IBM InfoSphere DataStage

Enterprise-scale data integration, transformation and delivery

Every day, torrents of data inundate IT groups and overwhelm business managers. Businesses need to sift through data to glean insights that can help increase company revenues and optimize profits. Yet even after pouring millions of dollars into business automation projects, such as enterprise resource planning (ERP), customer relationship management (CRM), supply chain management (SCM), business intelligence (BI) and data warehousing solutions, many companies are still plagued with disconnected, “dysfunctional” data. They are stuck with a massive, expensive sprawl of disparate information silos and nonintegrated, redundant systems that fail to deliver a single trusted view of the enterprise.

As a first step in deploying new enterprise applications, integrating business acquisitions or tackling IT business process initiatives, organizations need to ensure that data is reliable, relevant and readily available—in a consistent, condensed form—across the enterprise.

IBM InfoSphere Information Server can help organizations understand, cleanse, transform and deliver data for critical business initiatives. It tightly integrates a full range of enterprise information across a wide variety of data sources and targets. Whether building an enterprise data warehouse, implementing a BI platform or integrating dozens of source systems to support CRM, SCM or ERP, InfoSphere Information Server helps ensure that organizations have information that is trustworthy and actionable. With InfoSphere Information Server, you can design an infrastructure for data integration that is reliable, scalable and flexible enough to accommodate today’s dynamic business environments.
Integrate, transform and deliver trustworthy information with IBM InfoSphere DataStage

IBM InfoSphere DataStage® is a core product module of the InfoSphere Information Server information integration platform. It offers world-class data integration and transformation capabilities for unsurpassed levels of productivity.

Key features of IBM InfoSphere DataStage

- Offers unsurpassed connectivity to operational systems, databases and enterprise applications spread across mainframe and distributed systems
- Uses a common metadata repository for seamless integration with other InfoSphere Information Server modules with data profiling and data quality capabilities
- Provides an easy-to-use, top-down, work-as-you-think design interface
- Includes a comprehensive library of transformation components for easily defining common integration processes
- Incorporates a complete tool set to administer, deploy and update the data flows throughout the data integration life cycle
- Supports a Service Oriented Architecture (SOA) approach, facilitating reuse of complex integration data flows
- Builds on a dynamic parallel processing infrastructure that enables users to design once and deploy as needed at run time without changing any integration jobs
- Delivers real-time data feeds to optimize extract, transform, load (ETL) processes

Supported sources, targets and applications include:

- Almost any database, including partitioned databases such as IBM DB2® on any platform (with and without the Data Partitioning Feature), IBM Informix®, Netezza®, Oracle, Sybase (ASE and IQ), Teradata and Microsoft® SQL Server®, MySQL and Progress
- Web services
- Warehouse environments such as Teradata, Netezza and IBM InfoSphere Warehouse
- Analytic tools such as SAS and IBM SPSS
- Messaging and enterprise application integration (EAI) products including IBM WebSphere® MQ
- Public and private clouds

InfoSphere DataStage is a core component of the InfoSphere Information Server platform and leverages a wide array of capabilities that are native to the platform. Some of those key capabilities are:

Extensive enterprise connectivity

Successful enterprise-class information integration requires access to a full range of data sources—structured, semistructured or unstructured—within and outside of the enterprise. InfoSphere DataStage provides connectivity to a virtually unlimited array of heterogeneous data sources, targets and applications, which can be combined within a single job (see Figure 1).

Advanced development and maintenance

InfoSphere Information Server uses a powerful architecture that helps developers maximize speed, flexibility and effectiveness in building, deploying, updating and managing their data integration infrastructure. InfoSphere DataStage leverages the productivity-enhancing features of InfoSphere Information Server to help reduce the learning curve, simplify administration and optimize the use of development resources. The result is an accelerated development and maintenance cycle for data integration applications. With InfoSphere DataStage, organizations can achieve strong ROI by gaining access to trustworthy information and sharing it across applications and databases.
Integrated design interfaces and common metadata repositories

InfoSphere Information Server has a single design interface that is shared by both InfoSphere DataStage and IBM InfoSphere QualityStage™ modules, enabling designers to use any combination of data quality and data transformation capabilities to help ensure that the right data is brought together at the right time. InfoSphere Information Server also provides a unified metadata repository for InfoSphere DataStage and all other modules. Users can immediately access technical and process metadata developed during data profiling, cleansing and integration processes to speed development and reduce the chance for errors.

Ease of use

InfoSphere DataStage employs a work-as-you-think design interface (see Figure 2). Developers use a top-down model of application programming and execution, which allows them to create a visual data flow. A robust graphical palette helps developers diagram data flow through their environments with simple, GUI-driven, drag-and-drop design components. To maximize productivity, InfoSphere DataStage includes more than 50 prebuilt components and hundreds of transformations. Developers also benefit from powerful debugging capabilities and an open application programming interface (API) for leveraging external code.

Productivity and reuse

InfoSphere DataStage helps shorten the development cycle by promoting the reuse of existing data integration business logic. It employs a container concept that enables jobs and metadata created in one container to be shared and reused by other jobs. Quick Find and Advanced Find capabilities make it easy to locate objects for reuse across different projects. Robust job specification reporting provides documentation so other developers can easily understand job design and provide additional support.

Right-time data integration

The InfoSphere Information Server architecture enables InfoSphere DataStage to operate in real time, capturing messages or extracting data at a moment’s notice on the same platform that integrates bulk data and using the same transformation rules. Data integration jobs can be deployed with Java™ Message Services, web services and other methods. This SOA approach enables numerous developers to share complex data integration processes without requiring them to understand the steps contained in the services. The result: data can be used in more ways—without costly hand coding—to respond to an organization’s information integration needs on demand.

Figure 1: InfoSphere DataStage enables access to multiple source systems, integrating and transforming selected data to deliver trustworthy information to critical business functions.
**Market-leading flexibility and scalability**

InfoSphere Information Server facilitates high-performance integration of large amounts of data. By leveraging the parallel processing capabilities of multiprocessor hardware platforms, InfoSphere DataStage enables businesses to linearly scale the speed of data throughput. Organizations can scale transformation jobs to address the demands of ever-growing data volumes and ever-shrinking batch windows. Development is done using sequential logic—the deployment configuration automatically adds the desired degree of parallelism. An organization could take the application from 2-way processing in the morning to 32-way in the afternoon to 128-way processing at night—all with a simple change to the configuration file.

**The secret: Partitioning and dynamic repartitioning**

InfoSphere Information Server parallel technology operates using a divide-and-conquer technique, splitting the largest integration jobs into subsets (partition parallelism) and flowing these subsets concurrently across all available processors (pipeline parallelism). This combination of pipeline and data partition parallelism delivers true linear scalability for InfoSphere DataStage. Performance increases proportionally to the number of processors—hardware is the only limiting factor to performance.

*Figure 2: InfoSphere DataStage makes it easy to design enterprise data flows using a top-down, work-as-you-think metaphor*
Consider a transformation based on customer last name, where the enriching needs to occur on zip code for householding purposes and on credit card number for loading into the data warehouse database parallel loader. With dynamic data repartitioning, data is repartitioned on the fly between processes—without landing the data to disk, a slow and costly step required by many other integration products.

**Wide-ranging parallel support for SMP, MPP and grid deployments**
IBM InfoSphere Information Server scales effortlessly from symmetric multiprocessor (SMP) systems and SMP clusters to massively parallel processing (MPP) servers with hundreds of processors. The same integration capability is available for grid deployment on low-cost servers. This wide-ranging support for parallel processing helps ensure critical enterprise information integration jobs will scale to keep pace with business requirements.

IBM InfoSphere DataStage is available on a wide variety of platforms. For a complete list, please visit [ibm.com/software/data/infosphere/datastage/requirements.html](http://ibm.com/software/data/infosphere/datastage/requirements.html)

**About IBM InfoSphere Information Server**
InfoSphere Information Server is a market-leading data integration platform that helps you understand, cleanse, transform and deliver trustworthy information to your critical business initiatives (see Figure 3). The platform provides everything you need to integrate heterogeneous information from across your systems, including capabilities for information governance, data quality, data transformation and data synchronization that help ensure information is consistently defined, accurately represented, reliably transformed and regularly updated. InfoSphere Information Server facilitates and promotes collaboration among business and IT professionals to make sure that strategic initiatives such as business analytics, master data management, application consolidation, migration and data warehousing projects use trusted information that is accurate, comprehensive, insightful and available in real time.

Additional details on the IBM InfoSphere Information Server portfolio can be found at [ibm.com/software/data/integration/info_server](http://ibm.com/software/data/integration/info_server)

**For more information**
For more information about IBM InfoSphere DataStage, please visit [ibm.com/software/data/infosphere/datastage](http://ibm.com/software/data/infosphere/datastage)