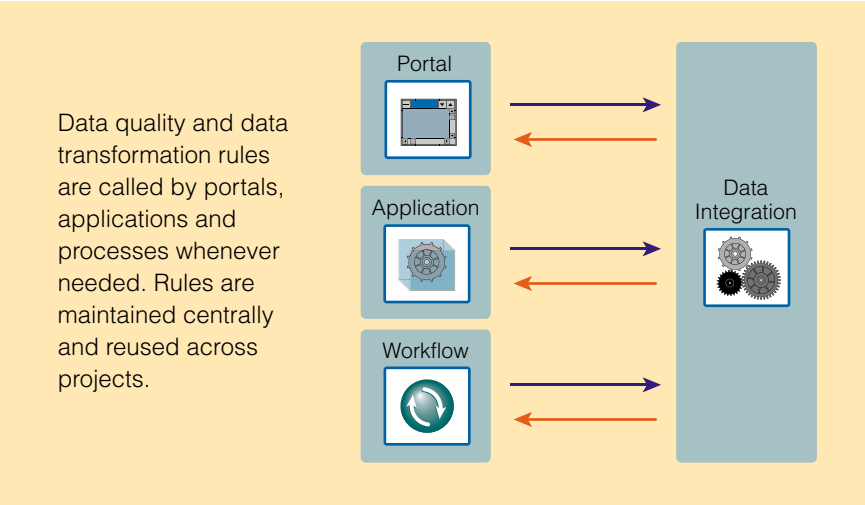
An excellent example of master data management can be found in one of the world's largest technology companies that uses the IBM platform to consolidate customer information from multiple internal systems. The goal of providing a single point of reference for the authoritative data about each customer was met by first rationalizing and profiling the source systems, and then merging the existing data into a single master store. For ongoing data entered into various systems, Information Services Director with QualityStage runs the same standardization, matching, and transformation processes to control information at the point of entry. With the new database in place, the company has a better understanding of its customers across all of its products, services and locations. This database can also lead to an improved level of support, and better tailored marketing and sales activities to the customer base.

In-flight enrichment

As companies develop new applications, the requirements for cross-system data validation, standardization, matching and transformation continuously arise. Much of the required logic has already been built into existing data warehousing applications. The cost to re-develop these processes in other technologies is very high, requiring high-level developers and producing another maintenance point when data structures eventually change.

In-flight enrichment publishes a more granular set of Information Server capabilities as services that can be leveraged by development efforts. This expansion on the master data management theme offers standard services for validation, standardization, matching, transformation and enrichment. Organizations publish these services to foster reuse and encourage consistent data standards. Development projects can then reuse these services rather than developing them from scratch. For example, to create a portal application, a development team can simply reuse an address validation routine published as an in-flight enrichment routine through the Information Services Director rather than re-create a routine, shown in Figure 3.

Figure 3: In-Flight enrichment services with IBM Information Server



Information Services Director creates in-flight enrichment services to publish discreet Information Server jobs for specific data. These services can be either granular in nature like a simple address validation or they can be more comprehensive, such as a service that validates, standardizes and inserts a customer record. Much of this logic already exists in analytical applications; however, it is not in a form that it can be leveraged by external applications. With Information Server, this logic can be exposed as services, allowing it to be utilized universally.

In-flight enrichment helps reduce the development burden and increases data consistency. Since these services are based on the common standards of a service-oriented architecture, they are very easy to incorporate into applications. From a development perspective, the services are available in a searchable directory, effectively creating a library of data integration components that can be reused from project to project. For ongoing maintenance, changes to these routines do not necessitate changes to the calling applications; therefore, these services can to be maintained separately without involving the applications teams.

Data consistency is improved because all applications share the same mechanisms for validation and standardization. All data entering the database goes through the same set of rigorous requirements, so the chances of errant or mismatched data are reduced.

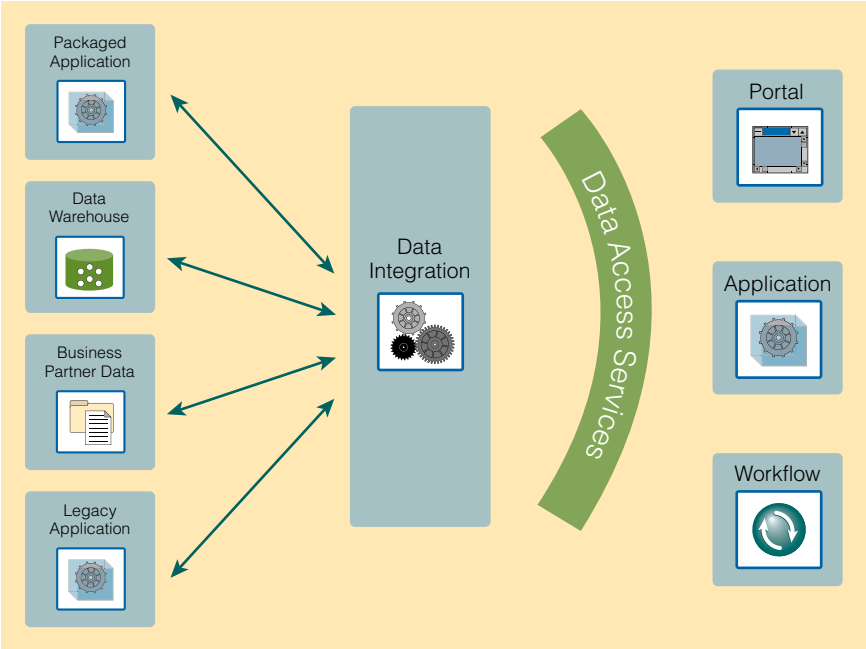
An example of in-flight enrichment in practice is illustrated by an international insurance data services company that uses the IBM platform to provide Web services–based validation and enrichment of property addresses. As insurance companies submit these address lists to the data services company for underwriting, real-time services standardize the addresses based on their rules, validate the accuracy of the address, match the addresses to a list of known addresses and enrich the addresses with additional information to assist in underwriting decision making. As a result, the company is able to automate 80 percent of the process of property address research and eliminate errors typically associated with it. This process was made easy using the publishing capabilities of Information Services Director and the standardization and matching capabilities of QualityStage.

Enterprise data integration services

Within many organizations, the number and complexity of applications that require access to accurate and trusted information is continuously on the rise. In most cases, limited reuse among these applications means that much of the effort, particularly in data access mechanisms, is reproduced in each project. The result is high development and maintenance costs and inconsistencies in accessing the data. Data managers quickly lose control of the data and can no longer enforce data standards or ensure data quality.

In the enterprise data integration services scenario, businesses create a standard data integration layer within their enterprise architectures to improve consistency, reuse and control (see Figure 4). This simple extension of the logic behind in-flight data enrichment expands the scope to all data, which helps to ensure consistent delivery of information as a service to

Figure 4: Data integration services with IBM Information Server



the business. The mechanisms and benefits are essentially the same, only magnified to encompass a larger data set. The primary difference is true separation of data management and application development. Each group can focus on its core competencies, using familiar tools and technologies without restricting or inhibiting the abilities of the other group. Each group also can make changes independently within its scope, without affecting the other group. When changes do need to cross the boundaries, metadata linkages provide an end-to-end understanding of the impact of change, which helps to ensure that nothing is missed.

Enterprise data integration services truly delivers on the promise of service-oriented architecture, taking full advantage of the inherent capabilities to deliver information as a service to businesses. It also leverages the capabilities of Information Server. The net result is better standardization and consistency, reduced development cycles and project risk, lower cost and risk in maintenance, and overall better and more available information for business users.

Enterprise data integration services are particularly pertinent to the burgeoning area of regulatory compliance and data governance. Many regulations involve understanding the content and lineage of data, and controlling how the data is used. For example, Sarbanes-Oxley includes a strong focus on how financial data was calculated and where it came from. This presents an ideal scenario for applying enterprise data integration services. Not only do these services help ensure a consistent enforcement of logic to data, but they also provide a metadata audit trail for the data—where it came from and what happened to it along the way. In addition, policies can be applied to the services to help ensure that the usage and security of those services is appropriately governed throughout the enterprise.

An IBM customer—one of the largest automotive manufacturers in the world—is already implementing enterprise data integration services. The automaker is consolidating all of its data integration logic into a common services layer, which can be reused across its EAI infrastructure and ETL (extract, transform, load) processing and can be called directly from corporate applications. This services layer simplifies data integration tasks to a simple service call and provides a library of reusable components, thereby reducing the development burden on individual projects. It also improves data consistency because the same integration logic is applied, regardless of the project. Information Server greatly simplifies the task of accessing and integrating these data sources using a visual development paradigm rather than hand-coding. The result is a tangible reduction in project development effort and a higher level of data consistency across the enterprise.

Getting started: an incremental approach

Although the examples described may seem like large-scale undertakings, each was implemented on an incremental basis, with iterative project goals that produced short-term results and reusable services. In most cases, these iterations produced tangible results within days or weeks, quickly showing a positive return on investment.

A successful incremental approach is predicated on selecting the appropriate starting point. Many customers begin by adding real-time capabilities to an existing data warehouse implementation, leveraging existing ETL logic. Other customers focus on specific opportunities related to master data management or in-flight enrichment. In any case, it is important to choose a manageable unit of work that can quickly demonstrate business value aligned with your ultimate objectives.

A good place to start is to simply ask the question: "Would my operational processes benefit from having better access to pieces of my analytical data?" Often, multiple pieces of analytical data would be valuable to operational systems, but in the past they were not easy to access. Business users can quickly identify these data elements, so they become great resources to identify the starting point. Publishing services that provide this data to operational processes whenever they need it can be extremely valuable to the business. Managers of analytical systems, business intelligence systems and data warehouses can approach application integration groups with these services and offer them as valuable components to help in operational processing. Once the initial service-oriented relationship begins, additional areas of synergy will naturally flow from it.

QualityStage can offer another good starting point, particularly for in-flight enrichment scenarios. For example, existing validation and standardization routines built into QualityStage can often be reused in Web applications, portal projects or application development projects.

The first project should focus on a discreet number of related capabilities that can be delivered in a short time, usually within 30 days. It should also attempt to leverage existing transformation, and deliver data cleansing or data federation logic as much as possible. Subsequent projects will expand on this baseline to provide a more complete set of functionality.

The advantages of improving your business agility are clear, while the means may sometimes seem difficult to achieve. Business agility can only be improved by reducing both the time it takes to recognize business events and the time it takes to respond with an appropriate or better defined reaction. It requires coordination among analytical, operational and transactional systems, which historically has been hindered by both technological and organizational obstacles. IBM Information Server can assist in overcoming these obstacles by providing the technological capabilities required to bridge these systems and by publishing the results as services that application developers will want to use.

Next steps

IBM DataStage, QualityStage and Information Services Director are available as product modules of IBM Information Server.

For more information about upgrading to Information Server or for information on other IBM Information Management product or services offerings, contact your local IBM representative and ask about our IBM Workshop programs for SOA, or visit our Web site at **ibm.com**/software/data/integration

IBM Software Group
Route 100
Somers, NY 10589

U.S.A.
Produced in the United States of America
June 2006

[1] Flint, David and Mark Raskino. *The Real-Time Enterprise: Key Issues for 2003*. Gartner. January 2003

[2] Beyer, Mark. *Designing and Implementing the Multiuse Data Warehouse.* Gartner, Inc. Gartner Symposium ITXpo. May 2006.

GC18-9725-01