Delivering information you can trust
June 2008

Information Management software

IBM InfoSphere DataStage Balanced Optimization

Maximize your investment in integration and database technologies

Contents

IBM® InfoSphere™ Information Server helps organizations derive value from the complex information spread across systems. It is a revolutionary software platform that profiles, cleanses and integrates information from heterogeneous sources to drive greater business insight faster and at lower cost. The IBM InfoSphere Information Server platform also:

- *Provides a comprehensive, unified foundation for enterprise information architectures with simplified scalability to help organizations manage current and future data requirements*
- *Delivers metadata-driven integration, providing breakthrough productivity and flexibility for integrating and enriching information*
- *Offers data quality and data governance capabilities to ensure consistent and accurate delivery of information for greater trust and compliance with information-centric regulations and requirements*
- *Accelerates time-to-value with proven, industry-aligned solutions and expertise combined with consistent, reusable information services*
- *Maximizes value and flexibility of IT investments by leveraging existing mainframe resources to perform information integration directly on the mainframe with no impact to your IBM z/OS® software costs*
- *Provides broad and deep connectivity to information across diverse sources, including structured, unstructured, mainframe and application sources*

IBM InfoSphere DataStage parallel architecture

As today's companies try to make sense of massive amounts of corporate information, they face the logistical challenges of managing, storing and sorting through rapidly expanding volumes of data. They need to gather all of their corporate data and deliver it to end users as quickly as possible to maximize its value. They must also integrate data at a more granular level—working with individual transaction data rather than general summary data.

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

- *Optimal parallelism and pipelining to complete increasing volumes of work in decreasing windows of time*
- *A data-flow architecture that allows data to be processed from input to output without landing it to disk—in both batch and real-time scenarios*
- *Dynamic data partitioning and in-flight repartitioning of data*
- *Support for scalable hardware environments, including symmetric multiprocessing (SMP), clustering, grid and massively parallel processing (MPP) platforms, without requiring modification of the underlying data integration process*
- *Support for leading parallel databases, including IBM DB2® Universal Database (UDB), Oracle and Teradata, in parallel and partitioned configurations*
- *Optimized file and queue processing to enable the system to deal with huge files which cannot fit into memory all at once, or with large numbers of small files*
- *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 bounds and should be able to scale linearly with the hardware environment.

Increasing performance should be as simple as adding processors or nodes to the hardware environment. In fact, these upgrades should be able to occur with no change to the underlying data integration application.

IBM InfoSphere Information Server and IBM InfoSphere DataStage® are built on this scalable software architecture, which delivers high levels of throughput and performance.

IBM InfoSphere DataStage Balanced Optimization

Traditional data integration platforms have long provided very robust connectivity capabilities to enterprise applications, mainframe data repositories and non-relational data sources, such as complex flat files. A key component in data integration platforms is the extract, transform and load (ETL) engine. ETL uses the high-performance, scalable integration architecture discussed above to extract data from one or more sources and then perform data transformations and enrichment before loading it into one or more targets. This method takes advantage of the high-performance, scalable engine but can also reduce the system impact on data sources and targets.

In contrast to the ETL approach, extract, load and then transform (ELT) platforms rely on the underlying database to provide connectivity and data transformations. By leveraging the relational database management systems'

(RDBMS) engine hardware for scalability, data transformations can efficiently handle large volumes of data optimizing disk I/O at the engine level for faster throughput. However, the RDBMS cannot optimize the complex processing of data that resides outside the database. This requires an external engine to do the processing, which negates the advantage of utilizing the database engine for extracting or replicating data from other data sources.

A truly complete data integration platform provides not only data replication and change data capture capabilities, but also the functionality to navigate enterprise applications, automatically generate the most efficient data extraction method and manage metadata. Users really want not only ETL and ELT, but also transform, extract, load and transform (TELT) and ultimately transform, extract, transform, load and transform (TETLT). To achieve this, users can specify that the processing be performed in the source and/or target as well as the InfoSphere DataStage engine.

This is the goal of IBM InfoSphere DataStage Balanced Optimization. To address customer demands for increased flexibility in handling data transformations as well as maximize their investment in parallel database systems, IBM has extended the functionality of IBM InfoSphere Information Server to incorporate alternative data transformation approaches that accommodate ETL as well as ELT, TELT and TETLT.

With InfoSphere DataStage Balanced Optimization, users can continue to express the logic of their integration processes using the current, natural, flow-oriented InfoSphere DataStage conventions, and then automatically or semi-automatically optimize their designs to enhance the performance of ELT

and structured query language (SQL) pushdown. InfoSphere DataStage Balanced Optimization will not require users to manually rewrite their queries and processing to achieve greater flexibility and throughput; users get the extended capabilities of ETL, ELT and even TETLT without sacrificing the benefits of the IBM InfoSphere Information Server platform. The rest of this white paper describes the IBM approach to balanced optimization.

Key optimizations can improve job performance

There are several places in a data integration flow where optimization can improve performance. Minimizing I/O is one of the first things that can help improve performance. Moving processing to the sources may reduce the amount of source data that is extracted; similarly, moving part of the processing to the target may help avoid target extractions. Minimizing data copying can further improve performance, as can the use of indices, native optimizations and database-specific features in conjunction with the InfoSphere DataStage parallel engine.

The core idea is to maximize parallelism in all three aspects of a data integration flow: in the I/O to and from databases; in the InfoSphere DataStage parallel engine; and inside the databases. By enabling parallel processing in all of these areas, IBM InfoSphere DataStage Balanced Optimization helps users achieve an optimal data integration flow.

Why do we need performance optimizations?

Optimization serves many purposes, including overall performance improvement and shifting computation to source or target platforms. Also, it can be useful to compare and contrast the results of different optimizations in specific data situations. InfoSphere DataStage Balanced Optimization gives users several levels of control over which optimizations are performed:

- *Database-specific optimization packages*
- *Overall control options, such as "push to source" and "push to target"*
- *Specific options, such as whether to use target staging tables*

Using InfoSphere DataStage Balanced Optimization

The InfoSphere DataStage developer begins the process as usual by designing the job using IBM InfoSphere DataStage Designer. The InfoSphere DataStage Designer client offers a much simpler way of writing ETL flows compared to complex SQL designer tools. The developer does not need to understand complex SQL nor database-specific SQL—InfoSphere DataStage translates the semantics expressed in the job into the appropriate database-specific SQL. Just as the developer does not need to understand the details of the parallel hardware on which the engine runs, he does not need to know what optimization needs to be performed while he is designing his job.

Once the job is designed, the developer optimizes it. The basic choices include pushing processing to the sources and/or the targets; pushing joins, lookups and other functions to the source/target; bulk loading where possible; and a few other database-specific options (see Figure 1).

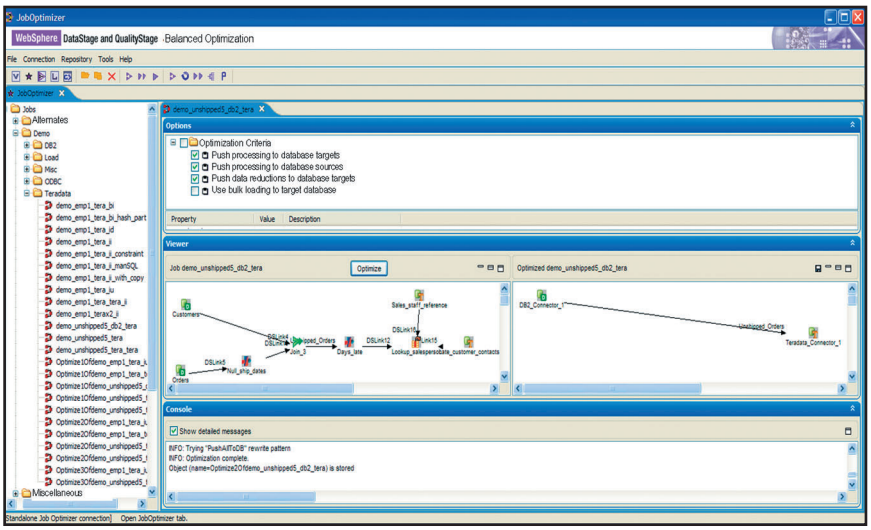

Figure 1: IBM InfoSphere DataStage Balanced Optimization offers developers several optimization choices and rewrites the job to reflect their selections.

The tool then rewrites the job based on the developer's selected options. The Balanced Optimization feature works under the covers: proprietary, patent-pending IBM algorithms rewrite the job to push as much functionality as possible into database targets and/or database sources. This includes potentially pushing all of the job functionality into the target database and the option to use bulk loading and unloading. The developer can iterate through the choices and run the optimization until he is satisfied with the result. Once he verifies the final output, the job is ready to compile and run.

Based on the databases and the specific functions of the databases that the developer wants to leverage, InfoSphere DataStage can help create an optimal job that leverages the power of the InfoSphere DataStage parallel engine, the parallel I/O and the databases that may be involved.

Table 1 contains a subset of potential optimizations. These optimizations depend on support from the databases and the location of the data.

	Push to source	Push to target
Processing (transformations, modify, filter, etc.)	✔	✔
Aggregation	✔	*
Joins, lookups and merges	✔	✔
Drop unnecessary processing or data	✔	✔
Use bulk I/O operations	✔	✔
Use temporary staging tables	✔	✔

* Pushing aggregation and other significant data-reduction processing to the target may not make sense if it requires loading a large amount of data into the target only to reduce it substantially. It does make sense when the aggregation or data reduction involves data already inside the target database.

Table 1: Potential optimizations can push job functionality to source or to target.

Job design and optimization example

This example illustrates pushing work into a data source and a data target, and the compositional power of the rules-based engine.

Figure 2 displays the initial job design to identify late orders. The job extracts orders from an Orders table, filters out orders that have already shipped, enriches the unshipped orders by joining with a Customers table, computes the number of days each order is late, modifies the data to comply with company standards for data representation, further enriches the list with the selling salesperson data and produces the desired output.

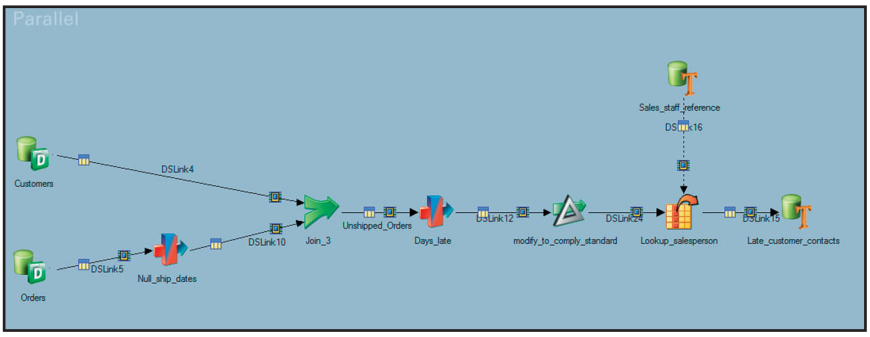

Figure 2: The initial InfoSphere DataStage job design does not include optimizations.

Figure 3 shows the same job with optimizations applied by IBM InfoSphere DataStage Balanced Optimization. This example uses four optimizations:

- *Push transformations into a source SQL statement*
- *Push joins of tables in the same data source into a SQL join*
- *Execute the data modification to company standards in the ETL engine*
- *Push lookup and transformations into the target*

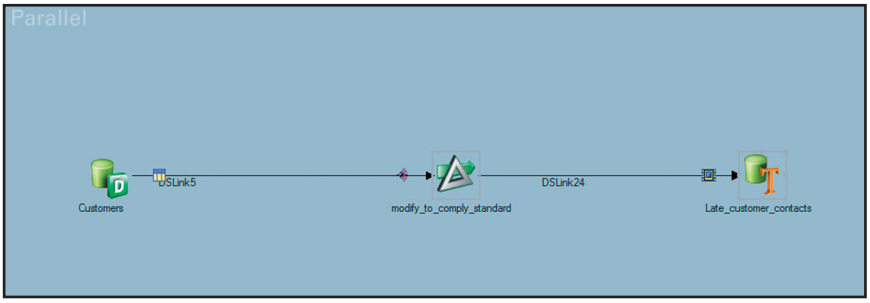

Figure 3: The completed job includes four optimizations.

The result is a simplified data integration flow that is optimized to take advantage of the strengths and efficiencies of the source and target databases as well as the InfoSphere DataStage engine. Also, with the rules-based approach provided by InfoSphere DataStage Balanced Optimization, this optimized flow can be created quickly and largely automatically, improving the efficiency of the developer.

Powerful, flexible data integration optimization tools can help organizations maximize the value of their corporate information while leveraging the significant investment made in the processing capacity of the RDBMS. IBM InfoSphere DataStage Balanced Optimization facilitates the creation of scalable, flexible, optimized data integration architectures while increasing the efficiency of the integration developers.

IBM InfoSphere Information Server Balanced Optimization specifications

Platform	• IBM InfoSphere DataStage v8.1 and later; parallel job types
Databases	• IBM DB2 UDB 9.1 • Teradata v2.5, v2.6, v12

For more information

To learn more about IBM InfoSphere Information Server and IBM InfoSphere DataStage Balanced Optimization, please contact your IBM marketing representative or IBM Business Partner, or visit **ibm.com**/software/data/integration/info_server_platform

IBM Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
June 2008

TAKE BACK CONTROL WITH **Information Management**

IMW14049-USEN-00