What's new in IBM InfoSphere Information Server 8.7

Additional features and enhancements to enable smarter information integration and governance, and help speed time to value

Contents

Organizations across industries have discovered that success depends on trusted information. Business leaders and developers alike must be able to access this information quickly, to understand and analyze it and to take well-informed action. The delivery of trusted information requires strong collaboration between business and IT and a unified platform that can manage enormous volumes of data and derive information from diverse and often conflicting sources.

IBM® InfoSphere® Information Server is a market-leading data integration platform that enables organizations to meet their trusted information requirements. It helps integrate heterogeneous information from across an organization's IT environment, making it possible to consistently and reliably understand, cleanse, transform and deliver trusted information for critical business initiatives (see Figure 1).
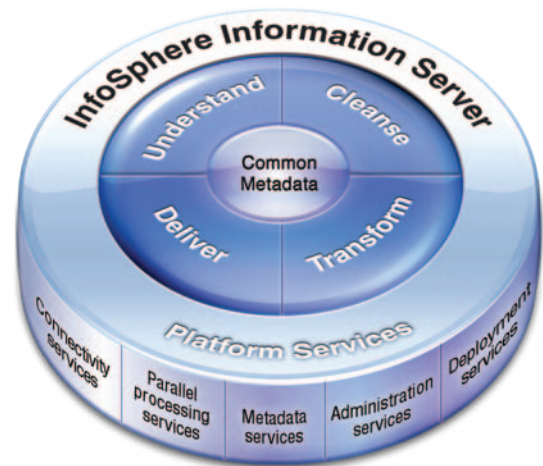

Figure 1: InfoSphere Information Server provides a foundation of parallel processing, shared metadata and other services that enables critical functions for creating trusted information.

InfoSphere Information Server facilitates the necessary collaboration between business and IT professionals so that strategic initiatives—such as business analytics, data warehousing, master data management and application consolidation and migration—utilize information that is accurate, comprehensive, insightful and available in real time.

With the introduction of InfoSphere Information Server 8.7, IBM helps organizations take the delivery of trusted information to the next level through the addition of more comprehensive information integration and governance capabilities.

InfoSphere Information Server 8.7: A more comprehensive and tightly integrated platform

InfoSphere Information Server 8.7 provides a smarter approach to information integration and governance. It makes high-quality information accessible, usable and manageable in new ways within a highly scalable, secure and robust data integration platform. InfoSphere Information Server 8.7 helps improve the operational management and governability of integration projects and information infrastructure—allowing organizations to turn insight into action in several ways:

- **Capitalize on productivity enhancements.** Obtain better business results and reduce implementation time and cost through features that enable the IBM data integration platform to do the heavy lifting instead of your developers and employees. Examples of this functionality include the smart management of metadata, new operational intelligence capabilities and new interactive parallel job debugging capabilities that enhance the productivity of developers.
- **Leverage comprehensive information governance capabilities.** Establish smarter, sustainable governance practices—including enriched data quality management— that meet your business requirements and support consistent business outcomes. Organizations can gain more control of their information and benefit from improved automation and streamlined workflows when managing business terms through a review and approval process; increased insight with holistic, personalized views of business terms, data models, assets, data flows and their relationships; and enhanced task management integration.
- **Harness the power of new and stronger integration.** Reuse assets and skill sets for increased efficiency; promote collaboration and streamline the process of delivering enterprise business solutions through tight and deep metadata integration with key systems. InfoSphere Information Server 8.7 includes enhanced integration with many data source systems, as well as integration with IBM Netezza® data warehouse appliances.
- **Use big data support to address challenging data volumes.** Address growing data volumes with big data tooling. InfoSphere Information Server 8.7 supports big data and Apache Hadoop, enabling organizations to take the results of big data analysis and leverage them across the enterprise.

This white paper describes what's new in the product modules that comprise InfoSphere Information Server 8.7.

Joint InfoSphere DataStage 8.7 and InfoSphere QualityStage 8.7 enhancements help increase productivity

The IBM InfoSphere DataStage® and IBM InfoSphere QualityStage® products provide core transformation and data cleansing capabilities, respectively, offered with InfoSphere Information Server. InfoSphere Information Server 8.7 adds two exciting capabilities that are shared by InfoSphere DataStage and InfoSphere QualityStage.

Interactive debugger for parallel jobs through the InfoSphere DataStage and QualityStage Designer client

InfoSphere Information Server 8.7 extends existing debug capabilities for parallel jobs by introducing an interactive debugger in the IBM InfoSphere DataStage and QualityStage Designer client integrated development environment (IDE). With this debugger, developers can establish breakpoints on the data flow links of their jobs to interrogate values as they are transformed at job runtime. The debugger fully supports the scalable parallelism of the InfoSphere Information Server engine, making this a unique technology to support a developer's investigative efforts.

The process is straightforward. The developer right-clicks any link to display the context menu, which has an option to toggle the breakpoint on or off (see Figure 2). Once a breakpoint is established, the default mode will cause a debug run to pause at every row passing through that link. A developer can edit the breakpoint and change this default behavior so that the job pauses at a different frequency or upon encountering a specified data condition, such as account_id = '12345'.

When the debug run pauses at a break condition, the developer can view the individual nodes where job stages are running, the
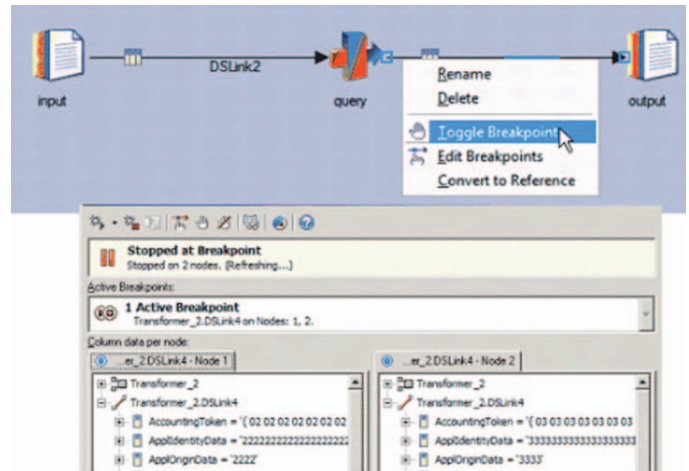

Figure 2: Establishing a breakpoint in the data flow link for debugging parallel jobs.

data that is on each node and the number of rows that have already passed through that link. During investigation, the developer can examine the properties of any stage of the job design and then alter the breakpoints based on new information that the data reveals. Developers can also add specific fields to a watchlist to focus on only fields of particular interest.

Because a job's debug mode executes using the same configuration file that the standard runtime uses, the debug run will catch specific data conditions that exist based on how data is being partitioned and distributed across the infrastructure. Consequently, the debug results will mimic the way a job will be deployed in an organization's production environment, regardless of whether the job is running on a single server or across an enterprise-class grid, or whether the databases accessed by the job are partitioned out to 1 node or 128 nodes.

InfoSphere DataStage and QualityStage Operations Console

Production operators, development teams and other stakeholders need direct access to the operational metadata of their data integration environments. Whether they are confirming that the environment is live, reviewing job run history or troubleshooting job failures, they must have relevant information to perform these tasks efficiently.

The new IBM InfoSphere DataStage and QualityStage Operations Console provides a view-only client for simple, web-based access to the runtime environment of InfoSphere Information Server. High-level graphs provide a summary view of the current state of affairs—including running processes, resource consumption and health of the environment (see Figures 3 and 4).

With a simple click or two, users can drill down into the supporting details they need to answer their questions and achieve higher levels of productivity. This interface includes the following key features:

- Graphical access that enables operators, developers and other stakeholders to view and analyze the runtime environment
- Simple, web-based, read-only access that is highly optimized for answering common questions
- Dashboard-style graphs representing current job activity, successful/failed jobs status, system health and resource consumption, including memory, processor and disk usage across the operational environment
- Visual cues that alert users to potential issues
- Server-wide and project-specific views based on user privileges
- High-level summary and detailed information for both current and historical activity
- Utilities to compare multiple job runs and filter relevant differences, which help accelerate investigation
- Fully documented schema to support integration of operational metadata with other applications

Figure 3: A summary view in the InfoSphere DataStage and QualityStage Operations Console.

Figure 4: Job performance and resource consumption over time.

To facilitate console access, InfoSphere Information Server 8.7 introduces the *operations viewer* role, which allows access to only the operations console. Organizations can assign this role to users and groups very generously, even in production environments, and trust that objects within InfoSphere Information Server will not be altered or executed. Additionally, in situations where operational metadata needs to be shared with other applications in the enterprise, organizations can use the console's fully documented metadata schema to define those integration points.

InfoSphere DataStage 8.7 enhancements for more intelligent and automated transformation

Using InfoSphere DataStage, organizations can design data flows that extract information from multiple heterogeneous source systems and transform it in ways that make it more valuable. The information can then be delivered to one or more target databases or applications through batch, real-time or event-based (messaging) processes, regardless of data complexity or volume. InfoSphere DataStage provides an intuitive graphical development interface and a highly scalable runtime environment to handle very demanding enterprise data integration requirements.

In addition to the enhancements below, InfoSphere DataStage 8.7 contains new connectivity, security and administration features that are discussed later in this white paper.

Balanced optimization for IBM Netezza data warehouse appliances

Balanced optimization enables data transformation in the optimal location for any given point in time: in the source database, in the high-performance engine of InfoSphere DataStage, in the target database or in any combination of these locations. The developer expresses the logic of the data integration processes in the current flow-oriented convention
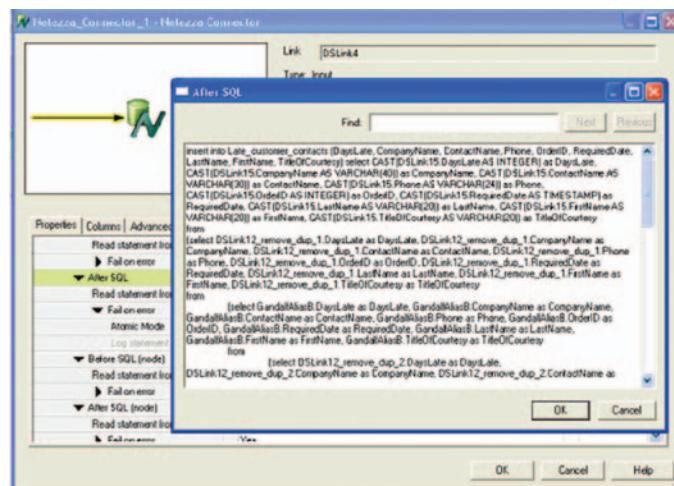

Figure 5: Custom SQL after balance optimization of a job through the IBM Netezza connector.

of InfoSphere DataStage, which then automatically optimizes the logic to benefit from pushing desired parts of the processing to the database.

With InfoSphere DataStage 8.7, the balanced optimization paradigm is extended to interoperate with Netezza data warehouse appliances through the new Netezza connector (see Figure 5). In a job that uses this connector, transformation processing may be pushed to the Netezza appliance, where the job takes advantage of the appliance's Snippet Processing Units (SPUs) and parallel processing capability. By keeping data in the appliance, data round-trip time is decreased. This functionality is particularly useful when data reduction can be achieved at the source database management system (DBMS) or when the source and target of a data integration task derive from the same DBMS. InfoSphere DataStage Balanced Optimization is an individually licensed feature.

Performance optimization

IBM continues to optimize the performance of the InfoSphere Information Server parallel engine. In InfoSphere DataStage 8.7, the partition and sort insertion algorithm has been improved for several design patterns. Additionally, the dataset sort information helps optimize any subsequent integration task that leverages that data. These improvements help reduce the sorting resource requirements in many jobs, allowing organizations to make better use of existing hardware resources.

InfoSphere QualityStage 8.7 enhancements for smarter, more automated data quality

InfoSphere QualityStage enables enterprises to create and maintain an accurate view of master data entities, such as customers, vendors, locations and products. It can be deployed in transactional, operational and analytic applications in batch and real-time environments. Its core capabilities for managing and maintaining data quality include:

- **Investigation:** Enables understanding of the nature and extent of data anomalies for effective cleansing and matching
- **Standardization:** Creates a standardized view of customer, partner or product data, which enables global address cleansing, validation and certification (for significant postal discounts in select localities), as well as geo-location
- **Probabilistic matching:** Provides an industry-leading matching engine to help ensure the best match results possible; built on a platform designed for high connectivity and scalability
- **Survivorship:** Helps ensure optimum consolidation, householding and linked views of record information; enables consolidated and accurate views of customers, partners and products

InfoSphere QualityStage 8.7 contains the following new features and enhancements that help establish smarter, more automated data quality.

Match Specification Setup Wizard

Developers can use the new Match Specification Setup Wizard to create match specifications for US name and address data and very rapidly build solutions for the most common customer use cases. The specifications created with the wizard provide criteria for matching records from one or two sources (see Figure 6). The traditional "from scratch" approach to match specification development still remains in the Match Designer. After defining a match specification using the wizard, developers should use the Match Designer to perform any testing and additional development for that match specification.
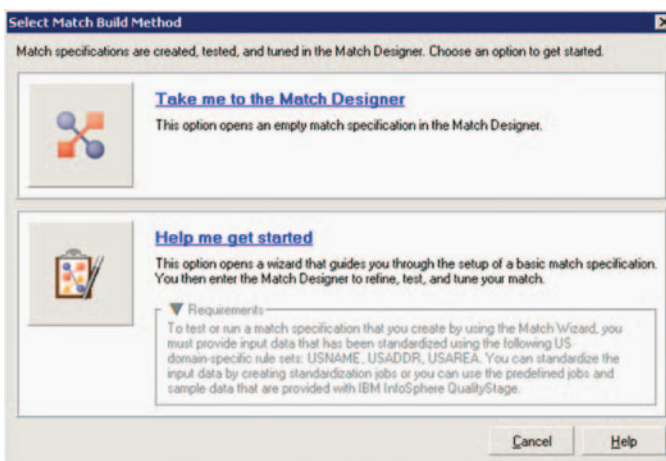

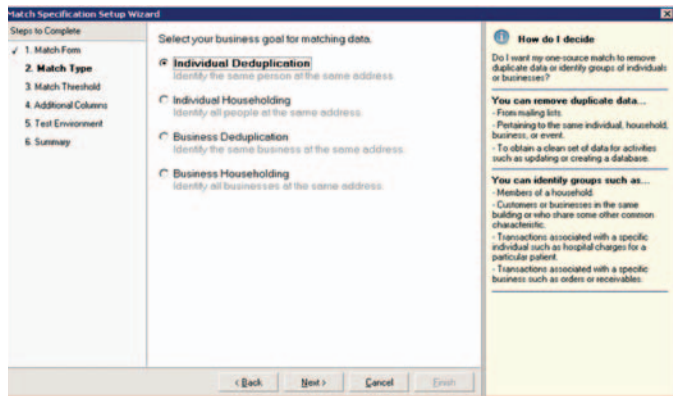*Figure 6*: Selecting the Match Specification Setup Wizard for building match specifications.

Figure 7: Selecting the appropriate match type in the Match Specification Setup Wizard in InfoSphere QualityStage.

The specifications support common matching goals for individuals and businesses, such as deduplication and grouping of records with similar values (see Figure 7). Developers can use the specifications in the new predefined matching jobs (see the next section of this white paper) or as models for specifications that meet an organization's matching requirements.

Predefined standardization and match jobs

To accelerate time to value, InfoSphere QualityStage 8.7 introduces predefined standardization and match jobs for US name and address data. The predefined standardization jobs generate standardized data and frequency information for that data. The predefined match jobs match records from one or two sources of input data (see Figure 8). These predefined jobs help newer users quickly and easily acclimate themselves to the platform's design canvas.
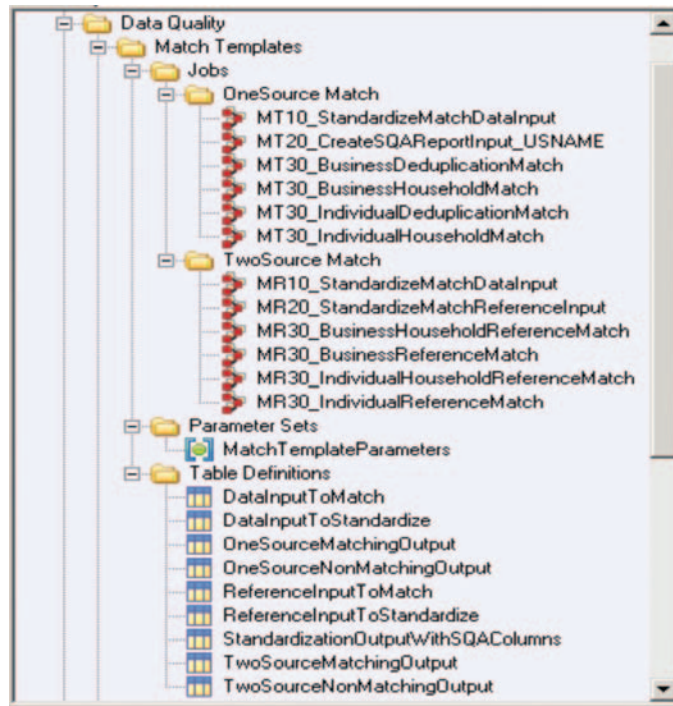

Figure 8: New predefined match jobs in InfoSphere QualityStage.

Sample data is provided for use with the predefined jobs. When opening a predefined job, the user will be presented with a complete job design. All the user needs to do is define the data sources and targets and test the data flow within the job.
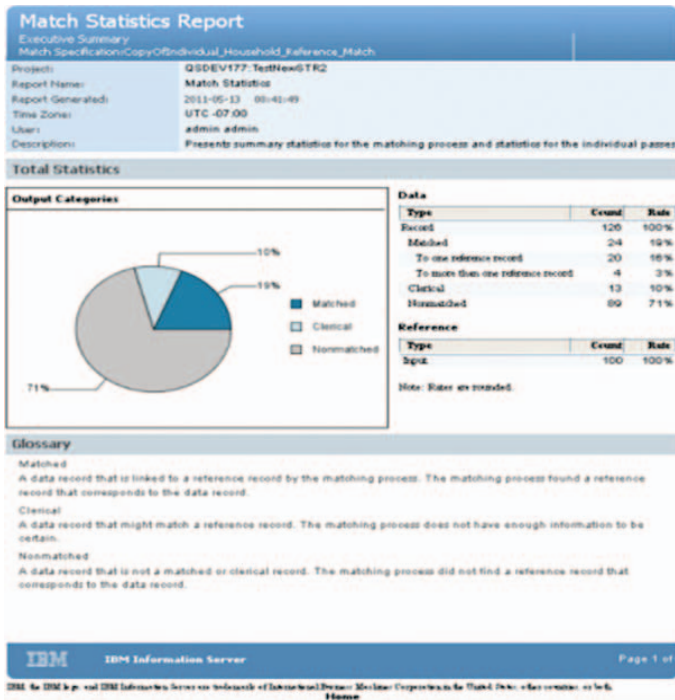
Figure 9: Summary view provided by the Match Statistics report in InfoSphere QualityStage.

Match Statistics report

The new Match Statistics report presents summary-level statistics about the matching results and statistics about the matching process for each match pass (see Figure 9). Although this information was previously provided as part of the match specification design process within the Match Designer, it is now available in production Match stage jobs for enhanced ease of use.

India rule set

Developers can use the new India rule set in jobs that standardize Indian address and area data. A shared container is imported with the rule set to help developers design a job that uses the rule set. This rule set is located in the same folder as the other country-level rule sets offered out of the box with InfoSphere QualityStage.

InfoSphere Information Analyzer 8.7 enhancements for more proactive and automated control of data quality

IBM InfoSphere Information Analyzer delivers data quality analysis functions within the context of a complete information integration platform, helping increase data accessibility and consistency throughout the enterprise. Active metadata in InfoSphere Information Server simplifies sharing and managing the metadata (that is, the quality analysis results) in the integration environment. Within InfoSphere Information Analyzer, quality analysis results are stored in the common metadata repository.

By validating the data with InfoSphere Information Analyzer in the early phases of data integration projects, developers can:

- Expedite the timely delivery of data-driven projects and trusted information
- Utilize business-driven rules analysis with reusable construction and application across multiple data sources to offer quick time to value
- Help minimize costs and resources of critical data integration projects
- Avoid the risk and impact of proliferating incorrect and inaccurate data

The following new capabilities in InfoSphere Information Analyzer 8.7 provide more proactive and more automated control of data quality than previous versions.

InfoSphere Information Analyzer Data Rule stage for DataStage and QualityStage

InfoSphere Information Analyzer 8.7 data rules are more pervasive within the data integration environment than in previous releases. The new Data Rule stage extends existing capabilities to develop, execute and monitor information that is persisted in a data store, and apply those rules to data that is in transit through a data integration task. The new Data Rule stage allows developers to use the InfoSphere DataStage and QualityStage Designer client to access existing rules (ones already created in InfoSphere Information Analyzer) and embed those rules into their jobs.

Using the Designer client to view a data rule in an InfoSphere DataStage job, a developer immediately sees the number and validity of analyzed rows (see Figure 10). The developer can then choose to process the invalid records differently than the valid records for reporting or proceed to automated remediation.

The user interface for this Rule stage is integrated into the Designer client. It allows developers to view and modify an existing rule definition or to define a new rule from scratch. For example, developers can determine which rule variables should be mapped to which input links (see Figure 11).
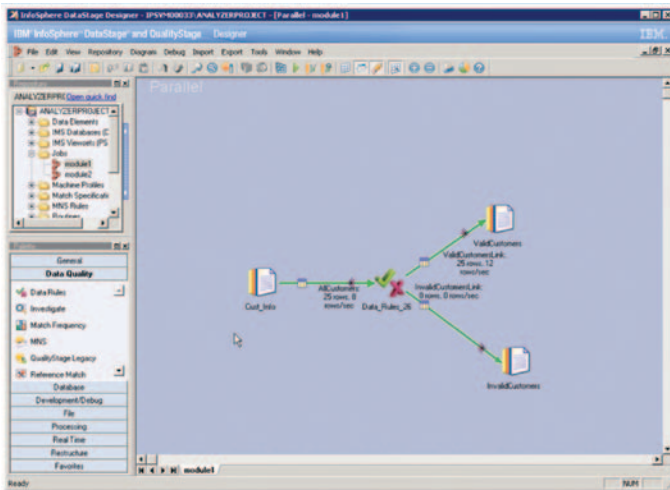

Figure 10: A data rule in an InfoSphere DataStage job, as viewed from the InfoSphere DataStage and QualityStage Designer client.
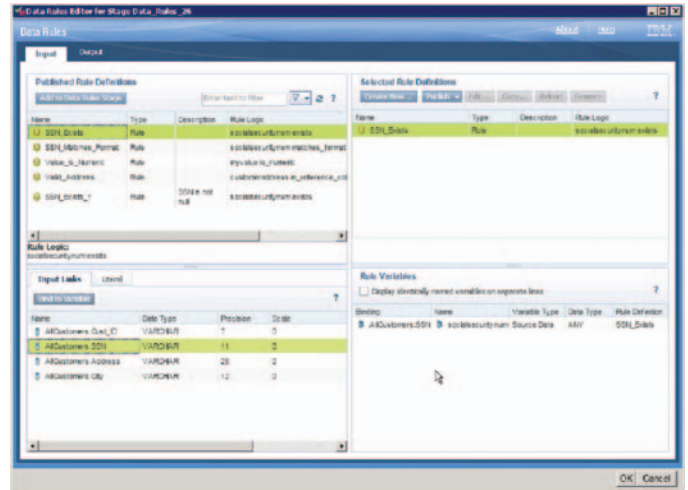

Figure 11: Mapping rule variables to input links.

The same rule definition—that is, the logical representation of a business rule such as "age > 18"—now can be applied to data stored in a repository or data being processed in an InfoSphere DataStage or InfoSphere QualityStage job. It can even be applied as a service. This functionality gives developers additional flexibility in determining how best to monitor data quality. For example, developers may choose to execute the rule first in the warehouse. If bad data causes business disruptions, they can take the same rule, embed it in a job that loads the warehouse and avoid that bad data before it enters the warehouse.

API enhancements for flow-through processing

InfoSphere Information Analyzer 8.7 adds several application programming interfaces (APIs) to its set of interfaces to support flow-through processing:

- Folder support
- Metrics and benchmarks
- Key and cross-domain analysis
- Public rules

Data rules output enhancement

This enhancement allows distinct records to be generated in rule output when the rule contains only aggregation functions. Only one record is output per grouping value.

Support for separation of logical, physical and implementation data models

As part of InfoSphere Information Server, InfoSphere Information Analyzer supports the use of physical and logical model assets in rule definitions and a facilitated approach to discover and bind associated assets. In inferred model generation, physical model generation is enhanced for an integrated approach. Developers can also use logical data model elements when specifying data rules to further enhance the reusability and consistency of data quality validation rules across the enterprise.

Performance enhancements

InfoSphere Information Analyzer 8.7 delivers multiple performance enhancements:

- Rules using join conditions are accelerated by pushing the join operations to the SQL select statement executed on the analyzed source whenever possible, instead of loading all the data and joining it in InfoSphere DataStage.
- Rules using a sampling are accelerated by reducing the number of rows that must be loaded from the source database to InfoSphere DataStage.
- When using the IBM DB2® database as the analysis database, developers can configure DB2 compression to reduce the storage footprint of the InfoSphere Information Analyzer database. Although query performance is not affected, updates and index management incur performance overheads with DB2 compression enabled.
- The rule execution engine was reengineered and optimized for speed. In benchmarking work done by the IBM team, rule execution performance in InfoSphere Information Analyzer 8.7 improved by 16 to 90 percent when compared against InfoSphere Information Analyzer 8.5.
- Performance enhancements for key user interface areas include open project, open project properties and the home page.

InfoSphere Blueprint Director for creating smarter data governance processes

IBM InfoSphere Blueprint Director is delivered with InfoSphere Information Server and IBM InfoSphere Optim™ products. InfoSphere Blueprint Director helps business and IT professionals create a collaborative project lifecycle, starting from the vision and moving through execution to completion. Using InfoSphere Blueprint Director, the team can collaborate on actionable information blueprints that connect the business vision for the information project with the corresponding business and technical artifacts. InfoSphere Blueprint Director then guides the team to apply recommended practices based on reference architectures and methodologies.

InfoSphere Blueprint Director delivers several new features and enhancements in the InfoSphere Information Server 8.7 release for creating smarter data governance processes.

Integration with task management

InfoSphere Blueprint Director now enables users to take an architecture diagram and understand not only the associated method (that is, the best practice for how to develop the architected solution), but also the associated tasks and their status. The team can select the relevant scope of a project's entire method content and export it to the IBM Rational Team Concert™ collaborative lifecycle tool. Note that method steps are associated with elements in a blueprint and that this association will be leveraged in subsequent steps. For each selected method element, InfoSphere Blueprint Director generates corresponding entries in Rational Team Concert. InfoSphere Blueprint Director also associates the tasks in Rational Team Concert back to a blueprint.

For example, a team member associates data source elements Source A and Source B in a blueprint diagram with method step "assess source system data quality" and selects this method step to be exported to Rational Team Concert. After Rational Team Concert generates the task that corresponds to the method step, the team member will see a link from the warehouse element in the blueprint to the task that was just created. This process allows the team to understand the progress of the project in two ways: from a pure project planning perspective and in the context of the larger architecture diagram. The team can also access the traditional project tracking dashboards provided by Rational Team Concert.

Once the tasks in Rational Team Concert are assigned to specific users, team members can use InfoSphere Blueprint Director to see which tasks are assigned to them and to identify and follow the associated best practice (see Figure 12).
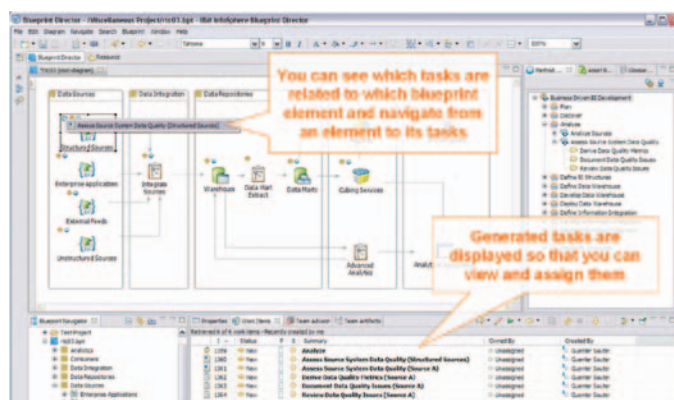

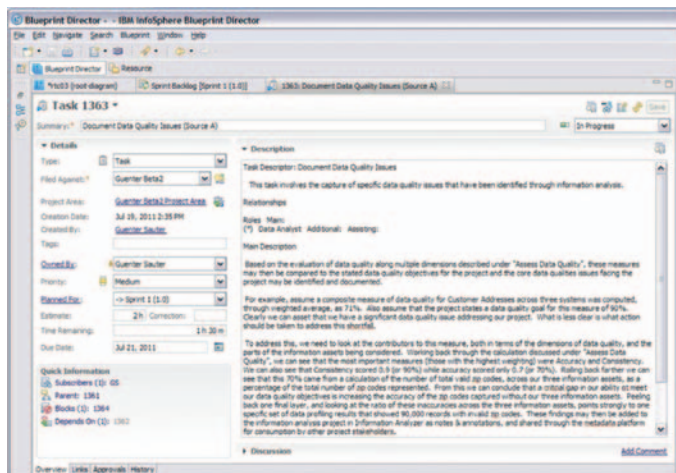*Figure 12*: Viewing tasks and their related blueprint elements in InfoSphere Blueprint Director.

Figure 13: Editing task details in InfoSphere Blueprint Director.

Figure 14: Using new milestone planning capabilities in InfoSphere Blueprint Director.

Users can also update their status easily within the system—for example, when they complete a task (see Figure 13).

Any changes—such as task status update, task delegation or addition of notes—are immediately visible to all other team members. This functionality helps ensure that the team has access to the latest status without the need for additional communication.

Task management integration can increase the level of governance being applied, including the ability to audit compliance with best-practice methods for developing projects. These capabilities also help increase decision-making agility because project status is readily available.

Milestone planning

InfoSphere Blueprint Director introduces additional capabilities in InfoSphere Information Server 8.7 to help teams explicitly define the evolution of their architectures over time. Many projects deliver the future state over time in multiple phases, with milestones for each phase. For example, over time a warehouse or master data management solution may add various sources and more consumers, such as data marts or reports. In this release, team members are now able to specify which elements become effective—or should be decommissioned—at which milestone (see Figure 14).

After defining the milestones and associating the blueprint elements with the milestones, team members can use a time slider to view the evolution of a blueprint. By simply selecting the milestones or sequence of milestones, the team can see the changes in the architecture over time. This capability gives the team enhanced visibility into the project road map and demonstrates how the information supply chain will be affected by the project. The increased level of awareness can help the team agree more quickly on both the vision and the path to realizing the vision.

Support of versioning through integration with source code control systems

InfoSphere Blueprint Director can interface with source code control systems (SCCSs) such as Rational Team Concert to help teams collaborate and share blueprints. Rather than sending blueprints by email, users check blueprints into an SCCS, where versions can be created and tracked. This functionality is highly valuable when teams are developing and discussing blueprints in a collaborative environment. It allows teams to have a controlled, consistent and governed approach when modifying and viewing blueprints.

Additional templates for lifecycle management and data quality

In this release, InfoSphere Blueprint Director introduces two new templates: the Managed Data Cleansing template and the information lifecycle management (ILM) template. The Managed Data Cleansing template provides a reference architecture and an associated method for deploying an example data warehousing and business intelligence solution. This template assumes the presence of several existing information sources with unknown structure and integrity, and it details the patterns involved in integrating information from these sources into a unified data warehouse. Trusted information from this warehouse is then made available to a range of analytics and reporting patterns and to information consumers.

The ILM template provides a logical blueprint and associated project-level implementation methodology for the development of processes supporting ILM. Key ILM areas addressed include data growth management (archiving) and test data management (subsetting), which covers data masking and application retirement.

Integration with other Eclipse applications

InfoSphere Blueprint Director is now integrated with other Eclipse applications that support shell sharing and can run in the same application window. For example, InfoSphere Blueprint Director can be shell shared with Optim Designer, IBM InfoSphere Data Architect or Rational Team Concert. Shell sharing streamlines navigation between a blueprint and the other integrated applications.

InfoSphere Business Glossary 8.7 enhancements for sharing common language and understanding

IBM InfoSphere Business Glossary helps create, manage and share an enterprise-wide controlled vocabulary that acts as the common language between business and IT. Establishing a common language is a critical step in aligning technology with business goals. The InfoSphere Business Glossary hierarchy and classification systems provide additional business context.
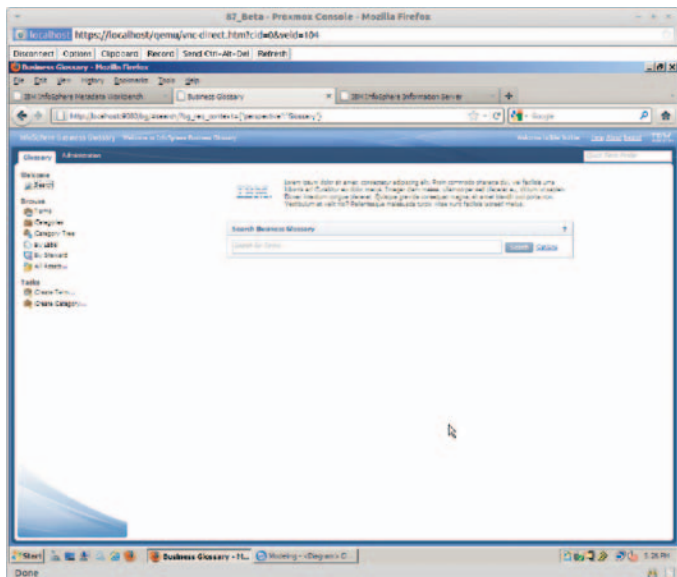
Figure 15: Suitable menu items displayed per user role in InfoSphere Business Glossary.

Figure 16: In-line editing and auto-suggest list for finding terms and repository objects.

InfoSphere Business Glossary 8.7 contains new and enhanced features to simplify usage, help business and IT professionals work intelligently and accelerate business results.

Enhanced user interface
This release of InfoSphere Business Glossary features a single URL entry point for all web-based activities, including authoring, administration and viewing. The single interface helps ensure a consistent experience for all InfoSphere

Business Glossary user roles (see Figure 15). In particular, this feature enables authors to create and edit glossary content in the context of how it will be viewed.

In-line content authoring
The enhanced user interface includes a new authoring experience that is based on in-line authoring and auto-suggest picker lists for quick content authoring and relationship building (see Figure 16).
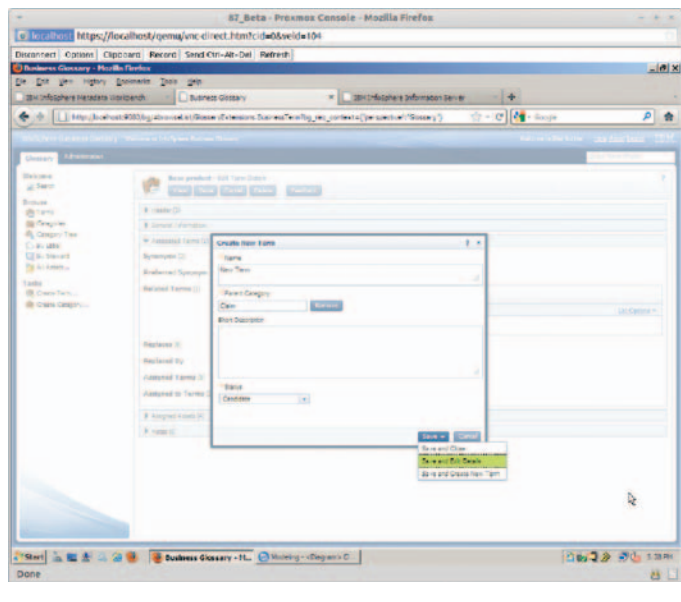
Figure 17: Quick create for categories or terms.

To create a new category or term, the author can simply perform a quick create, which enables new content to be created from within any other page (see Figure 17). This capability is designed to help authors define terms that are synonyms or are related to a term that an author is editing.
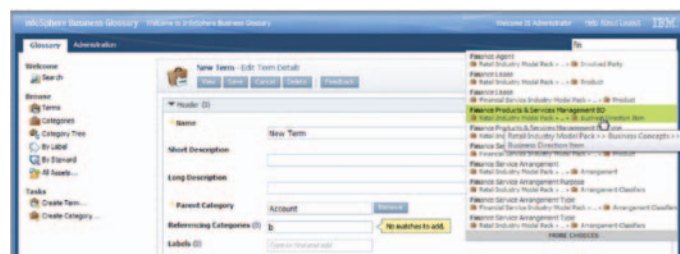

Figure 18: Auto-suggest for finding terms quickly.

Quick term finder

InfoSphere Business Glossary 8.7 includes a universal quick term finder for locating terms rapidly from any InfoSphere Business Glossary page (see Figure 18).

Development workflow with approvals and development glossary

This release supports a governed content-development workflow for the business glossary that includes the following new features:

- A separate development glossary where authors can work on the next version of the glossary without affecting the published glossary
- Separate editors and publishers, including authoring permissions per category (see Figure 19)
- Approval workflow with comments, accept, reject and publish actions (see Figure 20)
- Task list views of drafts, pending approval per author and approver (see Figure 21)
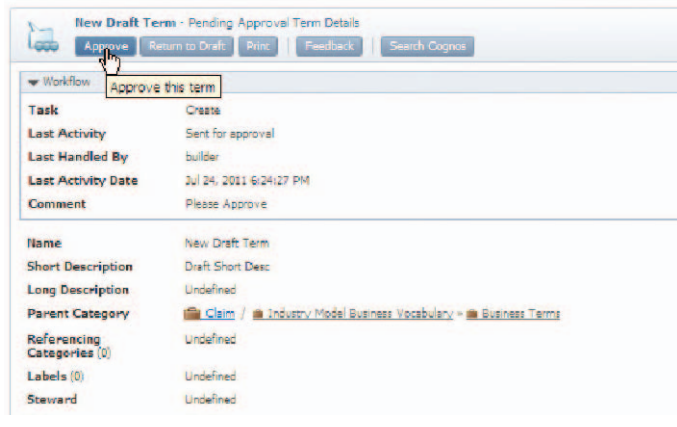
Figure 20: Approval or rejection from the details page.
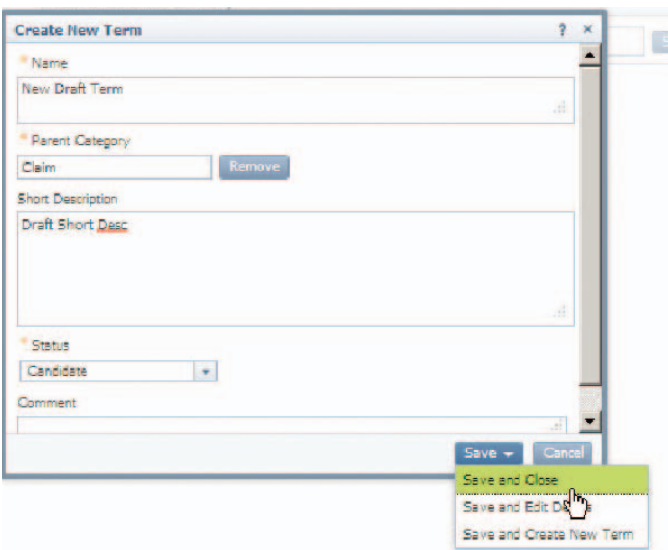
Figure 19: Modification of the development glossary by authors with permission to edit or create content in a category.
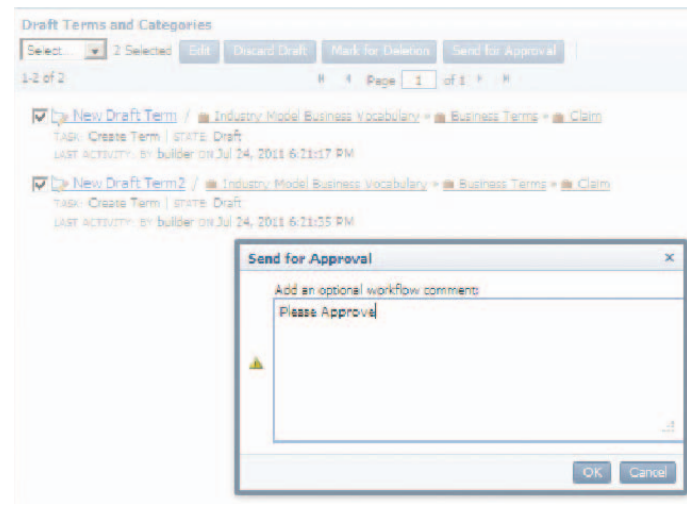
Figure 21: View of draft and pending approvals.

Figure 22: Adding criteria to narrow a list.

List filter mechanism

Large lists can be navigated easily by using the new filter mechanism. These filters, available from any list that exceeds a single display page, enable the author to add multiple criteria for narrowing the list (see Figure 22).

Business labels

Business labels are additional classification mechanisms that enable authors to assign meaning to repository assets through administrator-defined labels. These labels appear on the details page of any repository asset in InfoSphere Business Glossary and IBM InfoSphere Metadata Workbench and can be used in the advanced search, filter and metadata workbench query mechanisms for finding and reporting on particular items. Authors simply select a label to view all repository assets that are assigned to that label (see Figure 23).
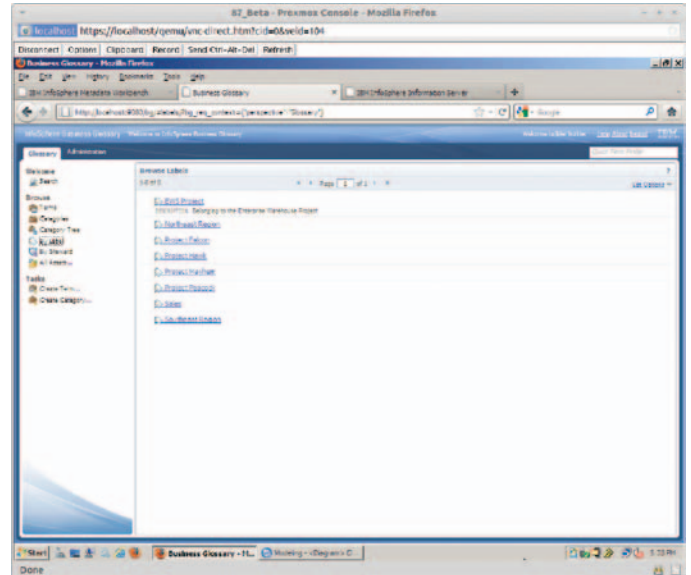
Figure 23: Viewing a labels list.

Logical model–to–business glossary content builder utility

In addition to supporting logical models in the metadata repository, the InfoSphere Business Glossary 8.7 release includes the ability to transform imported logical models into business glossary categories and terms. Through the InfoSphere Metadata Asset Manager feature, InfoSphere Business Glossary supports the import of logical models from InfoSphere Data Architect, Sybase PowerDesigner, Embarcadero ER/Studio and CA ERwin.

Separation of logical, physical and implementation data models

InfoSphere Business Glossary displays the new logical models available in InfoSphere Information Server 8.7, as well as distinct representations of physical models and implemented data resources. These three distinct layers are displayed in the new All Assets page of InfoSphere Business Glossary (see Figure 24), and relationships between these layers are displayed in each asset's details page (see Figure 25).

Fine-grained export

Whereas the 8.5 release supported export of only top-level categories, the 8.7 release includes the ability to export any category of InfoSphere Business Glossary content. This enables administrators to work with scoped areas of content for export and re-import.

The export procedure is now enabled through an export wizard, which helps simplify the overall process of content and file format selection (see Figure 26).

Figure 25: Implementation details about an object.

Figure 24: Different models and implanted schemas in InfoSphere Information Server as displayed by the All Assets page.

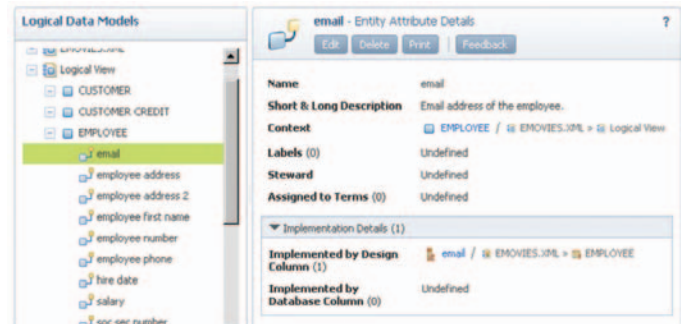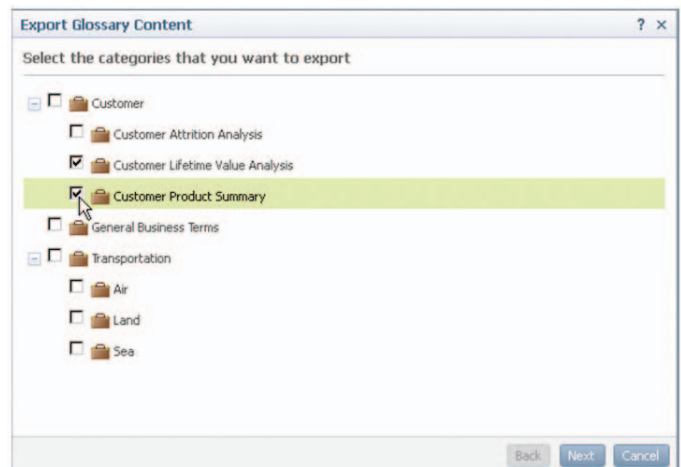*Figure 26*: Selecting content type through the export wizard.

Support for reports on third-party usage tracking

Also new in the 8.7 release, InfoSphere Business Glossary web pages now enable third-party web analytics to capture usage statistics (see Figure 27). These analytics tools can report visitor trends, geographic distribution, pages viewed, browsers used, navigation paths and time spent on pages.

InfoSphere Metadata Workbench 8.7 enhancements for a more comprehensive understanding of data

InfoSphere Metadata Workbench provides a powerful metadata management interface that supports not only InfoSphere Information Server metadata but also other metadata that plays critical roles in data integration processes. A centralized and holistic view across the entire landscape of data integration processes, with visibility into data transformations that operate inside and outside of InfoSphere Information Server, arms businesses with critical information that can lead to better decisions.

Organizations can gain control of information and benefit from automation when managing metadata with InfoSphere Metadata Workbench 8.7. This version contains several new features and enhancements to help organizations gain a comprehensive and integrated understanding of their data.

Enhanced data lineage and impact analysis

Upon invoking data lineage or impact analysis from the InfoSphere Metadata Workbench, developers can display the lineage graph with a single click to render all downstream and all upstream assets. In the previous release, developers were prompted to select the direction—"Where did data come from" or "Where is data going"—and then select a single target asset for analysis. Enhancements in this release allow an information asset to be displayed with its full name and identity, making it easier to identify (see Figure 28).
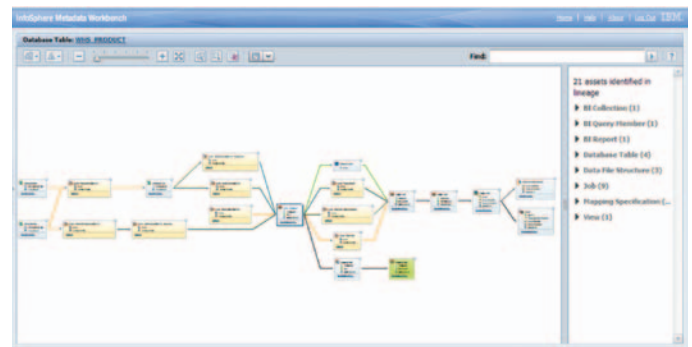

Figure 28: Display of information assets with full name and identity.

Figure 27: Third-party usage statistics enabled in InfoSphere Business Glossary web pages.

Initially, only InfoSphere DataStage and InfoSphere QualityStage job assets are shown in the lineage, which masks the complexity and integrated data flow from within a job (see Figure 29). Developers and advanced users may opt to expand a job so that they can view the data flow within the selected job.

Developers may also opt to refine the data lineage graph and select individual columns or fields of a table, file or report (see Figure 30). This feature allows them to further investigate the data flow for those specific assets.

The Inventory pane of the data lineage report continues to display additional information for each asset, which now includes custom attributes, logical or physical model relationships, column types and expressions (see Figure 31). In addition, the inventory can now be published as a PDF report for distribution.
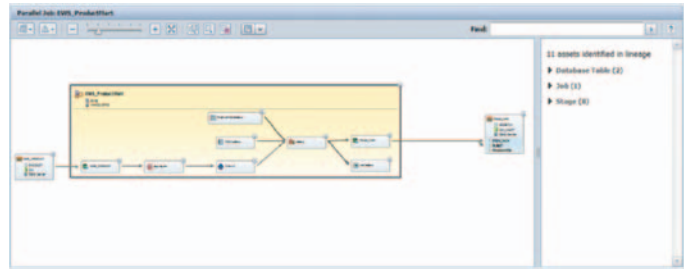

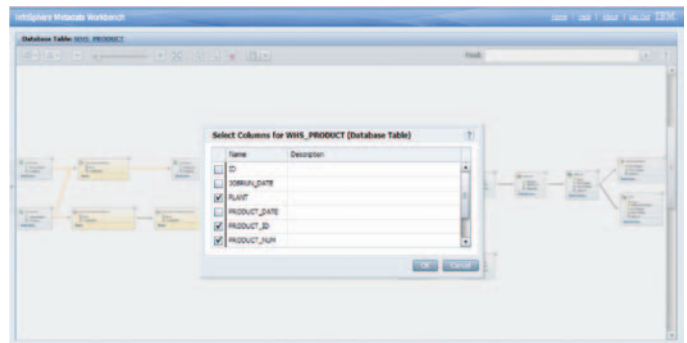*Figure 29*: Initial masking of job complexity.

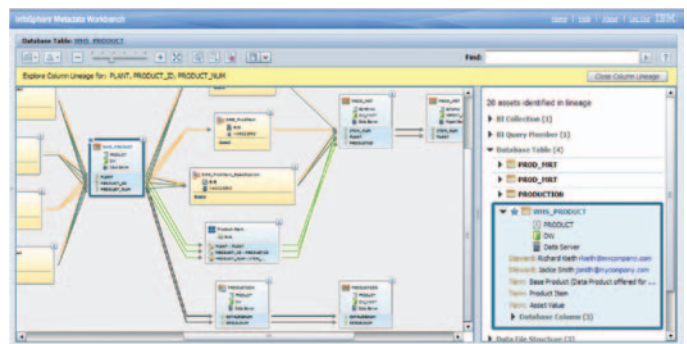*Figure 30*: Refining the data lineage graph.

Figure 31: Viewing additional information for an asset within the data lineage report.

Custom attributes

InfoSphere Metadata Workbench 8.7 enables the declaration of custom attributes for physical assets, so developers can document additional information such as deployment, technical owner and security status. Custom attribute definitions may be of type text, date or number; include multiple values; be restricted to a set of defined values; and be copied to child assets. Custom attribute definitions are managed within InfoSphere Metadata Workbench and may be imported, exported, edited or deleted (see Figure 32).

Custom attribute values can be declared within a comma-separated value (CSV) text file and imported through InfoSphere Metadata Workbench, so developers can easily script and apply the values to their imported physical assets. By editing physical assets within the workbench, developers can set descriptions or custom attribute properties.

Generic lineage framework: Integration with IBM InfoSphere Warehouse

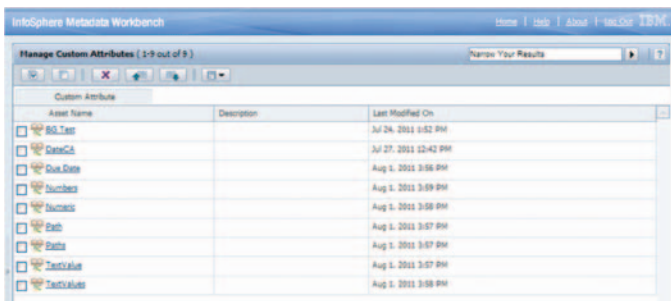InfoSphere Metadata Workbench 8.7 supports the use of IBM InfoSphere Warehouse, which delivers mapping documents depicting and documenting data flow between tables designed and implemented within InfoSphere Warehouse to InfoSphere Metadata Workbench. InfoSphere Metadata Workbench allows developers to generate lineage reports accounting for this data flow and to browse and search across all related mapping documents.

Expanded information asset types

While business labels are created and managed using InfoSphere Business Glossary, developers can now view, query or assign those labels using InfoSphere Metadata Workbench 8.7 (see Figure 33).
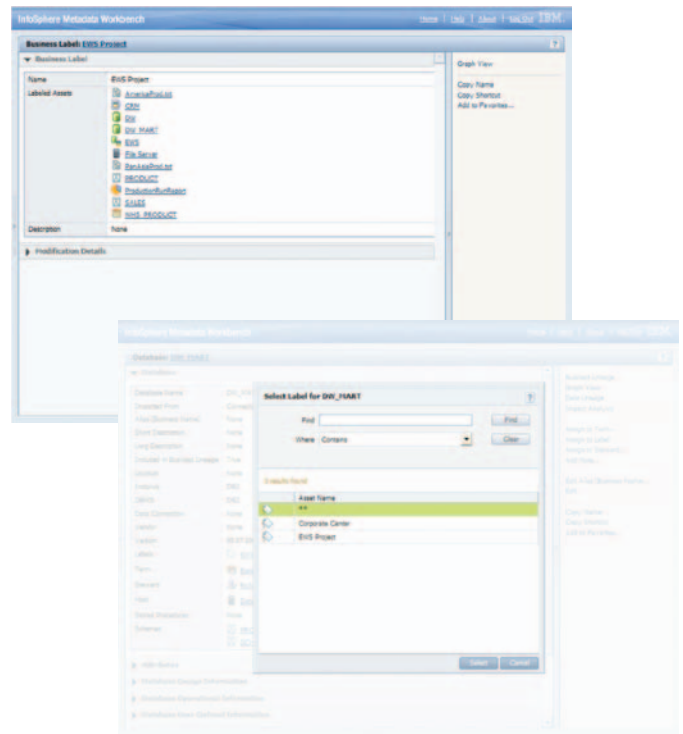

Figure 32: Managing custom attributes.

Figure 33: Using InfoSphere Metadata Workbench to view labels created in InfoSphere Business Glossary.

Display of logical and physical models

In InfoSphere Metadata Workbench 8.7, imported logical and physical data models are displayed with the implemented relationships between them. In addition, this release differentiates between physical data models and implemented physical data resources.

The addition of logical models and the distinction between physical data models and implemented physical data resources provide a consistent, meaningful hierarchy of database design and implementation. The increased visibility into how data sources should be designed and implemented helps create a strong chain of understanding and enforcement from design through implementation.

Published data rules from InfoSphere Information Analyzer

When defined and published from InfoSphere Information Analyzer, data rules now can be viewed or queried from InfoSphere Metadata Workbench (see Figure 34). Users can view rule expressions to obtain insight into the types of data quality processes that have been defined.
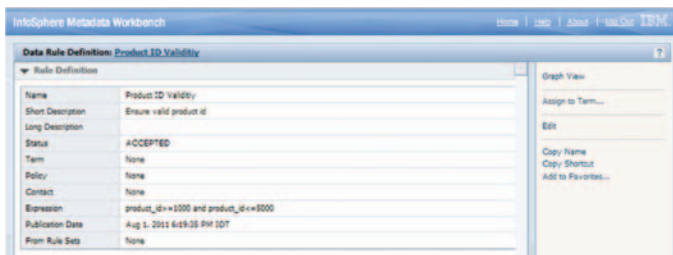

Figure 34: Using InfoSphere Metadata Workbench to view data rules defined and published in InfoSphere Information Analyzer.

Furthermore, querying this type of information provides key governance metrics and even regulatory reports that can display the data quality properties for data elements—both policy definitions and analysis results in a single report.

Display of BI Server and Folder

In the 8.7 release, InfoSphere Metadata Workbench displays the BI Server and Folder containment elements, providing intuitive and meaningful navigation of imported BI metadata.

Dialog to manage data lineage

InfoSphere Metadata Workbench now includes an updated manage dialog for selecting InfoSphere DataStage projects and invoking automated services, stage binding or data alias mapping (see Figure 35).
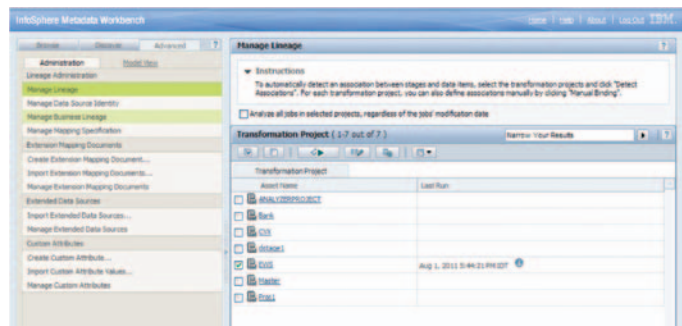

Figure 35: Selecting InfoSphere DataStage projects to manage data lineage.

InfoSphere Metadata Asset Manager 8.7 for more proactive and automated control of metadata

The new InfoSphere Metadata Asset Manager in the 8.7 release helps organizations manage shared suite metadata (including database and data file assets, logical and physical data models, and business intelligence assets) in a more automated and controlled manner.

InfoSphere Metadata Asset Manager is part of the InfoSphere Information Server platform and provides capabilities that include importing, reviewing and publishing metadata from a wide range of sources and managing metadata that already resides in the Information Server repository.

Logical model support

The suite tools and metadata repository of InfoSphere Information Server now support logical data models and physical data models and their contents, including logical entities and attributes, design tables and columns. If possible, relationships between logical and physical data models and the schemas that implement them are set automatically upon import. You can use the Repository Management tab of InfoSphere Metadata Asset Manager to set additional implementation relationships.

Import

InfoSphere Metadata Asset Manager centralizes the metadata import in InfoSphere Information Server into a web interface. You can import metadata to a staging area and inspect it before you share it to the metadata repository.

This centralized import functionality allows you to:

- Use both bridges and connectors for metadata imports
- Analyze and manually compare, re-import and preview the effect the import will have on existing metadata assets before you share it to the metadata repository
- Benefit from one-click access to InfoSphere Metadata Workbench (if installed) to view specific existing assets that might be affected by the import
- Maintain metadata throughout its life cycle, keeping a record of the contents of all imports and re-imports to the staging area and to the metadata repository
- Eliminate any problem assets, such as duplicates, before they are shared to the metadata repository
- Run express imports that automate the analysis and preview and automatically import directly into the metadata repository

Repository Management

The Repository Management tab has been moved into InfoSphere Metadata Asset Manager from its former location in the web console. In addition to the ability to browse and search common metadata in the metadata repository, the Repository Management tab has new functionality that allows you to:

- Experience improved management of duplicate assets, including the ability to merge or delete duplicates.
- Display and delete any assets that have been disconnected from their parent assets.
- Set implementation relationships between logical data models, physical data models, and database schemas or data files. These relationships let you trace a column that is used by an InfoSphere DataStage and InfoSphere QualityStage job back to the design column and logical entity that originally defined the column.

InfoSphere Information Server 8.7 integration with databases and files for comprehensive connectivity

Integration with databases and files is expanded and enhanced in InfoSphere Information Server 8.7. The new integration options promote collaboration and streamline the process of delivering enterprise business solutions.

Support for big data

InfoSphere Information Server 8.7 adds the Big Data File stage to help organizations address the challenges of increased data volumes through the user-friendly InfoSphere DataStage development paradigm. InfoSphere Information Server supports the Hadoop Distributed File System (HDFS), so users can access data directly on HDFS (see Figure 36).
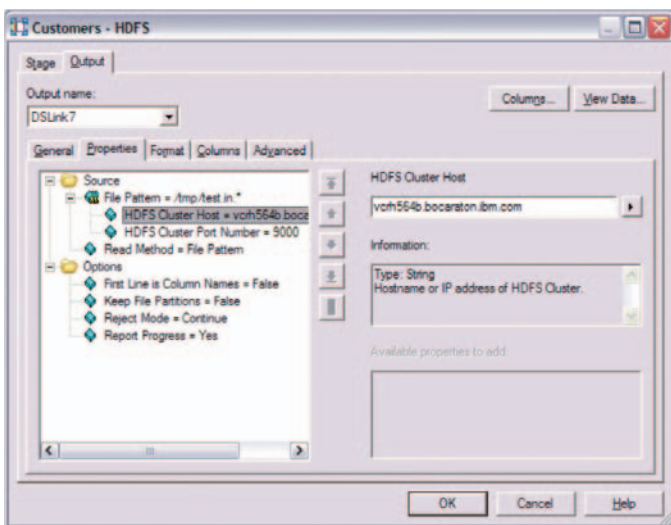
Figure 36: Accessing data on HDFS.

The Big Data File stage mirrors the Sequential File stage experience, allowing users to ramp up rapidly. The Big Data File stage enables multiple files to be read in parallel, either through a specific list or through file patterns, which simplifies how data is merged into a common transformation process.

Additionally, the InfoSphere DataStage parallel engine can mimic the same degree of partitioning as the big data file, or the engine can dynamically repartition that data on the fly based on business requirements. These features, which are also available in InfoSphere Information Server 8.5, Fix Pack 1, offer flexibility and optimization as organizations address their big data challenges.

Connector for Netezza appliances

The new InfoSphere Information Server connector for IBM Netezza Performance Server data warehouse appliances enables high performance and functionality by providing InfoSphere Information Server with native connectivity to IBM Netezza appliances (see Figure 37). The Netezza connector uses the massively parallel, highly scalable and dynamically configurable framework of InfoSphere DataStage to access each SPU in the IBM Netezza data warehouse appliance for optimum performance. For example, the connector supports performance loading by using the external table load method, which can be on the same database or on an external database.
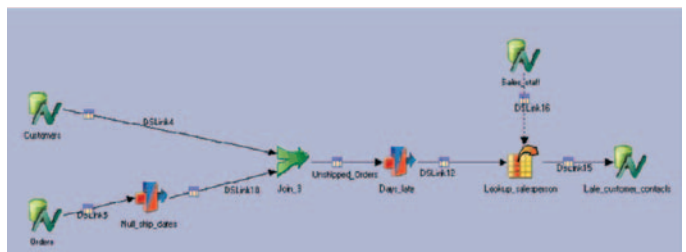

Figure 37: A view of the IBM Netezza connector as part of the InfoSphere DataStage framework.

The connector also includes the following features:

- Support for update-then-insert operation to speed writing to disk
- Creation of tables using random or user-defined distribution keys
- Row-selected SQL operation, which dynamically chooses the SQL operation at runtime based on the actual data
- Support for two-level and three-level reference object naming
- Use of various write modes, table actions and transaction modes
- Support for user-defined SQL; user-defined UDX can be used in a SQL statement by loading from a flat file, or SQL that is predefined can be used
- Use of properties for before and after SQL statements that can run across all runtime nodes or one node; various post actions available if the statement fails
- Strong out-of-the-box performance rates for loading and unloading data; IBM internal lab performance tests measured rates comparable to Netezza command-line bulk load utilities

Additional connectivity features

InfoSphere Information Server 8.7 includes several additional connectivity features:

- **Dynamic Relational Stage (DRS) connector:** Allows the database type to be dynamically chosen at runtime, which enables jobs to be reused in various deployment scenarios. InfoSphere Information Server supports this flexible connectivity between Oracle, IBM DB2 and Open Database Connectivity (ODBC) interfaces.
- **Change Data Capture (CDC) Transaction stage:** Allows information that is captured from log-based sources using InfoSphere Change Data Capture to be delivered in near-real time to InfoSphere DataStage. By using innovative bookmarking technology, this solution architecture is designed to provide guaranteed data delivery and transaction integrity through a complex set of data transformation and cleansing tasks to one or more targets.

- **Teradata connector:** Provides native support to the Teradata Multi-System Manager (TMSM) and dual load capability.
- **Enhanced SQL logging for connectors:** Enables connectors to write executed SQL statements to the job log, which helps support exception research.
- **DB2 connector performance enhancement:** Helps improve performance of the DB2 connector when loading to DB2.

Other connectivity enhancements include support for:

- IBM DB2 for Linux, UNIX and Windows 9.7 and 9.8, including the IBM DB2 pureScale® feature
- IBM DB2 10 for the IBM z/OS® operating system
- IBM WebSphere® MQ 7.0
- IBM Informix® 11.7
- Teradata Database versions 13.10 and 14.0
- Oracle Database 11*g* Release 2
- Sybase Adaptive Server Enterprise 15.5
- Sybase IQ 15.2
- IBM Netezza 4.6, 6
- DataDirect ODBC driver 6.1

InfoSphere Information Server 8.7 administration features for accelerating time to value

InfoSphere Information Server 8.7 includes new features designed to simplify administration, which helps accelerate time to value.

Maintenance mode

A new command-line tool allows the suite administrator to put InfoSphere Information Server into maintenance mode. When the system is configured in this way, any user who does not have the suite administrator role will be prevented from logging in. This feature simplifies the administrative steps required to quiesce the system for any activities that require exclusive administrator access to the program.

Command-line administration of user and roles

Certain organizations standardize the way in which they establish their InfoSphere Information Server environments for new project work. To support this process from beginning to end, InfoSphere Information Server includes expanded command-line interfaces, which permit user and role assignment to InfoSphere DataStage and InfoSphere QualityStage projects.

IPv6 support

Many organizations are migrating their networks to the expanded set of IP addresses. InfoSphere Information Server 8.7 is fully compatible with Internet Protocol version 6 (IPv6) addresses and can support dual-stack protocol implementations—for example, a mixed IPv4 and IPv6 environment.

InfoSphere Information Server 8.7 security features for protecting information assets

Enterprises are increasingly concerned with protecting their information assets. Software solutions must work alongside an organization's policies to enable a strong-security, low-risk posture. InfoSphere Information Server 8.7 addresses these requirements with several security enhancements.

Strongly encrypted credential files for command-line utilities

InfoSphere Information Server 8.7 introduces a unified means for handling credential arguments to various command-line tools, such as istool and dsjob. This new option, called *authfile*, provides the ability to reference an external file that maintains credentials required for login to InfoSphere Information Server. Administrators can then assign system permissions for the authfile as appropriate for users who need to execute InfoSphere Information Server commands. The parameters in this file can be maintained in the clear or strongly encrypted— an important capability for organizations concerned with protecting information at rest.

Strongly encrypted job parameter files for the dsjob command

Organizations use job parameters to provide flexibility in a variety of places within a job design, but particularly for connectivity to databases and enterprise applications. InfoSphere DataStage and InfoSphere QualityStage have provided encryption for these parameters within job flows for many years. InfoSphere Information Server 8.7 extends this ability by permitting the dsjob command-line paramfile option to specify a file that contains strongly encrypted parameter values. This feature provides encryption at rest for job parameters in the same way that the authfile provides it for login credentials.

Encryption algorithm customization

InfoSphere Information Server offers the Advanced Encryption Standard (AES)-128 symmetric-key encryption algorithm, which is compliant with many government and industry standards. Although AES-128 satisfies many organizations' security requirements, other organizations may wish to use their own encryption variant. In these instances, InfoSphere Information Server 8.7 provides the means to override the out-of-the-box algorithms.

Mixed LDAP and operating system support through PAM

Many organizations provide software access to end users through accounts defined in their Lightweight Directory Access Protocol (LDAP) servers and to functional users who are defined through the local operating system. These functional users represent a role that may be shared by multiple individuals within the organization or may simply be assigned to a scripted process. For organizations that have this mixed-mode security requirement, InfoSphere Information Server 8.7 has added support for pluggable authentication modules (PAM) to span LDAP and local operating system authentication mechanisms.

Automated audit trail for all DBMS access with IBM InfoSphere Guardium

Organizations are subject to a growing number of regulatory mandates to protect sensitive information such as financial records and personally identifiable information. These mandates require the establishment of mechanisms to detect, record and remediate unauthorized access or changes to sensitive data, even by privileged users. The IBM InfoSphere Guardium® database security platform provides a simple means of automating and centralizing compliance controls through a secure, centralized repository containing a fine-grained audit trail of all database activities. InfoSphere Information Server 8.7 connectors can be easily configured to automatically register these audit trail entries so that an organization has a comprehensive view of the origin of any data integration activity that affects the database.

For more information

For more information about IBM InfoSphere Information Server, including specifics about supported platforms, please contact your IBM marketing representative or IBM Business Partner, or visit: **ibm.com**/software/data/integration/info_server