Delivering information you can trust
May 2007

IBM Information Management software

# The hybrid approach to data warehousing for maximum flexibility

Contents



## Hybrid data warehousing: Flexible on demand access to business information

The concept of a hybrid data warehousing environment challenges the most fundamental tenet of data warehousing: The world of information processing must always be split into two parts—operations and information. In the early days of data warehousing, this split was clearly advantageous for IT because it protected the operational systems from the performance and security impact of ad hoc queries. On a broader scale, it also suited a business model in which different business functions operated in distinct and largely independent silos. However, current business trends and the introduction of new data integration techniques make a different approach both necessary and possible.

As illustrated in Figure 1, the strategy for managing data in a data warehouse has evolved in recent years. The information goals of early analytic systems were to provide information that could help companies better understand the overall position and performance of the business at a given point in time. Today, users look to bolster business processes within the context of knowing how to react to real-time events.

*Figure 1: Historical perspective of the data warehouse*

| Yesterday | Today |
|---|---|
| Point-in-time business intelligence | Real-time business intelligence |
| Batch data warehousing | Dynamic data warehouse environment |
| Separation of data warehouse and transaction systems | Consolidation of data warehouse and transaction systems |
| Self-contained historical data warehouse | Information integrated with other sources |
| Latency in development and deployment of business applications | Speed of delivery and development is critical |

In the on demand world, the phrase "what have you done for me lately" captures the prevailing attitude of customers toward providers of goods and services. Increasingly, information systems and automated business processes are literally running businesses. The future of business systems lies in the ability to address real-time data needs and to deliver that data as reusable packages in which data has been thoroughly cleansed, reconciled, transformed and delivered to be consumed by these business processes. A key component in this Information On Demand revolution is the *hybrid data warehousing environment*.

### The challenge: Provide reliable, on demand data access to decision makers

Decision makers today are increasingly impatient for information required to support their business decisions. Monthly reports are now weekly, daily and in some cases, hourly. It is unacceptable to have delays of weeks—or even days—in gathering, massaging and consolidating information. Meanwhile, the volume of data flowing into the warehouse continues to grow as the business is tracked with increasing regularity. In addition, around-the-clock business activity is shrinking or eliminating the batch windows that were typically used to load data and repopulate the data warehouses.

Now obvious, traditional data warehousing architecture cannot meet all of the business needs and data access required by increasingly cross-functional business initiatives. The architecture is particularly stressed by requirements for closer to real-time decision support, which can easily overwhelm the data warehouse with immediate, voluminous, but seldom-used details.

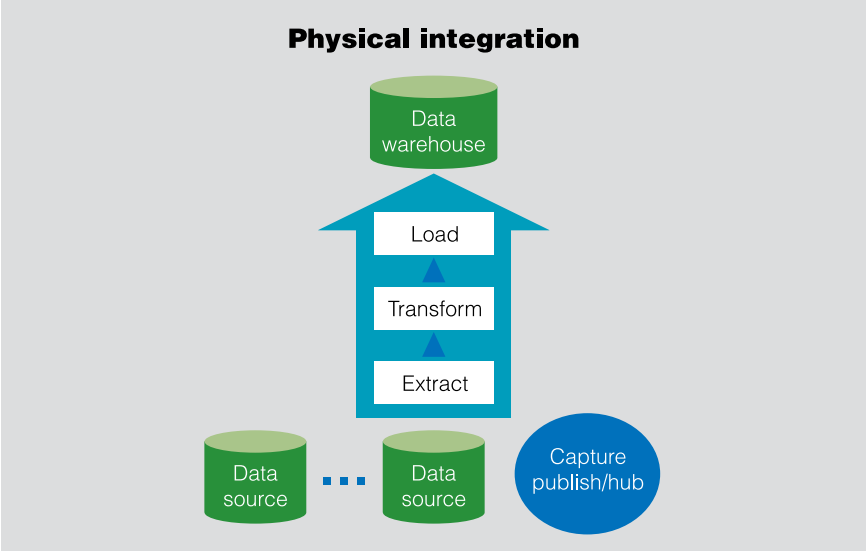
### The solution: IBM WebSphere Federation Server

To meet this challenge, companies must devise a strategy to make the right data available for use in a more accurate, actionable and timely manner. Inevitably, this means accessing data from various operational systems, reconciling the differences that exist among systems and bringing data together using physical or virtual integration techniques. Previous integration approaches offered by vendors have been largely an either/or proposition: physical integration or virtual integration. With the introduction of IBM® Information Server, including the IBM WebSphere® Federation Server module, a third alternative using the combination is now a practical reality.

### Physical integration consolidates data into single source

Physical integration enables users to consolidate data into a single data source. By moving data from multiple sources to one or more target databases, users can support the business intelligence function within their organizations. Replication helps synchronize copies of the data. The extract, transform and load (ETL) technique is the most common way to create data warehouses. As shown in Figure 2, ETL enables new data sources to be created and existing databases to be populated on an ongoing basis. These tools are designed to perform very complex transformations in volume, assimilating data from multiple data sources. ETL is often coupled with changed-data capture and event publishing tools to enable users to detect changes and apply only those changes since the last load activity.

*Figure 2: Physical integration creates new data sources*



A key benefit of this approach is the ability to predict performance and to create specific physical data layouts, pre-aggregations and/or restricted volumes of data. Figure 3 shows some of the attributes of physical integration.

ETL forms the backbone of today's business intelligence infrastructure. A central data warehouse based on a physical integration alternative is likely to provide the best performance for predictable queries. Another key driver for physical integration is that many organizations or departments believe they need to control local copies of important data to assure access or quality.

*Figure 3: Physical integration attributes*

| Characteristics | Strengths |
|---|---|
| • Single container for data<br>• Data cleansed and optimally structured<br>• Periodic, batch-oriented (not intended for real time)<br>• Specialized GUI tools that enable complex tasks<br>• Fixed scheduling | • Scheduling enables reliable point-in-time reporting<br>• ETL process helps eliminate data quality issues and standardize data formats<br>• Complex data transformations requiring calculations, aggregations or multiple stages can be performed<br>• Query performance is predictable and separation from production databases eliminates contention for machine resouces<br>• GUI-based tools help increase productivity |

| Barriers to physical integration | Data challenges |
|---|---|
| • Budget<br>• Resources<br>• Time<br>• Ownership<br>• Organizational issues | • Too big<br>• Too ad hoc<br>• Too temporary<br>• Too proprietary<br>• Too recent |

Physical integration is often not an option or requirement. For example, the creation of a new information database requires significant cooperation from the owners of the data sources—and they may be unwilling or unable to provide this level of cooperation. Another factor may be the prohibitive cost of either new databases or additional data placed into the data warehouse that increases disk space requirements, maintenance and other related costs.

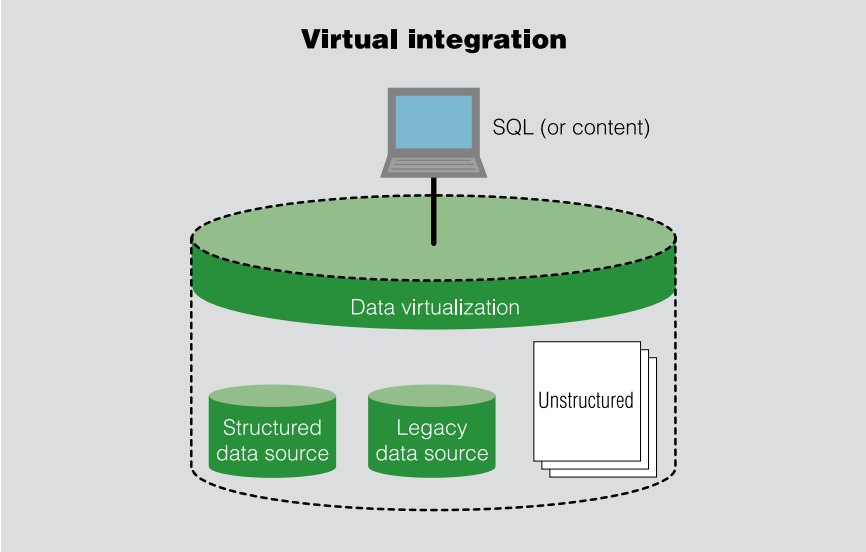
When physical integration is not a requirement or option, yet simultaneous data access from multiple sources—heterogeneous or homogeneous—is necessary, virtual integration is a viable and flexible option.

### Virtual integration enables simultaneous data access across the enterprise

Virtual integration, sometimes referred to as database federation or enterprise information integration (EII), provides virtualized integration across all enterprise data. It enables simultaneous access to diverse data and content sources as if they were a single source—regardless of where the information resides—while retaining autonomy and integrity of the data and content sources. Using IBM WebSphere® Federation Server, access to data is available anywhere in your enterprise—no matter where it resides and regardless of vendor—without creating new databases and without disruptive changes to existing ones. IBM WebSphere Federation Server is a virtualization technology that enables users to access multiple, diverse data sources simultaneously as if they were a single source.

Virtual integration, shown in Figure 4, uses logical views of locally stored data or federated views of remote data sources. To end users and client applications, data sources appear as a single collective database. Users and applications interface with the federated database managed by WebSphere Federation Server.

*Figure 4: Virtual integration enables access to enterprise data*



Query performance is an important element of federated technology. WebSphere Federation Server employs a powerful caching technology that provides a way to bridge the performance gap when directly accessing and joining remote data sources. Caching enables remote data to be available locally without the need to access the remote data source over the network. While not a panacea for all data access, the virtual integration approach is viable with acceptable performance under reasonable conditions.

The fundamental difference between virtual integration and physical integration (ETL) is that virtual integration is applied at the point where the data is being consumed, not as a preparatory step. It interfaces directly with information applications such as report writers, portals and analytical tools.

Data virtualization has several strengths, as noted in Figure 5; however, it is important to understand some of the issues associated with federated access before it is deployed. For example, it relies on joining tables based on common identifiers such as customer number or product ID. ETL products excel at reconciling differences that exist among descriptions of business entities and can use transformation and cleansing techniques to standardize these identifiers. Therefore, successful virtual integration often requires that data be prepared for consumption at some stage using physical integration techniques.

*Figure 5: Virtual integration attributes*

| Characteristics | Strengths |
|---|---|
| • Accesses data where it resides<br>• Offers dynamic joining of data<br>• Can compensate for weaknesses in data source capability<br>• Enables mixed relational and non-relational data<br>• Delivers data to the consuming application | • Relational access to non-relational sources and compensation for data sources not having certain database features<br>• Ability to explore data before a formal data model and metadata are created<br>• Rapid deployment<br>• Ability to be directly used by many tools and applications<br>• No unnecessary data duplication |

A virtual data warehouse alternative allows data to remain with the owner, improving organizational buy-in.

### Hybrid data warehouse leverages best of physical and virtual integration

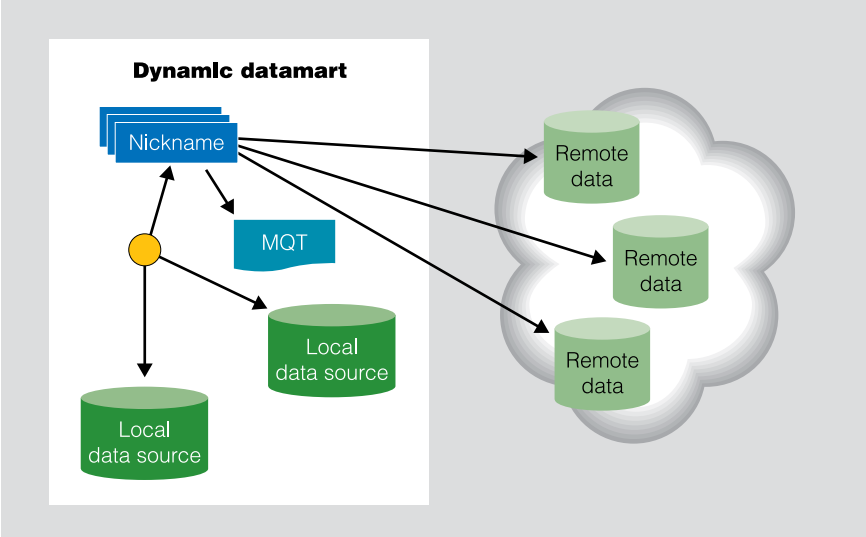
The concepts of a hybrid data warehousing environment include:

- ***Data warehouse***—*A database geared toward the business intelligence requirements of an organization. The data warehouse integrates data from various operational systems; data is typically loaded from these systems at regular intervals. Data warehouses contain historical information that enables analysis of business performance over time.*
- ***Datamart***—*A scaled-down version of a data warehouse that focuses on a particular subject area. The datamart is usually designed to support the unique business requirements of a specific department or business process.*

- ***Virtual datamart***—*A set of views defined on an enterprise data warehouse. These views appear to the user as a separate, self-contained database organized for one specific purpose.*
- ***Hybrid data warehouse/datamart environment***—*An environment that combines physical integration with virtual integration techniques. The hybrid data warehouse enables dynamic views of data sources that focus on a particular subject area.*

The resulting product—a hybrid data warehouse or datamart—shown in Figure 6 is a hybrid of a physical data store combined with federated views over remote data sources. It leverages the reliability of physical integration with real-time access characteristics of virtual integration.

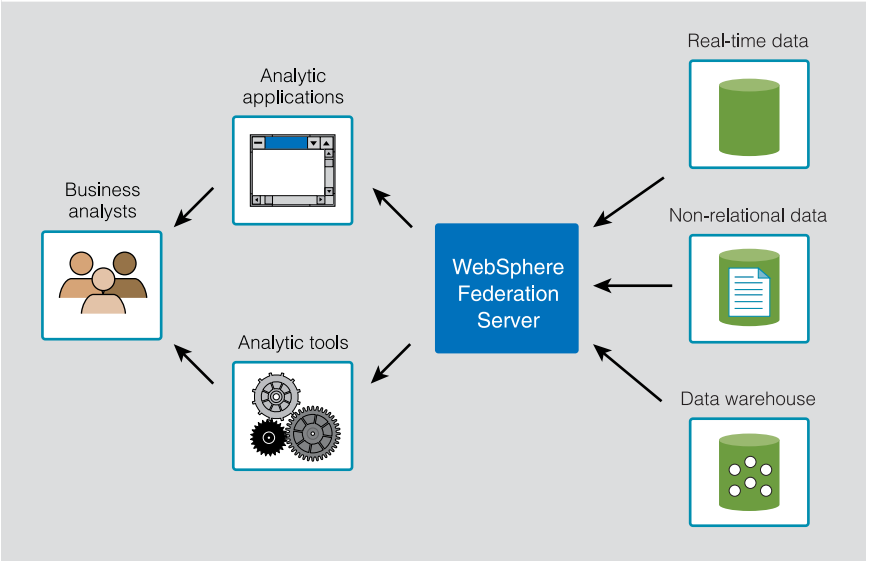
*Figure 6: Hybrid data warehouse*



The success of data warehousing and ETL processes has contributed to the rise in the number of data warehouses and operational data stores. This has created options for virtual integration of solutions within the post-ETL world of cleansed and reconciled data stores, shown in Figure 7.

*Figure 7: Extending the data warehouse*

| Customer challenge | Customer value |
| --- | --- |
| • Integrating information from multiple data and content stores for more comprehensive reporting<br>• Performing analysis that combines historical and real-time information<br>• Rapidly deploying new applications that require data outside existing data warehouses, such as legacy stores | • Better decisions from more comprehensive and more up-to-date information<br>• More responsive to business needs that require access to additional data sources<br>• Increased ROI from existing data warehouses |



While potentially providing the best of both worlds, the hybrid data warehouse is not without limitations. The combination of physical integration (ETL) and virtual integration (federation) requires employing the best practices of two different technical disciplines to work out the right trade-offs in how data is best prepared and delivered to the user. See Figures 8 and 9.

*Figure 8: Customer data integration*

| Customer challenge | Customer value |
| --- | --- |
| • Provide a holistic view of information for customer-facing or customer-supporting applications<br>• Reduce high development and maintenance costs to access diverse data sources<br>• Maximize value of customer data for customer satisfaction/retention and increasing sales | • Reduce coding and skills requirements when integrating two or more sources<br>• Reduce redundant data by consolidating only frequently accessed data<br>• Help reduce application maintenance costs<br>• Extend customer data warehouse with document and other content data |

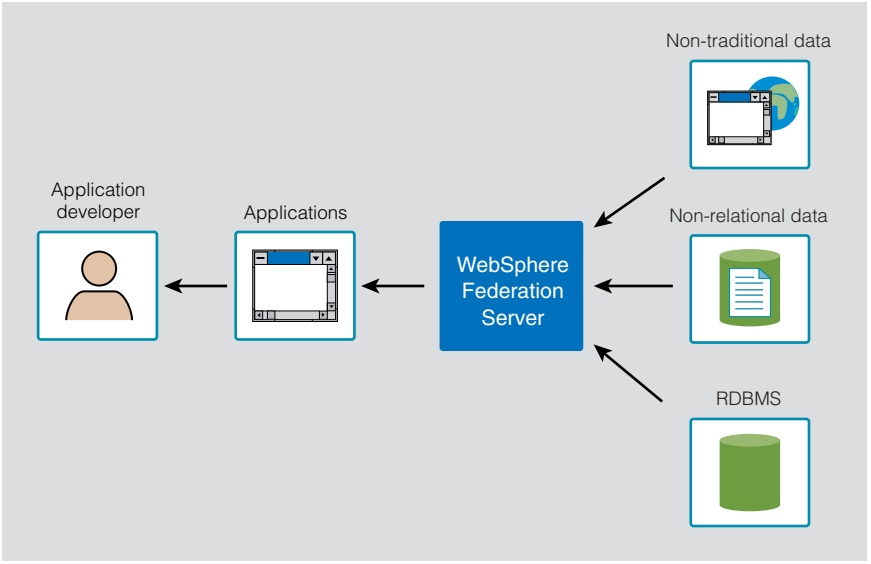
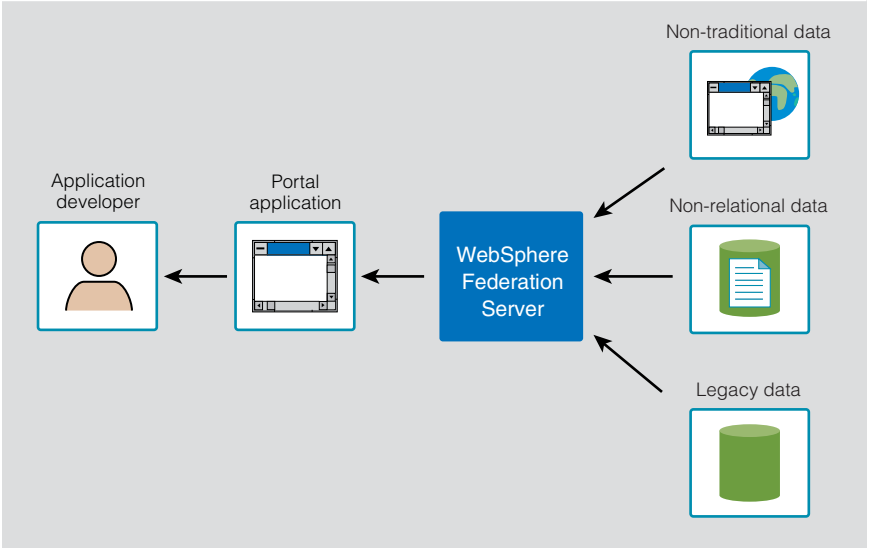


*Figure 9: Speeding portal application development*

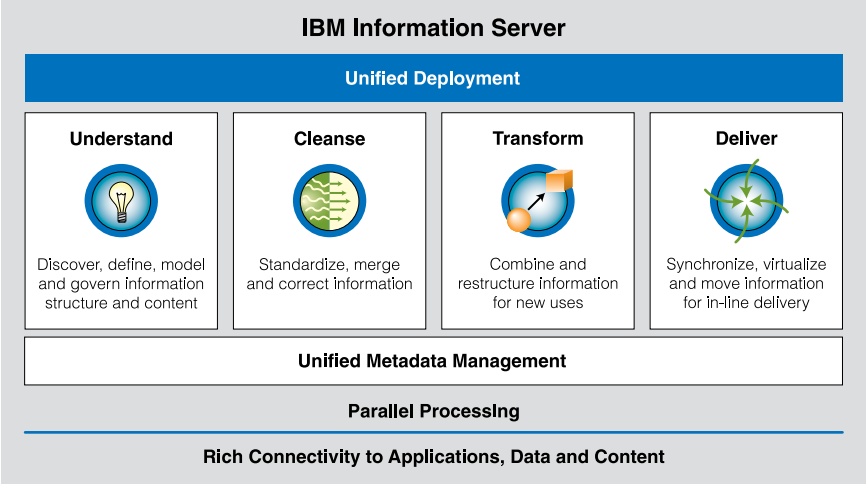
| Customer challenge | Customer value |
| --- | --- |
| • Integration of multiple data sources into a single application is complex and costly<br>• Accessing non-traditional sources is too impractical to leverage their benefit<br>• Time pressure to deploy new applications<br>• Scarcity of the skillset required to work with legacy, non-traditional data sources<br>• Extending built-in search to new domains | • Reduce amount of integration coding by 40–65 percent<br>• Use existing SQL tools to access all data<br>• Give applications access to all the relevant data sources<br>• Help reduce application maintenance costs<br>• Deploy existing skills over wider range of integration projects |



### IBM Information Server enables hybrid data warehousing

Sophisticated data warehousing customers understand that virtual and physical integration techniques can be combined. But in the past, business intelligence tools vendors or data connectivity specialists have focused on virtual integration. Likewise, ETL has been considered a separate discipline distinct from the job of delivering data to the user. Today, IBM Information Server breaks through these artificial barriers to provide a single toolset that makes the hybrid data warehousing environment possible. See Figure 10.

*Figure 10: IBM Information Server*

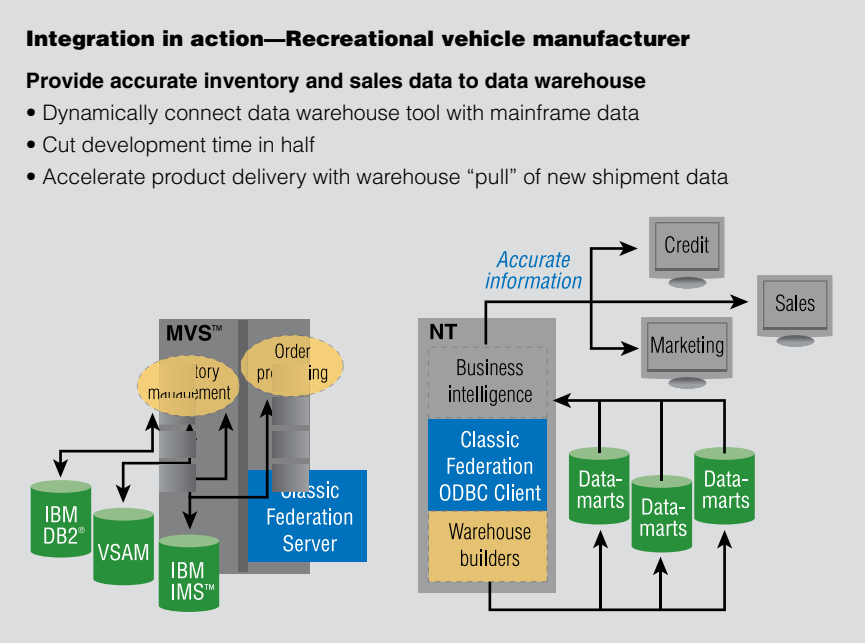


IBM Information Server features a unified set of product modules that help solve multiple types of business problems. Information validation, access and processing rules can be reused across projects, leading to a higher degree of consistency, stronger control over data and improved efficiency in IT projects. As shown in Figure 10, IBM Information Server transforms and enriches information to help ensure that it is in the proper context for new uses.

Hundreds of prebuilt transformation functions combine, restructure and aggregate information. Transformation functionality is broad and flexible to meet the requirements of varied integration scenarios. Likewise within the same platform, IBM Information Server offers the ability to virtualize data access using federation.

Within IBM Information Server, products can interoperate in several ways to allow customers the freedom to design their data warehouses in ways that best suit their business needs. ETL jobs can directly call WebSphere Federation Server for real-time access to data and federated queries can make calls to Service Oriented Architecture (SOA)-enabled ETL jobs to gain access to the powerful ETL techniques. More importantly, real-time access afforded through data federation gives users the ability to have current views of rapidly changing data sources.

*Figure 11: Federation in action*




## Conclusion

When implemented properly and applied to the right problems, the hybrid data warehouse can provide enormous business benefits and flexibility, both in cost efficiencies and revenue opportunities. Companies that can find and capitalize upon these opportunities often achieve an immediate competitive advantage, as shown in Figure 11.

IBM Information Platform and Solutions has helped companies achieve success in their on demand data warehousing efforts by providing a wide range of data integration technologies, methodology and best practices for quickly and successfully deploying these projects.

### For more information

For more information about IBM Information Server, contact your IBM marketing representative or visit **ibm.com**/software/data/integration

## Bibliography

IBM. *IBM Federated Database Technology*. 2002. **ibm.com**/developerworks/db2/library/techarticle/0203haas/0203haas.html

**Two-part series on using data federation technology:**
IBM. *Using data federation technology in IBM WebSphere Information Integrator: Data federation design and configuration*. Part 1. 2005. **ibm.com**/developerworks/db2/library/techarticle/dm-0506lin/

IBM. *Using data federation technology in IBM WebSphere Information Integrator: Data federation usage examples and performance tuning*. 2005. **ibm.com**/developerworks/db2/library/techarticle/dm-0507lin/

IBM. *Maximize the performance of WebSphere Information Integrator with Materialized Query Tables*. 2006. **ibm.com**/developerworks/db2/library/techarticle/dm-0605lin/

**IBM WebSphere Federation Server Version 9.1:**
IBM. *Use federated procedures in WebSphere Federation Server*. 2006. **ibm.com**/developerworks/db2/library/techarticle/dm-0605bhatia

IBM. *Maximizing your query result in WebSphere Federation Server V9.1 with Error Tolerant Nested Table Expression*. 2006. **ibm.com**/developerworks/db2/library/techarticle/dm-0609huang



IBM Software Group
Route 100
Somers, NY 10589

Printed in the United States of America
May 2007


***TAKE BACK CONTROL WITH*** **Information Management**

IMW11841-USEN-00