IBM InfoSphere Information Server
Version 8 Release 7

Integration Scenario Guide

SC19-3474-00

IBM InfoSphere Information Server
Version 8 Release 7

Integration Scenario Guide

SC19-3474-00

Note

Before using this information and the product that it supports, read the information in “Notices and trademarks” on page 25.

Contents

InfoSphere Information Server integration scenarios 1
Building a data integration application scenario . . 4
Modernizing a data warehouse with a focus on data quality scenario 5
Data quality and monitoring integration 6
IBM Industry Models 7
Common questions to address 7
Key scenario inputs 8
Key scenario input processes 9
Key scenario outputs 14

Contacting IBM 19

Accessing product documentation. . . . 21

Product accessibility 23

Notices and trademarks 25

Index 29

InfoSphere Information Server integration scenarios

Information integration is a complex activity that affects every part of an organization. To address the most common integration business problems, these integration scenarios show how you can deploy and use IBM® InfoSphere® Information Server and the InfoSphere Foundation Tools components together in an integrated fashion. The integration scenarios focus on data quality within a data warehouse implementation.

Data integration challenges

Today, organizations face a wide range of information-related challenges: varied and often unknown data quality problems, disputes over the meaning and context of information, managing multiple complex transformations, leveraging existing integration processes rather than duplicating effort, ever-increasing quantities of data, shrinking processing windows, and the growing need for monitoring and security to ensure compliance with national and international law.

Organizations must streamline and connect information and systems across enterprise domains with an integrated information infrastructure. Disconnected information leaves IT organizations unable to respond rapidly to new information requests from business users and executives. With few tools or resources to track the information sprawl, it is also difficult for businesses to monitor data quality and consistently apply business rules. As a result, information remains scattered across the enterprise under a multitude of disorganized categories and incompatible descriptions.

Some key data integration issues include:

- Enterprise application source metadata is not easily assembled in one place to understand what is actually available. The mix can also include legacy sources, which often do not make metadata available through a standard application programming interface (API), if at all.
- Master reference data, names and addresses of suppliers and customers, part numbers and descriptions, differ across applications and duplicate sources of this data.
- Hundreds of extract, transform, and load (ETL) jobs need to be written to move data from all the sources to the new target application.
- Data transformations are required before loading the data so it will fit into the new environment structures.
- The ability to handle large amounts of data that can be run through the process, and finish on time, is essential. Companies need the infrastructure to support the running of any of the transformation and data-matching routines on demand.

InfoSphere Information Server integration solution

InfoSphere Information Server and InfoSphere Foundation Tools components are specifically designed to help organizations address the data integration challenges and build a robust information architecture that leverages existing IT investments. The solution offers a proven approach to identifying vital information; specifying how, when, and where it should be made available; determining data management processes and governance practices; and aligning the use of information to match an organization's business strategy.

InfoSphere Foundation Tools components help your organization profile, model, define, monitor, and govern your information. By integrating the solutions provided by the InfoSphere Foundation Tools components, your organization can discover and design your information infrastructure and start building trusted information across the organization.

The IBM InfoSphere Information Server platform consists of multiple product modules that you can deploy together or individually within your enterprise integration framework, as shown in Figure 1. InfoSphere Information Server is designed to flexibly integrate with existing organizational data integration processes to address the continuous cycle of discovery, design, and governance in support of enterprise projects.

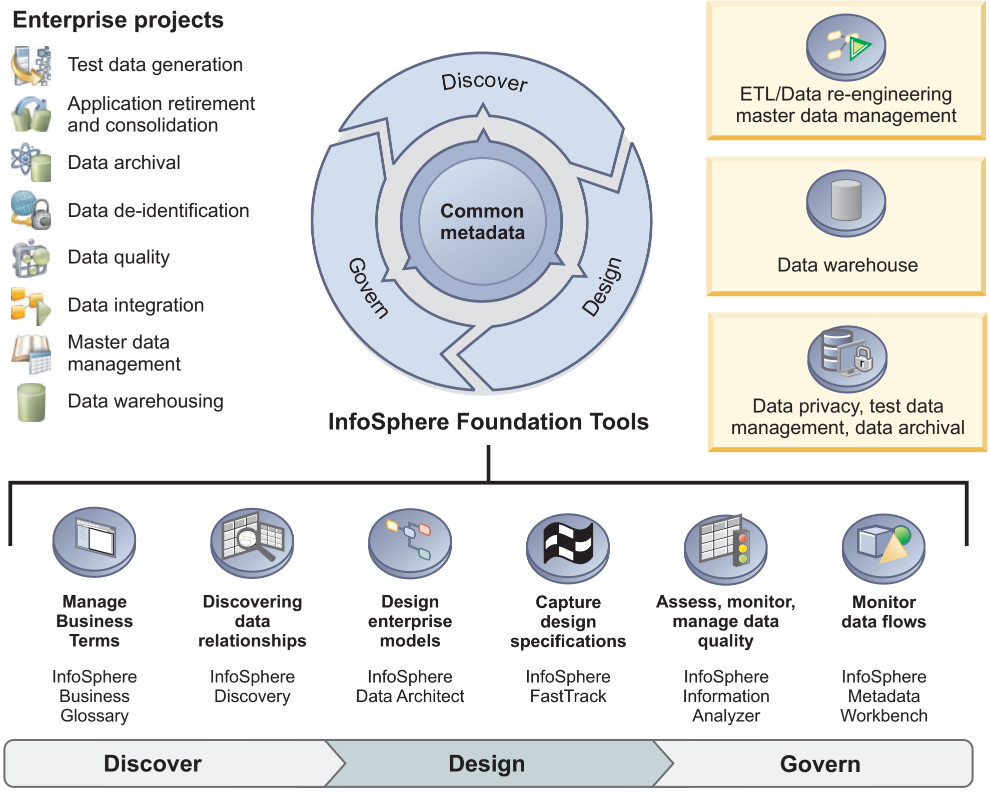

Enterprise projects
Test data generation
Application retirement and consolidation
Data archival
Data de-identification
Data quality
Data integration
Master data management
Data warehousing
Discover
Common metadata
Govern
Design
InfoSphere Foundation Tools
ETL/Data re-engineering master data management
Data warehouse
Data privacy, test data management, data archival
Manage Business Terms
InfoSphere Business Glossary
Discovering data relationships
InfoSphere Discovery
Design enterprise models
InfoSphere Data Architect
Capture design specifications
InfoSphere FastTrack
Assess, monitor, manage data quality
InfoSphere Information Analyzer
Monitor data flows
InfoSphere Metadata Workbench
Discover
Design
Govern

Figure 1. The InfoSphere Information Server platform supports your data integration processes.

Figure 2 on page 3 illustrates the components and the metadata they generate, consume, and share.

Typically, the process starts with defining data models. An organization can import information from IBM Industry Data Models (available in InfoSphere Data Architect), which includes a glossary, logical, and physical data model. The glossary models contains thousands of industry-standard terms that can be used to pre-populate IBM InfoSphere Business Glossary. Organizations can modify and extend the IBM Industry Data Models to match their particular business requirements.

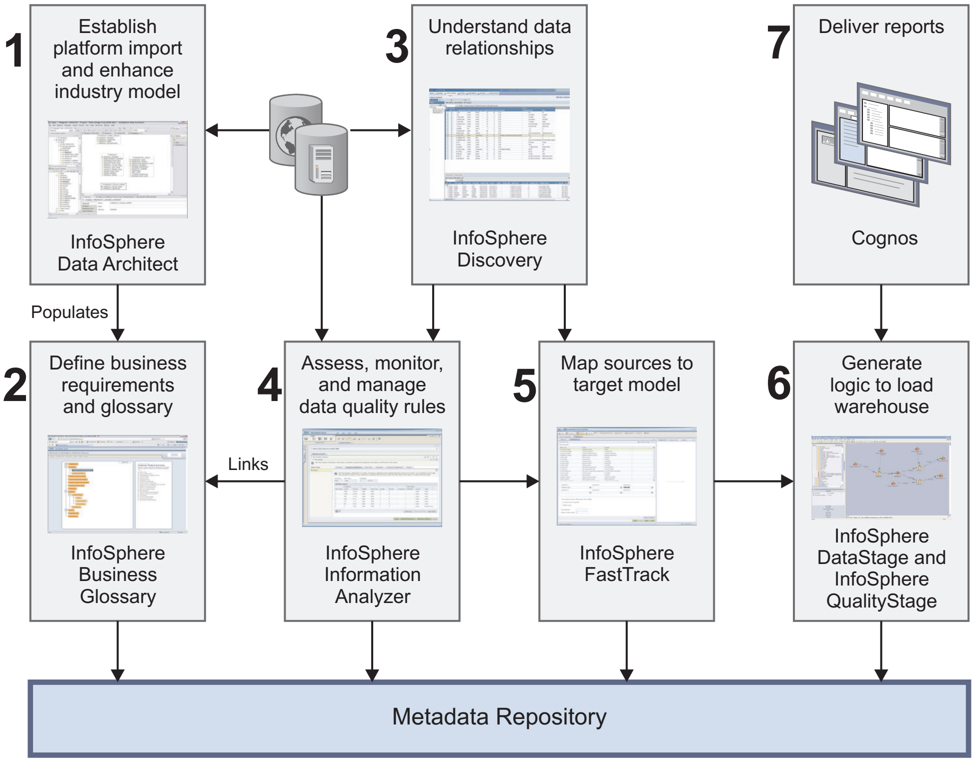

1
Establish platform import and enhance industry model
InfoSphere Data Architect
Populates
2
Define business requirements and glossary
InfoSphere Business Glossary
Links
3
Understand data relationships
InfoSphere Discovery
4
Assess, monitor, and manage data quality rules
InfoSphere Information Analyzer
5
Map sources to target model
InfoSphere FastTrack
6
Generate logic to load warehouse
InfoSphere DataStage and InfoSphere QualityStage
7
Deliver reports
Cognos
Metadata Repository

Figure 2. InfoSphere Information Server product modules

After the data models are defined and business context is applied, the analyst runs a data discovery process against the source systems that will be used to populate the new target data model. During the discovery process, the analyst can identify key relationships, transformation rules, and business objects that can enhance the data model, if these business objects were not previously defined by the IBM Industry Data Models.

From the discovered information, the analyst can expand the work to focus on data quality assessment and ensure that anomalies are documented, reference tables are created, and data quality rules are defined. The analyst can link data content to established glossary terms to ensure appropriate context and data lineage, deliver analytical results and inferred models to developers, and test and deploy the data quality rules.

The analyst is now ready to create the mapping specifications, which are input into the ETL jobs for the new application. Using the business context, discovered information, and data quality assessment results, the analyst defines the specific transformation rules necessary to convert the data sources into the correct format for the IBM Industry Data Model target. During this process, the analyst not only defines the specific business transformation rules, but also can define the direct relationship between the business terms and their representation in physical structures. These relationships can then be published to IBM InfoSphere Business Glossary for consumption and to enable better understanding of the asset relationships.

The business specification now serves as historical documentation as well as direct input into the generation of the IBM InfoSphere DataStage® ETL jobs. The defined business rules are directly included in the ETL job as either code or annotated to-do tasks for the developer to complete. When the InfoSphere DataStage job is ready, the developer can also decide to deploy the same batch process as an SOA component by using IBM InfoSphere Information Services Director.

Throughout this process, metadata is generated and maintained as a natural consequence of using each of the InfoSphere Information Server modules. The InfoSphere Information Server platform shares relevant metadata with each of the user-specific roles throughout the entire integration process. Because of this unique architecture, managing the metadata requires little manual maintenance. Only third-party metadata requires administration tasks such as defining the relationships to the InfoSphere Information Server metadata objects. Administrators and developers who need to view both InfoSphere Information Server and third-party metadata assets can use IBM InfoSphere Metadata Workbench to query, analyze, and report on this information from the common repository.

Building a data integration application scenario

IBM InfoSphere Information Server features a unified suite of product modules that are designed to streamline the process of building a data integration application.

The InfoSphere Information Server platform offers a comprehensive, integrated architecture built upon a single shared metadata repository allowing information to be shared seamlessly among project data integration tasks. You can use information validation, access, and business processing rules across multiple projects, leading to a higher degree of consistency, greater control over data and improved efficiencies. Figure 3 illustrates the capabilities: understand, cleanse, transform, deliver, and perform unified metadata management.

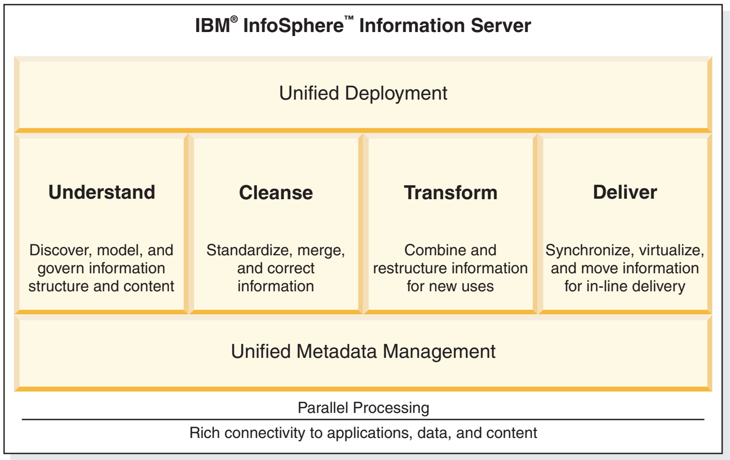

IBM® InfoSphere™ Information Server
Unified Deployment
Understand
Discover, model, and govern information structure and content
Cleanse
Standardize, merge, and correct information
Transform
Combine and restructure information for new uses
Deliver
Synchronize, virtualize, and move information for in-line delivery
Unified Metadata Management
Parallel Processing
Rich connectivity to applications, data, and content

Figure 3. InfoSphere Information Server integration functions

InfoSphere Information Server enables you to perform five key integration functions:

- *Understand the data*. InfoSphere Information Server helps you to automatically discover, model, define, and govern information content and structure, as well as understand and analyze the meaning, relationships and lineage of information. With these capabilities, you can better understand data sources and relationships and define the business rules that eliminate the risk of using or proliferating bad data.
- *Cleanse the data.* InfoSphere Information Server supports information quality and consistency by standardizing, validating, matching, and merging data. The

platform can help you create a single, comprehensive, accurate view of information by matching records across or within data sources.

- *Transform data into information*. InfoSphere Information Server transforms and enriches information to help ensure that it is in the proper context for new uses. It also provides high-volume, complex data transformation and movement functionality that can be used for stand-alone extract, transform, and load (ETL) scenarios or as a real-time data processing engine for applications or processes.
- *Deliver the right information at the right time.* InfoSphere Information Server provides the ability to virtualize, synchronize, or move information to the people, processes, or applications that need it. It also supports critical service-oriented architectures (SOAs) by allowing transformation rules to be deployed and reused as services across multiple enterprise applications.
- *Perform unified metadata management.* InfoSphere Information Server is built on a unified metadata infrastructure that enables shared understanding between the different user roles involved in a data integration project, including business, operational, and technical domains. This common, managed infrastructure helps reduce development time and provides a persistent record that can improve confidence in information while helping to eliminate manual coordination efforts.

Modernizing a data warehouse with a focus on data quality scenario

This scenario describes approaches to leveraging the IBM InfoSphere Information Server and Foundation Tools software to address data quality within a data warehouse environment.

There are three typical use cases where data quality is assessed or monitored in association with data warehouses.

Greenfield

Builds a new warehouse with activities that include discovery, terminology, lineage, data quality, data modeling, mapping, data transformation, and cleansing.

Modernization

Modifies or adds to an existing warehouse with activities that include terminology, discovery, impact analysis, data quality, data modeling, mapping, data transformation, and cleansing.

Governance

Manages and governs your existing warehouse with activities that include data quality, stewardship, terminology, lineage, and impact analysis.

In each use case, there are a range of activities that contribute to the overall solution. Each activity utilizes one or more product modules in the context of a broader methodology and process that makes up the initiative. For each activity or phase, there are certain inputs to the process, certain tasks to perform both within and outside a product, and certain outputs from the process that are utilized in subsequent activities or phases. Data quality is but one of those activities.

The activities with these use cases are not necessarily a rigid sequence of events. Often these are iterative activities, frequently occurring in parallel, where findings from one activity will influence another, and then require additional work.

For example, a data warehouse contains customer and account data. However, a greater view is desired into the impact of sales and marketing on customers and their buying habits. A group of sales management sources are targeted for addition to the existing data warehouse. Initial discovery work finds and maps four sales management systems for inclusion. However, data quality review finds significant issues when validating domains in all four systems, indicating that many fields are unpopulated or contain comments, rather than usable data. A review of the business terminology finds that there is a misunderstanding of the system's use and that two other systems are needed. The business terms are brought through the discovery process to map relationships to the previous four systems. Data quality review then validates that these are in fact the needed tables. Inferences from the data quality review are then provided to the data architects to improve the modeling of the new data warehouse tables.

There are a number of common *pain points* that you might experience in these use cases. These include:

- *Unclear terminology* when managing warehouse information. For example, where is revenue information and does the warehouse content match the business' expectation?
- *Unknown impact of change* that can break existing processes and disrupt the business.
- *Unknown lineage of information* that negatively impacts the trust in data. For example, am I using the right data sources for the requested information?
- *Unknown data quality* is one of the primary reasons why business users don't trust their data.
- *Unknown stewardship* where it is unclear who understands the data, who ensures the quality is maintained, and who governs access to the data.

In this scenario, a data warehouse, which could be an IBM warehouse or any other mainstream warehouse vendor such as Teradata or Oracle, is being expanded or *modernized*. This particular warehouse already contains a variety of financial information to enable effective reporting, but now needs to add customer data to provide broader analytical information. As with most organizations, their warehouse becomes an important place to maintain and manage the combination of financial, customer, and sales information for analysis.

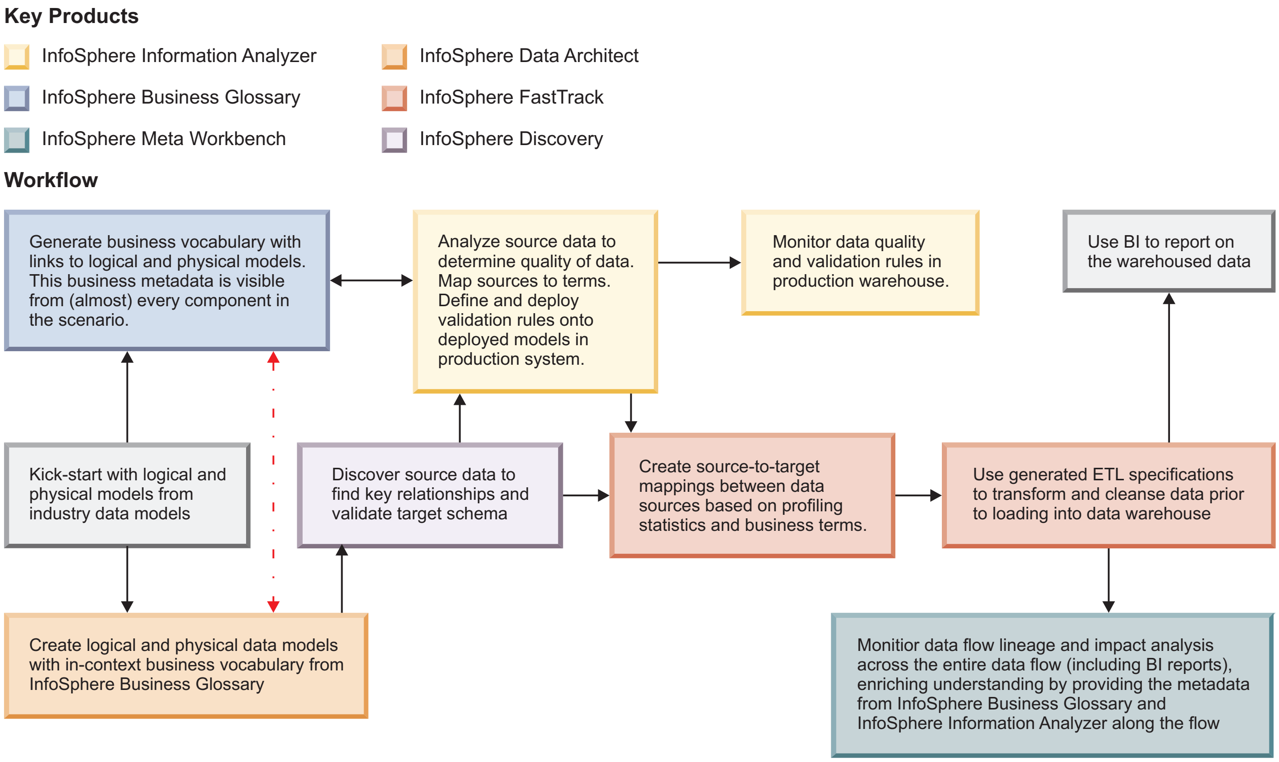
Data quality and monitoring integration

A data quality assessment and monitoring strategy addresses the issues surrounding the quality and integrity of information. Additionally, data quality procedures must be established to address these issues in a data warehouse.

IBM InfoSphere Business Glossary, IBM InfoSphere Discovery, IBM InfoSphere Data Architect, IBM InfoSphere Information Analyzer, IBM InfoSphere FastTrack, and IBM InfoSphere Metadata Workbench use existing information assets to feed a data warehouse through information integration based on a number of business intelligence requirements, potentially based on an industry model or standard.

Data warehouse use case

InfoSphere Information Analyzer can identify the issues surrounding the quality and integrity of information and the creation of data quality procedures or rules in a multi-user environment to monitor data quality over time. This scenario can exist on a stand-alone basis or be part of a broader initiative, such as data warehousing, that incorporates data quality.


Key Products
InfoSphere Information Analyzer
InfoSphere Data Architect
InfoSphere Business Glossary
InfoSphere FastTrack
InfoSphere Meta Workbench
InfoSphere Discovery
Workflow

Generate business vocabulary with links to logical and physical models. This business metadata is visible from (almost) every component in the scenario.

Analyze source data to determine quality of data. Map sources to terms. Define and deploy validation rules onto deployed models in production system.

Monitor data quality and validation rules in production warehouse.

Use BI to report on the warehoused data

Kick-start with logical and physical models from industry data models

Discover source data to find key relationships and validate target schema

Create source-to-target mappings between data sources based on profiling statistics and business terms.

Use generated ETL specifications to transform and cleanse data prior to loading into data warehouse

Create logical and physical data models with in-context business vocabulary from InfoSphere Business Glossary

Monitior data flow lineage and impact analysis across the entire data flow (including BI reports), enriching understanding by providing the metadata from InfoSphere Business Glossary and InfoSphere Information Analyzer along the flow

Figure 4. Integration workflow for the data warehouse

IBM Industry Models

To help organizations achieve results faster, IBM has packaged the knowledge from years of experience in working on information projects within specific industries into the IBM Industry Models.

These models provide a complete fully attributed enterprise data model along with Reporting Templates that outline the key performance indicators, metrics, and compliance concerns for each industry. IBM provides these industry models for six industries: banking, financial markets, health care, insurance, retail and distribution as well as telecommunications.

These models act as key accelerators for migration and integration projects, providing a proven industry-specific template for all project participants to refer to. Source system data can be loaded directly into InfoSphere Information Server, providing target data structures and pre-built business glossaries to accelerate development efforts. In addition, the Business Solution Templates provide templates for reports and data cubes within Cognos® 8 Business Intelligence. By using the IBM Industry Models, organizations can dramatically accelerate their projects and reduce risk, and also overcome traditional organizational issues typically faced when integrating information by providing a proven, neutral base model.

Common questions to address

As with any initiative, there are a number of common questions to ask in relation to data quality within a data warehouse initiative.

These questions include:

Business and metadata definition versus data reality

What is understood by the business user or subject matter expert and the IT resources (architects, modelers, and developers) can be significantly different. There can also be significant differences between different business groups in how they understand common business terminology.

- Is there clear understanding of business terminology?
- Is there a clear understanding of how business terminology relates to actual source data quality that needs to be validated or certified?
- Are there gaps between the business terminology and the metadata (either source metadata or target warehouse model)?
- Is actual metadata and data used to discover and verify business semantics and data quality?

Impact of change

When modernizing an existing system, it is important to understand the implication of the various changes that are planned.

- What are the upstream systems, meaning those that feed into the system to be changed, and downstream systems, meaning applications or systems that consume the information, that are impacted by the change?
- Who are the stewards?
- Are there any processes or applications such as business intelligence reports that could break and that need to be modified?

Data focus

Information added to a data warehouse typically comes from multiple divergent data sources or data sources divided among many tables.

- Is attention focused on core systems or tables, specific types or classes of data, or specific attributes or domains?
- Are any systems, sources, or entities considered the "source of record"?
- How will cross-source consistency or conflicts be addressed or resolved?

Validation and information delivery

Initiatives often leave little time for considerations of data quality up-front. However, when data is delivered and data quality is not achieved, the cost of correction and rebuilding trust is high.

- When is data quality analyzed? Is it addressed only after initial discovery and review, or throughout the data integration lifecycle?
- What metrics are critical for validating data quality?
- How will information results for data quality be delivered?

Key scenario inputs

When the data warehouse is new or is being modified to add new business objects (such as customer data) or data sources (such as customer information from the sales order system), there are requirements to expand the glossary to incorporate new terminology, the model to include all new entities and attributes, and the metadata to include the new sources with associated analytical information (both of relationships and data quality).

One of the founding principles of data integration is the ability to populate and store information about each process as stored metadata.

Models

To help you get started, IBM offers what are referred to as *third-normal form* data models for integrating large sets of detail data as well as predefined analytical models called TDW Reports or Business Solution Templates (BSTs) consisting of measures and dimensions for summaries and reporting. IBM also develops metadata models that contain dictionaries of business terms used to establish glossaries and scope large projects. All of the IBM models are mapped to each other for design acceleration and lineage. The models are deliverable in both InfoSphere Data Architect as well as CA ERwin Data Modeler.

Through existing metabroker technology, the physical models from InfoSphere Data Architect can be loaded as metadata content into the IBM InfoSphere Information Server metadata repository.

Terminology

Business terms are key in describing the types of data you are working with and in the language that makes sense to your business. This type of terminology definition could include not only terms about the target systems, but also key information that drives the business such as key performance indicators (KPIs) or expected benefits. An example for this warehouse would be profitability or purchase history. Understanding what that means drives collaboration, so keep in mind that everything connects at this common language.

After the information is imported, it is shared with the business to take advantage of it and share what it means through collaboration. You can manage metadata to capture corporate-standard business terms and descriptions that reflect the language of the business users. Organizations institutionalizing formal data management or data governance can publish these terms as a way to ensure that all business users have a consistent understanding of the organization's available information based on standard business definitions.

IBM InfoSphere Business Glossary provides the foundation for creating business-driven semantics, including categories and terms.

If you use other tools to import assets into the business glossary, you can use InfoSphere Business Glossary or IBM InfoSphere Metadata Workbench to assign an asset to a term. Typically, InfoSphere Business Glossary is used to assign large numbers of assets to terms. Because glossary content is stored in the InfoSphere Information Server metadata repository, you can interact with glossary content with the other components of the InfoSphere Information Server suite.

Key scenario input processes

The following topics describe the key scenario input processes.

Discovering source metadata

For this scenario, you now understand what the business requires and the terminology used to describe the business requirements. You understand terms such as customer or location, and you can leverage this across the organization. You also know the structure of your target data warehouse. You need to increase or improve your understanding of the structure and content of the actual incoming data.

IBM InfoSphere Discovery

IBM InfoSphere Discovery is used to identify the transformation rules that have been applied to a source system to populate a target such as a data warehouse or operational data store. Once accurately defined, these business objects and transformation rules provide the essential input into information-centric projects like data integration, IBM InfoSphere Master Data Management (MDM), and archiving.

InfoSphere Discovery analyzes the data values and patterns from one or more sources, to capture these hidden correlations, and bring them clearly into view. InfoSphere Discovery applies heuristics and sophisticated algorithms to perform a full range of data analysis techniques: single-source and cross-source data overlap and relationship analysis, advanced matching key discovery, transformation logic discovery, and more. It accommodates the widest range of enterprise data sources: relational databases, hierarchical databases, and any structured data source represented in text file format.

Using InfoSphere Discovery to understand data relationships by finding data values and patterns

In this scenario, you are combining three distributed data sources to discover related information for customer names, addresses, and tax identifiers.

Use InfoSphere Discovery to:

1. Create a source data discovery project. In Source Data Discovery projects, you review data value overlaps between tables within and across data sets. In addition, you can create a unified schema and discover unified schema primary-foreign keys. Results include a summary of the total number of tables and columns in the data sets, number of exclusive columns, percent of value overlap, number of tables and columns containing overlapping data, and more detailed statistics, including views of the actual overlapping data itself.
2. Import a set of tables from each data source.
3. Create data sets. A data set is a collection of database tables and text files to be processed, analyzed, or mapped. It can contain as many database tables and delimited or positional text files as needed, from as many ODBC connections as needed.
4. Run data discovery. Some database tables contain correct data type definitions in metadata. However, other database tables contain incorrect or incomplete metadata, and text files do not define the data types.

   In this step, InfoSphere Discovery calculates and displays statistics about the data in the data sets, along with displaying the metadata information. InfoSphere Discovery also examines all VARCHAR strings to determine if they contain date/time or numeric values, and changes them to the appropriate data type.

   After you run data discovery, you need to review the column analysis data. Verify the data types and make any necessary modifications, such as changing the length of a column or correcting a wrong data type defined in the metadata. Mark columns that are important to your project as Critical Data Elements (CDEs). Whenever you modify the tables in a data set, you must to re-run column analysis.
5. Discover primary-foreign (PF) keys. InfoSphere Discovery automatically imports PF key relationships when they are defined in a table's metadata. When the relationships are not defined, InfoSphere Discovery finds column

matches by examining the actual data. Column matches with the highest hit rates and selectivity are automatically designated as PF keys.

After you perform the Discover PF Keys task, verify the accuracy of the results. If you define or modify discovered column matches or PF keys, run Discover PF Keys again.

6. Discover data objects. A data object is a conceptual way of looking at table relationships within a data set. A data object represents a group of tables that are related by PF Keys. A data set can contain many data objects, with each data object consisting of many tables or just one table. A table can be both a parent and a child, so the same table might appear in several data objects. Typically, these data objects will relate to key concepts in the glossary or the new entities you are adding or modernizing in the data warehouse.
7. Discover overlaps. The Overlaps step provides a clear picture of overlapping data in your sources. Review the column data to verify that the discovered overlaps are useful and valid. Delete any obvious mismatches in the Value Overlap Details. If there is any doubt about the data in a particular overlap, use Column Summary and Column Overlaps to display the actual data. Define new overlaps and use InfoSphere Discovery to determine the statistics.

   Mark columns that are important to your project as Critical Data Elements (CDEs).

   When the overlaps discovery task is complete, you can start defining a unified schema. Apply this approach when the current source data model structures should be and could be re-used or when fast prototyping is required or when there are no specific requirements related to the target data model that would require a very unique design. Even in the latter case, it can be advantageous to create a unified schema as a staging point for standardized and aligned data from the data sources prior to consolidation and load to the data warehouse.
8. Create a target table for a unified schema. Define the target table schema. Populate the table with as many target columns as needed. You can modify data types and rearrange the column sequence.
9. Define source mapping. Map table columns to target columns and create filters for each table column until each target column contains appropriate data from the relevant tables. If necessary, return to the Target Table Schema tab to refine the target column characteristics.
10. Perform unified schema analysis. Consolidate data from different sources into the aggregated target table. The statistics and data previews in this screen are used to define matching conditions for use in the next step, match and merge.

    Review the resulting statistics. Statistics are shown for the aggregated columns in the target table, as well as for each individual table column used in the target table.

    As needed, refine the target table by returning to the Target Table Schema window and adding or removing target table columns. You can also change the source mapping.
11. Perform match and merge analysis. Define matching conditions for each target column. Also define conflict detection rules that determine whether the values in each group are considered conflicts or not. Merge the results and define conflict resolution rules that select the best value from the duplicate record alternatives. The unified schema is finished.
12. Export data from InfoSphere Discovery into the metadata repository. Now you can export the results from your unified schema into the metadata repository by using the Import/Export Manager in the DBM Metabroker to make the results available to InfoSphere Information Analyzer.

Data quality assessment

Understanding how the tables and columns of data are related is not sufficient to ensure effective utilization in the data warehouse. You must assess the data to deliver it downstream effectively.

During this assessment, you will thoroughly analyze the content to reveal all of the anomalies within the information. This quality assessment allows you to annotate the errors on the data to take action on it downstream. These annotations are shared at the metadata level. When you find and annotate these data problems you can put in data monitors to continue to look for these patterns within the data. This process ensures that you continue to have high quality data moving from the ‘as-is' source systems to the ‘to-be' target.

While data discovery provides a starting point for the effort, to ensure that data quality criteria are met and addressed both in the initial load of the data warehouse with this new information and in ongoing governance of the data warehouse, detailed data quality assessment is advised.

The focus is two-fold:

- First, you review and communicate information about anomalies, key reference tables for valid values, alignment of sources with business terminology, and requirements for data validation conditions.
- Second, you design and test data validation rules, benchmarks, and measures that will be applied to the data warehouse load process and the data warehouse itself after it is loaded.

Analyzing source data quality

This is the initial step to assess, review, and communicate core data quality information about the sources to feed into the data warehouse.

Using IBM InfoSphere Information Analyzer to run data quality and define sensitive data

The schema from the InfoSphere Discovery analysis can be consumed by InfoSphere Information Analyzer to provide accelerated deployment and better accuracy of data quality rules for analysis, monitoring, and management of quality over time.

For this scenario, the results of the discovery process provide the initial data source information with metadata about customer names, addresses, and tax identifiers.

Use InfoSphere Information Analyzer to:

1. Create or open an existing InfoSphere Information Analyzer project.
2. Add the data sources for which InfoSphere Discovery performed analysis into the Information Analyzer project.
3. Import the InfoSphere Discovery schema from the previous scenario, Discovering source metadata, from the metadata repository into InfoSphere Information Analyzer as a target for data quality rules.
4. Review the summary analysis results from InfoSphere Discovery under the Column Analysis task. Ignore fields that are irrelevant for use in the data warehouse. For fields where data quality is critical for evaluation, run column analysis. If validation of sensitive data is important, check the enhanced data classification option when running.

5. Verify data classes for accurate data classification. When viewing and verifying column analysis results, you can accept or override the system-inferred data class for the selected column. These data classes are key components to drive subsequent analysis.
6. Analyze the detail analysis results based on data classifications with attention to particular conditions and appropriate annotations of fields.
   - **Identifiers** - Check for duplicates and invalid formats
   - **Indicators** - Check for invalid and inconsistent flag values or data skews; generates valid value reference tables.
   - **Codes** - Check for invalid and default code values; generate valid value reference tables.
   - **Quantities** - Check for data outside valid ranges; generate valid range reference tables.
   - **Dates** - Check for data outside valid ranges; generate valid range reference tables.
   - **Text** - Identifies text fields that are necessary for the data warehouse; looks for problematic formats; utilizes data rules to explore for known conditions.
   - **Sensitive data** - Looks for invalid conditions.
   - **Other conditions** - Utilizes data rules to extend analysis for other business-defined conditions such as valid value combinations, ordering of data, computational expressions, complex data formatting, and so forth.
7. During analysis, assign columns to identified business terms as established in IBM InfoSphere Business Glossary. This data-centric view helps to validate and ensure the correct data will be linked with the right business concepts.
8. Report and review detail analysis results. This can be an iterative review cycle with subject matter experts. Typical conditions found in this step include:
   - Gaps in data (for example, the data found does not align with what the data warehouse requires or no data exists to meet a particular requirement. The gaps might require additional sources to be brought into the discussion and necessitate an additional cycle through IBM InfoSphere Discovery).
   - Gaps in knowledge (for example, no one knows what a particular code means or what is considered valid)
   - Issues in data (for example, problems exist in the data that need to be addressed and documented, or rules need to be put in place)
9. Publish the analysis results. You can view an analysis result and publish it to the metadata repository. This publication updates the metadata repository with the additional details found through data quality assessment and can be directly used by developers working with IBM InfoSphere DataStage or IBM InfoSphere QualityStage to establish the correct load processes from the identified data sources to the target schema defined by IBM InfoSphere Discovery or the data warehouse.
10. Design new or reuse existing data quality validation rules and metrics based on the findings of the data analysis. Not all fields will be selected for ongoing validation.
11. Test and review the new data rules and metrics by using the planned source data. These rule validations will be targeted to the data warehouse needs. Results will be part of the iterative review cycle with subject matter experts and will be used to establish initial benchmarks. The rule validations will expand or resolve items noted in the "Report and review detail analysis results" step above.

12. Publish or deploy the data rules and metrics. These data rules and metrics will provide the foundation for ongoing data quality monitoring.

InfoSphere Information Analyzer fulfills a critical role in the integration process. To facilitate the most complete and accurate mapping specifications, profiling results from InfoSphere Information Analyzer are directly accessible from within IBM InfoSphere FastTrack, where the specifications are defined and documented. These specifications then become the input requirements for the InfoSphere DataStage and IBM InfoSphere QualityStage™ ETL and cleansing jobs that support the business application being developed. The more information the analyst has about the true data structures and content, the more accurate the requirements are for the downstream developers.

Related information

Rules approaches and methods

Key scenario outputs

Relationship analysis with IBM InfoSphere Discovery and data quality analysis with IBM InfoSphere Information Analyzer produces a number of outputs that facilitate data warehouse modernization including better mapping and integration to the new data structures, key reference tables for load validation, and quality controls for ongoing data quality monitoring.

The four key outputs provided from the data quality assessment process are: shared metadata including analytical results and reference tables; shared metadata including linkages to business terminology; and data validation rules for ongoing quality monitoring.

Using shared metadata to understand scope and impact of change

IBM InfoSphere Metadata Workbench provides impact analysis over the systems that are in the scope of the warehousing initiative. Affected upstream and downstream systems as well as the routines that connect them are easy to identify. The impact analysis helps to reduce the risk that any required changes are missed and that data flows break.

Modernization of a data warehouse can result in changes to the existing data warehouse structure, including updated tables, facts, or new data relationships. Where the changes are additive, meaning they did not previously exist, the downstream impact should be minimal. Where the changes modify the data warehouse structure, impact analysis is critical both at the business and the integration levels.

Business lineage allows downstream consumers to understand impacts to data marts and reports fed from the data warehouse, particularly if terminology or data content is changing.

Data lineage allows developers to understand impacts and necessary changes to the data integration jobs that feed downstream data marts, systems, or applications.

Using shared metadata in mapping and development

You fully understand the information you have in your environment. You know the source and targets shape, size as well as business terms. You have a firm understanding of the assessment and you have discovered your data relationships.

Next, you must decide how to map the new or updated data sources to the data warehouse target. Start with the knowledge you just learned and use it to help deliver the right information to the business. The approach here is to take the source data and ensure that it fits the shape and size of the target.

To facilitate the most complete and accurate development work, analytical results from both InfoSphere Discovery and InfoSphere Information Analyzer are directly accessible from within InfoSphere FastTrack, where the specifications are defined and documented, as well as within InfoSphere DataStage, InfoSphere QualityStage, and InfoSphere Metadata Workbench.

The mapping specifications in InfoSphere FastTrack can directly incorporate the data sources identified to feed the data warehouse as Source information. The unified schema, if in use, or a physical model generated from the data warehouse models, can be incorporated as the target information. By reviewing the analytical information, including annotations and identified values, you can identify the specific transformation rules to use to appropriately align the data from the source to the target. Annotated reference tables can be denoted for lookup processes.

Where you find gaps or missing information between the sources and the target, you might have to do the following actions:

- Validate differences in terminology that is associated with source and target data
- Request additional data sources
- Go back to the initial discovery cycle
- Review additional information regarding data quality requirements
- Go back to the data quality assessment cycle

These specifications then become the input requirements for the InfoSphere DataStage and InfoSphere QualityStage ETL and cleansing and data re-engineering jobs that support the data warehouse modifications being developed. The more information you have about the true data content, the more accurate the requirements are for the downstream developers.

Using shared metadata in the business glossary

As the new sources and information are brought into the data warehouse, you might have to consider questions about how to use that information.

The questions might include:

- What happens in this environment if I add a system and then have to change or deliver some information to the business as quickly as possible?
- What if the executives want to understand where all the financial risk data came from, or when it was last updated?

The process requires reporting and data governance to manage any changes, deliver any reports, and keep everyone sharing the same information. Through the common shared metadata where business terminology is linked to actual data assets (both data sources and the data warehouse), information can be quickly communicated with regard to these types of questions.

Whether you are using the IBM InfoSphere Business Glossary to understand the business lineage from the source to the warehouse or the IBM InfoSphere Metadata Workbench to review the details of the metadata environment, you can find shared information linking terms to data sources to analytical information, maintaining your new warehouse components, and providing information to those working with the new contents in the warehouse.

Monitoring ongoing data quality

With the data quality rules and metrics developed and deployed from the data quality assessment, you can review and track the quality of information in the data warehouse over time. Delivery and availability of information are key factors in decisions at this stage. Different delivery options exist from which to track and monitor data quality.

Using IBM InfoSphere Information Analyzer to monitor data quality

In this scenario, you are setting up your data quality monitoring activities.

1. Identify the preferred approach to delivery and monitoring of data quality around these identified rules and metrics. The options include the user interface, reports, or publication of results through command line interfaces (CLI) or application programming interfaces (API). The API can be used to incorporate your organization's specific reporting solutions, including the data warehouse itself.
2. Identify the appropriate execution schedule for the identified rules and metrics and their associated reports. Where processes are driven through the interface or reports, these processes can be scheduled through the IBM InfoSphere Information Analyzer job scheduler or the IBM InfoSphere Information Server reporting console. If processes are driven through CLIs or APIs, you must script or schedule those externally.
3. Assuming that some level of monitoring will occur through the InfoSphere Information Analyzer user interface and reports, you review summarized rule and metric results as measured against established benchmarks (thresholds).
4. Where variances are seen, drill into details on the specific rules or metrics. These results include summarized statistics for each execution of the rule or metric from which you can go into exception details.
5. Deliver data exports or reports to subject matter experts. New and unexpected conditions will arise in the data. These conditions will require review and remediation. Source data or processes might need modification. Information in the data warehouse might require correction, or the rules might require subsequent updates and modification.
6. As necessary, return to the data quality assessment process to re-analyze source data or modify rules and metrics to address changing conditions.

Ultimately, the additions to the data warehouse are intended to provide broader knowledge and greater insight into the business itself. In the case of adding customer information on top of financial information in the data warehouse, this might provide insights in customer buying or spending habits, high-value or high-revenue sales efforts, and so on. However, this information is only as good as the quality of the data provided to the data warehouse. Even where the data quality of the sources is considered high, the lack of consistent terminology, well-defined models, problems with data alignment, inconsistent values, or improperly mapped relationships can all generate downstream issues. Highlighting the resulting lineage of data is not sufficient to ensure effective utilization and information quality. By putting appropriate focus on data discovery, data quality assessment, and ongoing quality monitoring, trust in the resulting information and analysis can be significantly increased.

Related information

Rules approaches and methods

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 1. IBM resources

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

Providing feedback

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

Table 2. Providing feedback to IBM

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

Table 2. Providing feedback to IBM (continued)

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

Accessing product documentation

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

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

Accessing the information center

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

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

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

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

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

Obtaining PDF and hardcopy documentation

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

Providing feedback about the documentation

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

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

Product accessibility

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

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

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

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

Accessible documentation

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

IBM and accessibility

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

Notices and trademarks

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

Notices

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

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

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

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

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

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

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

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

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

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

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

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

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

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

Trademarks

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

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

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

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

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

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

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

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

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

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

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

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

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

Index

C

common questions
  data warehouse scenario 8
customer support
  contacting 19

D

data integration application 4
data quality
  data warehouse scenario 5
data quality and monitoring integration
  data warehouse 6
data quality assessment process 12
  data warehouse process 12
data quality monitoring process 16
  data warehouse scenario 16
data warehouse process
  data quality assessment scenario 12
  discovering source metadata process 10
data warehouse scenario
  data quality 5
  data quality monitoring process 16
  key scenario inputs 8
  questions to ask 8
  shared metadata in mapping and development 15
  shared metadata in the business glossary 15
discovering source metadata process
  data warehouse process 10

I

IBM Industry Models 7
integration
  data warehouse 6
integration scenarios overview 1

K

key scenario input processes 9
key scenario inputs 8
  data warehouse scenario 8

L

legal notices 25

P

product accessibility
  accessibility 23
product documentation
  accessing 21

S

shared metadata in mapping and development 15
  data warehouse scenario 15
shared metadata in the business glossary 15
  data warehouse scenario 15
software services
  contacting 19
support
  customer 19

T

trademarks
  list of 25

W

warehousing change impact 14
warehousing scope 14

Printed in USA

SC19-3474-00