The advantages of upgrading to InfoSphere DataStage 8.7

Maximize your InfoSphere DataStage investment with high-performance parallel processing

Highlights

- Enables high-performance batch and real-time data extraction, transformation and loading
- Provides built-in scalability to future-proof your architecture
- Helps developers be more efficient and productive through automation and reuse of common development tasks

Industry experts agree that data and analytic processing demands are on the rise: for example, IDC predicts that “the amount of information managed by enterprise data centers will grow by a factor of 50”[1] over the next decade. At the same time, new requirements for information governance and data mining are expanding the role of analysis and solution delivery within IT departments.

As a result, data integration needs and organizational service-level agreement (SLA) requirements for timely information and project delivery are placing unprecedented stress on existing infrastructure—significantly driving up costs as IT attempts to keep pace with these rapid project deployments and exploding data growth. Collaboration and reuse is critical during these hyper-development cycles to maintain cost control and preserve solution quality.

IBM® InfoSphere® DataStage® responded to those forecasted demands in 2001 by adding a unique, high-performance parallel processing engine built upon a single integrated repository for maximum integration and collaboration between team members. The coupling of the “build as you think” graphical design environment with the parallel engine provides scalability and speed focused on automation of common functions and features for all developers. During the past decade, thousands of customers have taken advantage of the advanced capabilities for their data integration processing requirements.

Many significant advantages persuaded these customers to upgrade from InfoSphere DataStage Server to the automated, parallel technology in InfoSphere DataStage. The powerful, industry-leading parallel engine provides built-in scalability to future-proof your architecture using a design-once-and-deploy-anywhere approach (see Figure 1). In addition, the new advanced transformation capabilities will help developers be more efficient and productive through automation and reuse of common development tasks.

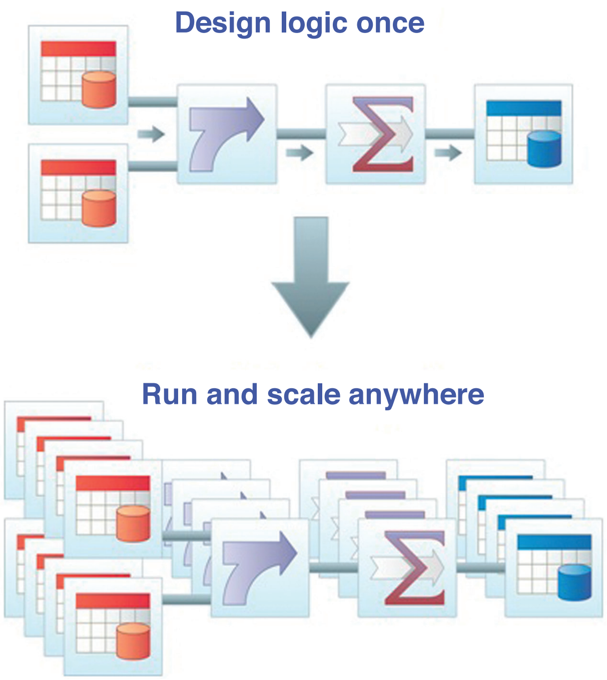

Figure 1: InfoSphere DataStage 8.7 supports a flexible and scalable runtime from the connectivity layer through all transformation tasks to scale with massive data volumes, shrinking batch windows and reducing hardware requirements.

What is InfoSphere DataStage?

InfoSphere DataStage is the flagship IBM data integration product that enables high-performance batch and real-time data extraction, transformation and loading in parallel between multiple sources and targets with automated development capabilities. The current version is InfoSphere DataStage 8.7.

What is InfoSphere DataStage Server?

InfoSphere DataStage Server is the original IBM extract, transform and load (ETL) product. It performs basic data extraction, transformation and loading through a straightforward graphical design environment.

Why upgrade from InfoSphere DataStage Server to InfoSphere DataStage 8.7?

The automated, parallel processing capabilities of InfoSphere DataStage 8.7 future-proof your organization's architecture so as your data volumes grow, your data integration environment can support those needs on an ongoing basis. It provides many functions and features that significantly enhance developer productivity and enable higher degrees of automation compared to its predecessor, InfoSphere DataStage Server. In addition, InfoSphere DataStage contains a variety of additional stages, functions and add-on modules not available in InfoSphere DataStage Server.

How does upgrading to InfoSphere DataStage 8.7 save my organization time and money?

First, your architecture will scale more efficiently because you will be able to take full advantage of your hardware's capabilities, including various high-availability configurations to help ensure maximum uptime when primary hardware fails. Second, processing more data in less time will help your organization maintain compliance with SLAs and other performance-based agreements. Third, the enhanced functionality provides standardized and configurable options for many complex data integration challenges, helping developers meet project timelines and contain cost overruns caused by custom coding. Lastly, InfoSphere DataStage works intuitively with the other components of IBM InfoSphere Information Server, providing customers with extensive capabilities to accelerate requirements gathering, jumpstart their job design, monitor and manage data quality, and more.

Enhance developer productivity

InfoSphere DataStage 8.7 offers features not available in InfoSphere DataStage Server that enhance developer productivity and deliver rapid time to value for information-centric initiatives. Key features include:

- Advanced stages for complex data integration requirements
  - Slowly changing dimension stage
  - Vertical pivot stage
  - Enhanced surrogate key management stage
  - Range lookup for data validation
  - Checksum stage
  - Array processing stages
- Transformation stage enhancements
  - Looping and caching functionality for custom multi-row processing logic
  - Robust utilities for string and date/time math functions
- State-of-the-art debugging features
  - Interactive debugger supporting SMP, MPP and grid environments running in multiple degrees of parallelism
  - Sample data generation
  - Data sampling for head and tail of the data stream
  - Directed output of data flowing through a link to the job log
- Extensible components that enable existing scripts, routines and other code to be reused and deployed in parallel
- Five powerful data joining methods: Join, Merge, Funnel, Database Lookup and In-Memory (equality and range) Lookup—all supporting multiple partition processing

> *"My clients have seen tremendous value in the new features and products associated with 8.5. I have upgraded multiple customers to release 8 so they could migrate from InfoSphere DataStage Server to InfoSphere DataStage Parallel and its family of products. Developers easily picked up the new parallel development techniques with a few days of training and quickly delivered 5x to 10x performance improvements on mission-critical projects."*
>
> —Andy Sorrell, Independent Consultant

Deliver advanced connectivity with superior performance

InfoSphere DataStage 8.7 offers improved connectivity designed for greater performance and better exploitation of new hardware than previous options available in InfoSphere DataStage Server. Examples include:

- Native partitioned loaders and readers for database management systems (DBMS) such as IBM DB2®, IBM Netezza®, Oracle and Teradata
- Exchanging information with big data sources such as the Hadoop Distributed File System

- Connectors for analytical sources like Hyperion and SAS
- Extensive automation features for enterprise application systems such as SAP, PeopleSoft, salesforce.com and more
- Integrated connectors with other IBM solutions including change data capture (CDC) for log-based data sourcing
- Distributed transaction stage with real-time guaranteed delivery from MQ or CDC sources to various DBMS targets via a two-phase commit
- Expanded support for multi-format flat file, reading files in parallel, IBM z/OS® file stage and iWay enterprise stage

Easily scale to meet demanding workloads

The powerful parallel engine in InfoSphere DataStage helps increase data throughput and reduce batch processing windows.

- Parallelism is controlled at runtime, not during design, to minimize development and operational complexity.
- Use of parallel data sets is designed to avoid inefficient temporary tables and sequential files for storage.
- Linear scalability means organizations can take advantage of new hardware as business requirements grow.
- Significant performance improvement is included for data sorts, joins and aggregations leveraging parallel capabilities.
- Increased computing capacity and flexibility at a lower cost in a grid environment with configuration options available through IBM.

> *"We recently upgraded and are seeing a huge boost in performance. Our ETL processes are running two to four times faster without any tuning."*
>
> —Retail Customer, Project Lead

What happens to my old InfoSphere DataStage Server jobs?

InfoSphere DataStage 8.7 supports InfoSphere DataStage Server jobs in the same installation, so you can introduce parallel capabilities to your most challenging processing requirements without having to rip-and-replace your complete environment. Also, InfoSphere DataStage 8.7 supports compatibility features such as database connector stages for your existing InfoSphere DataStage Server jobs. Migrating now ensures you will be able to take advantage of all current and upcoming InfoSphere DataStage functionality enhancements.

Automate management and tuning

InfoSphere DataStage 8.7 offers several capabilities to proactively manage and tune data integration processes and help make sure systems are running optimally at all times. Some key highlights are:

- The Operations Console, which provides a web-based dashboard and detailed views of the correlated job runtime and machine resource statistics
- Automated job performance analysis
- Resource estimation that models machine requirements as jobs scale
- Run-time optimization features that maximize job throughput
- IBM Support Assistant (ISA) Lite provides system health analysis to help ensure the machine is running optimally

Data matching and cleansing with InfoSphere QualityStage

IBM InfoSphere QualityStage® is IBM's leading data standardization, cleansing and probabilistic matching engine.

InfoSphere QualityStage is an add-on module for InfoSphere DataStage that shares the same user interface, data integration design and parallel runtime paradigm for easy adoption and deployment. Its capabilities include:

- Thorough data investigation and analysis processing for any kind of free-form data
- A single set of standardization, cleansing, matching and survivorship rules for your core business entities as stages on the Designer canvas
- A matching engine that leverages probabilistic technology to achieve superior match rates
- Ability to execute in batch, real-time or as a web service
- Worldwide address standardization verification and enrichment capabilities, including multiple postal certification modules
- Support for the creation and maintenance of high-quality master data for enterprise initiatives, including a single view of customer, supplier, product, location and more
- Serving as a foundational component of data quality programs and data governance initiatives

Integrated data rules with InfoSphere Information Analyzer

InfoSphere DataStage is also fully integrated with the IBM InfoSphere Information Analyzer data profiling and validation engine.

InfoSphere Information Analyzer helps users quickly and easily understand data by offering data quality assessment and monitoring, as well as flexible data rules design and analysis capabilities. These insights help users derive more value from enterprise data and accelerate information-centric projects. Users and developers benefit from the following capabilities:

- Shared metadata of data profiling results is available in the Designer interface to help developers understand how to best use sources of information.
- A common rules framework provides the ability to evaluate, analyze and address multiple data issues by record rather than in isolation
- Rules can be executed directly against an information source, via the InfoSphere Information Analyzer user interface or as part of an InfoSphere DataStage job for in-flight data validation via a stage on the Designer canvas.

Additional add-on module support available only with InfoSphere DataStage 8.7

There are several other modules designed to work specifically with InfoSphere DataStage 8.7:

- Balanced optimization adds complex ELT support for efficient database-centric transformations and utilization of database resources.
- InfoSphere DataStage Pack for Data Masking protects personally identifiable information (PII) through obfuscation that is specific to the class of data (for example, distinct algorithms for Social Security Numbers, phone numbers, names and so on).
- IBM InfoSphere FastTrack supports creation of mapping specifications and turns these directly into InfoSphere DataStage data integration jobs—or supports reverse engineering of parallel data integration jobs and creating documentation.

Making the migration to InfoSphere DataStage 8.7

IBM will work with you to determine the best option for your organization, and make your upgrade as fast and seamless as possible.

Please contact your sales representative for more details on how to migrate from InfoSphere DataStage Server to InfoSphere DataStage 8.7.

For more information

For more information about IBM InfoSphere Information Server and related services, visit: **ibm.com**/software/data/infosphere/datastage

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. We'll partner with credit-qualified clients to customize an IT financing solution to suit your business goals, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward. For more information, visit: **ibm.com**/financing

IBM Corporation
Software Group
Route 100
Somers, NY 10589 U.S.A.

Produced in the United States of America
December 2011

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

[1] Gantz, John and Reinsel, David. “Extracting Value from Chaos." IDC View. June 2011.

Please Recycle