Delivering information you can trust

December 2007

Information Management software

Delivering information integration with IBM DB2 Warehouse

Contents

Introduction

IBM has been providing industry-leading platforms for delivering analytics to business users since the first relational databases entered the market. Although the term we use to describe these systems has changed over time from "executive information systems" to "enterprise data warehouses," one thing has remained the same: IBM's ability to deliver enterprise-class solutions to both small and large businesses. This white paper describes how you can use IBM® Information Server and IBM DB2® Warehouse software to help meet and exceed these challenges with industry-leading integrated solutions.

The age of dynamic data warehousing and its demands of real-time data access and user consolidation require a robust, integrated integration and warehouse architecture. Dynamic data warehousing also requires a common information framework that defines common terms and models of how business information relates across the warehouse and source systems, and how it can be used to deliver business insight. IBM helps you achieve this with the Information On Demand (IOD) framework, a solution built around IBM Information Server and DB2 Warehouse capabilities. The IOD framework includes comprehensive business models and master data management (MDM) capabilities to provide a complete solution for delivering truly dynamic capabilities to your enterprise data warehouse.

The changing role of the data warehouse

During the early days of reporting solutions, much reporting was done within applications, which often sat on various technology platforms. As time and technology progressed, organizations began to consolidate data from different platforms into centralized, purpose-built reporting solutions or datamarts. As these solutions began to rise in popularity and number throughout organizations, the issues of reporting consistency and spiraling cost of ownership resulted in the advent of centralized data reporting environments. Those environments evolved into today's data warehouses.

Today we are seeing further changes: businesses no longer want to wait until information filters into the warehouse overnight. They would rather see the data arrive in the warehouse in near real time so they can make informed, on-the-spot decisions that better enable them to perform in the current dynamic business environment. This includes the ability to embed warehouse-driven analytical services into day-to-day business operations.

To achieve the vision of a centralized data warehouse, businesses increasingly must invest in an information integration strategy to address data timeliness and data quality. Some analysts estimate that poor data quality may cost U.S. businesses as much as US$600 billion a year. Problems include duplication of work efforts, erosion of customer confidence, lost business opportunities and poor decision making because of inaccurate or incomplete information.

IBM Dynamic Warehousing capabilities, based on IBM Information Server and DB2 Warehouse, provide a cohesive, integrated, scalable, high-performance platform for your business warehouse.

A scalable approach to data consolidation, harmonization and integration

Deploying an enterprise data warehouse is not a trivial task; it relies heavily on the quality and availability of data that feeds and informs the system. The systems generally require a high volume and wide variety of data from a range of established and current systems—systems that often span multiple departmental functions and data formats.

Large corporations need a clear and detailed data integration strategy that enables a business to focus on improving data quality for current initiatives while also maintaining data quality at the enterprise level for the long term. A well-conceived enterprise data integration strategy delivers a clean, trusted source of data upon which new and existing data warehouse applications can rely. Without a solid integration strategy, organizations risk significant and costly consequences:

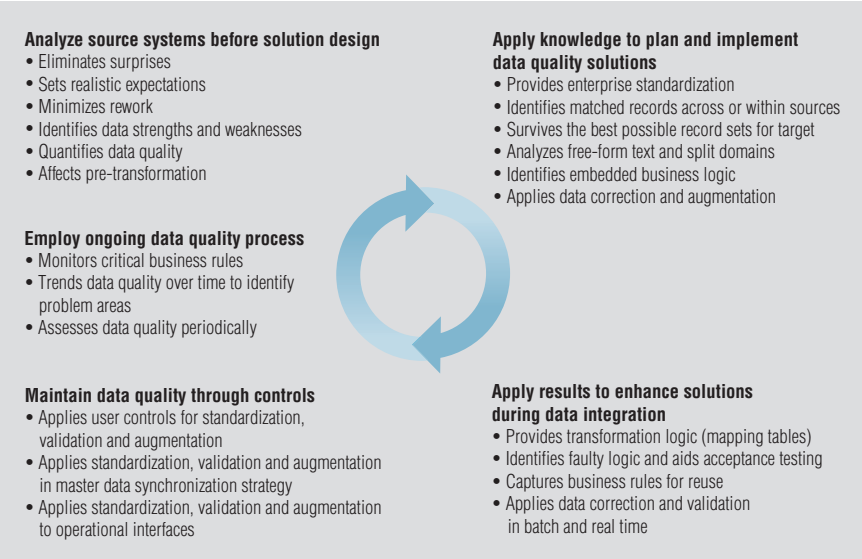
- *Difficulty scoping project phases due to lack of visibility into related data challenges*
- *Stalled projects and roadblocks caused by unforeseen data quality issues*
- *Excessive hand-coding and work-arounds to compensate for data quality problems*
- *Compromised project timelines and milestones*
- *Increased project burn rate after data quality issues are uncovered*

In the long term, the consequences of not implementing a thorough data consolidation, harmonization and integration strategy—or opportunities to create value from such a strategy—fall into four categories:

- ***Lost revenue opportunities*** *because systems do not share timely and accurate data*
- ***Excessive operational cost*** *generated by excessive business and IT workloads related to managing data-related problems*
- ***Complex business risk management*** *because data associated with risk factors is not easily accessible—making risk difficult to predict, manage and track*
- ***Difficulty in meeting government compliance requirements*** *because systems lack adequate processes for gaining complete and timely access to accurate and relevant compliance data*

The IBM data consolidation, harmonization and integration strategy takes advantage of its more than 25 years of reporting systems and data warehouse experience and its methodologies, tools and best practices to transform enterprise organizational data into a strategic corporate asset delivered on the DB2 Warehouse platform. The IBM strategy positions an organization to use this corporate asset across all applications and data integration initiatives. By leveraging the strategic enterprise data quality process shown in Figure 1, IBM can help an enterprise establish the fundamental processes for improving and maintaining the quality and accessibility of one of its most valuable assets—corporate data.

Figure 1: Strategic enterprise data quality process

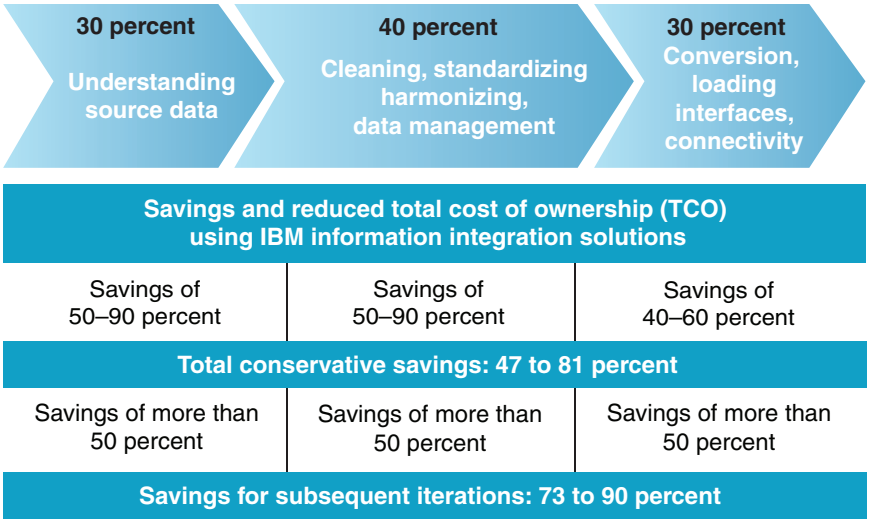


A proven methodology for enterprise data integration

The work effort and resulting value derived from data integration projects generally falls into three categories (see Figure 2):

- *Understanding the current state of the source data (30 percent of project effort)*
- *Cleaning, standardizing and harmonizing source data (40 percent of project effort)*
- *Converting, loading, creating interfaces and establishing connectivity (30 percent of project effort)*

Figure 2: Typical work effort and savings in information integration initiatives



The bulk of data integration effort and the areas that provide the greatest opportunities for positive impact of the system implementation come from the first two categories: understanding source data, and data alignment and harmonization. Information about source data is invaluable to the data warehouse project team. The data extraction and analysis stages should be initiated as early in the data warehouse implementation project as possible, which helps eliminate data harmonization issues before they become much larger and more difficult to resolve.

Operational source data intelligence gathered during the data analysis stage enables your project team to define an overall data integration strategy, beginning with the approach to cleansing data, defining harmonization rules, defining transformation logic and forming strategies for loading the DB2 Warehouse platform in parallel. An integral aspect of this approach is identifying business data owners to manage and assist in implementing these activities. The overall data consolidation approach can vary depending on specific data warehouse applications. For example, human resources (HR) data conversion in Europe and Asia-Pacific regions is not necessarily the same as HR data conversion in the U.S. However, a global data integration project template can guide the process and methodology so that it can be implemented consistently in each region.

The operational source data analysis helps subject matter experts and customer, product and supplier data owners provide critical input into the blueprint process by identifying challenges in the key data and validating the data against business rules and known requirements. During the analysis, valuable technical and business metadata and business rules are collected, which form the basis of the data preparation and solution specifications. The core project team can use the information to help determine whether a system can be integrated with the data warehouse by identifying gaps in the current data set based on data warehouse requirements and estimating the work effort necessary to complete the data integration process.

After the data analysis stage is complete, project teams perform a thorough data cleansing process. Data cleansing takes place within the data alignment stage—where it reconciles divisional requirements and aligns data by source—and within the data harmonization stage. The data harmonization stage aggregates enterprise requirements and harmonizes data across sources to produce a single, best source of operational data.

Data quality represents one of the most significant potential risks for a successful data warehouse implementation. Unclean data can have a direct and immediate impact on users' satisfaction with the data warehouse, as well as its functional effectiveness and accuracy.

Data analysis and cleansing functions are designed to maximize efficiency by automatically matching different data types and then standardizing them to follow the strategic and scalable data quality process. Data quality functions can be plugged into the user interface to control data quality from the moment it is entered by using service-oriented objects or direct application programming interfaces (APIs). These capabilities enable standardization functions in a DB2 Warehouse implementation to be used on a long-term basis to maintain data quality across other enterprise source systems.

Data cleansing functions include standardizing areas such as name and address, fields and free-form text, as well as validation and augmentation using reference sources, duplicate matching, records management and best information survivorship.

After data cleansing, you can design and implement alignment and harmonization processes and functions and design and construct data transformation and delivery functions. Using IBM methodology and a parallel-processing platform throughout the early stages of the data integration project can help improve the success of data delivery functions and align them with the shared-nothing parallel processing approach of the DB2 Warehouse platform. Transformation logic and business rules captured during data analysis, alignment and harmonization provide a clear roadmap for data conversion and solution implementation, as well as providing the foundation for ongoing data quality assurance.

An enterprise data integration platform

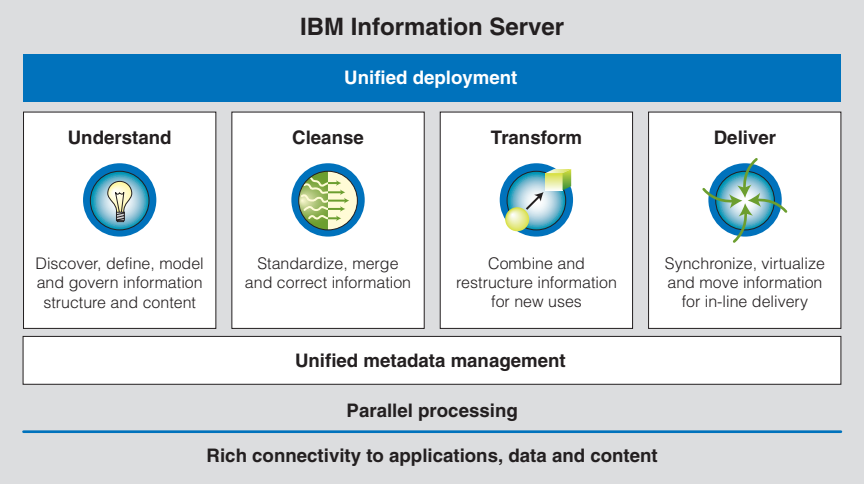
Today's complex data environment requires a comprehensive, scalable data integration solution for your DB2 Warehouse. Simply extracting the data, loading it into the database and reusing it for an analytical application is not adequate for the needs of most organizations. Creating a comprehensive data integration solution for DB2 Warehouse is not a simple project, but making the investment in the right solution can maximize efficiency and productivity—and help deliver long-term savings—by enabling manual processes to be automated and scaled.

Essential components for this kind of forward-thinking data integration solution for DB2 Warehouse include:

- *Integrated tooling across your information integration and DB2 Warehouse development toolset*
- *Connectivity to all relevant data sources, whether mainframe or distributed, internal or external, structured or unstructured*
- *Insight into the content, quality and structure of data sources to completely understand the data before it is integrated into the data warehouse*
- *The ability to standardize, cleanse and harmonize data to provide access to authoritative and consistent views of any individual or business entity across the extended enterprise*
- *The ability to effectively collect, transform and enrich high volumes of data to match performance and scalability requirements*

The IBM Information Server platform, shown in Figure 3, provides these fundamental capabilities. A high-performance, highly scalable data integration solution built on a parallel framework, it was designed and developed with

Figure 3: IBM Information Server platform

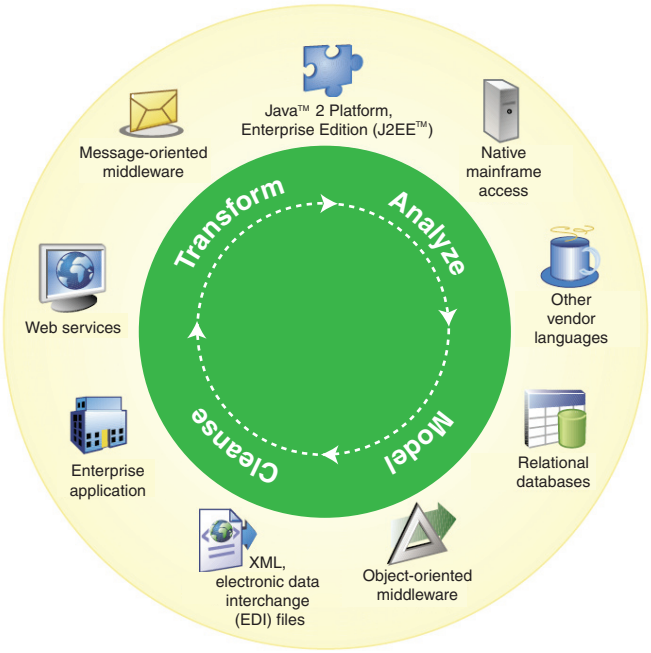


the aid of IBM's years of delivery experience and long-standing investment in high-performance computing. IBM Information Server can integrate and transform almost any data and content to deliver reliable information to your DB2 Warehouse. IBM Information Server provides breakthrough productivity, flexibility and performance to help ensure that you have the right information to run and grow your business.

With IBM Information Server, you can:

- ***Connect*** *to practically any data or content, wherever it resides.*
- ***Understand and analyze*** *information, including its meaning, relationships and lineage.*
- ***Cleanse*** *information to help ensure its quality and consistency.*
- ***Transform*** *information to meet the specific needs of DB2 Warehouse.*
- ***Meet the needs*** *of the most demanding enterprise data warehouses.*

Figure 4: Sources for IBM Information Server connectivity capabilities



Connect

IBM Information Server offers an extensive array of connectors and adapters that provide native access to relevant sources for both mainframe and distributed-computing environments. A representative list of the sources that connect to the platform is shown in Figure 4.

Figure 5: IBM enterprise data warehouse architecture

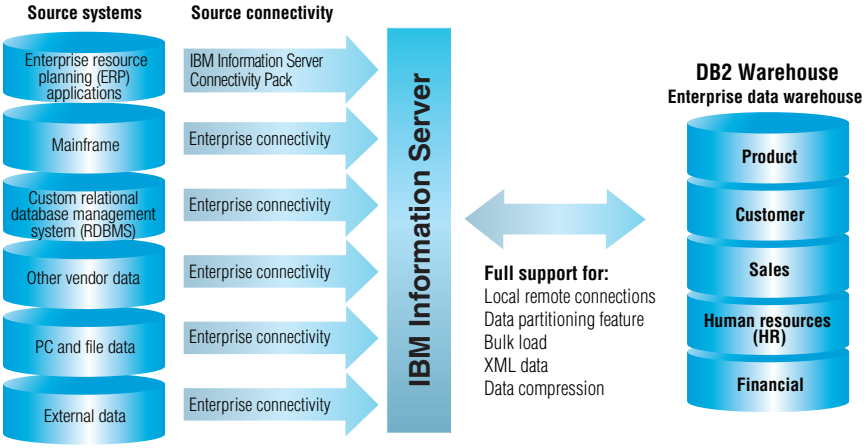


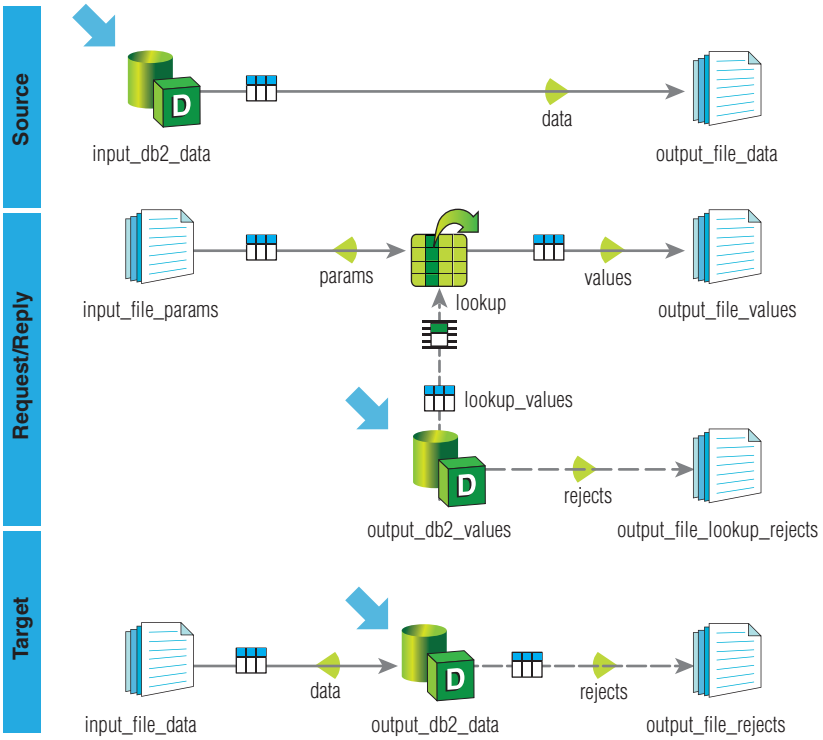
Figure 5 displays the technical architecture of an IBM DB2 enterprise data warehouse implementation using IBM Information Server. The ability of IBM Information Server to source from multiple enterprise repositories and load to DB2 with prebuilt interfaces enables quick development and deployment.

IBM Information Server DB2 Connector

Figure 6 displays the flexibility available with IBM Information Server DB2 Connector, which includes the following features:

- *Source, target and lookup context; reject links; passing line of businesses by reference, arrays, SQL Builder, pre/post run statements and metadata import*
- *Connector is CLI client interface–based and can connect to any database cataloged on the DB2 client; the DB2 client must be collocated with DB2 Connector, but the actual database may be local or remote to the DB2 Connector*

Figure 6: IBM Information Server DB2 Connector



- *Separate sets of connection properties for job setup phase (conductor) and execution phase (player nodes), which enable the same database to be cataloged differently on conductor and player nodes*
- *Support for specifying DB2 instance dynamically through connection properties which overrides the default environment settings (DB2INSTANCE environment variable)*
- *Support for the XML data type in DB2 9.1*

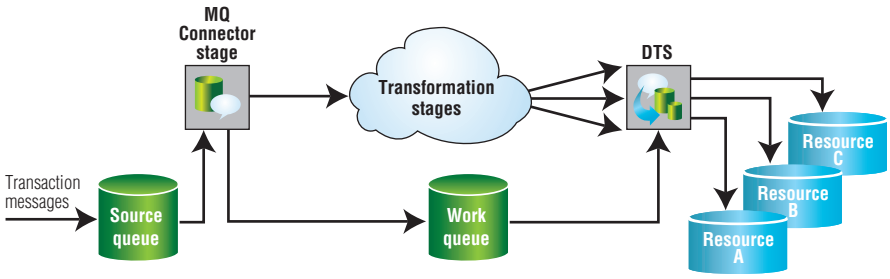
DB2 Connector also provides support for DB2 data partitioning feature (DPF). A job with the DB2 Connector target stage can be configured to assign one execution player node with each DB2 partition, and to write data to the partitioned database in parallel. This can provide dramatic performance improvement over sending the data to the same partition node and demanding DB2 to redirect data to corresponding partitions.

Distributed transaction stage for DB2

The connector framework is being enhanced to provide support for distributed two-phased XA transactions in IBM Information Server jobs. The XA capability is unique to DB2 and IBM Information Server.

The transaction data is carried by MQ messages that arrive to the source queue. Each message can encompass multiple database operations, and multiple messages may be grouped in a single physical transaction. The MQ Connector stage provides properties for configuration of transaction boundaries. Users can specify the number of source messages to include in each transaction or the time interval for collecting source messages in each transaction.

Figure 7: Distributed transaction stage for DB2

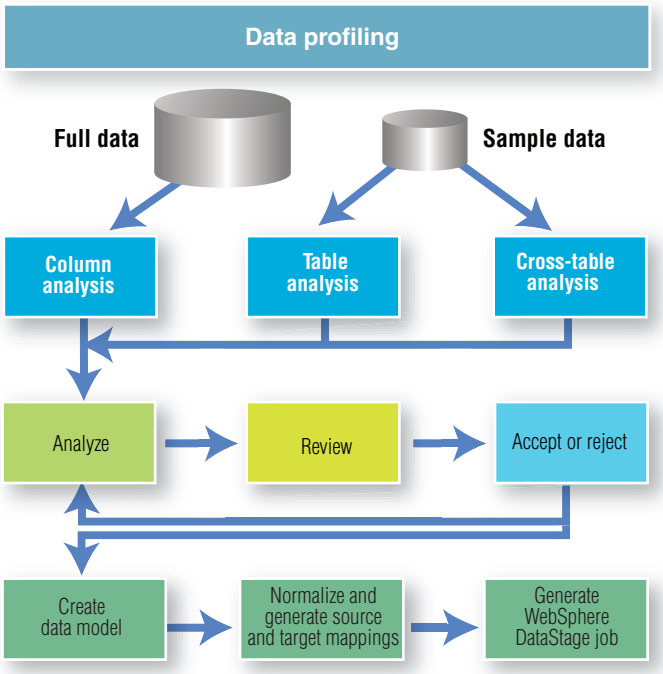


The MQ Connector stage uses a specially designated work queue as a temporary buffer storage for source messages that participate in transactions (see Figure 7). The retrieved messages are processed by any number and combination of transformation stages, chosen from a rich palette of stage types provided by IBM Information Server. The processed messages result in rows of data that arrive on the Distributed Transaction Stage (DTS) on one or more input links. Each input link on the DTS is associated with one external resource. The rows on each link are sent to the designated resource as insert, update or delete operations. The DTS reads messages from the work queue and updates external resources with the data rows corresponding to those work queue message. Reading of messages and writing to external resources is done in atomic distributed transactions using two-phase XA protocol.

Understand

The IBM Information Server platform provides the necessary automated data profiling analysis and monitoring capabilities to unlock the mystery of source content, quality and structure (see Figure 8). Data-driven table and column analysis and reporting help identify missing, inaccurate, redundant and inconsistent data. Automated discovery and relationship and dependency analysis help establish the true metadata of the source systems. Continuous data quality monitoring helps maintain the health of your data throughout its life cycle.

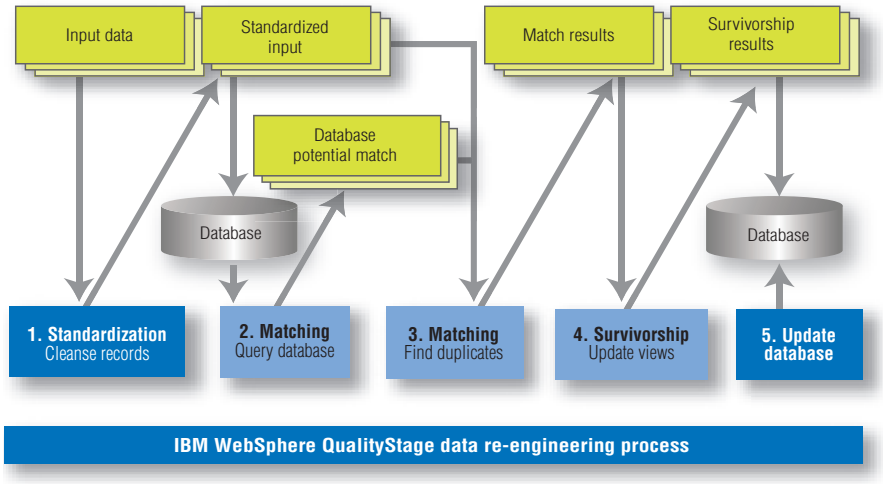
Figure 8: IBM Information Server data profiling capabilities



Cleanse

The diversity of business information systems across the extended enterprise requires business information to be clean before it enters DB2 Warehouse. The information must be identified, standardized, matched, reconciled and free of redundancies to help ensure quality and consistency, as shown in Figure 9.

Figure 9: IBM Information Server data quality process



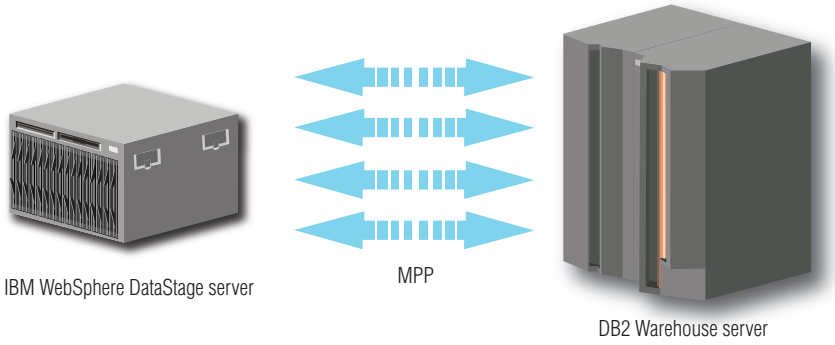
Data cleansing helps establish a single logically correct view of core business entities across the enterprise. To accomplish this, IBM Information Server solutions provide:

- *Standardized data fields to help establish information consistency*
- *Validation, certification and enrichment of common data elements using trusted data*
- *Matching of records across or within data sources, providing assurance that duplicate data is removed and enabling common entities to be linked together*
- *Loading of information records into a repository–in a standardized, cleansed format–based on the best information available from each unique entity, which results in the creation of a single comprehensive, accurate view of information that spans source systems*

Transform

IBM Information Server provides high-volume, complex data transformation and movement functionality. This functionality in either batch or real-time environments helps satisfy the needs of both the enterprise and the dynamic data warehouse. As a parallel platform, IBM Information Server is highly complementary to the DB2 Warehouse parallel architecture (see Figure 10). In addition, DB2 Warehouse tooling integrates with IBM Information Server. The close ties enable DB2 Warehouse developers to interact with IBM Information Server during development and testing to provide a more cohesive and team-driven development approach and help reduce time to delivery and the possibility of deployment errors—further reducing the cost of ownership for your implementation.

Figure 10: IBM Information Server parallel architecture synchronized with DB2 Warehouse parallel technology



IBM Information Server is a scalable application environment used to build parallel-processing applications. It offers dozens of prebuilt stages for performing the most common data integration tasks, such as sort, merge, join, filter, transform and aggregate. Each IBM Information Server stage is a complete functional component that has been engineered to run in parallel with high throughput and performance on multiprocessor computing systems. The prebuilt stages in IBM Information Server help eliminate the need to perform custom coding for common data-processing applications. Figure 11 shows some of the design stages available with IBM Information Server.

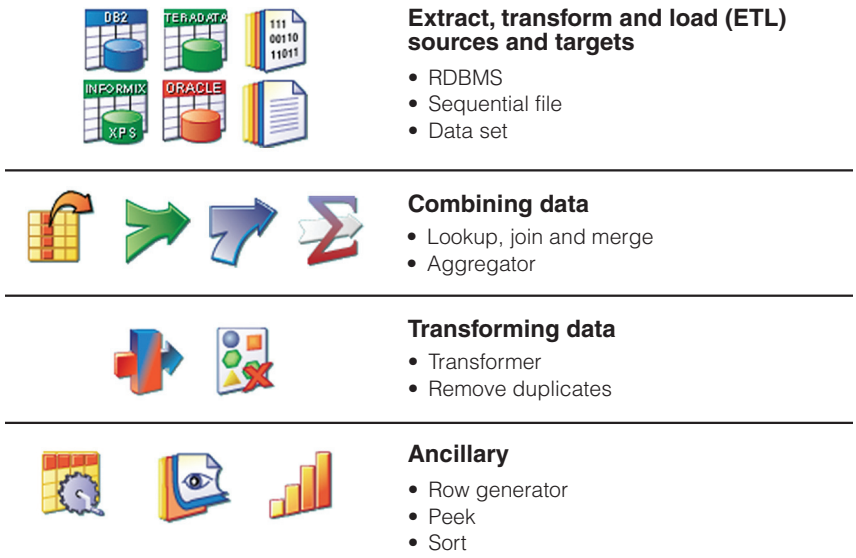


Figure 11: Popular IBM Information Server design interfaces

IBM Information Server enables developers to quickly encapsulate a set of steps in a data integration process as reusable components. Using this capability, organizations can achieve breakthrough productivity and establish a library of standardized components that embrace best practices and can result in higher quality and consistent information across all projects.

Deliver real-time information

Providing access to real-time operational data is a critical requirement of a dynamic warehouse. IBM Information Server includes high-performance change data capture (CDC) capabilities with industry-leading IBM DataMirror technology. DataMirror provides:

- ***Leading technology*** *including optimized log-based capture that enables fast end-to-end performance, can reduce impact on operational systems and can be less invasive to data sources than alternative solutions*
- ***Broad range of supported sources and targets****, both distributed and mainframe*
- ***Data delivery integrity*** *with use of fault tolerance and high-reliability features*

In addition, IBM Information Server is built as a foundation for Service Oriented Architecture (SOA). Any of the data integration or data quality jobs that are designed in IBM Information Server may quickly be deployed as services and made available to any SOA-enabled enterprise application.

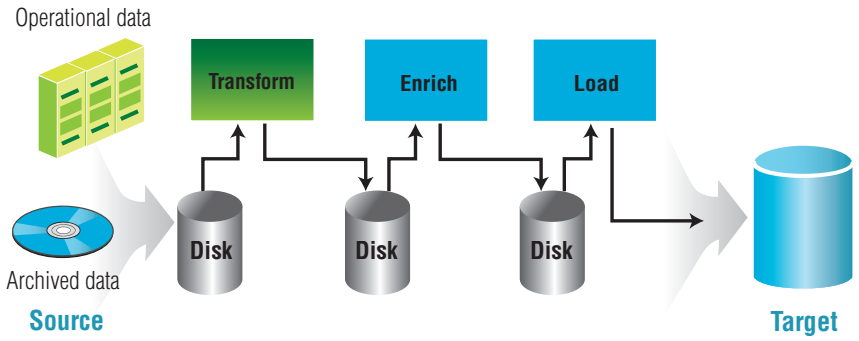
Extending infrastructure

A scalable infrastructure should provide native, high-performance parallel components—particularly sorting, aggregation and joins. But because any large enterprise has special and custom needs, a scalable infrastructure should be extensible to integrate existing programs and other vendors' tools as part of the data integration process. Programs originally written to run sequentially should be able to run in parallel on a partition of data, regardless of the programming language used (C, C++, COBOL and so on). A key requirement to integrate existing software code is the capability to operate only on the data (columns or fields) of each record and the ability of the infrastructure to simply pass the rest of the data not used (not touched or changed) through a component to the next downstream component in the data flow. The capability has been referred to as column or schema propagation. It is a critical aspect of integrating existing applications without change, making them more portable and usable. With column or schema propagation, software can be integrated and parallelized.

Parallel processing

As companies try to make sense of their massive amounts of corporate information, they face the logistical challenges of managing, storing and sorting through rapidly expanding volumes of data. To maximize the value of corporate data, they must gather and deliver it to end users as quickly as possible. As they manage these huge amounts of data, they also need to integrate it at a more granular level—dealing with individual transaction–level data rather than general summary data. The traditional way of handling large amounts of data was to stage it intermediately before each processing step could take place (see Figure 12).

Figure 12: Traditional batch approach to data management



To address these challenges, organizations need a scalable data integration architecture that has:

- ***Data-flow features*** *that enable data to process from input to output—without landing to disk—in batch and real-time scenarios*
- ***Dynamic data partitioning*** *and in-flight data repartitioning*
- ***Scalable hardware environments*** *that support symmetric multiprocessing (SMP), clustering, grid and massively parallel processing (MPP) platforms and do not require modification of the underlying data integration process*
- ***Support for leading parallel databases****, such as IBM DB2®, in parallel and partitioned configurations*
- ***An extensible framework*** *to incorporate in-house and third-party software*

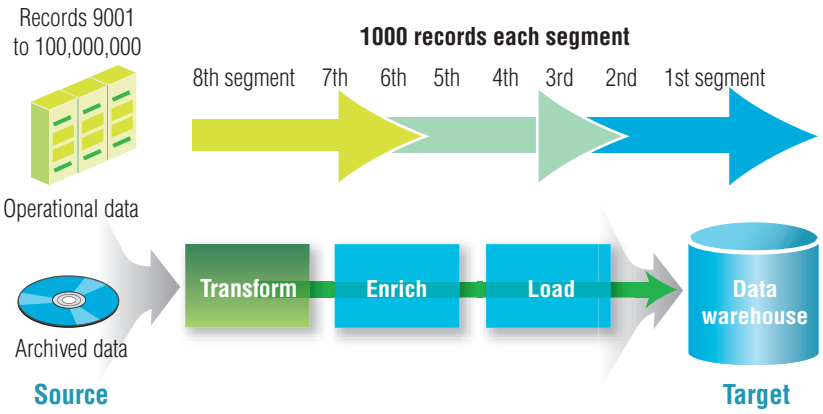
The architecture must be able to grow with the organization as data volumes and performance requirements increase. Most importantly, the architecture should not have any upper boundaries and should be able to scale linearly with the hardware environment. Also, increasing performance should be as simple as adding processors or nodes to the hardware environment. In fact, this type of hardware upgrade should be able to occur with no change to the underlying data integration application.

All components of the IBM Information Server portfolio take advantage of both pipeline parallelism and partition parallelism to achieve high throughput, high performance and scalability. This section explains why these features are required for a high-performance, scalable architecture.

Data pipelining

As the application begins to run in a data pipelining approach, records are pulled from the source system and moved through the sequence of processing functions defined in the data flow—also called the job. Because the records are flowing through the pipeline, they can be put through the sequence of processing functions without landing records to disk, as illustrated in Figure 13.

Figure 13: Data pipelining



Data is (or can be) buffered in blocks so that each process does not negatively affect the system when running one component or the next. This approach helps avoid deadlocks and can greatly accelerate performance by enabling both upstream and downstream processes to run concurrently.

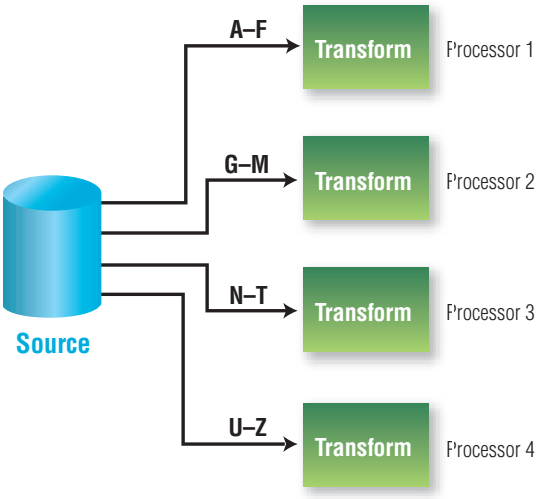
Implications of a data flow architecture that does not support data pipelining include:

- *Data must be landed to disk between processes, which can severely degrade performance, increase storage requirements and create almost insurmountable disk-management issues.*
- *The developer must manage the input/output (I/O) processing among components.*
- *The process becomes impractical for large data volumes.*
- *The application runs slower as disk use, management and design complexities increase.*
- *Each process must be completed before downstream processes can begin, which can limit performance and does not make full use of hardware resources.*

Data partitioning

Data partitioning is an approach to parallelism that involves segmenting the record set into partitions or subsets of records. Data partitioning generally provides good quality, linear increases in application performance. Figure 14 shows an example in which data is partitioned by customer last name before flowing into the next stage (the transform component).

Figure 14: Data partitioning



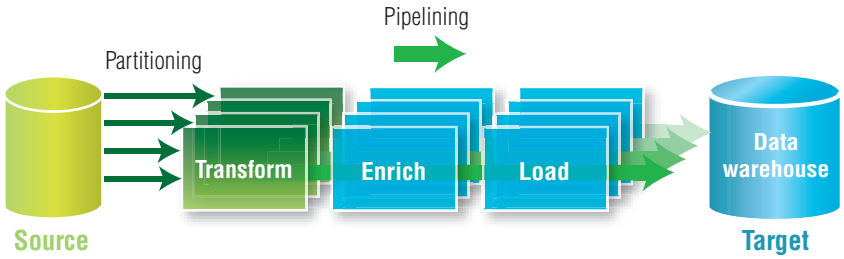
A scalable architecture should support many types of data partitioning:

- *Hash key (data) values*
- *Range*
- *Round robin*
- *Random*
- *Entire*
- *Modulus*
- *Tiered database partitioning (shared nothing, table, dimensional)*

IBM Information Server automatically partitions data based on the type of partition required by the stage. Other packaged tools lack this capability and require developers to hardwire data partitions, resulting in costly and time-consuming rewriting of applications or data partitions whenever the administrator wants to use more hardware capacity. The extra effort can consume many weeks or months of development and testing time prior to production.

In a well-designed, scalable architecture, the developer should not be concerned about the number of partitions that will run, the ability to increase the number of partitions or—most importantly—data repartitioning (see Figure 15).

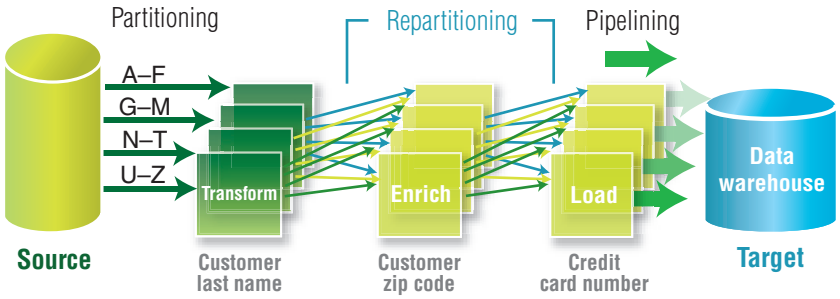
Figure 15: Data partitioning and parallel execution



Dynamic repartitioning

The examples in Figures 14 and 15 show data partitioned based on customer last name and the data partitioning was maintained throughout the flow. In many cases, this is impractical; for example, a transformation that requires data partitioned by last name when loading into the warehouse is based on customer account number. Dynamic data repartitioning is a more efficient and accurate approach. With dynamic data repartitioning, data is repartitioned on the fly between processes—without landing the data to disk—based on downstream process data partitioning needs (see Figure 16). Data is also being pipelined to downstream processes when it is available.

Figure 16: Dynamic data repartitioning—a more practical approach



Without partitioning and dynamic data repartitioning, the developer must:

- *Create separate flows for each data partition, based on the current hardware configuration*
- *Land data to disk between processes*
- *Manually repartition the data*
- *Start the next process*

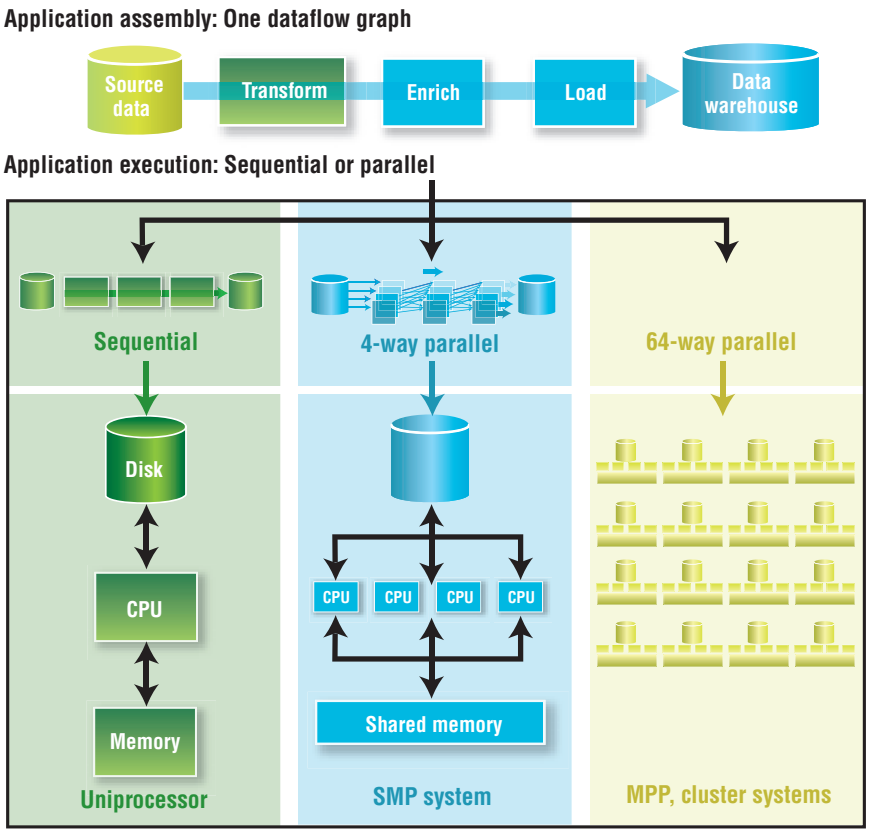
Consequently, the resulting application may be slower, disk use and management may increase and the design may be much more complex. The dynamic repartitioning feature of IBM Information Server helps developers overcome these issues.

Scalable hardware

For maximum scalability, it is not enough that integration software simply runs on SMP and MPP computer systems. If the data integration platform does not truly saturate all of the nodes of the MPP box or system in the cluster or grid, you cannot maximize scalability. IBM Information Server optimizes the use of all available hardware resources.

Users of IBM Information Server can create a simple sequential data flow graph using the designer canvas. When constructing the sequential data flow graph, users do not have to worry about the underlying hardware architecture or number of processors because the solution components make full use of SMP, clustered, grid and MPP environments. A separate configuration file defines the resources (physical and logical partitions or nodes, memory and disk) of the underlying multiprocessor computing system. As shown in Figure 17, the configuration provides a clean separation between the creation of the sequential data flow graph and the parallel execution of the application, helping to greatly simplify the development of scalable data integration systems that run in parallel.

Figure 17: A complete integration configuration



Without support for scalable hardware environments, clients can face the following problems:

- *Slower implementation because all available hardware resources are not maximized*
- *Manual intervention required for every hardware change because application design and hardware configuration are not decoupled*
- *Scaling on demand is not possible*

IBM Information Server is built on a highly scalable software architecture that delivers high levels of throughput and performance to address these challenges.

Conclusion

An enterprise data warehouse from IBM brings together all of your company data into a single repository for a completely integrated, 360-degree view of your business. Featuring high-performance parallel database technology, a full suite of data access and management tools and robust data-mining capabilities, an enterprise data warehouse from IBM delivers powerful performance.

The IBM Information Server platform offers true parallel processing with multiple processors running programs. Combined, DB2 Warehouse and IBM Information Server operate at a performance level that suits the proven, powerful DB2 analytic solution.

Together, IBM DB2 Warehouse and IBM Information Server offer a complete Dynamic Warehousing solution for any amount of data that takes advantage of IBM's integrated approach and investment in parallel processing. With the diverse IBM Information Server platform, DB2 clients can address specific data integration requirements including data discovery and data quality. Also, IBM Information Server addresses shrinking batch windows, expanding numbers of data sources, complex transformations and exponential data growth. IBM Dynamic Warehousing capabilities have the potential to energize your business initiatives—and provide an enterprise solution that enables employees to do their jobs with great precision and effectiveness, while producing a cost-effective return on your investment.

For more information

To learn more about IBM Information Server and IBM integration solutions, visit **ibm.com**/software/data/ips

IBM Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
December 2007

TAKE BACK CONTROL WITH **Information Management**

IMW11901-USEN-01