IBM InfoSphere DataStage
Version 9 Release 1

Introduction to IBM InfoSphere DataStage

IBM
IBM InfoSphere DataStage
Version 9 Release 1

Introduction to IBM InfoSphere DataStage

IBM
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to InfoSphere DataStage</td>
<td>1</td>
</tr>
<tr>
<td>Job lifecycle</td>
<td>1</td>
</tr>
<tr>
<td>Develop</td>
<td>3</td>
</tr>
<tr>
<td>Test</td>
<td>4</td>
</tr>
<tr>
<td>Deploy</td>
<td>6</td>
</tr>
<tr>
<td>Operate</td>
<td>8</td>
</tr>
<tr>
<td>Job designs</td>
<td>9</td>
</tr>
<tr>
<td>Extract and load data</td>
<td>9</td>
</tr>
<tr>
<td>Transform data</td>
<td>12</td>
</tr>
<tr>
<td>Enrich data</td>
<td>16</td>
</tr>
<tr>
<td>Cleanse data</td>
<td>17</td>
</tr>
<tr>
<td>Real-time processing</td>
<td>20</td>
</tr>
<tr>
<td>Big data processing</td>
<td>21</td>
</tr>
<tr>
<td>Combine jobs in a sequence job</td>
<td>23</td>
</tr>
<tr>
<td>InfoSphere DataStage integration in InfoSphere</td>
<td>23</td>
</tr>
<tr>
<td>product documentation</td>
<td>27</td>
</tr>
<tr>
<td>Accessing and providing feedback on the product</td>
<td>29</td>
</tr>
<tr>
<td>documentation</td>
<td>29</td>
</tr>
<tr>
<td>Notices and trademarks</td>
<td>31</td>
</tr>
<tr>
<td>Contacting IBM</td>
<td>37</td>
</tr>
<tr>
<td>Index</td>
<td>39</td>
</tr>
</tbody>
</table>
Introduction to InfoSphere DataStage

IBM® InfoSphere® DataStage® is a data integration tool for designing, developing, and running jobs that move and transform data.

InfoSphere DataStage is the data integration component of IBM InfoSphere Information Server. It provides a graphical framework for developing the jobs that move data from source systems to target systems. The transformed data can be delivered to data warehouses, data marts, and operational data stores, real-time web services and messaging systems, and other enterprise applications. InfoSphere DataStage supports extract, transform, and load (ETL) and extract, load, and transform (ELT) patterns. InfoSphere DataStage uses parallel processing and enterprise connectivity to provide a truly scalable platform.

With InfoSphere DataStage, your company can accomplish these goals:

• Design data flows that extract information from multiple source systems, transform the data as required, and deliver the data to target databases or applications.
• Connect directly to enterprise applications as sources or targets to ensure that the data is relevant, complete, and accurate.
• Reduce development time and improve the consistency of design and deployment by using prebuilt functions.
• Minimize the project delivery cycle by working with a common set of tools across InfoSphere Information Server.

Job lifecycle

You can develop, test, deploy, and run jobs that move data from source systems to target systems by using IBM InfoSphere DataStage.

InfoSphere DataStage works with the other IBM InfoSphere Information Server components to provide a unified information integration solution. Each component provides key capabilities through the phases of an information integration project:

• Plan
• Discover and analyze
• Design
• Develop
• Deploy
• Deliver

InfoSphere DataStage provides the capability to develop, test, deploy, and run the jobs that deliver the data. The users who complete the tasks in the job lifecycle include the developer, release coordinator, and operator. Your company might have different job roles or might call these roles by other names. One user can cover all the roles, or multiple users can share the roles.

The following figure shows the lifecycle of an InfoSphere DataStage job from job development through data delivery.
Develop

The developer creates the jobs that move the data from the source to the target by using the IBM InfoSphere DataStage and QualityStage® Designer client. The developer lays out the data flow by arranging the stages that are required for the data sources, the transformations, and the targets. The developer defines the links that connect the stages and configures the stages.

A stage describes a data source, a processing step, or a target system. The stage also defines the processing logic that moves the data from the input links to the output links. If the installation of InfoSphere Information Server includes IBM InfoSphere QualityStage, the stages that analyze and cleanse the data are also provided. If the installation includes IBM InfoSphere Information Analyzer, the stage that contains data rules is provided. Data rules validate data and check data quality.

When the developer saves the job, the details are written to the metadata repository. Other users who are authorized in the project can then view and edit the job designs.

A project is a container that organizes and provides security for the jobs. The administrator creates a project for the developers to work in by using IBM InfoSphere DataStage and QualityStage Administrator. The administrator defines security at the project level so that only users who are authorized for the project can access the jobs.

Test

The developer tests the job until it is ready for production by using the Designer client. The Designer client includes visual debugging tools and specific testing stages that support the process that developers follow as they design data integration jobs.

Deploy

The release coordinator verifies and approves the job and moves the job into the test environment or the production environment by using IBM InfoSphere Information Server Manager. Organizations that deploy artifacts through their enterprise source code control system can check artifacts directly in and out, and manage job promotion and versioning.

Operate

The operator runs and manages the jobs in production by using the IBM InfoSphere DataStage and QualityStage Operations Console. The operator monitors the project to evaluate how it is working with other parts of the...
system. To support analysis, the console provides charts about job performance and system performance, including CPU usage, memory usage, and disk usage.

Related concepts:
- Information integration phases
- Designing DataStage and QualityStage jobs

Develop

Develop the jobs that move the data from the source to the target.

You can create and develop jobs by using the Designer client. To develop a job, you define the job flow and configure the job as shown in the following figure. You can then save and compile the job, and review any compile errors.

![Figure 2. Developing a job](image)

Define the job flow

You can lay out the data flow for a job by arranging the stages on the Designer canvas and connecting the stages with links. The stages represent the data sources, the transformations, and the targets. The stages contain the processing logic for the job. The links carry metadata about the data that is moving between stages.
Configure the job

You can configure the job by adding table definitions, configuring the stages, and adding job parameters.

- You add table definitions to the links. Table definitions contain information about the structure of your data and about the location of the tables or files that contain your data. You can import a table definition from the source or target database, or you can create the table definition by using the Designer client.
- You configure the stages by specifying the properties that define how the stage processes the data.
- You can use parameters in your jobs to specify values at run time, rather than hardcoding the values.

Review compilation errors

You can review the stages that have errors when you compile a job. If a job has a compilation error, you can select an option to highlight the stage that contains the first error in the design. You can review information about the error, and then edit the stage to change any incorrect settings and recompile the job.

In this sample job, a column in the Transformer stage, Trim_Prep, is missing a required property, so the Transformer stage is highlighted.

![Diagram of job stages](image)

*Figure 3. Reviewing compile errors*

Related concepts:
- Designing DataStage and QualityStage jobs
- Sketching your job designs
- Defining your data
- Configuring your designs
- Alphabetical list of stages
- Job design tips

Test

Debug the jobs by using the debugging stages and the interactive debugging tool.

You can use the debugging stages early in the development cycle to generate test data and to sample data during processing. After you compile a job, you can run the job in debug mode and examine the column data at breakpoints to understand the job flow.
Generate test data

You can use the debugging stages to generate test data. For example, you can generate data based on the table definition for your source system, and then you can build and test downstream logic based on the generated data. You can use the data to test your job logic early in the job lifecycle before your source system data is available.

This sample job generates customer data by using the Row Generator stage. The job enriches the generated data with real customer data from an Oracle database by using the Lookup stage.

Sample the data

You can use the debugging stages to sample output data while you are developing a job design. For example, you can sample data in a data set that is the output of a
job by using the Peek stage. The Peek stage writes data to the job log. In this job, the original data is copied to another data set, and the sampled columns are written to the log file for you to review.

![Data Flow Diagram](image)

**Figure 6. Sampling data**

### Examine the column data at breakpoints

You can debug a job by setting breakpoints on the links where you want the job to stop when you run the job in debug mode. While the job stops at a breakpoint, you can examine the column data that is being processed. You can also change the breakpoints for other data conditions and examine stage properties and job parameter values. In this sample job, two breakpoints are set to stop the job after every five rows. The job is stopped at the breakpoints, and the Debug Window displays the column data for each breakpoint.

![Debug Window](image)

**Figure 7. Setting breakpoints**

**Related concepts:**
- Debugging parallel jobs

**Related tasks:**
- Debugging parallel jobs with the debugging stages
- Running parallel jobs in debug mode

### Deploy

Deploy jobs between the environments that support your software development lifecycle: development, testing, production, and other environments.
In a typical scenario, you promote your jobs from one environment to another environment as the jobs pass testing or other readiness requirements for your organization. You can move a single job or multiple jobs and their related assets between these environments by using IBM InfoSphere Information Server Manager.

You can complete other tasks that support the deployment process by using InfoSphere Information Server Manager. For example, you can search for assets that changed since a specified date or identify dependent objects. You can add assets from outside InfoSphere DataStage that should also be deployed with your jobs. You can manage your InfoSphere DataStage jobs and associated assets in your source control system.

The basic workflow for deploying a job by using the InfoSphere Information Server Manager includes the following tasks:

1. Define a package to specify the assets to deploy.
2. Build the package to produce a package file.
3. Copy the package file to the target system if necessary.
4. Deploy the package on the target system.

**Figure 8. Deploying a job**

**Define the package**

When you define the deployment package, you can select assets from the InfoSphere Information Server engines and projects that share a metadata repository. When the assets on the source system change, you can rebuild, recopy, and redeploy the package. You do not need to redefine the package.

**Build the package**

When you build the package, copies of the assets that are defined for the package are added to the package file. Items are pulled into the build in their state at the time of the build. If you rebuild a package, you can replace the entire package or update only the items in the package that changed since the last build. The package is then ready for deployment.
Deploy the package

You can deploy the package to different projects on the same InfoSphere DataStage and QualityStage system, such as when the development environment and test environment are on the same server. Or you can deploy the package to different InfoSphere DataStage and QualityStage systems.

Manage jobs in source control

You can check artifacts directly in and out of source code control systems, including IBM Rational Team Concert™, IBM ClearCase®, Subversion, Microsoft Team Foundation Server, and others. You can use your enterprise source code control system to manage job promotion and versioning if that is your company standard.

Related concepts:

Deploying jobs and accessing version control

Operate

Monitor and manage jobs in production.

You can access information about jobs, job activity, system resources, and workload management queues for InfoSphere Information Server engines by using the IBM InfoSphere DataStage and QualityStage Operations Console. You can run jobs on demand and manage system resources.

Find answers to common questions

The console dashboard provides quick answers when you view and analyze the runtime environment. In the dashboard charts, you can review the job run workload, recently completed job runs, status of engine services, and system resource usage. The dashboard also provides the count of successful and failed job runs and the status of key services that are running on the engine. The dashboard displays alerts when job runs fail and when metric thresholds are passed.

Access job details and history

From the dashboard, you can drill down to view specific job and project details. For example, you can drill down on the count of failed job runs to view a list of the failed jobs. You can then drill down on a specific job to see detailed properties, parameters, performance, and messages. From the Projects page, you can review high-level summary information and detailed information for current and historical activity. You can also retrieve and view the complete log directly from the server.

Optimize system resources

By using the workload management capabilities of the Operations Console, you can proactively manage system resources where multiple teams share a common hardware infrastructure. You can set system policies, monitor queued jobs, and
manage the workload management queues.

**Prioritize mission critical tasks**

You can throttle job activity where system resources exceed your specified thresholds and assign the priority of any submitted job. If you are a privileged user, you can use manual overrides to promote specific jobs to the top of the queue.

**Related concepts:**

- Monitoring jobs and job runs by using the Operations Console
- Administering workload management

**Job designs**

Design jobs that extract data from various sources, transform it, and deliver it to target systems in the required formats.

The following examples show some of the ways that you can design jobs by using IBM InfoSphere DataStage. The key elements of a job design are the stages and the links between the stages. A stage describes a data source, a processing step, or a target system. The stage also defines the processing logic that moves the data from the input links to the output links. Because the stages are flexible and configurable, you can link multiple stages together to satisfy complex business requirements.

These job designs include only a sample of the available stages. The stages that you use depend on your requirements and environment.

**Related information:**

- IBM InfoSphere DataStage Data Flow and Job Design (An IBM Redbooks publication)
- InfoSphere DataStage Parallel Framework Standard Practices (An IBM Redbooks publication)
- Job design tips

**Extract and load data**

You can develop jobs to move data between various sources and in various formats. These examples show how you can extract data, identify changes between files, and load data to a database or a data warehouse.

- “Extract data from a database to a file or data set”
- “Load data with high-speed native connectivity” on page 10
- “Load data in real time to a data warehouse” on page 10
- “Extract data from SAP” on page 11
- “Fetch data from a web service to enrich data” on page 11
- “Identify database changes” on page 12

**Extract data from a database to a file or data set**

You can extract bulk data from a database. For example, the Oracle Connector stage can use customized SQL in the Oracle database to extract the customer address, phone number, and account balance in parallel. You can develop a similar job to extract data from any relational database source.
In the first example, the extracted data is written to a sequential file to share with other departments in the company. In the second example, the extracted data is written to a data set for use downstream in another job. A data set stores data in a persistent form that uses parallel storage; that is, data partitioned storage. Data sets optimize I/O performance by preserving the degree of partitioning.

![Figure 10. Extract data from a database to a sequential file](image)

Load data with high-speed native connectivity

You can read data from a file, sort the data numerically to optimize storage, and bulk load the data to a database in native format. For example, you can load third-party data that your organization receives to a Netezza® table for analysis by using the Netezza Connector stage. Like each of the database connector stages, the Netezza Connector stage in this example uses the native API of the database to maximize performance.

![Figure 12. Load to a Netezza table](image)

Load data in real time to a data warehouse

You can develop a job to move data in real time to a data warehouse. For example, you can capture real-time information about sales, inventory, and shipments from your company’s transactional systems on disparate platforms. You can then standardize and integrate the information, and deliver it to a data warehouse that supports a single reporting framework.

In this example, the job loads data from orders in real time by using the WebSphere® MQ Connector stage. The job parses the data by using the XML stage and transforms the data by using the Transformer stage. The job then updates the reference table and loads data to a DB2® database by using the Slowly Changing Dimension stage. The Slowly Changing Dimension stage is a processing stage that
works within the context of a star schema database.

Extract data from SAP

You can extract data from SAP and load it to SAP Business Warehouse (SAP BW), to an existing enterprise data warehouse, or to a file or other target. For example, you can extract data by using the Advanced Business Application Programming (ABAP) Extract stage. The ABAP Extract stage generates an ABAP program and uploads it to the SAP system. The ABAP program runs an SQL query and extracts the data.

In this sample job, the extracted data is formatted for the target stage by a Transformer stage. The formatted data is loaded to an SAP business warehouse by using the SAP BW stage. The SAP stages are provided by and InfoSphere Information Server Pack for SAP BW.

Fetch data from a web service to enrich data

You can enrich data by using a web service to retrieve data from an application. For example, you can retrieve data from an HR application like Workday by using the Web Services Transformer stage. This sample job retrieves the data and then parses it by using the XML stage to generate rows and columns.

Figure 13. Load data to a data warehouse

Figure 14. Extract data from SAP

Figure 15. Fetch data from a web service
Identify database changes

You can identify database changes and apply the changes to a target database.

This sample job reads data that is captured by InfoSphere Change Data Capture (CDC) and applies the change data to a target database by using the CDC Transaction stage.

InfoSphere CDC is part of IBM InfoSphere Data Replication, which is a replication technology that captures database changes as they happen and delivers them to InfoSphere DataStage, target databases, or message queues. InfoSphere Data Replication collects data changes by using low-impact, low-latency database log analysis.

This solution provides several advantages. Overhead on the database system is lower because processing the data is log-based. Only changed data is moved across the network. The total processing on the ETL server is lighter because the server does not need to find the changes.

Transform data

You can transform data to satisfy both simple and complex data integration tasks. These examples show how you can join data, summarize data, use a pivot table to reorder data, and identify changes from a previous file.

- "Transform data with expressions" on page 13
- "Join data from multiple flat files" on page 13
Transform data with expressions

You can create transformations to apply to your data by using the Transformer stage. These transformations can be simple or complex and can be applied to individual columns in your data. For example, you can trim data by removing leading and trailing spaces, concatenate data, perform operations on dates and times, perform mathematical operations, and apply conditional logic. Transformations are specified by using a set of functions. You can define the transformations by using the expression editor.

This sample job includes several transformations: the customer number is trimmed; the country code is replaced with United States if it is US; the vendor and type codes are concatenated to create a vendor type code; and the current date is inserted.

Join data from multiple flat files

You can read data from multiple flat files and write the data to a single target file. You can combine the records from the files by using the Join stage and use techniques like key-based partitioning and in-memory sorting to accelerate processing. For example, you can combine customer records with order details and write the customer order to the target file.
Join heterogeneous sources

You can collect, standardize, and consolidate data from a wide array of data sources and data structures, and create a single, accurate, and trusted source of information. In this example, the job extracts data from two different types of databases and loads the results into another type of database in the appropriate format. The job uses a Join stage to combine the data and Transformer stages to trim and format the data.

Summarize data by common characteristics

You can summarize data by using the Aggregator stage. Records can be grouped by one or more characteristics. For example, you can group sales data both by day of the week and by month, and compute totals. These groupings might show that the busiest day of the week varies by season.
Pivot data in a table

You can map a set of columns in an input row to a single column in multiple output rows by using a horizontal pivot by using the Pivot Enterprise stage. You can also do vertical pivots, which combine related data from multiple rows into a single row with repeating field types.

In this example, the schema changes as shown in the two following tables. The data can be pivoted in either direction, horizontal from Table 1 to Table 2, or vertical from Table 2 to Table 1.

**Table 1. Original**

<table>
<thead>
<tr>
<th>OrderID</th>
<th>OrderDate</th>
<th>CustID</th>
<th>Name</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
</tr>
</thead>
<tbody>
<tr>
<td>123456</td>
<td>6-6-2013</td>
<td>ABCX</td>
<td>John Doe</td>
<td>notebook</td>
<td>monitor</td>
<td>keyboard</td>
</tr>
</tbody>
</table>

**Table 2. After horizontal pivot**

<table>
<thead>
<tr>
<th>PivotID</th>
<th>OrderID</th>
<th>OrderDate</th>
<th>CustID</th>
<th>Name</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>123456</td>
<td>6-6-2013</td>
<td>ABCX</td>
<td>John Doe</td>
<td>notebook</td>
</tr>
<tr>
<td>2</td>
<td>123456</td>
<td>6-6-2013</td>
<td>ABCX</td>
<td>John Doe</td>
<td>monitor</td>
</tr>
<tr>
<td>3</td>
<td>123456</td>
<td>6-6-2013</td>
<td>ABCX</td>
<td>John Doe</td>
<td>keyboard</td>
</tr>
</tbody>
</table>

Find deltas from yesterday's file

You can identify changes from a previous file. For example, you can compare yesterday's file with today's file, and make a record of the changes by using the Change Capture stage. The Change Capture stage takes two input data sets, denoted before and after, and outputs a single data set. The output records represent the changes that were made to the before data set to obtain the after data set. In this sample job, the changes between two sequential files are loaded to a Teradata database for analysis.
Related concepts:
- Transformer stage
- Join stage
- Aggregator stage
- Pivot Enterprise stage
- Change Capture stage

Enrich data
These examples show how you can enrich customer data with reference data and how you can apply business rules to determine an appropriate course of action.

- “Look up reference data for data enrichment”
- “Apply business rules” on page 17

Look up reference data for data enrichment
You can enrich data by adding data from a reference source. For example, you can add the country code to a list of customer addresses by using the Lookup stage. This sample job extracts the customer addresses from a DB2 database, references the country codes in an Oracle database, and combines the data by using the Lookup stage.
Apply business rules

In addition to the transformation rules and data rules that are included in InfoSphere Information Server, you can invoke complex business rules by using the ILOG® JRules connector stage. This stage accesses rule sets that are provided by IBM Operational Decision Manager. Organizations that develop a business rules framework can use those shared business rules in their data integration job flows.

For example, you can determine the appropriate course of action to take for each of your customers based on the applied business rules.

Figure 25. Apply business rules

**Cleanse data**

You can develop jobs to standardize, verify, and validate data. These examples show how you can standardize customer data, verify international addresses, identify duplicate records, and apply data rules by integrating stages from IBM InfoSphere QualityStage and IBM InfoSphere Information Analyzer.

- “Standardize data”
- “Verify international addresses” on page 18
- “Match records to identify duplicates” on page 18
- “Validate data with data rules” on page 19
the organization and any contact names are distinctly identified and that the
organization name has one representation. For example, the variations of the name
of the fictional Sample Outdoor Company, which might include SOC, S. O. C, and
Sample Outdoor Co, are all represented by Sample Outdoor Company. You can
also standardize other types of data, including email, product descriptions, or
addresses, so for example, 100 W. Main St and 100 West Main Street both become
100 W Main St.

The Standardize stage uses rule sets that are designed to standardize your data to
meet industry standards, to improve data matching, or to facilitate downstream
analytics. The Standardize stage is provided by IBM InfoSphere QualityStage. You
can develop the rule sets by using IBM InfoSphere QualityStage Standardization
Rules Designer, which is a web-based, data-driven classification tool.

Verify international addresses

You can validate and standardize address data to ensure that each address
conforms to postal standards and to determine whether the address is deliverable
by using the Address Verification stage, which is provided by IBM InfoSphere
QualityStage Address Verification Interface. For example, you can extract address
data from a customer relationship management system, verify the international
addresses, and load them to a Netezza data warehouse. The Address Verification
stage provides processes that can organize, verify, and transform address data for
postal standards and languages in multiple countries and regions.

Match records to identify duplicates

You can identify duplicate records for entities such as individuals, companies,
suppliers, products, or events by matching data. Matching is a probabilistic record
linkage system, provided by IBM InfoSphere QualityStage, that identifies records
that are likely to represent the same entity. The matching process improves the
integrity of your data. For example, you can identify duplicate customer records in
your customer relationship management system.

For this sample job, the match frequency data was generated in a previous job by
using a Match Frequency stage. The customer data is extracted from a DB2
database. The One Source Match stage identifies the matches, duplicates, and
non-matches, combines the data by using a Funnel stage, and loads the data to a
Netezza database. The clerical records, which require further review to determine whether they match the master records, are loaded to a DB2 database.

You can validate your data and ensure that the quality of your data conforms to business expectations for data cleanliness by using the Data Rules stage, which is provided by IBM InfoSphere Information Analyzer. For example, you check that a supplier has a supplier ID in the correct format, a supplier name, and a tax ID that is nine numeric characters. The data rules that are used in the Data Rules stage are typically created by a data analyst by using InfoSphere Information Analyzer.
Real-time processing

You can develop jobs for real-time data integration. These examples show how you can respond to a web service request and verify data between message queues.

- "Respond to a web service request"
- "Verify data between message queues"

Respond to a web service request

You can respond to a web service request in real time. For example, you can return an account number after a customer submits the required account information by using the stages that link to IBM InfoSphere Information Services Director. In this sample job, account data arrives in real time through an ISD Input stage. The data is enriched with the account number by a Lookup stage. The enriched data is returned in real time through an ISD Output stage. By using InfoSphere Information Services Director, you can deploy jobs that function as information services for service-oriented applications.

Verify data between message queues

You can develop a job to receive data from a messaging queue, verify the data, and load it to another message queue. In this example, the job uses the IBM WebSphere MQ Connector stage to connect to the message queues. The job standardizes the addresses and then validates the addresses by using a Data Rules stage. Invalid
addresses are sent to a table for rejected records, and valid addresses are loaded to the message queue.

**Figure 31. Verify data between message queues**

**Related concepts:**
- [Lookup stage](#)
- IBM WebSphere MQ connector
- [Making data consistent through standardization](#)
- [Data Rules stage](#)
- Designing InfoSphere DataStage and QualityStage jobs as services
- InfoSphere Information Services Director

**Big data processing**

You can develop jobs that exchange data with big data sources. These examples show how you can access files on the Hadoop Distributed File System (HDFS) and augment data with Hadoop-based analytics.

- "Access data on HDFS"
- "Augment data with Hadoop-based analytics" on page 22

**Access data on HDFS**

You can access files on HDFS. This sample job accesses orders from HDFS files by using the Big Data File stage. The job uses a Transformer stage to select a subset of the orders, combines the orders with order details, and writes the ordered items to subsequent HDFS files. You can deploy this job in the InfoSphere DataStage engine. You can also use IBM InfoSphere DataStage Balanced Optimization to process this logic within the Hadoop cluster. The job logic is then represented as MapReduce scripting.
Augment data with Hadoop-based analytics

You can augment data in a data warehouse with Hadoop-based analytical results. This sample job moves the analytical data from a Hive data warehouse system to a Netezza data warehouse.

The Hive stage runs on top of the Java™ Integration stage and provides a Hive connector for InfoSphere DataStage. The Hive stage is part of the Hive sample code for the Java Integration stage that is available from the InfoSphere Information Server and InfoSphere Discovery Exchange on IBM developerWorks®. Other connector accelerators that are available on developerWorks extend the Java Integration stage with customized logic for these data sources: HBase, Java Message Service (JMS), MongoDB, and Cassandra.

Figure 32. Access data on HDFS

Figure 33. Augmenting data with Hadoop-based analytics
Combine jobs in a sequence job

You can combine jobs in a sequence job to develop larger processing patterns.

You can develop sequence jobs to run multiple jobs in sequence and to integrate programming controls such as branching and looping into the job workflow. You can specify the control information, such as the different courses of action to take depending on whether a job in the sequence succeeds or fails. When jobs are combined in a sequence job, they can be easier to troubleshoot. Sequence jobs also provide checkpoint restart capabilities so that on restart, processing continues after the last successful step in the workflow.

This sample job runs three jobs in sequence by using Job Activity stages. The first job extracts the data. The second job transforms the data. The third job loads the data. The last Job Activity stage also sends warning messages to be distributed by using a Notification Activity. Exceptions from the job activities are processed by using an Exception Handler stage, which sends errors to be processed by using a Routine Activity stage.

Figure 34. Combine jobs in a sequence job

Related concepts:
- Building sequence jobs
- Sequence job activities

InfoSphere DataStage integration in InfoSphere Information Server

You can use the capabilities of other IBM InfoSphere Information Server components as you develop and deploy IBM InfoSphere DataStage jobs.
InfoSphere DataStage is integrated with the other InfoSphere Information Server components across the domains of data integration, data quality, and data governance. Here are some examples of how you can use the other components.

**Collaborate across projects by using the metadata repository**

The metadata repository serves as a collaboration point for users of the components in InfoSphere Information Server. You can investigate the metadata in the repository from across projects to support impact analysis and data lineage so that your organization understands how information is used.

For example, a data analyst in your company can profile data and discover that some of the data in a specific resource is bad. As a developer, you can review the data profile and develop a job to transform the data and correct the problem. A data steward in your company can run a report to review the data lineage and evaluate how the data is being used. Anyone in your company who is authorized to work on the project can review who profiled the data and who developed the job.

When you are developing a job in InfoSphere DataStage, you can use metadata that has been shared from multiple projects. You can import metadata from tables, views, and stored procedures to use in your job designs. Your jobs and metadata are automatically included in the metadata repository.

**Determine development requirements for a project by using IBM InfoSphere Blueprint Director**

An enterprise architect can create an information landscape or blueprint for a project by using InfoSphere Blueprint Director. As a developer on the project, you can review the blueprint, including the activities and tasks that are contained within each phase, to identify development requirements. When the blueprint is linked to actual metadata, you can identify potential issues in terminology, gaps in the information process, or areas for improving reuse and consistency.

**Collaborate with business analysts to jump-start job designs by using IBM InfoSphere FastTrack**

A business analyst can create specifications for source-to-target mappings by using InfoSphere FastTrack. The mappings in the specifications can contain data value transformations that define how to build applications. The business analyst can then generate draft jobs from the specifications. As a developer, you can use those jobs as the starting point for your InfoSphere DataStage jobs.

**Develop data rules by using IBM InfoSphere Information Analyzer**

You can use data rules from InfoSphere Information Analyzer to improve data quality in your InfoSphere DataStage jobs. For example, a data analyst can use InfoSphere Information Analyzer to profile data sources to develop data quality rules. You can then use these data rules in your InfoSphere DataStage jobs. As a developer, you can also create rules directly in the job designer, and then share those rules so that business analysts can also use them in InfoSphere Information Analyzer.
Deploy jobs as services by using IBM InfoSphere Information Services Director

You can deploy jobs that function as services for service-oriented applications by using InfoSphere Information Services Director. For example, you can develop a job that uses the InfoSphere Information Services Director input and output stages. You can then publish the data integration logic as shared services that can be invoked via REST, SOAP, or other bindings, and reused across the enterprise. InfoSphere Information Services Director processes the service requests from your InfoSphere DataStage jobs.

Related concepts:
- InfoSphere Metadata Workbench
- InfoSphere Blueprint Director
- InfoSphere FastTrack
- InfoSphere Information Analyzer
- InfoSphere Information Services Director
- IBM InfoSphere Information Server architecture and concepts
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 Blueprint Director
- IBM InfoSphere Discovery
- metadata workbench
- IBM InfoSphere Business Glossary
- IBM InfoSphere Business Glossary Anywhere
- IBM InfoSphere Information Analyzer
- IBM InfoSphere QualityStage
- IBM InfoSphere Information Services Director
- IBM InfoSphere DataStage
- IBM InfoSphere DataStage and QualityStage Designer
- IBM InfoSphere Data Click
- IBM InfoSphere FastTrack
- IBM InfoSphere Data Replication

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. Because the information center uses XHTML, you can set display preferences in your browser. This also allows you to use screen readers and other assistive technologies to access the documentation.

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 and providing feedback on the 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 InfoSphere Information Server information center. 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 menu item opens a local help system. Choose Help > Open Info Center to open the full InfoSphere Information Server 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/topics/ic_homepage_IS.html

  where 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 server1, that uses the default port, the web address is in the following format:


A subset of the information center is also available on the IBM website and periodically refreshed at http://pic.dhe.ibm.com/infocenter/iisinfsv/v9r1/index.jsp. This information center is the most up-to-date version and might include corrections, provided as comments.

Obtaining PDF and hardcopy documentation

- The 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
Providing comments on the documentation

Your feedback helps IBM to provide quality information. You can use any of the following methods to provide comments:

- To provide a comment about the information center that is hosted on the IBM website, sign in and add a comment. Comments submitted this way are viewable by the public. See more information.
- To send a comment about the information to IBM that is not viewable by anyone else, click the Feedback link on the top right side of any topic in the information center. You can do this from an information center that is installed with InfoSphere Information Server or from the information center that is available on the IBM website.
- **Send your comments by using the online readers' comment form at** [www.ibm.com/software/awdtools/rcf/](http://www.ibm.com/software/awdtools/rcf/)
- Send your comments by e-mail to comments@us.ibm.com. Include the name of the product, the version number of the product, and the name and part number of the information (if applicable). If you are commenting on specific text, include the location of the text (for example, a title, a table number, or a page number).
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.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, 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.

Privacy policy considerations

IBM Software products, including software as a service solutions, (“Software Offerings”) may use cookies or other technologies to collect product usage information, to help improve the end user experience, to tailor interactions with the end user or for other purposes. In many cases no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering’s use of cookies is set forth below.

Depending upon the configurations deployed, this Software Offering may use session or persistent cookies. If a product or component is not listed, that product or component does not use cookies.

Table 3. Use of cookies by InfoSphere Information Server products and components

<table>
<thead>
<tr>
<th>Product module</th>
<th>Component or feature</th>
<th>Type of cookie that is used</th>
<th>Collect this data</th>
<th>Purpose of data</th>
<th>Disabling the cookies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any (part of InfoSphere Information Server installation)</td>
<td>InfoSphere Information Server web console</td>
<td>• Session • Persistent</td>
<td>User name</td>
<td>• Session management • Authentication</td>
<td>Cannot be disabled</td>
</tr>
<tr>
<td>Any (part of InfoSphere Information Server installation)</td>
<td>InfoSphere Metadata Asset Manager</td>
<td>• Session • Persistent</td>
<td>No personally identifiable information</td>
<td>• Session management • Authentication • Enhanced user usability • Single sign-on configuration</td>
<td>Cannot be disabled</td>
</tr>
<tr>
<td>InfoSphere DataStage</td>
<td>Big Data File stage</td>
<td>• Session • Persistent</td>
<td>• User name • Digital signature • Session ID</td>
<td>• Session management • Authentication • Single sign-on configuration</td>
<td>Cannot be disabled</td>
</tr>
</tbody>
</table>
Table 3. Use of cookies by InfoSphere Information Server products and components (continued)

<table>
<thead>
<tr>
<th>Product module</th>
<th>Component or feature</th>
<th>Type of cookie that is used</th>
<th>Collect this data</th>
<th>Purpose of data</th>
<th>Disabling the cookies</th>
</tr>
</thead>
<tbody>
<tr>
<td>InfoSphere DataStage</td>
<td>XML stage</td>
<td>Session</td>
<td>Internal identifiers</td>
<td>• Session management • Authentication</td>
<td>Cannot be disabled</td>
</tr>
<tr>
<td>InfoSphere DataStage</td>
<td>IBM InfoSphere DataStage and QualityStage Operations Console</td>
<td>Session</td>
<td>No personally identifiable information</td>
<td>• Session management • Authentication</td>
<td>Cannot be disabled</td>
</tr>
<tr>
<td>InfoSphere Data Quality Console</td>
<td>Session</td>
<td>No personally identifiable information</td>
<td></td>
<td>• Session management • Authentication • Single sign-on configuration</td>
<td>Cannot be disabled</td>
</tr>
<tr>
<td>Information Governance Catalog</td>
<td>InfoSphere Blueprint Director, InfoSphere Business Glossary, InfoSphere Metadata Workbench</td>
<td>Session</td>
<td>No personally identifiable information</td>
<td>• Session management • Authentication • Single sign-on configuration</td>
<td>Cannot be disabled</td>
</tr>
<tr>
<td>InfoSphere Information Analyzer</td>
<td>Data Rules stage in the InfoSphere DataStage and QualityStage Designer client</td>
<td>Session</td>
<td>Session ID</td>
<td>Session management</td>
<td>Cannot be disabled</td>
</tr>
</tbody>
</table>

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.


**Trademarks**

IBM, the IBM logo, and ibm.com® are trademarks or registered 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](http://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.

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

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

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, LACSLink, ZIP, ZIP + 4, ZIP Code, Post Office, Postal Service, USPS and United States Postal Service. IBM Corporation is a non-exclusive DPV and LACSLink 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 4. IBM resources

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description and location</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Support Portal</td>
<td>You can customize support information by choosing the products and the topics that interest you at <a href="http://www.ibm.com/support/entry/portal/Software/Information_Management/InfoSphere_Information_Server">www.ibm.com/support/entry/portal/Software/Information_Management/InfoSphere_Information_Server</a></td>
</tr>
<tr>
<td>Software services</td>
<td>You can find information about software, IT, and business consulting services, on the solutions site at <a href="http://www.ibm.com/businesssolutions/">www.ibm.com/businesssolutions/</a></td>
</tr>
<tr>
<td>My IBM</td>
<td>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 <a href="http://www.ibm.com/account/">www.ibm.com/account/</a></td>
</tr>
<tr>
<td>Training and certification</td>
<td>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 <a href="http://www.ibm.com/training">http://www.ibm.com/training</a></td>
</tr>
</tbody>
</table>
# Index

## A
- ABAP Extract stage 9
- Address Verification stage 17
- addresses verify 17
- Aggregator stage 12

## B
- Balanced Optimization 21
- Big data jobs 21
- Big Data File stage 21
- Blueprint Director 24
- breakpoints 4
- business rules applying 16

## C
- Cassandra 21
- CDC Transaction stage 9
- Change Capture stage 12
- cleansing data jobs 17
- customer support contacting 37

## D
- data cleansing 17
- enrichment 16
- standardizing 17
- data integration tool 1
- data rules 24
- Data Rules stage 17
- Data Set stage 9, 12, 17
- DataStage Operations Console 8
- DB2 Connector stage 9, 16, 17, 20
- debugging 4
- deployment jobs 7
  - package 7
- developer 1
- developing jobs 3

## E
- enrich data jobs 16
- Exception Handler stage 23
- extract jobs 9

## F
- FastTrack 24

## G
- generating test data 4

## H
- Hadoop 21
- Hadoop Distributed File System 21
- HBase 21
- HDFS 21
- Hive stage 21

## I
- ILOG JRules stage 16
- Information Analyzer 24
- information integration 1
- Information Services Director 24
- InfoSphere Blueprint Director 24
- InfoSphere DataStage
  - job life cycle 1
  - overview 1
- InfoSphere DataStage Balanced Optimization 21
- InfoSphere FastTrack 24
- InfoSphere Information Analyzer 24
- InfoSphere Information Server
  - integration 24
- InfoSphere Information Services Director 20, 24
- ISD Input stage 20
- ISD Output stage 20

## J
- Java Integration stage 21
- Java Message Service (JMS) 21
- Job Activity stage 23
  - jobs combining 23
  - deploying 7
  - designs 9
  - developing 3
  - examples 9
  - life cycle 1
  - testing 4
- Join stage 12, 21
- JRules stage 16

## L
- legal notices 31
- load jobs 9
- Lookup stage 16, 20

## M
- Match Frequency stage 17
- message queues verify data between 20
- metadata repository 24
- MongoDB 21

## N
- Netezza Connector stage 9, 12, 17, 21
- Notification Activity stage 23

## O
- One Source Match stage 17
- operations console 8
- operator 1
- Oracle Connector stage 9, 16, 20
  - overview 1

## P
- Peak stage 4
- Pivot Enterprise stage 12
- product accessibility
  - accessibility 27
- product documentation
  - accessing 29

## R
- real-time processing jobs 20
- reference data 16
- release coordinator 1
- Routine Activity stage 23
- Row Generator stage 4

## S
- SAP 9
- SAP Business Warehouse 9
- SAP BW stage 9
- sequence jobs 23
- Sequential File stage 9, 12, 16, 17
- Slowly Changing Dimension stage 9
- software services
  - contacting 37
- Sort stage 9
- stages
  - ABAP Extract 9
  - Address Verification 17
  - Aggregator 12
  - Big Data File 21
  - CDC Transaction 9
  - Change Capture 12
  - Data Rules 17

© Copyright IBM Corp. 2011, 2013
stages (continued)
  Data Set  9, 12, 17
  DB2 Connector  9, 16, 17, 20
debugging  4
  Exception Handler  23
  Hive  21
  ILOG JRules  16
  ISD Input  20
  ISD Output  20
  Java Integration  21
  Job Activity  23
  Join  12, 21
  JRules  16
  Lookup  16, 20
  Match Frequency  17
  Netezza Connector  9, 12, 17, 21
  Notification Activity  23
  One Source Match  17
  Oracle Connector  9, 12, 16, 20
  Peak  4
  Pivot Enterprise  12
  Routine Activity  23
  Row Generator  4
  SAP BW  9
  Sequential File  9, 12, 16, 17
  Slowly Changing Dimension  9
  Sort  9
  Standardize  17, 20
  Teradata Connector  12
  Transformer  9, 12, 17, 20, 21
  Web Services Transformer  9
  WebSphere MQ Connector  9, 20
  XML  9
standardize
data  17
Standardize stage  17, 20
support
customer  37

T
Teradata Connector stage  12
testing
  jobs  4
trademarks
  list of  31
transformation
  jobs  12
Transformer stage  9, 12, 17, 20, 21

W
web services  20
Web Services Transformer stage  9
WebSphere MQ Connector stage  9, 20

X
XML stage  9