IBM InfoSphere DataStage
Version 8 Release 7

Introduction to IBM InfoSphere DataStage

GC19-3607-00

IBM InfoSphere DataStage
Version 8 Release 7

Introduction to IBM InfoSphere DataStage

GC19-3607-00

Note

Before using this information and the product that it supports, read the information in “Notices and trademarks” on page 29.

Contents

Chapter 1. Overview of InfoSphere DataStage

IBM® InfoSphere® DataStage® is a data integration solution that collects, transforms, and distributes large volumes of data, with data structures that range from simple to highly complex.

InfoSphere DataStage integrates data by using a high-performance parallel framework, extended metadata management, and enterprise connectivity. It also supports real-time data integration and offers a scalable platform that enables companies to solve large-scale business problems through high-performance processing of massive data volumes.

With InfoSphere DataStage, you can accomplish these goals:

- Create visual, sequenced data flows by using a top-down data flow model to build and run applications. A simple but robust graphical palette allows you to diagram the flow of data through your environment using drag-and-drop user interface design components.
- Design data flows that extract information from multiple source systems, transform that information in ways that make the data more valuable, and then deliver the data to one or more target databases or applications.
- Connect a wide variety of data sources and applications using a common set of tools and skills, enabling you to maximize speed, flexibility, and effectiveness in building, deploying, updating, and managing your data integration infrastructure.
- Leverage external code by using the adaptability and power of a versatile scripting language, powerful debugging capabilities, and an open application programming interface (API).

To start learning about InfoSphere DataStage, review the case studies, concepts, processes, and architecture.

Chapter 2. Case studies

Case studies are useful in learning about InfoSphere DataStage because they provide examples that show how the product is used in real situations. They also showcase some of the ways that companies rely on InfoSphere DataStage to accomplish their business objectives.

InfoSphere DataStage provides accurate data

Companies can rely on the broad range of connectivity offered by InfoSphere DataStage to deliver accurate data rapidly and in a standardized manner.

In the global economy, leveraging information has become key to competitive success. Yet trying to manually manage an exploding volume of data that is stored in silos has made it difficult for companies to unlock the value of their information for competitive advantage.

A large healthcare company understands this challenge all too well. The mission of the company is to improve healthcare delivery by making patient information available at the point of care. To do so, they need to rapidly consolidate, standardize, and manage information from a number of third-party partners that use a wide array of data sources and data structures. These partners include insurers, labs, prescription drug clearing houses, and health providers.

The staff that is responsible for data integration and business intelligence solutions had developed customized programs using COBOL to facilitate the integration process. However, manually coding applications to perform data integration and quality checks was time consuming.

In anticipation of nearly doubling the number of patients supported, the company needed a data integration platform that could quickly and cost-effectively profile, cleanse, and integrate information, regardless of the format or the source. Using InfoSphere DataStage along with other components of IBM InfoSphere Information Server allows the organization to create a single, accurate, and trusted source of information for populating its health record repository, clinician portal, and data warehouse.

The team was able to leverage the common metadata repository and user interface of InfoSphere Information Server to combine the extract, transformation, and load (ETL) processes of InfoSphere DataStage with the information analysis capabilities of InfoSphere Information Analyzer and the data quality features of InfoSphere QualityStage™. Doing so enabled them to optimize their results and implement a solution in just a few weeks.

Because this solution can execute processes in parallel, the staff can perform data analysis on an entire database table consisting of millions of rows and hundreds of columns in less than two hours. Previously, this task would have taken more than 24 hours to accomplish. The company then uses InfoSphere DataStage to collect, integrate, and transform data from its partners and make the data available to providers, leveraging the parallel processing capabilities of multiprocessor hardware platforms to rapidly handle large volumes of data.

Being able to effectively integrate information regardless of its source or structure is helping this healthcare company thrive. As a result, they expect to realize significant revenue growth - growth that their business intelligence team can handle without a huge investment in resources.

InfoSphere DataStage gives a complete picture

Companies can facilitate decision making by using InfoSphere DataStage to reconcile related information into a single and holistic view.

A clothing manufacturer who is a major player in the high-fashion apparel industry needed faster and more actionable information to speed decision making and to keep its processes in sync with the fast-changing market. To ensure that they have the right mix of products on the retail floor at any given time, high-fashion apparel manufacturers have to not only sense changes in selling patterns, but they also need to quickly translate that intelligence into a series of coordinated decisions that go right up the supply chain. These decisions range from knowing when and how much to ramp up or cut back on the production of some styles, sizes, and colors to choosing the right mix of transport modes to balance the urgency of delivery against cost.

Assembling the information that was needed to make key decisions was an arduous and time-consuming exercise. The primary sources of data were the five separate systems that the company relied on to run its business. Another key data source was the standardized product activity transaction reports that the company received from its wholesale channel. To unify this information into a coherent and complete picture of the situation, employees in various departments were required to manually integrate the data into spreadsheets. Only then could the managers make such basic decisions as which products to ship to each store, which products to order from suppliers, and how best to get new shipments in from overseas.

The inherent inefficiency of this approach was only the beginning of the problem. Limitations were placed on the company's ability to make decisions because after the data was generated by the company's core systems, it could take up to two days for managers to have the information in a form that they could act on. In addition to timeliness and transparency, the report deprived managers of the granularity that they needed to make decisions that could optimize their business operations.

Guiding many of these decisions was the overriding importance of meeting commitments to retailers. To minimize the risk of late deliveries, managers often reverted to air transport, which is three times the cost of water transport. Further, in-store replenishment decisions were hampered by lack of granularity, making it impossible to fine-tune the product and size mix that was shipped to the stores based on the differences in sales patterns from store to store or from region to region.

The company implemented a solution that uses InfoSphere DataStage and other IBM products to move data from its core applications to its data warehouse. Real-time information about sales, inventory, and shipments is captured directly into the company's core transactional systems, and the transactional information from five disparate core platforms is standardized and integrated into a single reporting framework.

The company now has faster and more intelligent decision-making through the availability of real-time sales, inventory, and logistics information. The reporting

cycle was reduced from as many as two days to a few minutes, and supply chain and logistics costs were reduced by 30%. Sales have increased due to the ability to provide an optimized mix of products on the retail floor, and the brand itself is strengthened by the company's increased responsiveness to changes in fashion trends.

InfoSphere DataStage brings context to data

Companies can eliminate redundant problem solving by using the simple drag-and-drop user interface of InfoSphere DataStage to deliver relevant information in real time, when and where it is needed.

A major computer services and technologies company provides a maintenance service that is the key offering in helping its clients achieve the highest level of network availability. Through this service, network problems are repaired in 2.5 hours or less following a failure notification.

One challenge in minimizing the mean time to repair network problems and maximizing staff productivity has been the difficulty in accessing past failure response information. Without insight into previous issues, staff often had to diagnose repetitive problems from scratch. Additionally, because parts management and the engineer dispatch system worked independently of each other and were not organically linked, the notations in each application could vary, causing further delays.

Aggregating information into a single repository would help this company arm its engineers with essential information that could accelerate repair times. To create a single data warehouse, the staff had to aggregate and manage a wide range of information from various sources. At the same time, executives planned to use the same data to streamline parts distribution processes with business partners.

Using InfoSphere Information Server as its integration platform and InfoSphere DataStage to implement the data warehouse, the company achieved its goals in just four months.

By using InfoSphere DataStage, information is provided to and shared among the applications used in each task. This information includes data about network failures, customer information, and components. Additionally, InfoSphere DataStage enables the sharing of maintenance component distribution instructions, distribution information, and other information with an external distribution partner. This exchange of data with the distribution partner is critical in helping ensure that the appropriate maintenance materials arrive at the customer site in a timely manner. For example, through the use of InfoSphere DataStage, the processing of maintenance material shipping instructions now occurs in near real time. As a result, delivery time of replacement components to the customer site has decreased from two hours to one hour, helping to reduce the overall mean time that is needed to repair network problems.

By using InfoSphere DataStage as part of a data integration platform that can aggregate information across multiple data sources and target applications, this company has reduced its mean time to repair network problems by ensuring that engineers have access to accurate and complete customer information.

InfoSphere DataStage offers insights into data

Companies can leverage the scalability and performance of InfoSphere DataStage to derive meaning from information as that information changes.

Imagine having a daily pulse on your business and knowing immediately which sales promotions were working, which products delivered the highest profitability, and which locations offered the most promise for new stores. Executives at a family-owned chain of grocery stores knew that this type of insight could help them grow the business into a $1 billion (USD) company. However, with more than 6 terabytes of product and customer data spread across different systems and databases, they could not easily assess operations at each store.

The company used InfoSphere DataStage to integrate data across its 15 stores and corporate systems to enable the sharing of trusted information and gain greater insight into operations. Doing so now enables corporate personnel to quickly review daily inventory levels, store sales, and cost of goods to see which products are selling, which are most profitable, and which promotions are most successful. For example, using InfoSphere DataStage, data from each point-of-sale system in every store is loaded daily into the company's IBM Informix-based sales consolidation system, helping company executives more quickly spot increasing demand for specific products.

Because the local government mandates prices for many grocery staples, such as milk, eggs, and bread, the company realizes another benefit from the seamless flow of information between their SAP and point-of-sale systems: they are able to quickly update prices across all stores, as needed, and more easily confirm compliance with government regulations.

By integrating information across the enterprise, the grocery chain has realized a nearly 30% increase in revenue and a $7 million (USD) increase in annual profitability. The CIO attributes these increases to better inventory management and the ability to more quickly adjust to changing market conditions. For example, the company has prevented losses for about 35% of its products now that it can schedule price reductions to sell perishable products before they spoil.

Savings are also being realized with improved staff productivity. Previously, it could take nearly a month for finance staff to manually compile sales tax information. Now, the information is immediately available in the SAP system with a simple query – a more than 98% improvement.

Additionally, new insight has helped corporate personnel better understand sales by location to determine where to build new stores. In fact, the company has successfully opened four new locations, including a new "supercenter," based on consumer behavior and buying patterns.

Chapter 3. Key concepts

InfoSphere DataStage provides the elements that are necessary to build data integration and transformation flows. These elements include stages, links, jobs, table definitions, containers, sequence jobs, and projects.

Stages

Stages are the basic building blocks in InfoSphere DataStage, providing a rich, unique set of functionality that performs either a simple or advanced data integration task. Stages represent the processing steps that will be performed on the data.

A stage describes a data source, a data processing step, or a target system and defines the processing logic that moves or transforms data. InfoSphere DataStage comes with a set of predefined stages, which are available on a palette from which you can drag them into the design canvas. These stages perform most common data integration tasks.

Each stage has properties that you use to specify the task that the stage will perform or how to process the data. A stage usually has at least one data input and one data output. Some stages, though, can accept more than one data input, and often, stages can output to more than one location.

Although the predefined stages that come with InfoSphere DataStage provide most of the application logic that you need in your enterprise, you might find that you need to create your own custom stages. InfoSphere DataStage provides a number of stage types for building custom stages.

Stages also offer configuration options that you can select to drive transformation tasks, alleviating tedious coding.

Related information

Alphabetical list of stages

Types of stages

Links

A link is a representation of a data flow that joins the stages in a job. A link connects data sources to processing stages, connects processing stages to each other, and also connects those processing stages to target systems. Links are like pipes through which the data flows from one stage to the next.

Links are used to specify how the data flows from one stage to another. Within InfoSphere DataStage, there are four types of links:

Input link

An input link is displayed as a solid line, which indicates the primary flow of data. An input link is used to connect a data source to a stage so that data can be processed.

Output link

An output link is also displayed as a solid line, again indicating the primary flow of data. An output link is connected to a stage and generally moves processed data from the stage.

Reference link

A reference link is a specific type of input link, and is displayed as a dotted line, which indicates that table lookups are being performed. It is an input link on a Transformer or Lookup stage that defines where the lookup tables exist.

Reject link

A reject link is displayed as a dashed line, which indicates that records were rejected. Records are rejected when they do not meet the business logic of the job. A reject link is an output link that identifies errors when the stage is processing records and then routes rejected records to a target stage.

Related information

Using links

Linking stages

Jobs

Jobs include the design objects and compiled programmatic elements that can connect to data sources, extract and transform that data, and then load that data into a target system. Jobs are created within a visual paradigm that enables instant understanding of the goal of the job.

A job is used to combine stages and links. A job represents the flow of data through InfoSphere DataStage. Within a job, stages represent data sources (input), the required transformations, and the destination of the data target (output). Links indicate the path of the data from the input, through each transformation, and to the output. A job can have multiple inputs, transformations, and outputs.

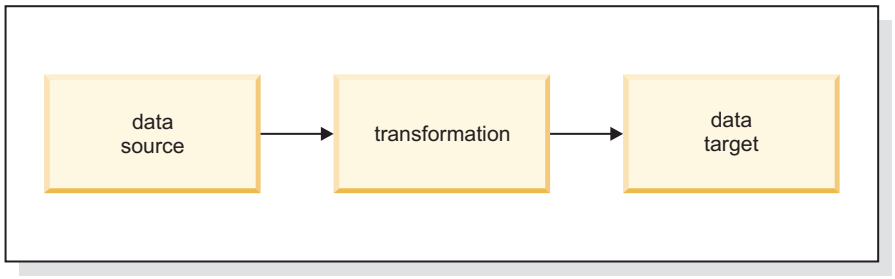

Figure 1. An example of a simple job

Related information

Getting started with jobs

Creating a new job

Creating a job from a template

Sequence jobs

A sequence job is a special type of job that you can use to create a workflow by running other jobs in a specified order. This type of job was previously called a job sequence.

By using sequence jobs, you can create more complex job designs. For example:

- You can build programmatic controls, such as branching or looping.
- You can specify different courses of action to take depending on whether a job in the sequence job succeeds or fails.
- You can run system commands or send emails.
- You can perform exception handling. For example, if any job in the sequence fails, you can transfer the sequencer control to a specific workflow branch where you define how the event will be handled.

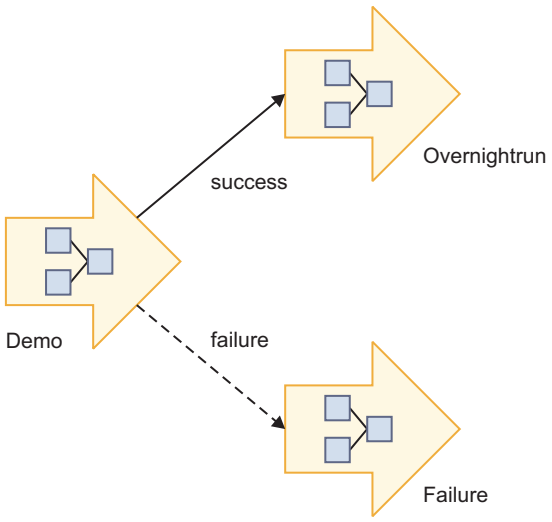

Figure 2. An example of a sequence job

Some of the components that you use in a sequence job differ from those components that you use in a job. For example, in a job, you use stages; however, in a sequence job, you use activities, such as parallel or server jobs. Similarly, you use links in a job, but in a sequence job you use triggers to define control flow.

You can also create a job control routine that controls other jobs from the current job. For example, you might create a job control routine that schedules two jobs, waits for them to finish running, tests their status, and then schedules a third job.

Related information

Building job sequences

Creating a job sequence

Activities

Triggers

Table definitions

Table definitions specify the format of the data that you want to use at each stage of a job. They can be shared by all the jobs in a project and between all projects in InfoSphere DataStage. Typically, table definitions are loaded into source stages. They are sometimes loaded into target stages and other stages.

Table definitions contain information about your source and target data. Links hold the table definitions and other metadata for the data that is moving between the stages. Table definitions include information such as the name and location of the tables or files that contain your data.

Table definitions also contain information about the structure of your data. Within a table definition are column definitions, which contain information about column names, length, data type, and other column properties, such as keys and null values.

Table definitions are stored in the metadata repository and can be used in multiple InfoSphere DataStage jobs. You can also use table definition metadata to facilitate data governance. For example, when business or IT users need to understand how data flows through the enterprise systems, you can perform impact analysis to identify relevant changes to the IT infrastructure and data lineage.

Related information

Setting properties for table definitions

Table definitions within parallel jobs

Containers

Containers are reusable objects that hold user-defined groupings of stages and links. Containers create a level of reuse that allows you to use the same set of logic several times while reducing the maintenance. Containers make it easy to share a workflow, because you can simplify and modularize your job designs by replacing complex areas of the diagram with a single container.

Containers help to simplify your job design. When business requirements need to be implemented by several stages, containers can help focus attention on subsets of the job design and then allow users to drill into that detail.

If the job has several stages and links, you might find it useful to create containers to describe a particular sequence of steps within the job. Containers are linked to other stages or containers in the job by input and output stages.

There are two kinds of containers:

Local container
A local container simplifies your job design. A local container can be used in only one job. However, you can have one or more local containers within a job.

Shared container
A shared container facilitate reuse. They can be used in many jobs. As with local containers, you can have one or more shared containers within a job.

You can use a mixture of local and shared containers within the same job.

You can use shared containers to make common job components available throughout the project. You can use a stage and its associated metadata to create a shared container. You can then add the shared container to the palette to make this pre-configured stage available to other jobs.

You can create a shared container from scratch, or you can place a set of existing stages and links within a shared container.

Related information

Local containers

Shared containers

Modular development

Projects

A project is a container that organizes and provides security for objects that are supplied, created, or maintained for data integration, data profiling, quality monitoring, and so on.

Projects are a method for organizing your work. You define data files, define stages, and build jobs in a specific project.

A project can contain one or more jobs. Any of the metadata objects in a project (such as jobs or table definitions, for example) can be grouped logically and organized into folders.

You can define security at the project level. Only users who are authorized for your project can access your jobs.

Related information

Projects page

Setting up a project

Chapter 4. Job design

You use a job to extract, transform, load, or check the quality of data. Building jobs in InfoSphere DataStage starts with a good design that is based on a strong understanding of your data integration requirements.

A job design is the metadata that defines the sources and targets that are used within a job and the logic that operates on the associated data. A job design is composed of stages and the links between those stages. That is, each data source and each transformation step are stages in the job design, and the stages are connected using links to show the flow of data.

The basic workflow in designing and developing a job includes the following steps:

1. Add the input and output stages.
2. Add transformation stages.
3. Use links to connect the stages.
4. Load table definitions into source stages and other stages, as necessary.
5. Add data source file properties.
6. Add data target file properties.
7. Edit transformation stages as needed, depending on their types.
8. Save and compile the job.
9. Run and monitor the job.
10. Review the log.

Before you begin to design a job, carefully consider the following points:

Understand the purpose of the job

To use InfoSphere DataStage, it is important to apply a structured methodology for gathering requirements.

A requirement might be as simple as loading a file into a database. A different requirement might be to remove any duplicate records from the file before loading it into the database. A more complex requirement might be to join data from three diverse databases, perform a series of data cleansing tasks on the data, reformat the data into a star schema, and then load the star schema along with three different aggregates into a set of tables.

As you can see, these requirements are very different and cause you to design jobs that are also very different. For example, for the more complex requirement, you might build a set of sequence jobs, whereas you might build just a single job for the simpler requirement. You must evaluate each requirement to determine how best to break up the work and then design logical subsets that best meet the requirement.

Understand the data structures

Before you start to design a job, consider:

- The number and type of data sources that you need to access in the job.
- The location of the data. You might choose to access the data differently depending on the type of system in which the data is stored.

- The content of the data. Think about the columns that are in your data, then determine whether you can import the table definitions or enter them manually. Keep in mind that table definitions might not be consistent among different data sources.

Understand the transformations

Determine what you expect the output data to look like after the transformations run and the data is loaded onto the target system. Decide whether you will work with some or all of the columns in the source data. Also, consider whether you need to aggregate or convert the data before moving on to the next stage.

Related information

Job design tips

Designing InfoSphere DataStage and QualityStage jobs

Data flow design

When designing the flow of your data, consider what data sources your job needs to use, what type of processing you want to perform on the data, and where you want to store the output data. This method helps you build and reuse components across jobs, minimizing the coding that is required to define even the most difficult and complex integration process.

When you start to build your job, begin by using stages and links to sketch out the data flow. You might find that jobs exist that have similarities to the job that you need to build. If that is the case, investigate whether elements of those existing jobs can be reused in your job, then plan how you can incorporate those elements.

Alternatively, you might need more than one job to accomplish your goal. You might find that you can more readily produce the desired outcome by combining existing jobs with new jobs that you create. You need to evaluate and determine which existing jobs could contribute to the goal and which jobs you need to create, then consider the most effective order, or sequence, in which to place the jobs.

Related information

Desiging for good performance

Parallel processing design

InfoSphere DataStage brings the power of parallel processing to the data extraction and transformation process. InfoSphere DataStage jobs automatically inherit the capabilities of data pipelining and data partitioning, allowing you to design an integration process without concern for data volumes or time constraints, and without any requirements for hand coding.

InfoSphere DataStage jobs use two types of parallel processing:

Data pipelining

Data pipelining is the process of extracting records from the data source system and moving them through the sequence of processing functions that are defined in the data flow that is defined by the job. Because records are flowing through the pipeline, they can be processed without writing the records to disk.

Data partitioning

Data partitioning is an approach to parallelism that involves breaking the

records into partitions, or subsets of records. Data partitioning generally provides linear increases in application performance.

When you design a job, you select the type of data partitioning algorithm that you want to use (hash, range, modulus, and so on). Then, at runtime, InfoSphere DataStage uses that selection for the number of degrees of parallelism that are specified dynamically at run time through the configuration file.

Related information

Parallelism basics

Parallel processing environments

Chapter 5. Job run processes

In InfoSphere DataStage, you run jobs in development to work through any issues before you schedule them or run them in a production environment.

When you run a job, the actual steps of extracting, transforming, and loading data take place. A job is typically run without any limits on the number of rows that are processed. Also, there typically are not any limits on the number of warnings that are displayed. However, you can set limits if you choose.

Related information

Running InfoSphere DataStage jobs

Scheduling jobs

You can schedule jobs to run one time or on a recurring basis.

You can schedule how often a job is run. For example, you can schedule it to run today, tomorrow, every day, or a specific day.

Scheduling in InfoSphere DataStage uses the functionality of the operating system, so scheduling is subject to the same rules that the operating system enforces. For example, on UNIX systems, only the root user is allowed to see other users' schedules; therefore, it is good practice to use a single ID to schedule all jobs.

Related information

Job scheduling

Scheduling a job

Monitoring jobs

You can monitor jobs in InfoSphere DataStage.

You can use either the InfoSphere DataStage and QualityStage Director or the Operations Console to access information about your jobs, job activity, and system resources. The Operations Console provides a great deal of analytical capabilities into the performance of the job run, system resources, and engine status.

You can perform other monitoring in InfoSphere DataStage. For example:

- Job monitoring provides a useful snapshot of the performance of a job at a moment of execution.
- Performance analysis provides deeper insight into runtime job behavior. This analysis is done by viewing charts that interpret job performance and computer resource utilization.
- You can estimate and predict the resource utilization of parallel job runs by creating models and making projections.
- You can create an audit trail of security-related events, including all security-related settings changes and user login and logout operations.

Related information

Introduction to monitoring jobs

Monitoring jobs and job runs by using the Operations Console

Resource estimation

Audit logging configuration

Resetting jobs

You can reset a job if there were problems when it ran.

Resetting is used when a job or sequence job has failed or ended unexpectedly. In these cases, the job will be left in an "aborted" state so that the developer or operator is aware that an issue occurred and needs to be addressed. The job must be reset before it can be run again. The reset action returns the job monitoring information to the state that it was in before the job run. After the job is reset, the job status is shown as "has been reset".

There is one exception to the requirement of resetting a job before you can run it again. When a sequence job that uses the checkpoint/restart feature fails, the status is shown as "aborted/restartable". That sequence job can be run again without being reset. Processing begins at the step following the last completed checkpoint. You can, however, choose to reset the sequence job. If so, the checkpoints are cleared and the next run of the sequence job begins at the first step of the workflow.

Related information

Resetting a job

Managing job performance

You can view the status of all the jobs in a project.

While a job is running, you can see the details of how the job is performing. After a job completes, you can see details about how it performed. You can view information for each active stage in the job and for each of the links into and out of a job.

You can view the following kinds of information:

- Whether a stage is compiled, running, finished, finished with warnings, or failed
- The number of rows that were processed by the stage or that were passed through a link
- The time that the stage began processing data
- The length of time that the stage has been actively processing data
- The number of rows that are being processed per second
- The percentage of CPU that is being consumed by a process

These details can be useful in understanding the ability of a job to process data efficiently. For example, by viewing the number of rows that were processed and the status of the stage, you can determine whether the job is running in the intended manner.

Related information

Job status details

Job status view

Troubleshooting jobs

Log files are generated when you run a job. You can use IBM InfoSphere DataStage and QualityStage Designer to access log files. You can use log files to troubleshoot problems that occur in jobs.

Logs vary between different jobs, depending on the types of stages that are used in a job. A typical job might contain messages about environment variables, NLS information, start and finish information, database information, and so on.

The log file might contain messages from many runs, validates, and resets, and therefore it might become large. Jobs that have multiple instances increase the log file even more, because each instance shares the same log file. You can purge the log occasionally to reduce the storage space that is needed. You can choose to auto-purge the log based on a certain number of runs or a predetermined number of days.

Related information

The job log

Job log view

Chapter 6. Architecture overview

InfoSphere DataStage is part of a larger suite of products called InfoSphere Information Server, which is a comprehensive, unified platform for enterprise information architectures.

InfoSphere Information Server is capable of scaling to meet any information volume requirement so that companies can deliver business results faster and with higher quality results. InfoSphere Information Server provides a single unified platform that enables companies to understand, cleanse, transform, and deliver trustworthy and context-rich information.

You install InfoSphere Information Server product modules, including InfoSphere DataStage, in logical tiers. A tier is a logical group of components within InfoSphere Information Server and the computers on which those components are installed. The tiers provide services, job execution, and metadata and other data storage for your product modules.

Each tier includes a subgroup of the components that make up the InfoSphere Information Server product modules. InfoSphere Information Server product modules also share many common components, such as administrative and security services; design, development, and deployment tools; metadata assets; and monitoring capabilities.

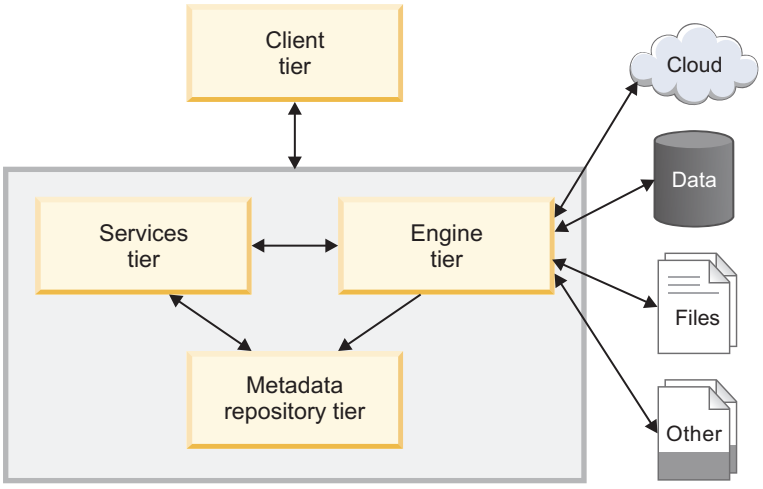

Figure 3. Tiered architecture of InfoSphere Information Server

The following table describes each tier.

Table 1. Tiers

Tier	Description
Client tier	The client tier includes the client programs and consoles that are used for development and administration, and the computers where they are installed.

Table 1. Tiers (continued)

Tier	Description
Engine tier	The engine tier includes the logical group of components (the InfoSphere Information Server engine components, service agents, and so on) and the computer where those components are installed. The engine runs jobs and other tasks for product modules.
Services tier	The services tier includes the application server, common services, and product services for the suite and product modules, and the computer where those components are installed. The services tier provides common services (such as metadata and logging) and services that are specific to certain product modules. On the services tier, WebSphere® Application Server hosts the services. The services tier also hosts InfoSphere Information Server applications that are web-based.
Metadata repository tier	The metadata repository tier includes the metadata repository, the InfoSphere Information Analyzer analysis database (if installed), and the computer where these components are installed. The metadata repository contains the shared metadata, data, and configuration information for InfoSphere Information Server product modules. The analysis database stores extended analysis data for InfoSphere Information Analyzer.

Related information

Introduction to IBM Information Server

Tiers and components

Tier relationships

Shared services

Chapter 7. Additional resources

Visit these deliverables to learn more about InfoSphere DataStage.

- *InfoSphere DataStage Data Flow and Job Design*: Describes the implementation of InfoSphere DataStage data flow and job design.
- InfoSphere DataStage Glossary: Terms and definitions for InfoSphere DataStage.
- InfoSphere DataStage information roadmap: Links to additional information resources that are available for InfoSphere DataStage.

Product accessibility

You can get information about the accessibility status of IBM products.

The IBM InfoSphere Information Server product modules and user interfaces are not fully accessible. The installation program installs the following product modules and components:

- IBM InfoSphere Business Glossary
- IBM InfoSphere Business Glossary Anywhere
- IBM InfoSphere DataStage
- IBM InfoSphere FastTrack
- IBM InfoSphere Information Analyzer
- IBM InfoSphere Information Services Director
- IBM InfoSphere Metadata Workbench
- IBM InfoSphere QualityStage

For information about the accessibility status of IBM products, see the IBM product accessibility information at http://www.ibm.com/able/product_accessibility/index.html.

Accessible documentation

Accessible documentation for InfoSphere Information Server products is provided in an information center. The information center presents the documentation in XHTML 1.0 format, which is viewable in most Web browsers. XHTML allows you to set display preferences in your browser. It also allows you to use screen readers and other assistive technologies to access the documentation.

For information about the accessibility features of the information center, see Accessibility and keyboard shortcuts in the information center.

The documentation that is in the information center is also provided in PDF files, which are not fully accessible.

IBM and accessibility

See the IBM Human Ability and Accessibility Center for more information about the commitment that IBM has to accessibility.

Accessing product documentation

Documentation is provided in a variety of locations and formats, including in help that is opened directly from the product client interfaces, in a suite-wide information center, and in PDF file books.

The information center is installed as a common service with IBM InfoSphere Information Server. The information center contains help for most of the product interfaces, as well as complete documentation for all the product modules in the suite. You can open the information center from the installed product or from a Web browser.

Accessing the information center

You can use the following methods to open the installed information center.

- Click the **Help** link in the upper right of the client interface.

  Note: From IBM InfoSphere FastTrack and IBM InfoSphere Information Server Manager, the main Help item opens a local help system. Choose **Help > Open Info Center** to open the full suite information center.
- Press the F1 key. The F1 key typically opens the topic that describes the current context of the client interface.

  Note: The F1 key does not work in Web clients.
- Use a Web browser to access the installed information center even when you are not logged in to the product. Enter the following address in a Web browser: http://host_name:port_number/infocenter/topic/com.ibm.swg.im.iis.productization.iisinfsv.home.doc/ic-homepage.html. The host_name is the name of the services tier computer where the information center is installed, and port_number is the port number for InfoSphere Information Server. The default port number is 9080. For example, on a Microsoft® Windows® Server computer named iisdocs2, the Web address is in the following format: http://iisdocs2:9080/infocenter/topic/com.ibm.swg.im.iis.productization.iisinfsv.nav.doc/dochome/iisinfsrv_home.html.

A subset of the information center is also available on the IBM Web site and periodically refreshed at http://publib.boulder.ibm.com/infocenter/iisinfsv/v8r7/index.jsp.

Obtaining PDF and hardcopy documentation

- A subset of the PDF file books are available through the InfoSphere Information Server software installer and the distribution media. The other PDF file books are available online and can be accessed from this support document: https://www.ibm.com/support/docview.wss?uid=swg27008803&wv=1.
- You can also order IBM publications in hardcopy format online or through your local IBM representative. To order publications online, go to the IBM Publications Center at http://www.ibm.com/e-business/linkweb/publications/servlet/pbi.wss.

Providing feedback about the documentation

You can send your comments about documentation in the following ways:

- Online reader comment form: www.ibm.com/software/data/rcf/
- E-mail: comments@us.ibm.com

Notices and trademarks

This information was developed for products and services offered in the U.S.A.

Notices

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785 U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web

sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003 U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to

IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_. All rights reserved.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at www.ibm.com/legal/copytrade.shtml.

The following terms are trademarks or registered trademarks of other companies:

Adobe is a registered trademark of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office

UNIX is a registered trademark of The Open Group in the United States and other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

The United States Postal Service owns the following trademarks: CASS, CASS Certified, DPV, LACS^{Link}, ZIP, ZIP + 4, ZIP Code, Post Office, Postal Service, USPS and United States Postal Service. IBM Corporation is a non-exclusive DPV and LACS^{Link} licensee of the United States Postal Service.

Other company, product or service names may be trademarks or service marks of others.

Contacting IBM

You can contact IBM for customer support, software services, product information, and general information. You also can provide feedback to IBM about products and documentation.

The following table lists resources for customer support, software services, training, and product and solutions information.

Table 2. IBM resources

Resource	Description and location
IBM Support Portal	You can customize support information by choosing the products and the topics that interest you at www.ibm.com/support/entry/portal/Software/Information_Management/InfoSphere_Information_Server
Software services	You can find information about software, IT, and business consulting services, on the solutions site at www.ibm.com/businesssolutions/
My IBM	You can manage links to IBM Web sites and information that meet your specific technical support needs by creating an account on the My IBM site at www.ibm.com/account/
Training and certification	You can learn about technical training and education services designed for individuals, companies, and public organizations to acquire, maintain, and optimize their IT skills at http://www.ibm.com/software/sw-training/
IBM representatives	You can contact an IBM representative to learn about solutions at www.ibm.com/connect/ibm/us/en/

Providing feedback

The following table describes how to provide feedback to IBM about products and product documentation.

Table 3. Providing feedback to IBM

Type of feedback	Action
Product feedback	You can provide general product feedback through the Consumability Survey at www.ibm.com/software/data/info/consumability-survey

Table 3. Providing feedback to IBM (continued)

Type of feedback	Action
Documentation feedback	To comment on the information center, click the Feedback link on the top right side of any topic in the information center. You can also send comments about PDF file books, the information center, or any other documentation in the following ways: • Online reader comment form: www.ibm.com/software/data/rcf/ • E-mail: comments@us.ibm.com

Index

Printed in USA

GC19-3607-00