# Optimizing Data Integration Solutions by Customizing the IBM InfoSphere Information Server Deployment Architecture

**IBM Redbooks Solution Guide**

IBM® InfoSphere® Information Server provides a unified data integration platform so that you can understand, cleanse, transform, and deliver trustworthy and context-rich information to critical business initiatives. InfoSphere Information Server includes several product modules and components that enable an integrated end-to-end information integration solution, as shown in Figure 1. You can deploy the InfoSphere Information Server components in specific patterns to achieve enterprise business objectives in an optimal manner. These patterns vary depending on several factors, including which product modules will be used, the size of the user community, and the needs for high availability.
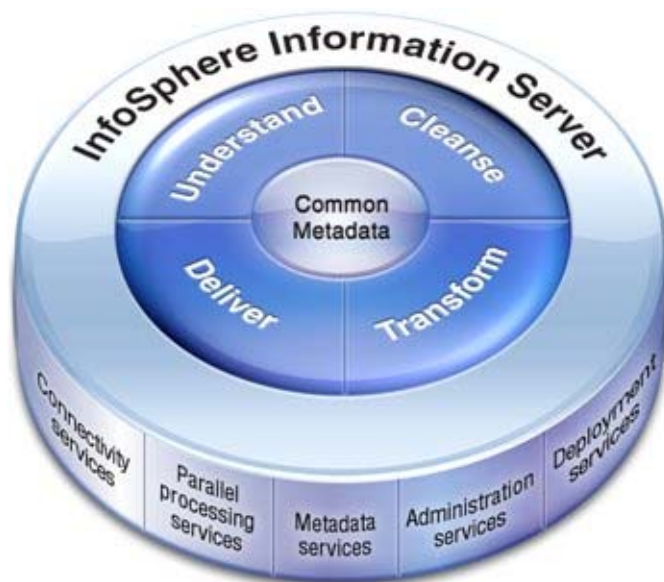


Figure 1. IBM InfoSphere Information Server: Solutions for optimized data integration

This solution also includes a facility for governance throughout the process. This key element of the InfoSphere Information Server platform is a rich, shared metadata layer that provides visibility into the business and the technical aspects of your information architecture.

## Did you know?

Enterprise data volumes are growing. In addition, the number of data sources, the number of consumers, and the number of integration points are growing, making information integration and governance more important than ever before. Integrating your data assets and managing them proactively with a strong data governance program are key enablers for achieving your business goals and for gaining a significant competitive advantage to keep your business growing and profitable.

## Business value

Thousands of organizations are employing the powerful components of the IBM InfoSphere Information Server. As a result, they have access to the current and accurate information they need for key business decisions that are delivered through this high performance data integration platform. In addition, InfoSphere Information Server generates key metadata as a by-product of development activities to populate the shared metadata repository and to enable organizations to maximize their reporting capabilities.

IBM InfoSphere Information Server provides a single unified platform that companies use to understand, cleanse, transform, and integrate their data sources to deliver trustworthy and context-rich information. It is built on a multitiered platform that includes a suite of product modules and components that focus on information integration. Furthermore, it can integrate with other third-party applications to enable the use of information, wherever it exists in the enterprise.

As examples, the following component modules can be part of an InfoSphere Information Server solution:

- IBM InfoSphere Blueprint Director
- IBM InfoSphere Business Glossary
- IBM InfoSphere DataStage®
- IBM InfoSphere FastTrack
- IBM InfoSphere Information Analyzer
- IBM InfoSphere Metadata Workbench
- IBM InfoSphere QualityStage®

In addition, a solution can include the following integrally related applications:

- InfoSphere Discovery
- InfoSphere Data Architect
- InfoSphere Data Replication

InfoSphere Information Server delivers trustworthy information and supports information governance objectives with advanced capabilities built on a highly scalable, security-rich, and robust data integration platform that is designed to simplify your data integration needs. It addresses the demanding information needs of organizations and improves the operational management and governability of integration projects and information infrastructure.

InfoSphere Information Server can provide the following benefits:

- Fully integrated platform (common metadata foundation)
- Proven performance and scalability
- Enablement of business and IT collaboration
- On-time delivery by using a reusable approach to delivery
- Implementation of proven practices across shared teams and workstreams
- Collaboration on the design before action is taken

InfoSphere Information Server supports a wide range of initiatives, including business intelligence, master data management, infrastructure rationalization, business transformation, and risk and compliance (governance).

## Solution overview

For successful deployment of InfoSphere Information Server components in typical logical infrastructure topologies that use the shared metadata capabilities of the platform, you must follow the guidelines and criteria. The guidelines support the requirements for development lifecycle, data privacy, information security, high availability, and performance. The objective is to help you evaluate your own information requirements to determine an appropriate deployment architecture. It can help you fulfill specific use cases and effectively use the functionality of the InfoSphere Information Server product modules and components.

InfoSphere Information Server consists of a robust, scalable, server architecture that is built on three distinct components or functional tiers:

●     Services tier. A Java Platform, Enterprise Edition (Java EE) application server.
●     Repository tier. A metadata database, known as the *Information Server repository tier*.
●     Engine tier. A parallel processing runtime framework.

In addition, InfoSphere Information Server has a client tier that consists of thin and rich client product modules. Both the application server and the database server are standard server applications. Figure 2 shows a high-level view of the InfoSphere Information Server architecture.
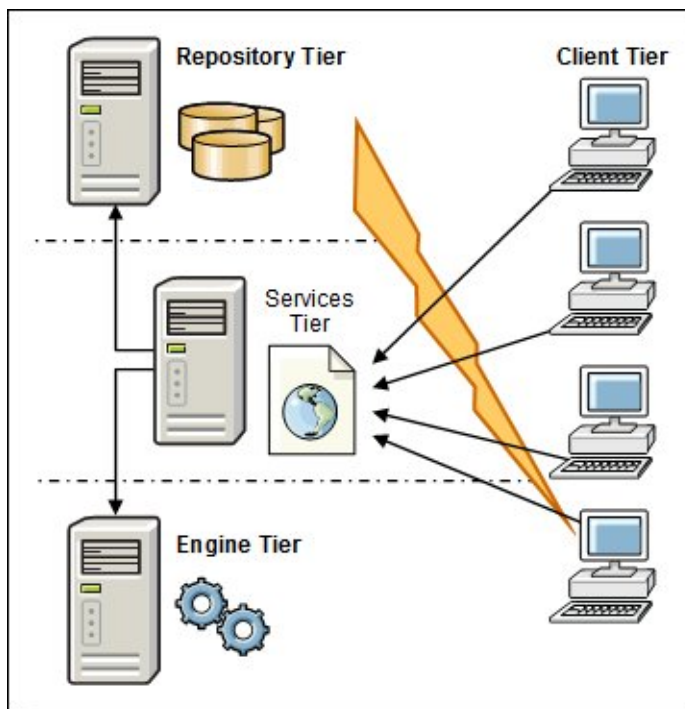


Figure 2. IBM InfoSphere Information Server architecture

One of the key advantages of InfoSphere Information Server is the shared infrastructure for all the individual InfoSphere Information Server product components. This shared infrastructure enables integration of all these components, minimizing duplicate effort and maximizing reuse. In addition, InfoSphere Information Server integrates metadata from external applications with other metadata in its repository. This way, the InfoSphere Information Server product modules can each use this external metadata to achieve the business objectives.

The product modules provide business and technical functionality throughout the entire initiative, from the planning through design phases to the implementation and reporting phases. Each product module uses the repository for its own persistence layer, yet its only access to the repository is through the IBM WebSphere® Application Server repository service. WebSphere Application Server is configured at installation time to access the repository through its Java Database Connectivity (JDBC) layer, based on the credentials that are passed to it during the installation. This approach is one way that InfoSphere Information Server ensures the integrity and security of its data and metadata.

InfoSphere Information Server has a unique, metadata-driven design that helps to align business goals and IT activities. It provides consistent terms and rules, and it captures business specifications. These specifications are used to automate development tasks, helping you gain greater insight to data by tracking its lineage. Some deployment architectures for these types of implementations might introduce challenges in fully using the shared metadata platform across products, environments, and servers. Furthermore, data privacy and information security requirements add other levels of complexity. Fortunately, InfoSphere Information Server supports a wide range of implementation topologies from a basic single computer to highly available clusters and grid configurations. They satisfy your requirements and enable the robust and high performing solutions for your particular environment.

## Solution architecture

IBM InfoSphere Information Server has several installation topologies that you can implement in any solution. They range from a basic single computer to highly available clusters and grid configurations. Each topology considers the deployment architectures for each of the major software subsystem components (tiers) of InfoSphere Information Server in combination and individually. However, the focus is constrained to the perspective of a single, overall installation environment, such as development, test, or production. The following topologies and related topics are available:

- Single computer
- Client server
- Dedicated computer per tier
- Dedicated engine
- Clusters
- Active-passive
- Massively parallel
- Grid
- Web services
- Disaster recovery

Figure 3 illustrates the InfoSphere Information Server solution architecture for implementing the topologies.
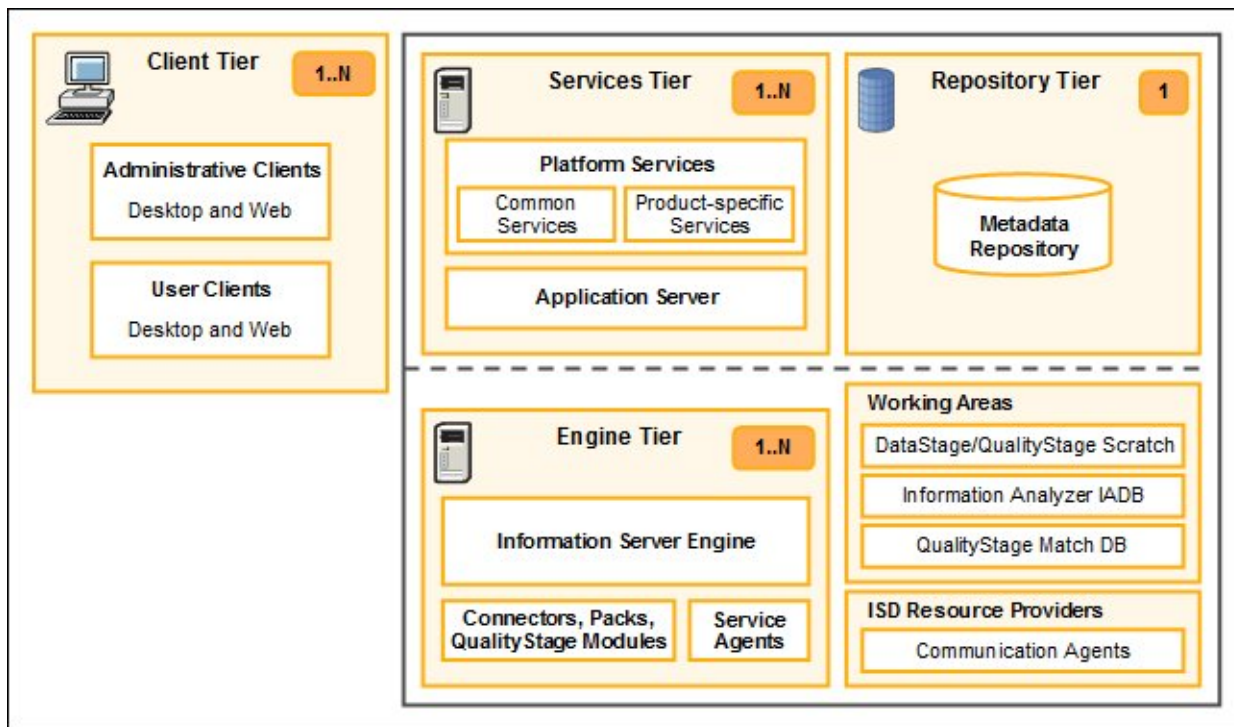


Figure 3. IBM InfoSphere Information Server solution architecture

All InfoSphere Information Server product modules generate and consume metadata. The generated metadata is persisted in the InfoSphere Information Server repository. Through this shared InfoSphere Information Server repository, InfoSphere Information Server provides collaboration between the product modules and users, such as analysts, data stewards, subject matter experts, and developers. This behavior is true for all InfoSphere Information Server instances, such as development, test, production, or other environments that a company might have. However, the development lifecycle varies for the InfoSphere Information Server assets, such as InfoSphere DataStage job design (extract, transfer, and load (ETL) program) and data source technical metadata.

The lifecycles of the InfoSphere Information Server assets have an impact on the deployment architecture for the InfoSphere Information Server product modules. Depending on business requirements, InfoSphere Information Server product modules might be deployed in some or all of the environments. The same product module can also be deployed differently, based on the feature and the level of metadata collaboration that you are using. Therefore, the typical development, test, and production deployment architecture might not be appropriate for each of the InfoSphere Information Server product modules and business use-case scenarios.

You need to consider various factors when you choose the deployment architecture for InfoSphere Information Server. To effectively use the InfoSphere Information Server product modules and achieve the business objectives, you can select one of the four deployment architectures. But which one should you choose? The following architectures are available:

- **Simple.** The typical development, test, and production environment architecture. In a typical deployment architecture, the source code is developed in the development environment, tested and deployed to the test environment, and then tested and deployed to the production environment.
- **Unified**. A combination of development, metadata reporting, and a collaboration environment.
- **Reporting.** Has a dedicated metadata environment to provide all stakeholders full-time access to the various types of metadata (business and technical) that exist within InfoSphere Information Server.
- **Governance**. External metadata can be analyzed and reported, and an enterprise glossary can be created or shared to increase governance in your company.

Development of InfoSphere DataStage and InfoSphere QualityStage fits into this typical development lifecycle. An important consideration is to size the production environment appropriately so that data integration processing is not negatively affected by any other processes because the main purpose of the production environment in this case is data integration. When you use other InfoSphere Information Server product modules, such as InfoSphere Information Analyzer, InfoSphere FastTrack, InfoSphere Business Glossary, and InfoSphere Metadata Workbench with InfoSphere DataStage and InfoSphere QualityStage, consider whether you need to dedicate separate capacity for these other modules.

## Deciding which architecture to choose

To help you decide which of the four deployment architectures is the best one for your environment, consider the following main determining factors:

- Which products are you using?
- Do you have metadata collaboration (reuse of metadata for other purposes) between multiple products?
- How many users will do lineage, impact analysis, and use the glossary?
- How intricate is your glossary?
- What is your environment capacity?

The following process assesses the architecture and deployment environments:

1.  Using the decision chart in Figure 4, determine an appropriate deployment architecture (how many and which kinds of environments are required to support the product mix).
2.  Determine in which environment to deploy each of the product modules.
3.  Evaluate the requirements for high availability.
4.  Design a topology that is based on business and IT requirements, availability, and product mix for each environment.
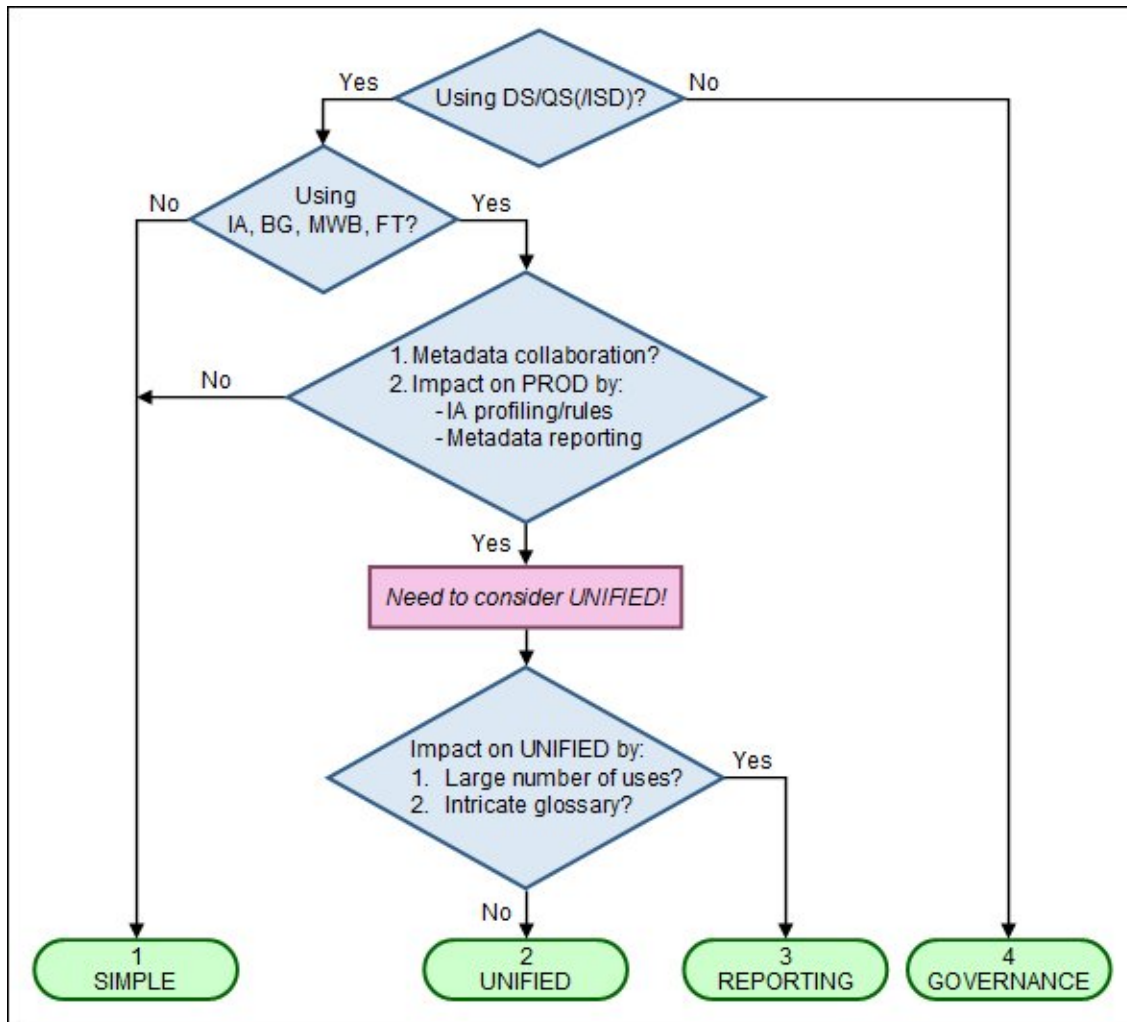


Figure 4. Architecture decision flow chart

The InfoSphere Information Server tiers are available on the following platforms:

- The installable client tier components, which provide the user interface, are available only on Microsoft Windows platforms. Business Glossary and Metadata Workbench require only a supported web browser; there are no client installable components for these two InfoSphere Information Server modules.

- The server tiers (services, engine, repository) are available on Linux, UNIX, and Windows platforms (Microsoft Windows Server, Red Hat Linux, SUSE Linux, IBM AIX®, Oracle Sun Solaris, and Hewlett Packard HP-UX). Although each services tier component can be deployed on a separate host, deploy the services and engine tiers on the same platform type.

- The database for the InfoSphere Information Server repository can be implemented by using IBM DB2®, Oracle, or SQL Server.

In deployment topologies, where clients are connected to the server tiers over communication links that have poor bandwidth, latency, or both (for example, because of long distance), consider using a remote desktop. An example is terminal services or Citrix installations for the client access architecture.

The system requirements for InfoSphere Information Server vary, based on the scope and scale of the system. The configuration that you need to support the environment with satisfactory performance can vary depending on multiple factors and requirements for system or environment availability. Such factors include server speed, memory, disk I/O, data volumes, and network and server workload.

## Usage scenario

To illustrate this solution, assume a small-to-medium sized industrial company recently purchased InfoSphere Information Server and is planning to use both InfoSphere DataStage and InfoSphere QualityStage for data integration. In addition, the company purchased InfoSphere Business Glossary, but the company does not intend to roll it out now to the entire enterprise. The company also purchased InfoSphere Metadata Workbench and intends to use it for the data lineage capability.

The company is allocating three or four developers during the initial phase of the project and will add more developers when more business units agree to participate in the data integration program. The current production requirements dictate that the company should have a high availability failover environment if the primary production server fails. After reviewing the various types of available options, the company determined that a simple active-passive configuration can meet the company needs. The company discussed the benefits of having the development environment be clustered to allow continuous development efforts, but could not justify the additional hardware costs. The only environment that will be configured for high availability is the production environment.

Now the company must decide which deployment architecture to use. The first step in determining the deployment landscape is to review the decision chart (Figure 4) to determine an appropriate deployment architecture. The answers to the decision-chart questions for this scenario are as follows:

1. Using InfoSphere DataStage and InfoSphere QualityStage? Yes
2. Using InfoSphere Information Analyzer, InfoSphere Business Glossary, InfoSphere Metadata Workbench, and InfoSphere FastTrack? Yes
3. Will related workload for metadata collaboration require more resource for InfoSphere Information Analyzer profiling and data rules or metadata reporting? No

Next is the key decision point that determines whether the *simple deployment architecture* (development, test, production) is sufficient or whether the scenario requires the advanced design features that are typical of the *unified environment*.

Because the use of InfoSphere Business Glossary is limited (to a few business units), and only data lineage reports are generated from the InfoSphere Metadata Workbench, there is currently not enough collaboration to justify a *unified development environment*. Furthermore, a significant enough load is not expected on the production environment that might necessitate a separate metadata reporting environment. (Production is a natural environment for deploying these product modules.)

The result from using this decision tree is to suggest a *simple deployment architecture* that will consist of development, test, and production environments.

The product modules can be deployed to the following environments:

- InfoSphere DataStage and InfoSphere QualityStage are deployed in all environments according to the standard software development lifecycle (SDLC) methodology.

- InfoSphere Business Glossary is deployed only in the production environment, for accessibility by authorized users. Glossary authorship is handled through the built-in workflow mechanisms, preventing unauthorized users from viewing preapproved content.

- InfoSphere Metadata Workbench is deployed in the production environment. However, there might be justification to also deploy InfoSphere Metadata Workbench in the development environment. More information about the requirements of the developers is needed to determine this deployment.

According to the business requirements for the industrial company in the example, the company requires active-passive high availability in the production environment alone. All other issues being equal, and in consideration of simplifying the high availability configuration, the company can install all three server tiers on the same server (physical or virtual) for the production environment.

If you are using only InfoSphere DataStage or QualityStage, the common practice is to choose this *simple architecture*. Information Services Director is handled in the same way as InfoSphere DataStage jobs. Even if you are using other InfoSphere Information Server product modules, there are also times when this simple architecture is the appropriate architecture. In this case, consider the following questions to help you identify whether this simple architecture might not be appropriate for your case:

- Are you reusing the metadata generated in any of the InfoSphere Information Server product modules for any other purpose? Is there collaboration? Consider the following examples:
  - o InfoSphere Information Analyzer profiling results in published and reused in InfoSphere DataStage job development.
  - o Technical assets are linked to business terms.

- Is your execution of data integration jobs in the production environment affected by any of the following items?
  - o Reporting metadata with InfoSphere Metadata Workbench
  - o Sharing a glossary by using InfoSphere Business Glossary
  - o Profiling data or using quality rules as factors in InfoSphere Information Analyzer

The main purpose of the production environment is data integration. Therefore, it should not be affected by any of the other processes. If any of these points match your case, the simple architecture is probably not the correct architecture for you. Consider using either the unified or reporting architecture.

For information about the deployment architectures and how to determine which to choose for your environment, see *IBM Information Server Deployment Architectures*, SG24-8028.

## Integration

The IBM InfoSphere Information Server product modules and components that work together to achieve business objectives within the information integration domain. The product modules provide business and technical functionality throughout the entire initiative from the planning through design phases to the implementation and reporting phases. IBM InfoSphere Data Architect, IBM InfoSphere Discovery, and IBM InfoSphere Data Replication are also integrally related applications.

In addition to the product modules, InfoSphere Information Server includes the following utilities for importing external assets into the repository, deploying internal assets from one environment to another, and managing these assets to remove duplicates and orphans:

- InfoSphere Information Server Manager
- The `istool` command-line utility (CLI)
- IBM InfoSphere Metadata Asset Manager

## Supported platforms

IBM InfoSphere Information Server tiers are available on the following platforms:

- The installable client tier components, which provide the user interface, are available only on Windows platforms. InfoSphere Business Glossary and InfoSphere Metadata Workbench require only a supported web browser. These two InfoSphere Information Server modules do not have any client installable components.

- The server tiers (services, engine, repository) are available on Linux, UNIX, and Windows platforms. Although each services tier component can be deployed on a separate host, the services and engine tiers should be deployed on the same platform type.

- The database for the InfoSphere Information Server repository can be implemented by using DB2, Oracle, or SQL Server.

For more information, see the Systems Requirements for InfoSphere Information Server at:
http://www.ibm.com/software/integration/information server/requirements

## Ordering information

IBM InfoSphere Information Server is available only through IBM Passport Advantage®. It is not available as a shrink wrapped product.

You can purchase it directly from BM or from an authorized IBM Business Partners for Software Value Plus. For more information about IBM Software Value Plus, see the
Authorized portfolio: Authorized products at:
http://www.ibm.com/partnerworld/page/svp_authorized_portfolio

To locate IBM Business Partners for Software Value Plus in your geographic region for a specific Software Value Plus portfolio, contact your IBM representative. For ordering information, go to the IBM Offering Information page at:
http://www.ibm.com/common/ssi/index.wss?request_locale=en

On this page, enter `InfoSphere Information Server`, select the information type, and then click **Search**. On the next page, narrow your search results by geography and language.

## Related information

For more information, see the following references:

- IBM announcement letters and sales manuals
  http://www.ibm.com/common/ssi/index.wss?request_locale=en

- *IBM InfoSphere Information Server Deployment Architectures*, SG24-8028
  http://www.redbooks.ibm.com/abstracts/sg248028.html

- *IBM InfoSphere Information Server Installation and Configuration Guide*, REDP4596
  http://www.redbooks.ibm.com/abstracts/redp4596.html

- IBM InfoSphere Information Server
  http://www.ibm.com/software/data/integration/info_server

- IBM InfoSphere Information Server Enterprise Edition
  http://www.ibm.com/software/data/integration/info_server/enterprise-edition

- IBM InfoSphere Information Server Information Center
  http://publib.boulder.ibm.com/infocenter/iisinfsv/v9r1/index.jsp

- The IBM Publications Center
  http://www.ibm.com/shop/publications/order

# Notices

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

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 give 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.*

**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. 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. 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.

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 measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

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.

This document was created or updated on January 11, 2013.

Send us your comments in one of the following ways:
- Use the online **Contact us** review form found at:
  **ibm.com**/redbooks
- Send your comments in an e-mail to:
  redbook@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

This document is available online at http://www.ibm.com/redbooks/abstracts/tips0964.html .

# Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®
DataStage®
DB2®
IBM®
InfoSphere®
Passport Advantage®
QualityStage®
Redbooks (logo)®
WebSphere®

The following terms are trademarks of other companies:

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

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

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

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

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