IBM InfoSphere Information Server
Version 9 Release 1

Introduction

IBM
IBM InfoSphere Information Server
Version 9 Release 1

Introduction
Note
Before using this information and the product that it supports, read the information in "Notices and trademarks" on page 79.

US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.
Contents

Introduction to InfoSphere Information Server .......... 1
Information integration phases ...................... 3
Plan ................................................. 4
Discover and analyze .................................. 5
Design ................................................. 6
Develop ................................................. 6
Deploy ................................................. 7
Deliver ................................................. 7
Components in the suite ................................ 8
InfoSphere Blueprint Director ......................... 9
InfoSphere Business Glossary .......................... 12
InfoSphere DataStage .................................. 16
InfoSphere Data Architect ............................... 20
InfoSphere Discovery .................................. 23
InfoSphere FastTrack .................................. 26
InfoSphere Information Analyzer ....................... 29
InfoSphere Information Services Director ................. 32
InfoSphere Metadata Workbench ....................... 35
InfoSphere QualityStage ................................ 38
Additional components in the InfoSphere Information Server portfolio ......................... 41
Mainframe integration .................................. 41
Companion components for InfoSphere Information Server ........................................ 43
IBM InfoSphere Information Server architecture and concepts ..................................... 52
Tiers and components ................................ 55
Parallel processing in InfoSphere Information Server ........................................ 55
Shared services in InfoSphere Information Server ........................................ 60
Integrated metadata management in InfoSphere Information Server ........................................ 64
High availability in InfoSphere Information Server ........................................ 68
Appendix A. Product accessibility ....................... 71
Appendix B. Reading command-line syntax ................ 73
Appendix C. Contacting IBM ........................... 75
Appendix D. Accessing and providing feedback on the product documentation ..................... 77
Notices and trademarks ................................ 79
Index .................................................... 83
Introduction to InfoSphere Information Server

InfoSphere® Information Server provides a single platform for data integration. The components in the suite combine to create a unified foundation for enterprise information architectures, capable of scaling to meet any information volume requirements. You can use the suite to deliver business results faster, while maintaining data quality and integrity throughout your information landscape.

InfoSphere Information Server helps your business and IT personnel collaborate to understand the meaning, structure, and content of information across a wide variety of sources. By using InfoSphere Information Server, your business can access and use information in new ways to drive innovation, increase operational efficiency, and lower risk.

Figure 1 on page 2 shows key capabilities of InfoSphere Information Server that your business can use to implement a complete data integration strategy. The core of these capabilities is a common metadata repository that stores imported metadata, project configurations, reports, and results for all components of InfoSphere Information Server. When you share imported data in the metadata repository, other users in your organization can interact with and use the imported assets in other InfoSphere Information Server components.
InfoSphere Information Server includes the following core capabilities:

**Understand and collaborate**
Create a blueprint of your information project to develop a unified view of your business. Use this blueprint to establish a common business language that helps align business perspectives and IT perspectives. Discover and define existing data sources by understanding and analyzing the meaning, relationships, and lineage of information.

Improve visibility and data governance by enabling complete, authoritative views of information with proof of lineage and quality. These views can be made widely available and reusable as shared services, while the rules inherent in them are maintained centrally.

**Cleanse and monitor**
Standardize, cleanse, and validate information in batch processing and real time. Load cleansed information into analytical views to monitor and maintain data quality. Reuse these views throughout your enterprise to establish data quality metrics that align with business objectives, enabling your organization to quickly uncover and fix data quality issues.

Link related records across systems to ensure consistency and quality of your information. Consolidate disparate data into a single, reliable record to ensure that the best data survives across multiple sources. Load this
master record into operational data stores, data warehouses, or master data applications to create a trusted source of information.

Transform and deliver
Design and develop a blueprint of your data integration project to improve visibility and reduce risk. Discover relationships among systems and define migration rules that integrate asset metadata across multiple sources and targets. Understanding relationships and integrating data reduces operating costs and promotes data quality.

Collect, transform, and distribute large volumes of data. Use built-in transformation functions that reduce development time, improve scalability, and provide for flexible design. Deliver data in real time to your business applications through bulk data delivery (ETL), virtual data delivery (federated), or incremental data delivery (change data capture).

Information integration phases
InfoSphere Information Server focuses on several phases that are part of an effective information integration project. These phases constantly evolve as the lifecycle of the project grows and changes. By providing key information integration capabilities, InfoSphere Information Server addresses each of these phases to ensure that your project is successful.

The following diagram shows how the suite components work together to create a unified data integration solution. A common metadata foundation enables different types of users to create and manage metadata by using tools that are optimized for their roles. This focus on individualized tooling makes it easier to collaborate across roles.

Enterprise Architects use InfoSphere Blueprint Director to plan and manage your project vision. After a blueprint of your information project exists, Data Architects can use InfoSphere Data Architect to discover the structure of your organization’s data, relate and integrate data assets, and create physical and logical models based on those relationships. This data can be input to InfoSphere Business Glossary, where Business Analysts and Data Analysts define and establish a common understanding of business concepts.
Data Analysts can also use InfoSphere Discovery to automate the identification and
definition of data relationships, feeding that information to InfoSphere Information
Analyzer and InfoSphere FastTrack.

Data Quality Specialists use InfoSphere Information Analyzer to design, develop,
and manage data quality rules for your organization's data to ensure data quality. As
your organization's data evolves, these rules can be modified in real time so
that trusted information is delivered to InfoSphere Business Glossary, InfoSphere
FastTrack, InfoSphere DataStage® and QualityStage®, and other InfoSphere
Information Server components.

Data Analysts can use InfoSphere FastTrack to create mapping specifications that
translate business requirements into business applications. Data Integration
Specialists can use these specifications to generate jobs that become the starting
point for complex data transformation in InfoSphere DataStage and QualityStage.
By using the InfoSphere DataStage and QualityStage Designer, Data Integration
Specialists develop jobs that extract, transform, load, and check the quality of data. SOA Architects use InfoSphere Information Services Director to deploy integration
tasks from the suite components as consistent, reusable information services.

InfoSphere Metadata Workbench provides end-to-end data flow reporting and
impact analysis of your organization's data assets. Business Analysts, Data
Analysts, Data Integration Specialists, and other users interact with this component
to explore and manage the assets that are produced and used by InfoSphere
Information Server. InfoSphere Metadata Workbench enables users to understand
and manage the flow of data through your enterprise, and discover and analyze
relationships between information assets in the InfoSphere Information Server
metadata repository. You use InfoSphere Metadata Asset Manager to import
technical information into the metadata repository, such as BI reports, logical
models, physical schemas, and InfoSphere DataStage and QualityStage jobs.

Plan

InfoSphere Information Server includes capabilities that you can use to manage the
structure of your information project from initial sketches to delivery. By
collaborating on blueprints, your team can connect the business vision for your
project with corresponding business and technical artifacts.

To enhance your blueprint, you can create a business glossary to develop and
share a common vocabulary between your business and IT users. The terms that
you create in your business glossary establish a common understanding of
business concepts, further improving communication and efficiency.

As your information landscape evolves, it is crucial to understand how information
assets are connected. To help users understand the origin of data, you can associate
the terms that you create to information assets in your blueprint.

For example, users can view data lineage reports to understand how data flows
between assets. Terms are stored in the common metadata repository so that they
can be shared and reused by users of other suite tools. When a user changes a
term, the change is made in every location where that term is used, ensuring that a
vocabulary is standardized throughout your enterprise.
Discover and analyze

InfoSphere Information Server can help you automatically discover the structure of your data, and then analyze the meaning, relationships, and lineage of that information. By using a unified, common metadata repository that is shared across the entire suite, InfoSphere Information Server provides insight into the source, usage, and evolution of a specific piece of data.

A common metadata foundation enables different types of users to create and manage metadata by using tools that are optimized for their roles. This focus on individualized tools makes it easier to collaborate across roles. For example, data analysts can use analysis and reporting functions to generate integration specifications and business rules that they can monitor over time. Subject matter experts can use web-based tools to define, annotate, and report on fields of business data.

By automating data profiling and data-quality auditing within systems, your organization can achieve the following goals:

- Understand data sources and relationships
- Eliminate the risk of using or proliferating bad data
- Improve productivity through automation
- Use existing information assets throughout your project
Design

InfoSphere Information Server can help you design and create information models based on specific requirements of your information project. Carefully designing your physical data models, logical data models, and databases ensures that your architecture can handle changes as they occur, rather than reacting to changes after they happen.

New data continuously enters your applications, data warehouses, and business analytic systems. By using InfoSphere Information Server, you can design sophisticated data quality rules that you can modify in real time as your data evolves. In addition, you can scan samples of your data to determine their quality and structure so that you can correct problems before they affect your project. This approach ensures reliability and integrity of your data by consistently monitoring changes and making modifications.

You can also design your architecture to move large quantities of data in real time from your source applications to your data warehouse or analytics dashboard. Poor design requires constant changes to adapt your environment as the size of data volumes fluctuate. InfoSphere Information Server helps you to design your architecture to handle these demands from the outset so that the information that you need in your warehouses and analytic systems is delivered quickly and reliably.

Develop

InfoSphere Information Server supports information quality and consistency by standardizing, validating, matching, and merging data. By using the suite of components, you can certify and enrich common data elements, use trusted data such as postal records for name and address information, and match records across or within data sources.

InfoSphere Information Server enables a single record to survive from the best information across sources for each unique entity, helping you to create a single, comprehensive, and accurate view of information across source systems.

In addition, InfoSphere Information Server transforms and enriches information to ensure that it is in the required context for new uses. Hundreds of prebuilt transformation functions combine, restructure, and aggregate information.

Transformation functions are broad and flexible to meet the requirements of varied integration scenarios. For example, InfoSphere Information Server provides inline validation and transformation of complex data types such as US Health Insurance Portability and Accountability Act (HIPAA), and high-speed joins and sorts of...
heterogeneous data. InfoSphere Information Server also provides high-volume, complex data transformation and movement functions that can be used for stand-alone extract, transform, and load (ETL) scenarios, or as a real-time engine for processing applications or processes.

Deploy

InfoSphere Information Server is built on a framework that enables information to be moved throughout the stages of your data integration project. These tools provide the capabilities necessary to integrate with your source code control system, move information assets throughout the enterprise, monitor the operational environment, and administer changes.

After you design and develop extract, transform, and load (ETL) jobs and other transformation jobs, you can monitor the activity, system resources, and workload management queues. An intuitive console aggregates this information so that you can troubleshoot problems with failed jobs, monitor and improve the performance of job runs, and actively monitor the status of your environment from a single location.

You can also import, export, and manage common metadata assets that are used by various components in the InfoSphere Information Server suite. After you share imports to the metadata repository, the imported assets are available to users of other suite tools. Other users can analyze these assets, use them in jobs, assign them to terms, or designate stewards for the assets. Deploying a common metadata repository ensures that your metadata is consistently applied and available to all users in your enterprise.

Core administration tasks, such as security, licensing, logging, and scheduling are centralized in a single console. Administrators can stop and restart services, manage user accounts, and back up and restore data across the enterprise. Changes are effective throughout the entire suite, simplifying administration and accelerating deployment to other InfoSphere components.

Deliver

InfoSphere Information Server includes the capabilities to virtualize, synchronize, and move information to the people, processes, and applications that need it. Information can be delivered by using federation-based, time-based, or event-based processing, moved in large bulk volumes from location to location, or accessed in place when it cannot be consolidated.
InfoSphere Information Server provides direct, local access to various information sources, both mainframe and distributed. It provides access to databases, files, services, and packaged applications, and to content repositories and collaboration systems. Companion products allow high-speed replication, synchronization, and distribution across databases, change data capture, and event-based publishing of information.

Components in the suite

The InfoSphere Information Server suite comprises numerous components, each providing distinct capabilities for data integration. Coupled together, these components form the building blocks necessary to deliver trusted information across your enterprise, regardless of the complexity of your environment.

InfoSphere Information Server solutions include different components to meet your needs. Each solution includes InfoSphere Blueprint Director, InfoSphere Discovery, and InfoSphere Metadata Workbench as the foundation components. The additional components in each solution provide distinct capabilities that focus on data quality, data integration, and connecting your business users and IT users.

- **InfoSphere Information Server Business Information Exchange** provides capabilities that help you to establish and maintain an understanding of your business as your information project changes.
- **InfoSphere Information Server for Data Integration** provides capabilities that enable transformation and ongoing delivery of data.
- **InfoSphere Information Server for Data Quality** provides rich capabilities that enable ongoing insight into data quality.
Table 1. Components that are included in each InfoSphere Information Server solution

<table>
<thead>
<tr>
<th>Component</th>
<th>InfoSphere Information Server Business Information Exchange</th>
<th>InfoSphere Information Server for Data Integration</th>
<th>InfoSphere Information Server for Data Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>InfoSphere Blueprint Director</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>InfoSphere Discovery</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>InfoSphere Metadata Workbench</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>InfoSphere Data Architect</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>InfoSphere Business Glossary</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>InfoSphere Business Glossary Anywhere</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InfoSphere Information Analyzer</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>InfoSphere QualityStage</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>InfoSphere Information Services Director</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InfoSphere DataStage and QualityStage Designer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InfoSphere Data Click</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InfoSphere FastTrack</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InfoSphere Data Replication</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For more information about InfoSphere Data Click, see the IBM InfoSphere Data Click User’s Guide.


**InfoSphere Blueprint Director**

You can use InfoSphere Blueprint Director to define and manage blueprints of your data integration project, from initial sketches through delivery.

InfoSphere Blueprint Director provides consistency and communication across your data integration project by linking the solution overview and detailed design documents together. Team members can understand the project as it evolves. Creating a well documented and complete blueprint of both the business vision and the technical vision helps your IT department align business requirements with your enterprise reference architecture.
Scenario for governing the project vision

This scenario shows how one organization used InfoSphere Blueprint Director to govern the vision for their data integration project.

Banking: Asset integration

A major banking institution acquired several smaller banks as part of a recent expansion. The enterprise architect sketched the vision for the integration project to illustrate how to integrate the new and existing data warehouses. However, the sketches were not linked to physical assets, leaving the business vision disconnected from the technical solution. Without a tool to bridge the gap between the IT vision and the business vision, project managers were unable to accurately track the multiple data sources that needed to be integrated.

The enterprise architect used InfoSphere Blueprint Director to create a blueprint for the warehouse project, turning the vision for the project into a governed artifact. The blueprint is linked to actual metadata, allowing the company to collaborate and build a unified vision for the project.

By drilling into the subdiagrams and predefined project methods, the team can view every part of the project, including the activities and tasks contained within each phase. The team can publish the blueprint and share it with the architecture team to review new projects and identify potential issues across projects.

InfoSphere Blueprint Director tasks

You can use InfoSphere Blueprint Director to synchronize the project vision with the actual solution by linking the blueprint to business and technical artifacts throughout the lifecycle of your data integration project. Understanding each phase in the project enables enterprise architects to incorporate change into the existing architecture.

Your organization can use InfoSphere Blueprint Director to complete the following tasks:

Define and manage the project architecture

By visualizing the project architecture in a blueprint, team members can understand which tasks are required at each phase of the project, how those tasks must be implemented, who is working on each task, and which assets are involved.

Enterprise architects create blueprints to form the reference architecture for your project. The main blueprint might depict the data sources, data integration points, data repositories, analytic processes, and consumers. Each of these domains represent the different phases of the project, and can be modified as the project changes.

Link the project design to existing metadata

By planning information integration projects and linking the designs to existing metadata, enterprise architects can incorporate existing architectures into new designs.

Users can develop their free-form designs into a layered, interactive blueprint that represents your project. Subdiagrams represent additional parts of each phase of the project and can be linked to the main blueprint. The overall solution connects the visions of each team member with real-world data and concepts, rather than developing a disconnected diagram.
Enterprise architects can develop blueprints based on proven methodologies. By linking a method to a blueprint, enterprise architects can create documentation that describes user roles, responsibilities, and tasks for each phase of the project to provide contextual guidance on the development of solution artifacts.

**Modify the environment as the project changes**
To highlight changes in each phase of the project, users can incorporate milestone planning into each blueprint.

Users can define the evolution of a blueprint over time, and view the assets that are active during each project milestone. To represent existing metadata and assets specific to your business, users can create custom elements to include in blueprints.

Users can then publish blueprints to a common repository so that the extended team can review blueprints across projects. This visibility helps team members to identify areas for improving reuse and consistency.

**Where InfoSphere Blueprint Director fits in the suite architecture**
Enterprise architects use InfoSphere Blueprint Director to create a reusable blueprint that encapsulates the vision of your information integration project.

For example, the IT team can connect blueprint elements to metadata and use InfoSphere Metadata Workbench to view the referenced metadata. The IT team can also connect elements to data models in InfoSphere Data Architect, ETL definitions in InfoSphere DataStage, cleansing functions in InfoSphere QualityStage, URLs in web browsers, or files in other applications.

By integrating with other InfoSphere Information Server components and linking information assets to actual metadata, the blueprint and subdiagrams provide an accurate representation of your information landscape. As the project changes, team members can collaborate to update blueprints, ensuring that the business vision and the technical project plan are synchronized.
InfoSphere Business Glossary

InfoSphere Business Glossary is an interactive, web-based tool that enables users to create, manage, and share an enterprise vocabulary and classification system. InfoSphere Business Glossary Anywhere, its companion product, augments InfoSphere Business Glossary by improving productivity, increasing collaboration, and assigning ownership of business data to data stewards.

InfoSphere Business Glossary provides a collaborative authoring environment that helps members of an enterprise create a central collection of enterprise-specific terminology, including relationships to technical information assets. Such a collection, called a business glossary, is designed to help users understand business language and the business meaning of information assets like databases, jobs, database tables and columns, and business intelligence reports.

From within InfoSphere Business Glossary, designated users can define terms, categories, information governance policies, and information governance rules.

By using InfoSphere Business Glossary and InfoSphere Business Glossary Anywhere, users can gain insight into common business terminology, descriptions of data, ownership of terms and metadata, and how terms relate to information assets.

Scenarios for creating a business glossary

These scenarios show how two different organizations used IBM® InfoSphere Business Glossary and IBM InfoSphere Business Glossary Anywhere to develop and manage their metadata to solve business problems.

Insurance: Risk information

A leading insurance company hired several new claims agents as part of a recent expansion. To process claims correctly, agents must know how the company defines levels of risk. After three months, the company realized that claims filed by these agents were being filed incorrectly, leading to longer processing time and customer dissatisfaction. For example, a new agent was instructed to file a claim for an Assigned Risk 3 after deductible. However, the agent did not know what Assigned Risk 3 was and did not know where to find more information, so the claim was filed in the wrong category.

The company deployed InfoSphere Business Glossary to define and categorize terms as they relate to the business. In addition to a definition, the business analyst for the company associated each term with an asset, such as a specific table or column in the company database. The business analyst then assigned the terms to individual stewards to maintain the terms according to the needs of the company. As the definitions change and as new terms are added, stewards can update the terms so that the definition for each term remains current.

When the company hires new claims agents, they can use InfoSphere Business Glossary to browse for terms that they are unfamiliar with, decreasing the time that it takes for them to understand how to file new cases. In addition, the agents can view related business terms, instructions on how to handle related claims, and the location in the database where data for each term is stored.

Manufacturing: Production information

The chief financial officer (CFO) of a leading manufacturing company wants to determine the profitability of a production line by using specific financial
parameters. The CFO contacted the head of the finance department to obtain information about each stage of production, but data was not available. Without information about each stage of production, the CFO could not determine where operating costs were highest or lowest.

The finance department used InfoSphere Business Glossary to categorize each stage of production. Members of the finance team assigned terms within each category to specific database tables and columns. These relationships enable business analysts at the company to create a business intelligence report on revenues and operating expenses at each stage of production.

Business analysts at the company used InfoSphere Business Glossary to view each category, check the meaning of its terms, and select which terms were needed for the report. The analysts contacted the database management team to build a report based on the selected terms. In addition to understanding the semantic meaning of these terms, the analysts used business lineage to view the data sources that populated the business intelligence report. The CFO was able to review financial parameters for each stage of production to analyze where changes were necessary. While viewing the report, the CFO used InfoSphere Business Glossary Anywhere to understand the definitions of the fields that were used in the report without having to contact the finance department.

InfoSphere Business Glossary tasks
You can use InfoSphere Business Glossary to create a business glossary, which is a central information collection that helps users in your enterprise understand business language and the business meaning of information assets.

Your organization can use InfoSphere Business Glossary to complete the following tasks:

**Browse and search for terms and categories**
To understand the meaning of business terms as defined in the enterprise, users can browse and search for terms and categories in the business glossary.

For example, an analyst might want to find out how the term “Accepted Flag Type” is defined in his business. The analyst can use InfoSphere Business Glossary or IBM InfoSphere Business Glossary Anywhere to find the definition of the term, its usage, and other terms related to it. The analyst can also find out what information governance rules and information assets are related to the term and can find out details about these assets.

**Develop and share a common vocabulary between business and technology**
A common vocabulary gives diverse users a common understanding of business concepts, improving communication and efficiency.

For example, one department in an organization might use the word “customer,” a second department might use the word “user,” and a third department might use the word “client,” all to describe the same type of individual. You can use InfoSphere Business Glossary to capture these terms, define their meaning, create relationships between them and consolidate terminology to achieve increased precision in communications. Other business glossary users can refer to this information at any time.

**Associate business information with other information assets**
Glossary authors can attach business meaning to technical assets and other information sources by assigning them to business terms and information governance rules.
Technical assets, such as database columns, tables, or schemas that reside in the InfoSphere Information Server metadata repository, can be shared with other InfoSphere Information Server products. As a result, users of InfoSphere Business Glossary can view information about these technical assets, and depending upon their access rights, associate the assets with terms or rules that are defined in the business glossary.

**Develop and share a common set of information governance policies and rules**

Information governance policies and rules give team members a common understanding of business requirements or objectives and an understanding of how to manage information assets to comply with those requirements or objectives.

By defining information governance policies and information governance rules in InfoSphere Business Glossary, users can easily share business requirements and objectives and describe how information assets comply with those requirements and objectives with the appropriate members of your enterprise. The integration capabilities between information governance rules, operational rules, and information assets close the loop between declared business objectives and implemented data governance.

Once defined, information governance rules can be associated with runtime data rules such as archiving rules, data quality rules, security rules, and data standardization rules. In addition, information governance rules can be associated with terms and with data assets such as databases, tables, columns and model objects. Within InfoSphere Business Glossary, you can specify the type of relationship that the information governance rule has to the information asset. For example, you can specify whether the information governance rule is implemented by, or should be implemented by, an information asset. You can also specify the inverse relationship to indicate that an information asset implements a particular information governance rule. In addition, you can specify that a rule governs an asset, or that an asset is governed by a rule.

**View relationships between business glossary assets and other assets**

In addition to learning about the relationships between business glossary assets and other information assets, users can also view business lineage reports to discover how data flows among information assets.

InfoSphere Business Glossary, when used in conjunction with InfoSphere Metadata Workbench, provides information about the flow of data among information assets. For example, a business lineage report might show the flow of data from one database table to another database table, and from there to another report.

**InfoSphere Business Glossary Anywhere tasks**

IBM InfoSphere Business Glossary Anywhere provides instant access to business glossary content from text files that are open on Microsoft Windows-based computers. In addition, InfoSphere Business Glossary Anywhere comes with the IBM InfoSphere Business Glossary REST API to extend functionality to other applications.

You can use InfoSphere Business Glossary Anywhere to complete the following tasks:

**Search the business glossary from the desktop**

With IBM InfoSphere Business Glossary Anywhere, users can perform searches of the business glossary while in the midst of other computer-based tasks with no loss of context or focus.
Extend functionality to other applications

IBM InfoSphere Business Glossary Anywhere provides licensing for the InfoSphere Business Glossary REST (Representational State Transfer) API. The InfoSphere Business Glossary REST API allows developers to build new client applications or extend existing applications that access and author business glossary content.

Where InfoSphere Business Glossary fits in the suite architecture

You can use InfoSphere Business Glossary to create, manage, and share an enterprise vocabulary and classification system. Because InfoSphere Information Server uses a shared metadata repository, the business glossary that you create can be shared with other components in the suite.

You use InfoSphere Metadata Asset Manager to import technical information assets into the metadata repository, such as BI reports, logical models, physical schemas, and InfoSphere DataStage and QualityStage jobs. You can interact with the assets in the metadata repository by using InfoSphere Metadata Workbench.

You can use InfoSphere Discovery to discover data relationships, and then use these relationships in InfoSphere Business Glossary to establish a common vocabulary and promote collaboration across your business teams and IT teams.

The terms and definitions that you create can be linked to assets in the blueprints that you develop with InfoSphere Blueprint Director. You can link terms and definitions to assets in the metadata repository, such as database columns, tables, or schemas that you analyze with InfoSphere Information Analyzer. You can access business glossary content from other components in the suite, such as InfoSphere FastTrack, InfoSphere Metadata Workbench, and InfoSphere Data Architect. By using IBM InfoSphere Business Glossary Anywhere, you can search your business glossary to understand the meaning of key terms in your business.

InfoSphere Business Glossary also supports integration with other software such as IBM Cognos® Business Intelligence products and IBM Industry Models. InfoSphere Business Glossary Packs for IBM Industry Models provide industry-specific sets of terms that you can use to quickly populate a business glossary.
**InfoSphere DataStage**

InfoSphere DataStage is a data integration tool that enables users to move and transform data between operational, transactional, and analytical target systems.

Data transformation and movement is the process by which source data is selected, converted, and mapped to the format required by target systems. The process manipulates data to bring it into compliance with business, domain, and integrity rules, and with other data in the target environment.

InfoSphere DataStage provides direct connectivity to enterprise applications as sources or targets, ensuring that the most relevant, complete, and accurate data is integrated into your data integration project.

By using the parallel processing capabilities of multiprocessor hardware platforms, InfoSphere DataStage enables your organization to solve large-scale business problems. Large volumes of data can be processed in batch, in real time, or as a web service, depending on the needs of your project.

Data integration specialists can use the hundreds of prebuilt transformation functions to accelerate development time and simplify the process of data transformation. Transformation functions can be modified and reused, decreasing the overall cost of development, and increasing the effectiveness in building, deploying, and managing your data integration infrastructure.

As part of the InfoSphere Information Server suite, InfoSphere DataStage uses the shared metadata repository to integrate with other components, including data profiling and data quality capabilities. An intuitive web-based operations console enables users to view and analyze the runtime environment, enhance productivity, and accelerate problem resolution.
Balanced Optimization

Balanced Optimization helps to improve the performance of your InfoSphere DataStage job designs that use connectors to read or write source data. You design your job and then use Balanced Optimization to redesign the job automatically to your stated preferences.

For example, you can maximize performance by minimizing the amount of input and output (I/O) that are used, and by balancing the processing against source, intermediate, and target environments. You can then examine the new optimized job design and save it as a new job. Your root job design remains unchanged.

You can use the Balanced Optimization features of InfoSphere DataStage to push sets of data integration processing and related data I/O into a database managements system (DBMS) or into a Hadoop cluster.

Integration with Hadoop

InfoSphere DataStage includes additional components and stages that enable integration between InfoSphere Information Server and Apache Hadoop. You use these components and stages to access and interact with files on the Hadoop Distributed File System (HDFS).

Hadoop is the open source software framework that is used to reliably manage large volumes of structured and unstructured data. HDFS is a distributed, scalable, portable file system written for the Hadoop framework. This framework enables applications to work with thousands of nodes and petabytes of data in a parallel environment. Scalability and capacity can be increased by adding nodes without interruption, resulting in a cost effective solution that can run on multiple servers.

InfoSphere DataStage provides massive scalability by running jobs on the InfoSphere Information Server parallel engine. By supporting integration with Hadoop, InfoSphere DataStage enables your organization to maximize scalability in the amount of storage and data integration processing required to make your Hadoop projects successful.

**Big Data File stage**

The Big Data File stage enables InfoSphere DataStage to exchange data with Hadoop sources so that you can include enterprise information in analytical results. These results can then be applied in other IT solutions.

**Oozie Workflow Activity stage**

The Oozie Workflow Activity stage enables integration between Oozie and InfoSphere DataStage. Oozie is a workflow system that you can use to manage Hadoop jobs.

Scenarios for data transformation

These scenarios show how organizations used InfoSphere DataStage to address their complex data transformation and movement needs.

**Retail: Consolidating financial systems**

A leading retail chain watched sales flatten for the first time in years. Without insight into store-level and unit-level sales data, they could not adjust shipments or merchandising to improve results. With long lead times for production and existing, large-volume manufacturing contracts, they could not change their product lines quickly, even if they understood the problem. To integrate the
company’s forecasting, distribution, replenishment, and inventory management processes, they needed a way to migrate financial reporting data from many systems to a single system of record.

The company deployed InfoSphere Information Server to deliver data integration services between business applications in both messaging and batch file environments. InfoSphere DataStage is now the company-wide standard for transforming and moving data. The service-oriented interface allows them to define common integration tasks and reuse them throughout the enterprise. This new methodology and the use of reusable components for other global projects will lead to future savings in design, testing, deployment, and maintenance.

**Banking: Understanding the customer**

A large retail bank understood that the more it knew about its customers, the better it could market its products, including credit cards, savings accounts, checking accounts, certificates of deposit, and ATM services. Faced with terabytes of customer data from vendor sources, the bank recognized the need to integrate the data into a central repository where decision-makers could retrieve it for market analysis and reporting. Without a solution, the bank risked flawed marketing decisions and lost cross-selling opportunities.

The bank used InfoSphere DataStage to automatically extract, transform, and load raw vendor data into its data warehouse, such as credit card account information, banking transaction details, and web site usage statistics. From a unified data warehouse, the bank can generate reports to track the effectiveness of programs and analyze marketing efforts. InfoSphere DataStage helps the bank maintain, manage, and improve its information management with an IT staff of three instead of six or seven, saving hundreds of thousands of dollars in the first year alone, and enabling it to use the same capabilities more rapidly on other data integration projects.

**InfoSphere DataStage tasks**

You use InfoSphere DataStage to develop *jobs*, which process and transform your data. You can administer, manage, deploy, and reuse these jobs to integrate data across many systems throughout your organization.

Your organization can use InfoSphere DataStage to complete the following tasks:

**Process and transform large volumes of data**

By handling the collection, integration, and transformation of large volumes of data, your organization can linearly scale the speed of data throughput. A scalable platform that includes parallel processing and incorporates flexible, reusable functions enables users to design logic once, and then run and scale that logic anywhere.

By using parallel processing capabilities of multiprocessor hardware platforms, you can scale transformation jobs to address any demands, large or small. During development, the deployment configuration automatically adds the degree of parallelism that you specify. By making a simple change to the configuration file, you can change your application from 2-way processing to 32-way processing to 128-way processing.

**Design reusable transformation jobs**

Reusable transformation functions enable data integration specialists to maximize speed, flexibility, and effectiveness in their designs.
Data integration specialists use the rich user interface for all design work, including workflow, data integration, and data quality. Prebuilt transformation functions can be dragged to a design, making it easy to determine the flow of information and the transformations that occur. Any portion of the design can be shared and reused across the data integration landscape, maximizing reuse and productivity.

**Extend connectivity to various objects**

By using common connectors, any data source that is supported by InfoSphere Information Server can be used as input to or output from InfoSphere DataStage, enabling your organization to integrate data effectively across the enterprise.

A nearly unlimited number of heterogeneous data sources and targets are supported, including text files, complex data structures in XML, enterprise resource planning (ERP) systems such as SAP and PeopleSoft, nearly any database, web services, and business intelligence (BI) tools like SAS.

**Manage operations and resources**

By operating in real time, your organization can capture messages or extract data at any moment on the same platform that integrates bulk data and uses transformation rules. This integration ensures that data can be used to respond to your data integration needs on demand.

Real-time data integration support captures messages from Message Oriented Middleware (MOM) queues using JMS or WebSphere® MQ adapters to combine data into operational and historical analysis perspectives. By using InfoSphere DataStage with InfoSphere Information Services Director, data integration jobs can be deployed with Java™ Message Services, web services, or other services. This service-oriented architecture (SOA) enables numerous developers to share complex data integration processes without having to understand the steps contained in the services.

You can use the InfoSphere DataStage Operations Console to access information about your jobs, job activity, and system resources for each of your InfoSphere Information Server engines. The Operations Console is useful for troubleshooting failed job runs, improving job run performance, and actively monitoring your engines.

**Where InfoSphere DataStage fits in the suite architecture**

In its simplest form, InfoSphere DataStage enables users to integrate, transform, and deliver data from source systems to target systems in batch and real time. By using prebuilt transformation functions, data integration specialists can reduce development time and improve the consistency of design and deployment.

Any part of a design can be shared and reused, so that logic can be designed once and scaled to run on any system that uses InfoSphere DataStage. Data integration specialists can import metadata from tables, views, and stored procedures to use in their designs. This flexibility and interoperability ensures full data lineage from the source, through transformations, to the target.

After data is transformed, it can be delivered to data warehouses, data marts, and other enterprise applications such as PeopleSoft, Salesforce.com, and SAP. InfoSphere Information Server Packs support these enterprise applications and many others so that metadata can be integrated with InfoSphere DataStage.

SOA architects use InfoSphere Information Services Director to publish data integration logic as shared services that can be reused across the enterprise.
Providing the transformed data from InfoSphere DataStage in a service-oriented architecture (SOA) ensures that accurate, complete, and trusted information is delivered to business users.

### InfoSphere Data Architect

By using InfoSphere Data Architect, your organization can design data assets, understand data assets and their relationships, and streamline your data integration project.

InfoSphere Data Architect provides tools to discover, model, visualize, and relate heterogeneous data assets. Users can design and develop logical, physical, and dimensional models. Traditional data modeling capabilities are combined with unique mapping capabilities and model analysis, all organized in a modular, project-based application.

InfoSphere Data Architect discovers the structure of heterogeneous data sources by examining and analyzing the underlying metadata. You can browse the hierarchy of data elements to understand their detailed properties and visualize tables, views, and relationships in a contextual diagram.

Impact analysis lists all dependencies on data elements so that users can manage changes without interruption. Advanced synchronization technology compares two models, a model to a database, or two databases. Changes from the comparison can be promoted within and across data models and data sources, ensuring that data remains current.

Integration with other InfoSphere components enables integration across the enterprise. For example, dimensional models designed with InfoSphere Data Architect reduce development time for InfoSphere Data Warehouse and Cognos business intelligence reporting. By using the shared metadata repository, InfoSphere Data Architect can share terms and definitions with InfoSphere Business Glossary, making corporate standards easier to implement.
**Scenario for designing data**

This scenario shows how one organization used InfoSphere Data Architect to simplify and accelerate the design of their data warehouse, dimensional models, and change management processes.

**Retail: Expanding metadata**

A successful retail company recently expanded their operations to offer online sales. The company needed to expand their customer metadata to include new entities such as tax ID, online payment method, address, and shipping method. Each new entity needed a specific term to distinguish it from existing entities so that naming, meaning, and values remained consistent. Without a way to understand and visualize the structure of their data, the data architect was faced with manually designing a new data solution to support their expanding customer metadata.

By using InfoSphere Data Architect, the data architect imported an existing logical data model, and transformed the logical model to a physical data model. To visualize and work with the logical model, the data architect created a diagram, which is a visual representation of the model. The diagram shows each of the entities that comprise the model, including relationships between data. The data architect added entities to the model to represent new customer metadata. All of the entities and attributes are tables and columns in the physical model, which can be published in a report. This report includes the diagram and all associated tables and columns so that the database administrator can accurately update the existing database.

As new entities were created, the data architect created new business terms for each entity. Each of these terms were exported to InfoSphere Business Glossary, where the data analyst added the terms to the existing enterprise vocabulary.

**InfoSphere Data Architect tasks**

You can use InfoSphere Data Architect to discover, model, relate, and standardize diverse and distributed assets.

Your organization can use InfoSphere Data Architect to complete the following tasks:

**Discover the structure of heterogeneous data sources**

By understanding the structure of your data sources, you can help your users to understand the detailed properties of each data asset, while increasing efficiency and reducing time to value for your data integration project.

InfoSphere Data Architect discovers the structure of heterogeneous data sources by examining and analyzing the underlying metadata. The tool requires only an established Java Database Connectivity (JDBC) connection to the data sources to explore their structures by using local queries.

After discovering the structure of your data, data architects can view peer relationships between databases, schemas, and tables. This topology diagram helps to visualize the connections between source objects and target objects in your enterprise.

**Develop data models**

By using InfoSphere Data Architect, you can create logical, physical, and domain models for DB2®, Informix®, Oracle, Sybase, Microsoft SQL Server, MySQL, and Teradata. Elements from logical and physical data models can
be represented in diagrams that use Information Engineering (IE) notation. Physical data models can also use Unified Modeling Language (UML) notation.

You can import and use IBM Industry Models in InfoSphere Data Architect to incorporate data models for your specific industry. By supporting IBM Netezza® data warehouse appliances, you can include transformations from the IBM Industry Models to IBM Netezza physical data models.

Dimensional modeling is also supported to extend logical and physical data models. Dimensional models map the aspects of each process within your business, reducing the time required to develop warehouse and business intelligence systems.

Implemet corporate standards
Defining and implementing corporate standards increases data quality and enterprise consistency for naming, meaning, values, relationships, privileges, privacy, and lineage. By using InfoSphere Data Architect, data architects can define standards once and associate them with diverse models and databases.

Data architects can also incorporate terms and definitions from InfoSphere Business Glossary into existing or new data models to align business and IT.

Where InfoSphere Data Architect fits in the suite architecture
InfoSphere Data Architect is a data design solution that helps you design and develop dimensional models, implement corporate standards, and facilitate collaboration across your teams.

Business analysts use InfoSphere Business Glossary to create, manage, and share an enterprise vocabulary and classification system. Data architects use InfoSphere Data Architect to incorporate the terms and definitions from that vocabulary into existing or new data models to align business and IT.

Data architects can extend the definitions with information about the mapping of those concepts to the physical databases so that business users can identify appropriate data sources for ad hoc queries or reports. Those definitions can also become the basis for InfoSphere Information Server to load data warehouses, consolidate databases for mergers, or establish and manage master data.

Data architects can use InfoSphere Data Architect to design dimensional models for InfoSphere Data Warehouse, IBM Netezza, and Cognos BI reporting. This support for dimensional modeling can help reduce development time of warehouse and business intelligence systems.

Team members across your organization can use InfoSphere Data Architect as a plug-in to a shared Eclipse instance, or share assets through standard configuration management repositories like Rational Team Concert™ and Subversion. This integration enables collaboration across your organization and ensures a clear division of responsibilities.
**InfoSphere Discovery**

InfoSphere Discovery provides innovative data exploration and analysis techniques to automatically discover relationships and mappings among structured data in your enterprise. The analysis is based on actual values in the data, rather than on just metadata.

This value-driven analysis means that InfoSphere Discovery can detect relationships between tables and columns whose names or metadata alone do not suggest any connection. InfoSphere Discovery can identify and generate highly complex transformations that you can use to describe the locations and formats of sensitive data, describe the relationships of data elements across applications, or output as SQL code or extract, transform, and load (ETL) code for use in data transformation jobs.

Before you implement any data-centric projects, you must know what data you have, where it is located, and how it relates between various systems.

InfoSphere Discovery is a data analysis tool that provides a full range of data analysis capabilities that include single source profiling, cross-system data overlap analysis, and matching key discovery. It provides a profiling solution with automatic primary-foreign key discovery and validation. In addition, InfoSphere Discovery can analyze the data overlap across multiple sources simultaneously.

**Scenarios for data discovery**

These scenarios show how organizations used InfoSphere Discovery to examine and better understand their data.

**Retail: Uncovering relationships**

A major clothing retailer watched sales plateau steadily for two continuous quarters. The retailer needed a new marketing strategy, but was unsure where to begin their campaign. With their current data profiling tool, data analysts could find potential primary keys, but were forced to manually instruct the tool to find that same key in another table. Determining relationships between tables was time
consuming, difficult, and not intuitive. Without a solution to discover relationships, data analysts could not make associations between massive amounts of sales data.

By using InfoSphere Discovery, data analysts analyzed the data values in all tables that contained customer data and automatically generated an entity relationship diagram. The diagram showed relationships between data, such as how customer age and demographic information related to specific clothing purchases. Data analysts reviewed existing primary keys and created new primary keys based on these new relationships.

Data analysts then used the primary-foreign key relationships to group tables into business entities comprised of related tables. These tables represented specific business objects, which are logical clusters of all tables in a data set that have one or more columns that contain data related to the same business entity. Using these smaller, targeted business objects, the clothing retailer focused on relationships in their sales data to help drive their marketing strategy.

**Healthcare: Discovering sensitive data**

A major hospital recently decided to convert all patient records from paper documents to digital records. Because the hospital has multiple branches, several databases were created to house the digital records. The resulting data was largely structured, but each database contained tables that were formatted inconsistently and contained empty cells. Sensitive data, such as social security number, blood type, and list of medications, was not labeled appropriately or was merged with other parts of the patient record. Without a unified solution, discovering and analyzing the digital records would require months of human involvement and manual manipulation of data.

The hospital used InfoSphere Discovery to discover statistics about the columns in each of the databases. These statistics were used to develop a detailed understanding of the structure and format of the patient records, which helped to normalize and standardize records. Using the built-in classification algorithms, data analysts identified patterns that matched data in each record, such as patient name, address, and date of birth. By using custom classifications, data analysts isolated sensitive data elements, and enforced enterprise-wide policies to protect these elements, masking them from unauthorized users.

**InfoSphere Discovery tasks**

You can use InfoSphere Discovery to identify and document what data you have, where it is located, and how it is linked across systems. You can create and populate data sets, discover data types, discover primary-foreign keys, and organize data into data objects within each data source.

Your organization can use InfoSphere Discovery to complete the following tasks:

**Run column analysis**

Column analysis automatically discovers many standard column statistics, such as cardinality, data type frequency, value frequency, and minimum and maximum values. You can drill down into statistics from column analysis to obtain different views of how your data is related. You run column analysis on each table within a data set.

You can also run overlap analysis on multiple data sources simultaneously. All columns are compared to all other columns for overlaps. The results are displayed in a graph and table format that you can use to drill down, view, sort, and filter the statistics. In addition to identifying overlapping
and unique columns, you can manage the process of tagging attributes that you consider critical to your analysis. These critical data elements, or CDEs, are the specific attributes that you want to include in your new target schema if you are migrating data or consolidating data into a new application, metadata management hub, or data warehouse.

**Discover and analyze primary-foreign keys**

InfoSphere Discovery can discover column matches, which are relationships between data in two columns in different tables, within the same data set. This relationship can be strong or weak. You can set minimum hit rate statistics and other criteria that must be met for a column pair to be considered a match.

You can automatically discover matching keys between any two data sources, even if the key is a composite key that involves many columns. You can then prototype different matching keys to determine the best matching key across multiple data sources.

**Organize data sets into data objects**

InfoSphere Discovery uses the primary-foreign key relationships to group tables into entities composed of related tables. These related tables, or data objects, are logical clusters of all tables in a data set that have one or more columns that contain data that is related to the same business entity. These business objects can be entered into IBM Optim™ for archiving data, and for creating consistent sample data sets for test data management.

Data objects are also useful when you start comparing data across sources. Each source might have different data structures and formats. Focusing on each source at the business object level and creating consistent samples of data that can be compared across sources helps to split large data sets into smaller related groups of tables and map those groups across data sources.

**Discover and analyze transformations and business rules**

InfoSphere Discovery can map two existing systems together to facilitate data migration, consolidation, or integration. InfoSphere Discovery automatically discovers complex cross-source transformations and business rules between two structured data sets.

After discovering substrings, concatenations, cross-references, aggregations, and other transformations, InfoSphere Discovery identifies the specific data anomalies that do not meet the discovered rules. These capabilities help to develop ongoing audit and remediation processes to correct errors and inconsistencies.

**Build unified schemas**

InfoSphere Discovery includes a complete workbench for analyzing multiple data sources and prototyping the combination of those sources into a consolidated, unified target, such as a master data management (MDM) hub, application, or enterprise data warehouse.

InfoSphere Discovery helps build unified data table schemas by accounting for known critical data elements and proposing statistic-based matching and conflict resolution rules. Data analysts use these rules to determine which data to consolidate for data migration, MDM, or data warehousing.

**Where InfoSphere Discovery fits in the suite architecture**

InfoSphere Discovery automates the data analysis process and generates results for other components in the InfoSphere Information Server suite.
By using InfoSphere Discovery, data analysts can identify and generate complex transformations that describe the locations and formats of sensitive data and then describe the relationships of data elements across applications.

After adding an InfoSphere Information Server data source in InfoSphere Discovery, data analysts can import and export physical tables to and from the common metadata repository. By connecting to an InfoSphere Information Server data source, data analysts can run InfoSphere Discovery tools on a wider range of information and connect with InfoSphere Information Analyzer. This connection can be used to analyze data and discover relationships that exist in the data source.

InfoSphere Business Glossary terms and assets can be imported into InfoSphere Discovery projects and automatically assigned to existing columns that match the imported assigned assets. The terms can also be manually assigned to other columns in the project. Similarly, InfoSphere Discovery classifications can be exported to InfoSphere Business Glossary.

Data analysts can also export database model files from InfoSphere Discovery to InfoSphere Information Server by using InfoSphere Metadata Asset Manager. The imported metadata includes implemented data resources and analysis information. Source and target mappings can be exported from InfoSphere Discovery to InfoSphere Metadata Workbench as extension mappings to view in data lineage reports. Target maps that are based on a source-to-target relationship discovery can be exported to InfoSphere FastTrack, helping business analysts connect relationships to enhance the integration of related data.

InfoSphere FastTrack provides capabilities to automate the workflow of your data integration project. Users can track and automate multiple data integration tasks, shortening the time between developing business requirements and implementing a solution.

Business analysts use InfoSphere FastTrack to translate business requirements into a set of specifications, which data integration specialists then use to produce a data integration application that incorporates the business requirements.
By automating the flow of information and increasing collaboration, development time is reduced. By linking information in a shared metadata repository, data is accessible, current, and integrated across the data integration project.

**Scenario for mapping data**
This scenario shows how an organization used InfoSphere FastTrack to consolidate, map, and transform data to solve a business problem.

**Financial services: Creating specifications**

A financial institution recently acquired several companies. The IT team was faced with supporting disparate data environments for its core banking system, which made customer data difficult to manage. The business analysts worked extra hours creating source-to-target mappings by manually creating specifications that map customer data across multiple targets. Because each of the acquired companies used different business terminology, the business analysts constantly revised existing term and definitions.

By using InfoSphere Information Server, the IT team consolidated the business category and category terms, customer metadata (such as account balance and credit history data fields), and multiple data models in the metadata repository. The IT team used InfoSphere Information Analyzer to analyze the data and InfoSphere Business Glossary to consolidate new and existing terms and definitions.

By using InfoSphere FastTrack, the IT team specified data relationships and transformations that the business analysts used to create specifications, which consist of source-to-target mappings. The specifications link the customer data fields to key business terms and transformation rules that are used to compute new data fields. As new customer data is added to the metadata repository, relationships can be discovered automatically and are linked to existing data records.

The business analysts can also use InfoSphere FastTrack to discover and optimize mapping by using existing data profiling results from the metadata repository. The business analysts can use these data profiling results to generate a job that a data integration specialist can modify by using the InfoSphere DataStage and QualityStage Designer client.

**InfoSphere FastTrack tasks**
You can use InfoSphere FastTrack to automate the workflow of data integration projects by maximizing team collaboration, facilitating automation, and ensuring that projects are completed successfully and on time.

Your organization can use InfoSphere FastTrack to complete the following tasks:

**Create mapping specifications for business requirements**
By capturing and defining business requirements in a common format, you can streamline collaboration between business analysts, data modelers, and data integration specialists.

You can import mapping specifications into InfoSphere FastTrack from existing spreadsheets or other CSV files. These mapping specifications show the relationship between source data and business requirements. You can annotate column mappings with business rules and transformation logic, and then export the mapping specification to the metadata repository so that it is accessible by other team members.
Discover and map relationships between data sources

By creating customized discovery algorithms, you can find exact, partial, and lexical matches on corresponding column names across source-to-target mappings.

As part of the mapping process, you can create new business terms by using InfoSphere Business Glossary, and then document relationships to corresponding physical columns. You can then publish these relationships to the InfoSphere Information Server metadata repository so that they can be shared across your teams.

Generate jobs from mapping specifications

Business analysts can take the business logic from one or more mapping specifications and transform them into an InfoSphere DataStage and QualityStage job.

The generated job includes all transformation rules that the business analyst created so that the business logic is captured automatically, without intervention from the data integration specialist. Any annotations and instructions that the business analyst creates are also included in the generated job.

Where InfoSphere FastTrack fits in the suite architecture

You can use InfoSphere FastTrack to track and automate efforts that span multiple data integration tasks from analysis to code generation, shortening the time from business requirements to solution implementation.

By using the shared metadata repository, InfoSphere FastTrack enables business analysts to discover relationships between columns in multiple tables, link columns to business glossary terms, and generate jobs that become the starting point for complex data transformation in InfoSphere DataStage and QualityStage. Source-to-target mappings can contain data value transformations that, as part of mapping specifications, define how to build applications.

Business analysts can use existing and newly created business terms by linking each business term to the physical structures to reveal a comprehensive relationship. The full lineage of business-to-technical metadata can then be published to InfoSphere Business Glossary and accessed by other users who have access to the metadata repository.

Business analysts can use the source-to-target relationships uncovered by using InfoSphere Discovery and the data analysis results from InfoSphere Information Analyzer to find and compare new column relationships. The business analyst can then connect these relationships to enhance the integration of related metadata.

The automated processes of discovery and analysis create reusable artifacts that are stored in the metadata repository and are accessible by authorized team members, enhancing the auditability and security of your data integration project. The audit trail can be followed and questions about historical decisions can be answered.
InfoSphere Information Analyzer

InfoSphere Information Analyzer provides capabilities to profile and analyze data to deliver trusted information to your organization.

Data quality specialists use InfoSphere Information Analyzer to scan samples and full volumes of data to determine their quality and structure. This analysis helps to discover the inputs to your data integration project, ranging from individual fields to high-level data entities. Information analysis enables your organization to correct problems with structure or validity before they affect your data integration project.

After analyzing data, data quality specialists create data quality rules to assess and monitor heterogeneous data sources for trends, patterns, and exception conditions. These rules help to uncover data quality issues and help your organization to align data quality metrics throughout the project lifecycle. Business analysts can use these metrics to create quality reports that track and monitor the quality of data over time. Business analysts can then use IBM InfoSphere Data Quality Console to track and browse exceptions that are generated by InfoSphere Information Analyzer.

Understanding where data originates, which data stores it lands in, and how the data changes over time is important to develop data lineage, which is a foundation of data governance. As part of the InfoSphere Information Server suite, InfoSphere Information Analyzer shares lineage information by storing it in the metadata repository. Other components in the suite can access lineage information directly to simplify the collection and management of metadata across your organization.

Scenarios for information analysis

These scenarios show how different organizations used InfoSphere Information Analyzer to facilitate data integration projects by understanding their data.
**Government: Tax collection**

The tax authority of a large government needed to modernize its tax collection system. The tax authority was responsible for collecting taxes of all types from eligible taxpayers, including individuals, organizations, and vehicle owners. Tax data was collected into a central repository, but after several decades, the details of taxpayers were in various formats such as flat files, spreadsheets, and relational databases. Taxpayer data often included multiple tax identifiers, variation in spellings of a name, or the same identifier was assigned to multiple taxpayers. Consolidating data resulted in multiple data quality issues that impeded progress and slowed development of an updated solution.

The tax authority implemented InfoSphere Information Analyzer to profile data and confirm potential data quality issues in taxpayer data. In phase one of the implementation, fifty percent of the data quality issues in the income tax segment were solved by using InfoSphere DataStage and QualityStage. In phase two, issues in sales tax data were resolved. In phase three, the income tax and sales tax databases were consolidated to form a single database. Fixing data quality issues and consolidating data helped the tax authority to identify individuals who were delinquent in paying taxes or who listed their assets under different names. This solution helps to detect fraud and avoid future exploitation of the tax code.

**Transportation services: Data quality monitoring**

A transportation service provider develops systems that enable its extensive network of independent owner-operators to remain competitive in the market. The owner-operators were exposed to competition because they were unable to receive data quickly, and executives had little confidence in the quality of the data that they received. Because the owner-operators had to manually intervene to reconcile data from multiple sources, productivity slowed excessively.

The owner-operators used InfoSphere Information Analyzer to better understand and analyze their legacy data. By increasing the accuracy of their business intelligence reports, the owner-operators restored executive confidence in their company data. Moving forward, the owner-operators implemented a data quality solution to cleanse their customer data and identify trends over time, further increasing their confidence in the data.

**Food distribution: Prepare infrastructure rationalization**

A leading US food distributor had more than 80 separate mainframe, SAP, and JD Edwards applications supporting global production, distribution, and customer relationship management (CRM) operations. This infrastructure rationalization project included planning for CRM, order-to-cash, purchase-to-pay, human resources, finance, manufacturing, and supply chain operations. The company needed to move data from these source systems to a single target system.

The company used InfoSphere Information Analyzer to profile its source systems and create master data for key business dimensions, including customer, vendor, item (finished goods), and material (raw materials). The company plans to migrate data into a single master SAP environment and a companion SAP business intelligence reporting platform.

**InfoSphere Information Analyzer tasks**

You can use InfoSphere Information Analyzer you to complete selected integration tasks as required or combine them into a larger integration flow.
Your organization can use InfoSphere Information Analyzer to complete the following tasks:

Data profiling and analysis
Completing data profiling and analysis helps you to understand the structure, content, and quality of data. Users can identify data anomalies, validate column and table relationships, and drill down to exception rows for more detailed analysis of data inconsistencies. Data profiling helps detect data quality rules and relationships that users can refine based on the needs of your organization.

InfoSphere Information Analyzer helps to complete the following data profiling functions:
• Column analysis
• Key analysis (primary, natural, or foreign keys)
• Cross-domain analysis

Data monitoring and trending
Data rules help InfoSphere Information Analyzer users assess data completeness, validity, and formats, and determine whether value combinations are valid. Rules can be simple column measures that incorporate data profiling results, or can include complex conditions that test multiple data fields.

Business users can develop additional rules to assess and measure content and quality over time to complete trending and pattern analysis, and establish baselines across uncommon data sources. Users can evaluate new results against existing benchmarks to track data quality improvements.

Facilitating integration
InfoSphere Information Analyzer facilitates information integration by using the available source and target metadata, defined data rules, and validation tables to design new data integration tasks.

By generating a set of values that data rules compare source data against, InfoSphere Information Analyzer users can generate reference tables that are used for mapping, range checking, and validity checking. Data rules can be invoked in-stream as part of InfoSphere DataStage and QualityStage jobs, or within an InfoSphere Information Services Director web service. Rules or rule set definitions can be reused as a unique rule validation stage to ensure that incoming content meets prescribed rules, and that the outgoing contents can be created based on defined rule logic.

To promote collaboration and metadata integration, InfoSphere Information Analyzer users can access the results of data rules from the metadata repository, and share the results with users of other components in the InfoSphere Information Server suite. By using the built-in application programming interfaces (APIs) and command-line interfaces (CLIs), InfoSphere Information Analyzer users can deliver results as custom dashboards or applications outside of the InfoSphere Information Server suite.

Where InfoSphere Information Analyzer fits in the suite architecture
You can use InfoSphere Information Analyzer to understand data by completing data quality assessments, data quality monitoring, and flexible data rule design and analysis. This component is supported by a range of shared services and integrates with several InfoSphere Information Server components.
Data analysts and business analysts use InfoSphere Discovery to uncover source-to-target relationships, which can be fed into InfoSphere Information Analyzer as part of data profiling. Relationships between data can be linked to existing data rules and other metadata such as patterns, ranges of values, lists of valid values, and number of null values.

Data analysts can also use InfoSphere Information Analyzer to uncover actual data and assign terms to the data, rather than guessing what the content is based on metadata definitions. Terms and policies can be linked to the data rules to help understand how and where data is evaluated.

Data quality specialists use InfoSphere Information Analyzer to develop data rules that evaluate data against benchmarks, existing rules, and specified values. By combining multiple rules, data quality specialists can provide a broad assessment of records and data sources to enable rules analysis at multiple levels.

Results that are generated by using InfoSphere Information Analyzer can be shared with other components such as InfoSphere DataStage and QualityStage. By accessing the shared metadata repository, data integration specialists can use these tools to integrate data quality results into their projects. Integration with InfoSphere Business Glossary enables all suite users to connect data sources to business terminology so that insights into data quality can be shared with a broader audience.

**InfoSphere Information Services Director**

InfoSphere Information Services Director provides an integrated environment that enables users to rapidly deploy InfoSphere Information Server logic as services.

The service-oriented architecture (SOA) infrastructure of InfoSphere Information Services Director ensures that data integration logic that is developed in InfoSphere Information Server can be used by any business process. The best data is available at all times, to all people, and to all processes.
InfoSphere Information Services Director defines services as true business objects, which are deployed locally in your application server. This integration completely hides the complexity of implementation from the service that consumes published data integration logic.

Invoking service-ready data integration tasks ensures that business processes such as quote generation, order entries, and procurement requests receive data that is correctly transformed, standardized, and matched across applications, partners, suppliers, and customers.

After an integration service is enabled, any enterprise application, .Net or Java developer, Microsoft Office application, or integration software can invoke the service by using a binding protocol such as web services, Representational State Transfer (REST), and Really Simple Syndication (RSS).

**Scenarios for improving efficiency**

These scenarios show how organizations used InfoSphere Information Services Director to improve efficiency and validate information in real time.

**Pharmaceutical: Improving efficiency**

A leading pharmaceutical company needed to include real-time data from clinical labs in its research and development reports. The company used IBM InfoSphere DataStage to define a transformation process for XML documents from labs. This process used SOA to expose the transformation as a web service, allowing labs to send data and receive an immediate response. Pre-clinical data is now available to scientific personnel earlier, allowing lab scientists to select which data to analyze. Only the best data is chosen, greatly improving efficiency and accuracy in clinical research.

**Insurance: Validating addresses in real time**

An international insurance data services company employed InfoSphere Information Server to validate and enrich property addresses by using web services. As insurance companies submit lists of addresses for underwriting, services standardize the addresses based on their rules, validate each address, match the addresses to a list of known addresses, and enrich the addresses with additional information that helps with underwriting decisions.

The project was simplified by using the standardization and matching capabilities of IBM InfoSphere QualityStage. The company then used InfoSphere Information Services Director to publish the enriched property data to their internal website. The property data is always current and available because changes are automatically applied. The company now automates 80 percent of the process and eliminated most errors.

**InfoSphere Information Services Director tasks**

By enabling integration tasks as services, InfoSphere Information Services Director is a critical component of your application development and integration environment.

Your organization can use InfoSphere Information Services Director to complete the following tasks:

**Enable real-time data warehousing**

You can complete analytical processing and loading of data based on transaction triggers, ensuring that time-sensitive data in the warehouse is
always current. Users can publish their existing data integration logic as services that can be called in real time from any process.

**Deploy matching services**

Matching services allow reference data such as customer, inventory, and product data to be matched to and kept current with a master store with each transaction. Users can package data integration logic as a shared service that can be called by enterprise application integration platforms.

**Enable in-flight transformation**

Enrichment logic can be packaged as shared services so that capabilities such as product name standardization, address validation, or data format transformations can be shared and reused across projects.

**Reuse enterprise data services**

InfoSphere Information Services Director enables the data access functions of many applications to be aggregated and shared in a tier of common services. Instead of each application creating its own access code, these services can be reused across projects, simplifying development and ensuring a higher level of consistency.

One of the major advantages of this approach is that you can combine data integration tasks with the leading enterprise messaging, enterprise application integration (EAI), and business process management (BPM) products by choosing the protocol binding that you want to use to deploy information services.

**Where InfoSphere Information Services Director fits in the suite architecture**

InfoSphere Information Services Director relies on other components of InfoSphere Information Server for understanding, cleansing, and transforming information and deploys those integration tasks as consistent and reusable information services.

SOA architects can search and browse services by category and view descriptions that are defined by InfoSphere Information Services Director. The directory of available services is linked to the shared metadata repository to enable centralized management of data integration services. If WebSphere Service Registry and Repository is installed, InfoSphere Information Services Director also provides a direct link to the WebSphere Service Registry and Repository.

Each service is published by defining operations completed by an operations provider. InfoSphere Information Services Director agents contain handlers to process service requests from InfoSphere DataStage, InfoSphere QualityStage, IBM DB2, InfoSphere Federation Server, Oracle, and IBM InfoSphere Master Data Management Server.

Service consumers can access information services by using multiple technologies for program interoperability. InfoSphere Information Services Director uses the same service to support multiple protocol bindings, all defined within the WSDL file. This implementation improves the utility of services and increases the likelihood of reuse and adoption across the enterprise. InfoSphere Information Services Director provides the ability to publish the same service by using SOAP over HTTP (web services), SOAP and text over JMS, EJB, REST, REST 2.0, and RSS.
Additional resources for InfoSphere Information Services Director

For more information about IBM InfoSphere Information Services Director and IBM InfoSphere Information Server components, see the following resources:

**IBM InfoSphere Information Server information center**
http://publib.boulder.ibm.com/infocenter/iisinfsv/v9r1/index.jsp

**DB2 as an Information Provider**

**Detailed tutorials using IBM InfoSphere Information Services Director and Information Consumers that can access InfoSphere Information Services Director**

**Using the IBM Information Service Call Builder for WebSphere Portlet Factory**

**InfoSphere Metadata Workbench**

InfoSphere Metadata Workbench provides end-to-end data flow reporting and impact analysis of the information assets that are used by InfoSphere Information Server components.

You use InfoSphere Metadata Workbench to understand information by showing the complete lineage, indicating where information comes from, and how that information changes as it moves through data integration processes. By tracing and maintaining relationships of information throughout your data integration project, InfoSphere Metadata Workbench improves transparency of IT to the business, and increases IT responsiveness.

Using the web-based navigation, users can view and interact with the metadata and information assets in the InfoSphere Information Server metadata repository. For example, users can import a business intelligence report and use the data flow
reporting capabilities of InfoSphere Metadata Workbench to understand how data is moved and transformed by IBM InfoSphere DataStage.

Understanding what, where, and who is impacted by changes to data helps to reduce operational costs for constantly changing data.

**Scenario for managing metadata**
This scenario shows how one organization used InfoSphere Metadata Workbench as part of their reporting solution.

**Financial services: Regulatory compliance reporting**

A major financial institution needed to produce accurate, up-to-date Basel II reports that would meet international regulatory standards. To meet this regulatory imperative, the company had to understand precisely where information concerning market, operational, and credit risk was stored.

Glossary terms with precise definitions were created in InfoSphere Business Glossary and assigned to physical data sources where the data was stored. InfoSphere Discovery was used to find existing data sources automatically, which eliminated much manual work.

InfoSphere FastTrack was used to document source-to-target mappings, which contribute to the full Basel II data flow. InfoSphere Information Analyzer was used to ensure the quality of the data flowing into the BI regulatory reports. InfoSphere Metadata Workbench was used to demonstrate the flow of information from legacy data sources, through the warehouse and into the BI reports.

The integrated metadata management of InfoSphere Information Server enabled the company to produce timely, accurate compliance reports, understand the location of critical data, and monitor data quality.

**InfoSphere Metadata Workbench tasks**
You can use InfoSphere Metadata Workbench to understand and manage the flow of data in your data integration project by discovering and analyzing relationships between information assets.

Your organization can use InfoSphere Metadata Workbench to complete the following tasks:

**Explore information assets**
Exploring and maintaining the relationships of information in your data integration project improves transparency of IT to your business and increases IT responsiveness.

You can create queries to discover specific assets that are used by different InfoSphere Information Server components. Queries can be published and made available to other users in your organization.

**Analyze dependencies and relationships**
Analyzing dependencies and relationships between assets helps users understand where data originates, how it relates to other data, and what types of processing the data undergoes throughout the project lifecycle.

You can create data lineage reports, business lineage reports, and impact analysis reports to visualize relationships. Data lineage reports help you understand where data comes from and where it goes. Business lineage reports show less detailed reports, excluding detailed information that
business users do not need. Impact analysis reports help you to understand the dependencies and the effects of changes between assets.

Manage metadata
You can create and edit descriptions for assets in the InfoSphere Information Server metadata repository. These changes proliferate through the metadata repository so that other suite users have access to the most current metadata.

You can also extend data lineage to external processes that do not write to disk, or ETL tools, scripts, and other programs that do not save their metadata in the metadata repository. You can create and import extended assets, and then use these assets in extension mapping documents to track the flow of information to and from the extended data sources and other assets.

Where InfoSphere Metadata Workbench fits in the suite architecture
InfoSphere Metadata Workbench provides a visual, web-based exploration of metadata that is generated, used, and imported by InfoSphere Information Server. Users across your organization can use InfoSphere Metadata Workbench to interact with information assets and metadata in the metadata repository.

Regardless of role, InfoSphere Metadata Workbench users can view the entire landscape of data integration processes, with visibility into data transformations that operate inside and outside of InfoSphere Information Server.

• Data integration specialists can explore InfoSphere DataStage and QualityStage jobs to understand the designed flow of information between data sources and targets, including the operational metadata resulting from job runs.

• Data quality specialists can view InfoSphere Information Analyzer summaries and run dependency analysis reports to show which assets or resources depend on one another.

• Data analysts can view InfoSphere FastTrack specifications to view and understand relationships between information assets, such as the mappings that exist between source data, metadata, and other information assets.

• Business users can view a simplified report of a data flow that contains only those assets and details that are pertinent to their job. Exploring business intelligence reports enables business users to understand how information assets are connected, which enables these users to better align with IT.
**InfoSphere QualityStage**

InfoSphere QualityStage provides capabilities to create and maintain an accurate view of data entities such as customer, location, vendors, and products throughout your enterprise.

InfoSphere QualityStage uses predefined, customizable rules to prepare complex information about your business entities for transactional, operational, and analytic applications in batch, real time, or as a web service. Information is extracted from the source system, measured, cleansed, enriched, consolidated, and loaded into the target system.

By completing analysis at the character or word level, InfoSphere QualityStage helps uncover data anomalies and inconsistencies before transformational processing occurs. Data from disparate sources is automatically standardized into fixed fields, such as given name, date of birth, gender, and phone number. Data quality rules then assign the correct semantic meaning to the input data to facilitate matching.

By ensuring data quality, InfoSphere QualityStage reduces the time and cost to implement master data management (MDM), business intelligence, enterprise resource planning (ERP), and other strategic customer-related IT initiatives.

**Scenarios for data cleansing**

These scenarios show how several organizations used InfoSphere QualityStage to cleanse their data.

**Banking: Consolidating information**

To facilitate marketing and mail campaigns, a large retail bank needed a single dynamic view of its customer households from 60 million records in 50 source systems. This process of grouping customer account records from multiple sources, known as householding, helps the bank to understand each customer by consolidating their information into a single record.
The bank used InfoSphere QualityStage to automate the process. Consolidated views are matched for all 50 sources, yielding information for all marketing campaigns. The result is reduced costs and improved return on marketing investments. Householding is now a standard process at the bank, which has a better understanding of its customers and more effective customer relationship management.

Pharmaceutical: Operations information

A large pharmaceutical company needed a data warehouse for marketing and sales information. The company had diverse legacy data with different standards and formats. Information was buried in free-form fields. The company also struggled with incorrect data values, discrepancies between field metadata and actual data in the field, and duplicate data. Obtaining a complete, consolidated view of an entity such as total quarterly sales from the prescriptions of one doctor was nearly impossible. Reports were difficult and time-consuming to compile, and the accuracy of the reports was suspect.

The company chose InfoSphere QualityStage to investigate fragmented legacy data at the level of each data value. Analysts can now access complete and accurate online views of doctors, the prescriptions that they write, and their managed-care affiliations for better decision support, trend analysis, and targeted marketing.

Insurance: Real-time data quality checks

A leading insurance company lacked a unique ID for each subscriber, many of whom participated in multiple health, dental, or benefit plans. Subscribers who visited customer portals on the web could not get complete information about their account status, eligible services, and other details.

Using InfoSphere QualityStage, the company implemented a real-time, in-process data quality check of all portal inquiries. InfoSphere QualityStage and WebSphere MQ transactions were combined to retrieve customer data from multiple sources and return integrated customer views. The new process provides more than 25 million subscribers with a real-time, complete view of their insurance services. A unique customer ID for each subscriber also helps the insurer move toward a single customer database for improved customer service and marketing.

InfoSphere QualityStage tasks

You can use InfoSphere QualityStage to establish a clear understanding of your data and to improve data quality.

Your organization can use InfoSphere QualityStage to complete the following tasks:

Data investigation

You use InfoSphere QualityStage to understand the nature and extent of data anomalies and enable more effective data cleansing and matching. Investigation capabilities give your organization complete visibility into the condition of data at any moment. Data problems in legacy sources can be identified and corrected before they corrupt new systems.

Investigation uncovers potential anomalies, metadata discrepancies, and undocumented business practices. Invalid values and default values are identified so that they can be corrected or added to fields that are proposed as matching criteria.

Data standardization

Creating a standardized view of your data enables your organization to
maintain accurate views of key entities such as customer, partner, or product. Data from multiple systems is reformatted to ensure that data has the correct, specified content and format. Standardization rules are used to create a consistent representation of the data.

With data standardization, IBM InfoSphere QualityStage Standardization Rules Designer provides capabilities to enhance standardization rule sets. You can add and modify classifications, lookup tables, and rules. You can also enhance information by completing global address cleansing, validation and certification, and geolocation, which is used for spatial information management. Longitude and latitude are added to location data to improve location-based services.

**Data matching**

The matching process ensures that the information that runs your enterprise is based on your business results, reflect the facts in the real world, and provide an accurate view of data across your enterprise.

Powerful matching capabilities detect duplicates and relationships, even in the absence of unique identifiers or other data values. A statistical matching engine assesses the probability that two or more sets of data values refer to the same business entity. After a match is confirmed, InfoSphere QualityStage constructs linking keys so that users can complete a transaction or load a target system with quality, accurate data.

**Data survivorship**

Survivorship ensures that you are building the best available view of related information. Business and mapping rules are implemented to create the necessary output structures for the target application. Fields that do not conform to load standards are identified and filtered so that only the best representation of the match data is loaded into the master data record.

Missing values in one record are supplied with values from other records of the same entity. Missing values can also be populated with values from corresponding records that have been identified as a group in the matching stage.

**Where InfoSphere QualityStage fits in the suite architecture**

InfoSphere QualityStage is a data reengineering environment that is designed to help cleanse and enrich data to meet business objectives and data quality management standards.

Data integration specialists use InfoSphere QualityStage to complete the preparation stage of enterprise data integration, often known as data cleansing. InfoSphere QualityStage and InfoSphere DataStage share an infrastructure for importing and exporting data, designing, deploying, and running jobs, and reporting.

InfoSphere QualityStage uses the source systems analysis from InfoSphere Information Analyzer, and supports the transformation functions of InfoSphere DataStage. InfoSphere QualityStage also incorporates data rules from InfoSphere Information Analyzer directly for in-stream or real-time data validation. Data analysts can then use InfoSphere Information Analyzer to assess the ongoing results and trends from InfoSphere QualityStage processing.

Data integration specialists can use InfoSphere QualityStage and InfoSphere DataStage to access the shared metadata repository to obtain access to current metadata and enterprise data for the integration project. Data that is imported by InfoSphere Metadata Asset Manager is accessible from the InfoSphere DataStage
and QualityStage Designer. Data integration specialists use the same design canvas to specify the flow of data from preparation to transformation and delivery.

Working together, these products automate data quality assurance, which was previously a manual or neglected activity within many data integration projects. The combined benefits help your organization avoid low return on investment (ROI) caused by working with poor quality data.

---

### Additional components in the InfoSphere Information Server portfolio

Taking an active role in defining your data integration strategy is crucial to achieving your business goals. Whether your goals are data quality, data integration, connecting business with IT, or some combination thereof, InfoSphere Information Server is scalable to meet your unique needs.

InfoSphere Information Server includes additional components that extend capabilities to incorporate mainframe data and enterprise metadata as part of your data integration project.

#### Mainframe integration

InfoSphere Information Server extends its capabilities to the mainframe with InfoSphere Information Server for System z® and InfoSphere Classic. These solutions enable your organization to deliver trusted information for key business decisions and provide the most current information to people, processes, and applications.

#### InfoSphere Information Server for System z

InfoSphere Information Server software extends scalability, information consistency, and performance to IBM System z, providing new flexibility to integrate information directly on the mainframe.

InfoSphere Information Server for System z provides a unified foundation for enterprise information architectures, delivering trusted information for key business initiatives such as business intelligence, real-time reporting, data migration, data warehousing, data governance, customer relationship management,
and service-oriented architecture (SOA) projects. The Linux software for System z also provides rich security features, stability, flexibility, interoperability, and reduced software costs.

InfoSphere Information Server for System z is a fully integrated software platform that profiles, cleanses, transforms, and delivers information from both mainframe and distributed data sources to drive greater insight for your business without added IBM z/OS® operational costs. Your organization can derive more value from the complex, heterogeneous information spread across systems.

With breakthrough productivity and performance for cleansing, transforming, and moving this information consistently and securely throughout your enterprise, InfoSphere Information Server for System z helps you access and use information in new ways to drive innovation, increase operational efficiency, and lower risk. InfoSphere Information Server for System z uniquely balances the reliability, scalability, and security of the System z platform with the low-cost processing environment of the Integrated Facility for the Linux specialty engine.

In a System z environment, all of the fundamental information components of the solution are hosted on the System z server, including the operational data store, the data warehouse, and any data marts utilized in the system. This model is well suited to address strategic business needs, centered around the demand for real-time or near real-time access to information to support key decision-making or business process requirements.

**InfoSphere Classic components**

InfoSphere Classic components support information management initiatives that require accurate, trusted mainframe data integration.

InfoSphere Classic components provide the latest information to the people, processes, and applications that need it. Use your mainframe data as you always have, while sharing it with the rest of your information infrastructure.

With the immediate integration of mainframe data that InfoSphere Classic components provide, you can meet demanding service level requirements, integrate your business and technical metadata, and standardize processes and tooling throughout your organization. In a service-oriented architecture (SOA) environment, you can package mainframe data integration as services and then reuse the services throughout your IT infrastructure.

Your organization can use InfoSphere Classic products independently, or with other components of InfoSphere Information Server, such as InfoSphere DataStage, InfoSphere QualityStage, InfoSphere Information Analyzer, and InfoSphere Information Services Director. These tools enable you to analyze, cleanse, and reuse your z/OS data, as well as package z/OS data processing as services across your enterprise.

InfoSphere Information Server tools dynamically import the logical table definitions, column definitions, and view definitions on your InfoSphere Classic data server into their own metadata stores or repositories. After the metadata is available, the tools access your z/OS data by using the built-in ability to access relational databases. This automated metadata and operational integration ensures reusability while maintaining the independence and integrity of each solution.

Each InfoSphere Classic component provides capabilities that integrate your data while capitalizing on your investments in mainframe technology.
InfoSphere Classic Data Event Publisher for z/OS
Captures changes to mainframe data as they occur and publishes the changes to applications and files. You can link changed-data events with business processes and drive change-only updating of data warehouses, data marts, and operational data stores.

InfoSphere Classic Federation Server for z/OS
Provides standards-based relational access to non-relational mainframe data by using applications and tools that use SQL. You can return SQL result sets to federated queries or update the data at the source.

InfoSphere Classic Replication Server for z/OS
Maintains a copy of your mainframe data by using a subscription paradigm that pulls data changes to the copy as they occur. You can replicate mainframe data to local and remote relational databases as necessary.

InfoSphere Data Integration Classic Connector for z/OS
Monitors source databases or file systems for changes to data that you want to publish or replicate. You can also publish change messages to files on your z/OS server for later retrieval and processing by applications.

Companion components for InfoSphere Information Server
Companion components for InfoSphere Information Server provide extended connectivity for enterprise applications, change data capture, and high-speed, event-based replication and publishing from databases.

InfoSphere Information Server Packs provide connectivity to widely used enterprise applications such as SAP and Oracle. These prebuilt packages enable your organization to integrate data from existing enterprise applications into new business systems.

InfoSphere Federation Server helps your organization to virtualize data and provide information in a form that applications and users need, while hiding the complexity of the underlying sources. Data virtualization allows information to be accessed through a common interface that centralizes the control of data access.

InfoSphere Data Replication integrates information across heterogeneous data stores in real time to support greater productivity, improve data security and availability, and integrate systems more efficiently.

InfoSphere Information Server Packs
InfoSphere Information Server Packs enable your organization to use InfoSphere QualityStage, InfoSphere DataStage, and SOA-based capabilities to create a complete data integration solution.

The InfoSphere Information Server Packs enable enterprise applications to benefit from multiple capabilities of InfoSphere Information Server, such as support for complex transformations, automated data profiling, data quality, and integrated metadata management. The InfoSphere Information Server Packs help your organization integrate data and create consistent, trustworthy information.

The following products provide InfoSphere DataStage connectivity for enterprise applications. This list is not exhaustive, but highlights the capabilities of commonly used InfoSphere Information Server Packs.
InfoSphere Information Server Pack for JD Edwards EnterpriseOne
Improve integration from EnterpriseOne applications by extracting application data and business views, including delta processing to capture incremental changes.

InfoSphere Information Server Pack for Oracle Applications
Extract data from the entire Oracle E-Business Suite of applications, including Oracle Financials, Manufacturing, and CRM to simplify integration of Oracle Applications data in target environments.

InfoSphere Information Server Pack for Salesforce.com
Increases visibility into lines of business by extracting, loading, cleansing, and synchronizing data hosted in Salesforce.com applications with other business data stored in enterprise resource planning (ERP) applications, data warehouses, and analytic applications.

InfoSphere Information Server Pack for SAP Applications
Optimize time to value and lower total cost of ownership by extracting and loading data from SAP Applications, SAP ERP, and all SAP Business Suite application modules.

InfoSphere Information Server Pack for SAP BW
Enhance the openness of SAP NetWeaver Business Intelligence by integrating non-SAP data into SAP Business Information Warehouse.

InfoSphere Information Server Pack for Siebel
Rapidly extract data from Siebel for use in other target environments. By using this pack, your organization can use customer relationship management (CRM) information throughout your enterprise.

Scenarios for integrating enterprise data:

The following scenarios demonstrate how two organizations used InfoSphere Information Server Packs to integrate their enterprise data with InfoSphere Information Server.

Life science: Integrating around SAP BW

A global leader in life science laboratory distribution implemented SAP BW for sales, supply chain, and contact center metrics, but still has a huge amount of data on non-SAP systems in areas of enterprise resource planning, financials, and custom applications.

The IT department needs to support the business by delivering key sales and revenue status reports and an analytical workspace to corporate and field staff in a timely way, meeting scalability requirements, easily managing metadata, and moving data from sources to targets.

The company uses the InfoSphere Information Server Pack for Oracle Applications to access financial and accounts receivable data, InfoSphere DataStage to transform it, and the InfoSphere Information Server Pack for SAP BW to load the transformed data into SAP BW.

The company now quickly assembles data sources, performs data transformations, and enforces referential integrity before loading data into SAP BW. Data is ready faster and the process is easy to use. The business users can react more quickly to changes in their marketplace.
Inventory projections: Incorporating SAP BW

One of the world’s leading producers and marketers of lawn care and gardening products needed to minimize inventory costs at its 22 distribution hubs. But without knowing their customers’ forecasted demand, the company was forced to carry excess inventory to protect against running out of stock and creating customer dissatisfaction because of lost sales.

Managers realized that if they could collect retail customer point-of-sale and sales forecast data from outside of their SAP applications, compare it to their internal SAP data, and properly load it into SAP BW, they could get a customized view of the data and properly plan shipments and inventory to meet demand.

The company implemented InfoSphere DataStage, InfoSphere Information Analyzer, and InfoSphere Information Server Packs for SAP Applications and SAP BW to collect sales data from customers and then cleanse, match, and load that data into SAP BW.

The resulting information helped the company lower inventory by 30 percent, which amounted to $99.4 million in one year. This savings contributed to a $158 million increase in free cash flow. The project also significantly reduced distribution and holding costs.

Where InfoSphere Information Server Packs fit in the suite architecture:

InfoSphere Information Server Packs provide high-speed connectivity to packaged enterprise applications that use the metadata capabilities of InfoSphere Information Server to help your organization integrate data and create consistent, trustworthy information.

To provide a complete data integration solution, InfoSphere Information Server Packs provide the following capabilities:

- Manage connections to application source systems
- Import metadata from source systems
- Integrate design and job control in InfoSphere DataStage
- Use InfoSphere DataStage to load data to target applications, including other enterprise applications and data warehouses or data marts
- Allow bulk extract, load, and delta processing

Figure 2 on page 46 shows how InfoSphere Information Server Packs fit within the InfoSphere Information Server architecture.
InfoSphere Federation Server

InfoSphere Federation Server enables your organization to access and integrate diverse data and content as if it were a single resource.

Federation, also known as enterprise information integration, provides an optimized and transparent data access and transformation layer with a single relational interface across enterprise data.

Applications that work with a federated server can interact with a single virtual data source, greatly reducing development time and maintenance costs. Your organization can develop a single, integrated view of diverse sources and reuse the view in multiple places to create a single point of change.

InfoSphere Federation Server uses the metadata of source systems to automate the building and compiling of federated queries. Metadata enables tracing and auditing throughout the federation process. Federated queries can scale to run against any volume of information by using the parallel processing engine of InfoSphere Information Server.

By using a federated system, you can send distributed requests to multiple data sources within a single SQL statement. For example, you can join data that is in an IBM DB2 table, an Oracle table, a web service, and an XML file in a single SQL statement.

Scenarios for data integration:

The following scenarios show how different organizations used InfoSphere Federation Server to solve their data integration needs.
Financial services: Risk management

A major European bank wanted to improve risk management across its member institutions and meet deadlines for Basel II compliance. The bank had different methods of measuring risk among its members. Without a consolidated view of risk management, the bank had to generate individual reports for each vendor.

The bank used InfoSphere Federation Server as part of their solution to enable compliance with Basel II by implementing a single mechanism to measure risk. A database management system stores a historical view of data, handles large volumes of information, and distributes data in a format that enables analysis and reporting. The bank can view data in operational systems that are spread across the enterprise, including vendor information.

Product development: Defect tracking

A major automobile manufacturer needed to quickly identify and fix defects in several lines of its cars. Traditional methods, such as data queries or reporting, were too complex and too slow to identify the sources of problems.

By installing InfoSphere Federation Server, the company was able to identify and fix defects by mining data from multiple databases that store warranty information and correlating warranty reports with individual components or software in its vehicles.

Government: Emergency response

A small government needed to eliminate storage of redundant contact information and simplify maintenance. The department had limited resource for any improvements, and had only one DBA and one manager.

The government chose InfoSphere Federation Server to join employee contact information in a human resources database on Oracle with information about employee skills in an IBM DB2 database. The information was presented to emergency personnel through a web portal that is implemented with IBM WebSphere Application Server. By interfacing with existing SQL tools, the small staff was able to merge employee contact information across multiple databases.

InfoSphere Federation Server tasks:

InfoSphere Federation Server helps your organization to virtualize data and provide information in a form that applications and users need while hiding the complexity of the underlying sources. Data virtualization allows information to be accessed through a common interface that centralizes the control of data access.

Your organization uses InfoSphere Federation Server to complete the following tasks.

Combine data from disparate sources

Combining data into federated databases enables you to access and integrate diverse data and content as if they were a single resource.

Your organization can correlate data from local tables and remote data sources as if the data is stored locally in the federated database. Data can be inserted, deleted, or updated in multiple data sources simultaneously, regardless of where the data is located.
Provide unrestricted access to your data
By using virtual data integration, InfoSphere Federation Server enables your organization to access data anywhere in your enterprise, regardless of what format it is in or what vendor you use.

You can access data by using standard SQL and any tool that supports JDBC or ODBC, without creating new databases and without disruptive changes to existing databases.

Extend data visualization
By adding InfoSphere Information Services Director to your solution, you can provide deploy federation logic as reusable services for a service-oriented architecture (SOA).

You can also extend DB2 pureScale to become a federated database server. By using advanced query processing, a federated server can distribute the workload among itself and the data sources that it works with. The federated server determines which part of the workload is most effectively run on a specific server to speed performance.

Where InfoSphere Federation Server fits in the suite architecture:
InfoSphere Federation Server provides a single view of your enterprise by virtually consolidating enterprise data from multiple disparate sources. InfoSphere Federation Server provides access to data from diverse sources and integrates with other InfoSphere Information Server components to simplify data integration.

By using InfoSphere Federation Server, users can access, join, and manipulate data of different formats such as relational, mainframe, XML, flat files, and web services.

As middleware, InfoSphere Federation Server presents enterprise data to end users as if they were accessing a single source. Wrappers enable almost any type of data source to be consolidated without interruption to existing applications and systems. Users can create a single query to incorporate data from multiple sources, such as sales, customer information, and reports. Security policies can be implemented and used throughout the unified view of the enterprise to simplify access and restriction to data.

By incorporating InfoSphere Information Services Director, users can deploy federation logic as reusable services for a service-oriented architecture (SOA). This integration provides a single point of access for multiple data sources and accelerates application development.

Each service is published by defining operations completed by an operations provider. InfoSphere Information Services Director agents contain handlers to process service requests from InfoSphere DataStage, InfoSphere QualityStage, IBM DB2, InfoSphere Federation Server, Oracle, and IBM InfoSphere Master Data Management Server.

Data integration specialists can access InfoSphere Federation Server directly in an ETL job by using InfoSphere DataStage and QualityStage Designer to incorporate data federation.
InfoSphere Data Replication

InfoSphere Data Replication is a high-volume, low-latency data replication solution that uses WebSphere MQ message queues for high availability and disaster recovery, data synchronization, and data distribution.

Data is replicated from source production systems to secondary databases, making current data available for live reporting without impacting source systems. By distributing and consolidating data across geographies, InfoSphere Data Replication enables your organization to synchronize data globally to increase visibility and maintain transactional integrity.

InfoSphere Data Replication detects and responds to data changes in source systems, publishes changes to subscribed systems, and feeds changed data into other modules for event-based processing. Changed-data events are formatted into XML messages and published as WebSphere MQ messages that can be used by other applications to drive subsequent processing. InfoSphere Data Replication can feed changed data from any source relational database management system (RDMS) and information management system into IBM Netezza directly, or by using InfoSphere DataStage and QualityStage. These monitoring capabilities increase visibility into the health and performance of your replication environment.
Active-active continuous high availability ensures that your secondary database is always ready as a backup to maintain continuity. Your organization can migrate applications and databases between platforms without interruption to your business.

**Scenarios for data replication:**

The following scenarios show how different organizations used InfoSphere Data Replication to securely integrate information across their businesses.

**Retail: Enhancing the customer experience**

A leading retailer wanted to expand their product inventory from 10 million to 20 million items. The retailer needed to synch their mainframe inventory application with distributed applications to ensure that items were in stock. Without the ability to synchronize their inventory and continually detect and respond to data changes in source systems, items might be shown as in stock, but not be available when customers try to purchase them. The retailer might lose sales and customer loyalty by not maintaining the most current data across their enterprise.

The retailer used InfoSphere Data Replication for DB2 for z/OS to ensure that the most current data is always available throughout their enterprise. Changes to the mainframe inventory are automatically detected and formatted into XML messages that are published to the retail website. The quantity of items available is continually updated so that customers know exactly how many units of an item are in stock. The website is updated to reflect the current quantity, and the retailer can notify customers when the quantity of an item that they viewed on the website is running low. The retailer can also use InfoSphere Data Replication to detect when inventory is low so that more units can be purchased. This solution ensures that customer satisfaction and loyalty remain high, and that customers return to the retailer for additional purchases.

**Financial services: Supporting continuous operations**

A large, organized financial services company wanted to grow their business to support international operations. The existing trading services were not prepared for globalization, and the current database performance did not have the capacity to meet projected load requirements. The company needed a solution that supported high scalability and constant availability before they moving forward with their plans to expand.

The company included InfoSphere Data Replication for DB2 for z/OS as part of their solution to enable continuous, integrated management of their data across time zones, and from multiple data sources. The software delivers high availability, scalability, and performance with improved security that the company required to expand their operations. Because of the active-active high availability configuration, the company can complete daily maintenance operations without stopping applications or experiencing downtime. If the primary system is affected by an outage, a secondary system becomes active so that trading can continue without interruptions.

**InfoSphere Data Replication tasks:**

InfoSphere Data Replication includes real-time data replication capabilities to help your organization support high availability, database migration, application
consolidation, dynamic warehousing, master data management (MDM), service oriented architecture (SOA), business analytics, and data quality processes.

Your organization uses InfoSphere Data Replication to complete the following tasks:

**Integrate information in real time**
Real-time data integration enables your organization to sense and respond to relevant business data changes throughout your enterprise.

InfoSphere Data Replication provides real-time feeds of changed data for data warehouse or master data management (MDM) projects, enabling your organization to make operational and business decisions based on the latest information. As your organization changes, data and applications can be consolidated without interruption. Data can be routed to various message queues to be consumed by multiple applications, ensuring accurate and reliable data across your enterprise.

**Deliver data continuously**
Continuous delivery of data ensures that your critical business operations are always available and contain the most current information.

With InfoSphere Data Replication, you can synchronize data between two systems to provide continuous availability. If the primary system is impacted by a planned or unplanned outage, a secondary system is available so that your business continues to run without interruption.

**Publish changed data to multiple targets**
InfoSphere Data Replication captures changed-data events from database logs and publishes them as WebSphere MQ messages that can be used by other applications to drive subsequent processing.

Changes to source tables or events are captured and converted to messages in an Extensible Markup Language (XML) format. This process provides a push data integration model that is suited to data-driven enterprise application integration (EAI) scenarios and change-only updating for business intelligence and MDM. These changes can be published to databases, web applications, message queues, data warehouses, and other information brokers that are central to your business.

**Where InfoSphere Data Replication fits in the suite architecture:**

In its simplest form, InfoSphere Data Replication securely integrates information across heterogeneous data stores in real time. By integrating with other InfoSphere and IBM products, you can extend data replication throughout your enterprise.

Integration with InfoSphere DataStage provides real-time data feeds to ETL processes with transactional integrity and no staging required, and enables real-time validity checks of changed data against defined data rules during the transformation process. Real-time data can also be fed into IBM Netezza for scalable, high-volume replication of data.

Change data capture can be implemented with InfoSphere Information Server. Source metadata is integrated into InfoSphere Metadata Workbench, enabling functions like impact analysis and data lineage to extend from the initial log-based replication task throughout the rest of the information management infrastructure. Changes are instantly replicated in the shared metadata repository so that all InfoSphere Information Server components have access to the most current data.
When combined with InfoSphere Data Replication for DB2 for z/OS, your organization can extend data replication capabilities to the mainframe. This integration enables your organization to replicate heterogeneous data to and from DB2 for z/OS in a secure environment that is always available.

**IBM InfoSphere Information Server architecture and concepts**

IBM InfoSphere Information Server provides a unified architecture that works with all types of information integration. Common services, unified parallel processing, and unified metadata are at the core of the server architecture.

The architecture is service oriented, enabling IBM InfoSphere Information Server to work within evolving enterprise service-oriented architectures. A service-oriented architecture also connects the individual suite product modules of InfoSphere Information Server.

By eliminating duplication of functions, the architecture efficiently uses hardware resources and reduces the amount of development and administrative effort that are required to deploy an integration solution.

[Figure 3 on page 53](#) shows the InfoSphere Information Server architecture.
Unified parallel processing engine

Much of the work that InfoSphere Information Server does takes place within the parallel processing engine. The engine handles data processing needs as diverse as performing analysis of large databases for IBM InfoSphere Information Analyzer, data cleansing for IBM InfoSphere QualityStage, and complex transformations for IBM InfoSphere DataStage. This parallel processing engine is designed to deliver the following benefits:

- Parallelism and data pipelining to complete increasing volumes of work in decreasing time windows
- Scalability by adding hardware (for example, processors or nodes in a grid) with no changes to the data integration design
- Optimized database, file, and queue processing to handle large files that cannot fit in memory all at once or with large numbers of small files

Common connectivity

InfoSphere Information Server connects to information sources whether they are structured, unstructured, on the mainframe, or applications.
Metadata-driven connectivity is shared across the suite components, and connection objects are reusable across functions.

Connectors provide design-time importing of metadata, data browsing and sampling, run-time dynamic metadata access, error handling, and high functionality and high performance run-time data access. Prebuilt interfaces for packaged applications called packs provide adapters to SAP, Siebel, Oracle, and others, enabling integration with enterprise applications and associated reporting and analytical systems.

**Unified metadata**

InfoSphere Information Server is built on a unified metadata infrastructure that enables shared understanding between business and technical domains. This infrastructure reduces development time and provides a persistent record that can improve confidence in information. All functions of InfoSphere Information Server share the same metamodel, making it easier for different roles and functions to collaborate.

A common metadata repository provides persistent storage for all InfoSphere Information Server suite components. All of the products depend on the repository to navigate, query, and update metadata. The repository contains two kinds of metadata:

- **Dynamic**
  - Dynamic metadata includes design-time information.

- **Operational**
  - Operational metadata includes performance monitoring, audit and log data, and data profiling sample data.

Because the repository is shared by all suite components, profiling information that is created by InfoSphere Information Analyzer is instantly available to users of InfoSphere DataStage and InfoSphere QualityStage, for example.

The repository is a J2EE application that uses a standard relational database such as IBM DB2, Oracle, or SQL Server for persistence (DB2 is provided with InfoSphere Information Server). These databases provide backup, administration, scalability, parallel access, transactions, and concurrent access.

**Common services**

InfoSphere Information Server is built entirely on a set of shared services that centralize core tasks across the platform. These include administrative tasks such as security, user administration, logging, and reporting. Shared services allow these tasks to be managed and controlled in one place, regardless of which suite component is being used. The common services also include the metadata services, which provide standard service-oriented access and analysis of metadata across the platform. In addition, the common services tier manages how services are deployed from any of the product functions, allowing cleansing and transformation rules or federated queries to be published as shared services within an SOA, using a consistent and easy-to-use mechanism.

InfoSphere Information Server products can access three general categories of service:

- **Design**
  - Design services help developers create function-specific services that can also be shared. For example, InfoSphere Information Analyzer calls a column analyzer service that was created for
enterprise data analysis but can be integrated with other parts of InfoSphere Information Server because it exhibits common SOA characteristics.

**Execution**
Execution services include logging, scheduling, monitoring, reporting, security, and Web framework.

**Metadata**
Metadata services enable metadata to be shared across tools so that changes made in one InfoSphere Information Server component are instantly visible across all of the suite components. Metadata services are integrated with the metadata repository. Metadata services also enable you to exchange metadata with external tools.

The common services tier is deployed on J2EE-compliant application servers such as IBM WebSphere Application Server, which is included with InfoSphere Information Server.

**Unified user interface**
The face of InfoSphere Information Server is a common graphical interface and tool framework. Shared interfaces such as the IBM InfoSphere Information Server console and the IBM InfoSphere Information Server Web console provide a common interface, visual controls, and user experience across products. Common functions such as catalog browsing, metadata import, query, and data browsing all expose underlying common services in a uniform way. InfoSphere Information Server provides rich client interfaces for highly detailed development work and thin clients that run in Web browsers for administration.

Application programming interfaces (APIs) support a variety of interface styles that include standard request-reply, service-oriented, event-driven, and scheduled task invocation.

**Tiers and components**

**Parallel processing in InfoSphere Information Server**
Companies today must manage, store, and sort through rapidly expanding volumes of data and deliver it to end users as quickly as possible.

To address these challenges, organizations need a scalable data integration architecture that contains the following components:

- A method for processing data without writing to disk, in batch and real time.
- Dynamic data partitioning and in-flight repartitioning.
- Scalable hardware that supports symmetric multiprocessing (SMP), clustering, grid, and massively parallel processing (MPP) platforms without requiring changes to the underlying integration process.
- Support for parallel databases including DB2, Oracle, and Teradata, in parallel and partitioned configurations.
- An extensible framework to incorporate in-house and vendor software.

IBM InfoSphere Information Server addresses all of these requirements by exploiting both pipeline parallelism and partition parallelism to achieve high throughput, performance, and scalability.
Parallelism basics in IBM InfoSphere Information Server

The pipeline parallelism and partition parallelism that are used in IBM InfoSphere Information Server support its high-performance, scalable architecture.

- **Data pipelining**
- **Data partitioning**
- **Dynamic repartitioning** on page 57

Data pipelining

*Data pipelining* is the process of pulling records from the source system and moving them through the sequence of processing functions that are defined in the data-flow (the job). Because records are flowing through the pipeline, they can be processed without writing the records to disk, as Figure 4 shows.

Data can be buffered in blocks so that each process is not slowed when other components are running. This approach avoids deadlocks and speeds performance by allowing both upstream and downstream processes to run concurrently.

Without data pipelining, the following issues arise:

- Data must be written to disk between processes, degrading performance and increasing storage requirements and the need for disk management.
- The developer must manage the I/O processing between components.
- The process becomes impractical for large data volumes.
- The application will be slower, as disk use, management, and design complexities increase.
- Each process must complete before downstream processes can begin, which limits performance and full use of hardware resources.

Data partitioning

*Data partitioning* is an approach to parallelism that involves breaking the record set into partitions, or subsets of records. If no resource constraints or other data skew issues exist, *data partitioning* can provide linear increases in application performance. Figure 5 on page 57 shows data that is partitioned by customer surname before it flows into the Transformer stage.
A scalable architecture should support many types of data partitioning, including the following types:
- Hash key (data) values
- Range
- Round-robin
- Random
- Entire
- Modulus
- Database partitioning

InfoSphere Information Server automatically partitions data based on the type of partition that the stage requires. Typical packaged tools lack this capability and require developers to manually create data partitions, which results in costly and time-consuming rewriting of applications or the data partitions whenever the administrator wants to use more hardware capacity.

In a well-designed, scalable architecture, the developer does not need to be concerned about the number of partitions that will run, the ability to increase the number of partitions, or repartitioning data.

**Dynamic repartitioning**

In the examples shown in Figure 5 and Figure 6 on page 58, data is partitioned based on customer surname, and then the data partitioning is maintained throughout the flow.

This type of partitioning is impractical for many uses, such as a transformation that requires data partitioned on surname, but must then be loaded into the data warehouse by using the customer account number.

*Figure 5. Data partitioning*
Dynamic data repartitioning is a more efficient and accurate approach. With dynamic data repartitioning, data is repartitioned while it moves between processes without writing the data to disk, based on the downstream process that data partitioning feeds. The InfoSphere Information Server parallel engine manages the communication between processes for dynamic repartitioning.

Data is also pipelined to downstream processes when it is available, as Figure 7 shows.

Without partitioning and dynamic repartitioning, the developer must take these steps:
- Create separate flows for each data partition, based on the current hardware configuration.
- Write data to disk between processes.
- Manually repartition the data.
- Start the next process.

The application will be slower, disk use and management will increase, and the design will be much more complex. The dynamic repartitioning feature of InfoSphere Information Server helps you overcome these issues.

**Scalability of the InfoSphere Information Server engine**

IBM InfoSphere Information Server is built on a highly scalable software architecture that delivers high levels of throughput and performance.

For maximum scalability, data integration software must use all available system resources to accomplish data integration tasks. This capability must extend beyond
Symmetric Multiprocessing (SMP) systems to include both Massively Parallel Processing (MPP) systems and grid systems.

InfoSphere Information Server components use grid, SMP, and MPP environments to optimize the use of all available hardware resources.

For example, when you use the IBM InfoSphere DataStage and QualityStage Designer to create a data-flow graph, the underlying hardware architecture and number of processors is irrelevant. A separate configuration file defines the amount and location of parallel processing the job should run with. This configuration is bound to the job at run time and determines the resources required from the underlying computing system.

As Figure 8 shows, the configuration provides a clean separation between creating the data-flow graph and the parallel execution of the application. This separation simplifies the development of scalable data integration systems that run in parallel.

Application Assembly: One dataflow graph

![Application Assembly Diagram]

Application Execution: Sequential or parallel

![Application Execution Diagram]

Figure 8. Hardware complexity made simple

Without dynamic support for scalable hardware environments, the following problems can occur:

- Processing is slower, because hardware resources are not maximized.
- Application design and hardware configuration cannot be decoupled, and manual intervention and possibly redesign is required for every hardware change.
- Scaling on demand is not possible.

InfoSphere Information Server utilizes powerful parallel processing technology to ensure that large volumes of information can be processed quickly. This technology ensures that processing capacity does not inhibit project results and allows solutions to easily expand to new hardware and to fully utilize the processing power of all available hardware.
Support for grid computing in InfoSphere Information Server

With hardware computing power a commodity, grid computing is a highly compelling option for large enterprises. Grid computing allows you to apply more processing power to a task than was previously possible.

Grid computing uses low-cost computing resources, processors, and memory that are available on the network to create a shared pool of computing resources that can be applied to various tasks. Grid computing software provides a list of available computing resources and a list of tasks. When a computer becomes available, the grid software assigns new tasks according to appropriate rules.

A grid can be made up of thousands of computers. Grid-computing software balances IT supply and demand by letting users specify processor and memory requirements for their jobs, and then find available machines on a network to meet those specifications.

The parallel processing architecture of InfoSphere Information Server simplifies the development of scalable integration systems that run in parallel and use the inherent computing power and flexibility of grid environments.

InfoSphere Information Server’s pre-bundled grid edition provides rapid implementation of grid scalability. It includes an integrated grid scheduler and integrated grid optimization. These capabilities help you deploy integration logic across a grid without affecting job design, which provides unlimited scalability.

Shared services in InfoSphere Information Server

IBM InfoSphere Information Server provides extensive administrative and reporting facilities that use shared services and a Web application that offers a common interface for all administrative and reporting tasks.

Administrative services

IBM InfoSphere Information Server provides administrative services to help you manage users, roles, sessions, security, logs, and schedules. The Web console provides global administration capabilities that are based on a common framework.

The IBM InfoSphere Information Server console provides these services:

- "Security services"
- "Log services" on page 62
- "Scheduling services" on page 62

Security services

Security services support role-based authorization of users, access-control services, and encryption that complies with many privacy and security regulations. As Figure 9 on page 61 shows, the console helps administrators add users, groups, and roles and lets administrators browse, create, delete, and update operations within InfoSphere Information Server.

Directory services act as a central authority that can authenticate resources and manage identities and relationships among identities. You can base directories on the InfoSphere Information Server internal directory, on external directories that are based on LDAP and Microsoft Active Directory, or Microsoft Windows and UNIX local operating systems.
Users use only one credential to access all the components of InfoSphere Information Server.

![Figure 9. Adding a new user to a group](image)

**InfoSphere Information Server** creates an audit trail of security-related events. This includes all activities that set or modify security-related settings and all user authentications and application logins. You can configure which audit events to log and how much information to include based on your auditing requirements. Security auditing trails assist in the detection of access to controlled information and application usage. Monitoring and analysis of the logged audit information can lead to improvements in the control of data access and helps to prevent malicious or careless unauthorized access to sensitive data or configuration settings. The monitoring of application and individual user access, including system administration actions, provides an historic record of activity. This information allows you to adjust user or group security roles to enable or prevent access to application features. This information can also assist in showing compliance with corporate security policies.

Auditing services create an audit trail of security-related events by logging the execution of the following types of activities:

- Creation and removal of users and groups
- Assignment or removal of a user from a group
- User password changes
- Changes to security roles assigned to users or groups
- Changes to user and group permissions on a project and the associated project-level security roles that are assigned
- Changes to engine credential mapping
- User login
- User logout
- Session termination
- Session timeout
- Changes to audit logging configuration settings
The creation and removal of users and groups, assignment or removal of a user from a group, and user password changes can be logged only if the User Registry Configuration is set to **InfoSphere Information Server User Registry**. This registry is also known as the InfoSphere Information Server internal user registry.

You can configure the location, size, name, and number of audit log files, as well as the events to log.

**Log services**

Log services help you manage logs across all of the InfoSphere Information Server suite components. The Web console provides a central place to view logs and resolve problems. Logs are stored in the common repository, and each InfoSphere Information Server suite component defines relevant logging categories.

You can configure which categories of logging messages are saved in the repository. Log views are saved queries that an administrator can create to help with common tasks. For example, you might want to display all of the IBM InfoSphere Information Services Director error events that were logged in the last 24 hours.

Figure 10 shows where logging reports can be configured in the IBM InfoSphere Information Server Web console. Logging is organized by server components. The Web console displays default and active configurations for each component.

![Figure 10. Administrative console for setting up logs](image)

**Scheduling services**

Scheduling services help plan and track activities such as logging and reporting and suite component tasks such as data monitoring and trending. Schedules are
maintained by using the IBM InfoSphere Information Server console, which helps you define schedules; view their status, history, and forecast; and purge them from the system.

**Reporting services**
Reporting services manage run time and administrative aspects of reporting for IBM InfoSphere Information Server.

You can create product-specific reports for IBM InfoSphere DataStage, IBM InfoSphere QualityStage, and IBM InfoSphere Information Analyzer, and you can create cross-product reports for logging, monitoring, scheduling, and security services.

You can also access, delete, and purge report results contents from an associated scheduled report execution.

You can set up and run all reporting tasks from the IBM InfoSphere Information Server Web console. You can retrieve, view, and schedule reports to run at a specific time and frequency. You can tag reports as favorites and restrict their access for security purposes.

The following figure shows the IBM InfoSphere Information Server Web console.

![Figure 11. Creating a logging report by using the Web console](image)

You define reports by choosing from a set of templates and setting the parameters for that template. You can specify a history policy that determines how the report will be archived. Additionally, you can set a time frame for the report expiration, if needed. Reports can be formatted as DHTML, HTML, PDF, RTF, TXT, XLS, and XML.

**Scalability of the common services tier**
IBM InfoSphere Information Server is built on a highly scalable software architecture that delivers high levels of throughput and performance.
IBM InfoSphere Information Server services are hosted by IBM WebSphere Application Server Network Deployment (ND), which is a J2EE-compliant application server. You can implement WebSphere Application Server cluster topologies to maximize the number of users who can use the system concurrently.

The term *cluster* used within a WebSphere Application Server context refers to a set of application servers that are managed together and participate in workload management.

The following WebSphere Application Server cluster topologies are supported:

- Multiple application server instances within the same host computer (*vertical clustering*). Vertical clustering allows the system to maximize the resource allocation of cluster members, which improves performance.
- Members of the same cluster on multiple host computers (*horizontal clustering*). Horizontal clustering allows the InfoSphere Information Server services tier to run on several nodes, but still operate with clients as a single system instance. This configuration makes the most effective use of hardware resources. Also, if a node becomes unavailable, its workload can be routed to other nodes in the cluster, which improves availability.

To implement a WebSphere Application Server clustering topology, you must also deploy a Web server or a load balancer as a front end to the cluster. A front-end Web server dispatches Web requests to cluster members according to preconfigured policies. A load balancer balances Web requests across cluster members. Either solution increases performance and security because the Web server or load balancer guarantees a unique HTTP entry point to the cluster.

You can implement an IP sprayer topology that consists of load balancers that are upstream of front-end Web servers. This topology allows intelligent load balancing among Web servers based on server availability and workload capacity.

To increase security, you can set up firewalls and create a DMZ to protect each tier within your WebSphere Application Server topology. To increase availability, you can also set up backup Web servers and load balancers.

**Integrated metadata management in InfoSphere Information Server**

Tools for data profiling, data modeling, data transformation, data quality, and business intelligence play a key role in data integration. The integrated metadata management capabilities of IBM InfoSphere Information Server enable these tools to work together to meet your enterprise goals.

Metadata management in InfoSphere Information Server offers many advantages:

- Sharing metadata throughout the suite from a single metadata repository creates accurate, consistent, and efficient processes.
- Changes that you make to source systems can be quickly identified and propagated throughout the flow of information.
- You can identify downstream changes and use them to revise information in the source systems.
- You can track and analyze the data flow across departments and processes.
- Metadata is shared automatically among tools.
- Glossary definitions provide business context for metadata that is used in jobs and reports.
• Data stewards take responsibility for metadata assets such as schemas and tables that they have authority over.

• By using data lineage, you can focus on the end-to-end integration path, from the design tool to the business intelligence (BI) report. Or you can drill down to view any element of the lineage.

• You can eliminate duplicate or redundant metadata to create a single, reliable, version that can be used by multiple tools.

Managing metadata
The metadata repository of IBM InfoSphere Information Server stores metadata from suite tools and external tools and databases and enables sharing among them. You can import metadata into the repository from various sources, export metadata by various methods, and transfer metadata assets between design, test, and production repositories.

The metadata repository
The single metadata repository provides users of each suite tool with a common understanding of the structure of the data that flows through the tools of the InfoSphere Information Server suite. With a shared repository, changes that are made in one suite tool are automatically and instantly visible throughout the suite.

The single repository ensures that you can use a database table that is imported from a database or design tool in the following ways, among others:

• For analysis in IBM InfoSphere Information Analyzer
• To create mappings in IBM InfoSphere FastTrack
• To create table definitions in an IBM InfoSphere DataStage and QualityStage job

The same table can also be assigned a term and a steward in IBM InfoSphere Business Glossary. The table can also be part of a data lineage report in IBM InfoSphere Metadata Workbench that links it to the original database design, to the job that uses the table, and to the business intelligence (BI) report that is based on the table.

The metadata repository shares, stores, and reconciles a comprehensive spectrum of metadata:

Business metadata
Provides business context for information technology assets and adds business meaning to the artifacts that are created and managed by other IT applications. Business metadata includes glossary terms, stewardship, and examples.

Operational metadata
Describes the runs of IBM InfoSphere DataStage and QualityStage jobs, including rows written and read, and the database table or data files that are affected. You can use IBM InfoSphere Metadata Workbench to create data lineage reports that combine design and operational information.

Technical metadata
Provides details about the following types of assets:

• Implemented data resources, including host computers, databases and data files, and their contents. The assets can be imported from a design tool, a database, or a BI tool.
The metadata repository is an IBM WebSphere J2EE application. The repository uses standard relational database technology (such as IBM DB2 or Oracle) for persistence. These databases provide backup, administration, scalability, transactions, and concurrent access.

**Importing and exporting metadata**

InfoSphere Information Server offers many methods of importing metadata assets into the metadata repository. Some methods include the ability to export metadata from the repository to other tools, files, or databases. InfoSphere Metadata Asset Manager imports assets into the metadata repository by using bridges and connectors.

**Connectors, operators, and plug-ins**

InfoSphere DataStage and QualityStage use connectors, operators, and plug-ins to connect to various databases to extract, transform, and load data. InfoSphere Information Analyzer and InfoSphere FastTrack use connectors to access databases. In all cases, metadata about the implemented data resources, including host, database, schemas, tables, and columns, is stored in the metadata repository for use by other suite tools.

**InfoSphere Metadata Integration Bridges**

Bridges let you import metadata into the metadata repository from external applications, databases, and files, including design tools and BI tools. Some bridges can also export metadata. You can import many types of metadata, including the following:

- Hosts, databases, schemas, stored procedures, database tables, database columns, and foreign keys
- Data files, data file structures, data file fields
- BI reports, models, and their contained assets
- Analysis information from IBM InfoSphere Discovery
- Logical data models and physical data models from design tools such as CA ERwin and IBM InfoSphere Data Architect
- Users and groups to designate as stewards for assets in the metadata repository

**Exchange of XML and CSV files**

Several suite tools provide interfaces for import and export of XML and comma-separated values (CSV) files that contain metadata of different types:

- You can use InfoSphere Metadata Workbench to import extension mapping documents and extension data sources that capture information about processes and data sources from tools, scripts, and other programs that do not save their metadata to the metadata repository.
- You can use InfoSphere FastTrack to import and export mapping specifications in CSV format.
- You can use InfoSphere Business Glossary to import glossary content, including categories, terms, and relationships to other assets.
Browsing, analyzing, and deleting repository metadata

Users of each suite tool can browse and select the types of metadata assets that the tool uses. For example, users of InfoSphere DataStage and QualityStage can select jobs and the table definitions and stages that are used by jobs. Several tools provide a wider view of the contents of the metadata repository:

- Users of InfoSphere Metadata Workbench can browse and query the full spectrum of assets in the repository and run data lineage and impact analysis reports.
- Users of InfoSphere Business Glossary can find and browse assets of many types to assign terms to the assets or designate stewards or the assets.
- By using the repository management functionality of InfoSphere Metadata Asset Manager, you can browse all implemented data resources, logical data model assets, physical data model assets, and BI assets in the metadata repository. You can delete or merge duplicate assets.

Moving assets between metadata repositories

After you have developed and tested your jobs and processes, you can move them to a production environment. You can use the istool command line to move assets from one InfoSphere Information Server repository to another. For example you can move assets from a development environment to a test environment, and from a test environment to a production environment.

By using the command line, you can move multiple types of assets and the relationships between them:

- Jobs and projects from InfoSphere DataStage and QualityStage
- Categories, terms, and stewards from IBM InfoSphere Business Glossary
- Analysis summaries, projects, and metrics from InfoSphere Information Analyzer
- Mapping specifications from IBM InfoSphere FastTrack
- Implemented data resources, including metadata for databases, schemas, tables, columns, and data files.
- Logical data model assets and physical data model assets.
- BI metadata, including BI reports, BI models and their contained assets.
- InfoSphere Information Server users, roles, and reports

The following tools also have user interfaces for moving assets between metadata repositories:

- InfoSphere DataStage and QualityStage
- InfoSphere Data Architect
- InfoSphere Business Glossary
- InfoSphere Information Analyzer

Scenario for metadata management

The comprehensive metadata management capability provides users of InfoSphere Information Server with a common way to deal with descriptive information about the use of data. The following scenarios describe uses of this capability.

Business analytics

A large, for-profit education provider needed to devise a strategy for better student retention. Business managers needed to analyze the student life
cycle from application to graduation in order to direct their recruiting
efforts at students with the best chance of success.

To meet this business imperative, the company designed and delivered a
business intelligence solution using a data warehouse. The warehouse
contains a single view of student information that is populated from
operational systems.

The IT organization uses InfoSphere Information Server and its metadata
repository to coordinate metadata throughout the project. Other tools that
are used include Embarcadero ER/Studio for data modeling and IBM
Cognos for business intelligence. The reports that are produced show an
accurate view of student trends over the lifecycle from application to
graduation.

The consumers are able to understand the meaning of the fields in their BI
reports by accessing the business definitions in InfoSphere Business
Glossary. This enables them to identify key factors that correlate student
characteristics and retention. They are also able to understand the origin of
data in the reports by using business lineage, which enables them to trust
the sources and flow of the data that they are looking at. The net result is
the ability to make better decisions with more confidence, allowing the
education provider to design and implement effective initiatives to retain
students.

High availability in InfoSphere Information Server

InfoSphere Information Server includes features that increase the availability of the
system.

A highly available system maximizes the percentage of time that the system services
are operational from the user point of view. To heighten availability, you
implement topologies and technologies that introduce redundancy. The aim is to
reduce or eliminate the number of single points of failure (SPOF): elements that, if
they fail, cause critical aspects of the system to stop operating.

IBM provides different high-availability solutions for each InfoSphere Information
Server tier. This arrangement optimizes the best available options for each major
component in your installation. Each solution helps you to design different highly
available configurations, from relatively simple setups to complex installations.
Most solutions involve clustering of hardware and software components to provide
redundancy.

In general, the higher the level of overall availability that you want to achieve, the
more complex the system you must design and maintain. Also, more highly
available systems generally require more hardware. For these reasons, consider the
level of availability that you require within each software tier in the system. You
might want a different level of availability within your development system than
you have within your production system.

For more information about designing a topology for InfoSphere Information
Server, see the Designing a topology for InfoSphere Information Server white paper
(http://www.ibm.com/developerworks/data/bestpractices/infosphere/
istopology/index.html).

Metadata repository options to increase availability

To increase the high availability of the metadata repository or the IBM InfoSphere
Information Analyzer analysis database, implement database clustering technology.
To address availability, you can create a clustered configuration of the metadata repository database. You can also create a clustered configuration of the IBM InfoSphere Information Analyzer analysis database.

The term cluster used within this context refers to a group of independent physical computers or logical partitions (nodes) that are interconnected and work together as a single system. To add capacity, you can add nodes to the cluster without interrupting service.

It is possible to implement the database clusters in an enormous array of topologies. IBM InfoSphere Information Server supports a subset of these configurations. For information about implementing database clustering within your installation, refer to the IBM InfoSphere Information Server Planning, Installation, and Configuration Guide.

**Common services tier options to increase availability**

To increase the availability of the common services tier, you can implement IBM WebSphere Application Server clustering or an active-passive topology.

In a WebSphere Application Server cluster, you can implement horizontal clustering, where members of the cluster reside on multiple host computers. If the primary node fails, the workload is routed to other nodes in the cluster. You configure backup web servers or load balancers to take over if the primary server fails.

In an active-passive topology, you install the services tier on both an active server and a passive server and cluster the computers by using high availability software. If the active server fails, the high availability software restarts services on the passive server.

**InfoSphere Information Server engine options to increase availability**

To increase the availability of the IBM InfoSphere Information Server engine, create an active-passive configuration.

To set up a robust engine tier, create an active-passive solution by using high-availability (HA) software such as Tivoli System Administration for Multiplatforms. This solution consists of at least two servers that share a common storage resource such as a storage area network (SAN). At run time, only one of the servers is active, running engine processes. The HA software maintains a heartbeat signal between the servers that indicates that the active server is operational. If the heartbeat signal stops, the HA software causes one of the passive servers to take over. While the passive server is initializing, the engine is not available to the system.

With this configuration, a floating IP address and virtual host name are created for the system at installation time. This information is associated with the active server. All client programs connect to the system by using this address or host name. If the active server fails, the information is automatically reassociated with the passive server. Client programs do not need to connect to a different address.
Appendix A. 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](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.

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.
Appendix B. Reading command-line syntax

This documentation uses special characters to define the command-line syntax.

The following special characters define the command-line syntax:

[] Identifies an optional argument. Arguments that are not enclosed in brackets are required.

... Indicates that you can specify multiple values for the previous argument.

| Indicates mutually exclusive information. You can use the argument to the left of the separator or the argument to the right of the separator. You cannot use both arguments in a single use of the command.

{} Delimits a set of mutually exclusive arguments when one of the arguments is required. If the arguments are optional, they are enclosed in brackets ([ ]).

Note:
- The maximum number of characters in an argument is 256.
- Enclose argument values that have embedded spaces with either single or double quotation marks.

For example:

wsetsrc{-S server} [-l label] [-n name] source

The source argument is the only required argument for the wsetsrc command. The brackets around the other arguments indicate that these arguments are optional.

wlsac [-l | -f format] [key...] profile

In this example, the -l and -f format arguments are mutually exclusive and optional. The profile argument is required. The key argument is optional. The ellipsis (...) that follows the key argument indicates that you can specify multiple key names.

wrb -import {rule_pack | rule_set}...

In this example, the rule_pack and rule_set arguments are mutually exclusive, but one of the arguments must be specified. Also, the ellipsis marks (...) indicate that you can specify multiple rule packs or rule sets.
Appendix C. 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

<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/software/sw-training/">http://www.ibm.com/software/sw-training/</a></td>
</tr>
</tbody>
</table>
Appendix D. 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 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/v9r1/index.jsp](http://publib.boulder.ibm.com/infocenter/iisinfsv/v9r1/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](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](http://www.ibm.com/e-business/linkweb/publications/servlet/pbi.wss)

© Copyright IBM Corp. 2006, 2012
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 comment on the information center, click the Feedback link on the top right side of any topic in the information center.
- 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).
- You can provide general product feedback through the Consumability Survey at [www.ibm.com/software/data/info/consumability-survey](http://www.ibm.com/software/data/info/consumability-survey)
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 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, LACS\textsuperscript{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\textsuperscript{link} licensee of the United States Postal Service.

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