Successful information governance through high-quality data

IBM InfoSphere® Information Server offers comprehensive data quality capabilities that improve outcomes

IBM

Contents

Executive summary

Only a decade ago, five million records would have been considered a large volume of data. Today, the volume of data stored by enterprises is often in the petabyte, or even exabyte, range. This explosion is not limited to structured data; in fact, most of the added volume flows from unstructured sources, such as email, images and documents. Lost, inaccurate or incomplete information also can generate high costs and lost productivity when having to hunt for information or reconcile data. Poor data quality can lead to failed business processes and erroneous decision making.

IBM InfoSphere® Information Server offers end-to-end data quality capabilities for information governance. With it, organizations can define a common business language to reduce miscommunication between business and IT, and understand data and their relationships to gain a complete picture before beginning a project. InfoSphere Information Server provides capabilities for analyzing and monitoring data quality continuously to reduce the proliferation of incorrect or inconsistent data. Firms can cleanse, standardize and match data to ensure its quality and consistency and to provide a single version of the truth, and maintain data lineage so end users can trace data back to original sources, establishing trust and confidence in the information received.

IBM InfoSphere Information Server provides a data-quality suite that can help transform a company into one leveraging comprehensive, trustworthy data for decision making. It is the foundation for successful data quality initiatives, helping derive optimal value from the complex, heterogeneous information spread across systems. InfoSphere Information Server provides a resilient, reliable, high-performance platform for mission-critical data.

The importance of effective information governance

An organization can have hundreds or even thousands of different systems. Information can come from many places — such as transactions, operationals, document repositories and external information sources — and in many formats, including data, content and streaming information. There are often meaningful relationships among the data, wherever it originates.

Table 1 illustrates the differences between two companies attempting to compile a complex report. Company A has no information governance system, and compiling the report requires the effort of multiple people over a long period of time, resulting in a report with untimely information. Company B, which has implemented effective information governance, needs only one person and can have a timely report within minutes. It is clear that decision making based on reliable, accurate data is more easily achieved through information governance.

Company A	Company B
Does not have information governance system	Has information governance system
Data exists only in disconnected silos	Data is consolidated into a data warehouse
Timely data cannot be accessed quickly	Data warehouse is refreshed nightly
Data from each source must be searched and compiled using a different process	One query can be designed to search and compile all data
Multiple analysts are needed for several days to consolidate data from all sources	One analyst can compile the needed data within an hour
Changes to the report take significant time to implement	Changes to the report can be made quickly through a simple refinement of the query

Table 1: Comparison of different companies attempting to create the same complex report, with and without an information governance system

An organization must be able to manage its supply chain of information, then integrate and analyze it to make business decisions, as illustrated by Figure 1. Unlike a traditional supply chain, an information supply chain has a many-to-many relationship. For example, data about the same person can come from many places—that person may be a customer, an employee and a partner—and the information can end up in many reports and applications. As well, various systems may define the same information differently. Given this complexity, integrating information, ensuring its quality and interpreting it correctly are crucial tasks that enable organizations to use the information for making effective business decisions. Information must be transformed into a trusted asset and governed to maintain quality over its lifecycle. The underlying systems must be cost-effective, easy to maintain and perform well for the workloads they need to handle, even as information continues to grow at exponential rates.

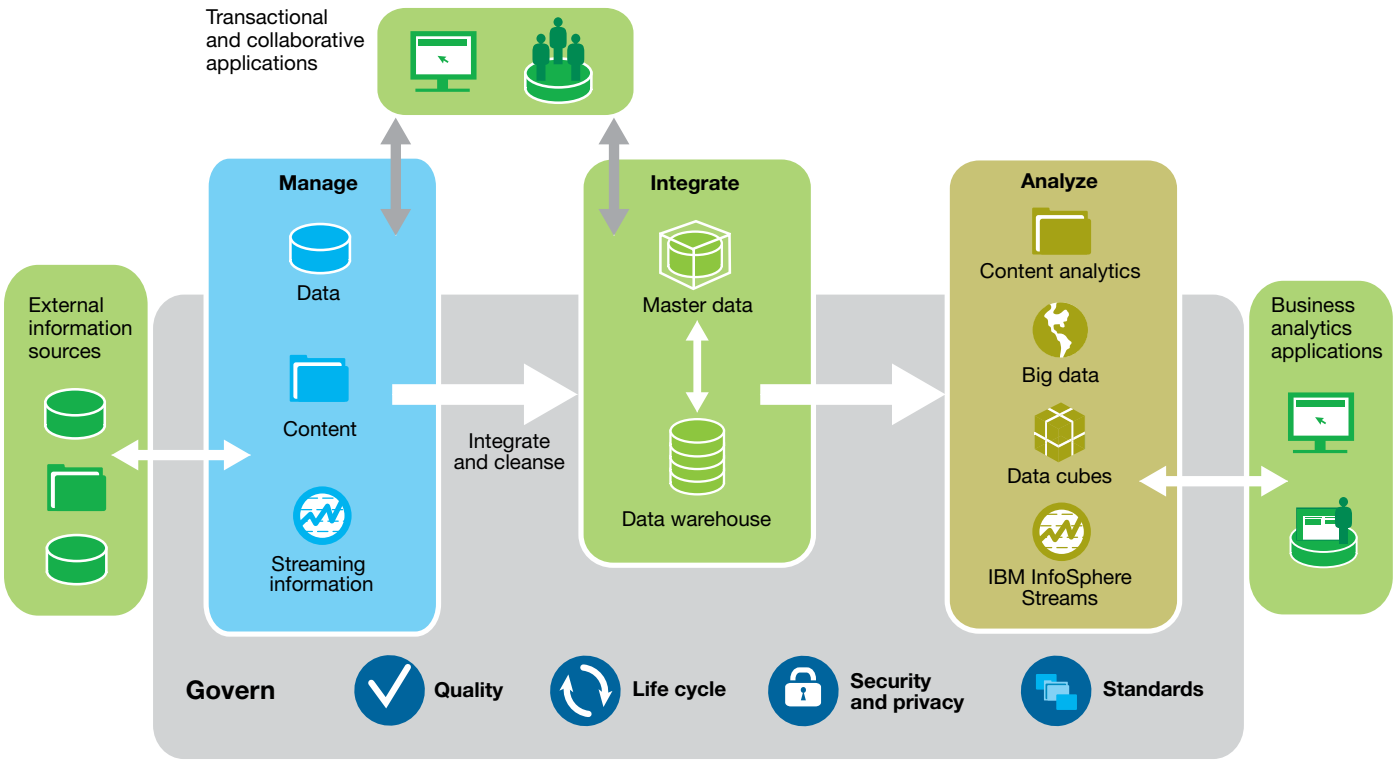

Figure 1: Governance enhances the quality, availability and integrity of the information supply chain

Effective information governance can enhance the quality, availability and integrity of an organization's data by fostering cross-organizational collaboration and structured policy making. It balances functional silos with enterprise-level oversight, directly affecting four factors critical to an organization: increasing revenue, lowering costs, reducing risk and increasing confidence.

Excellent data quality is achieved through several essential attributes:

- **Completeness:** All related data must be linked from all possible sources.
- **Accuracy:** Data must be correct and consistent, with common data problems remediated, such as misspellings or abbreviations..
- **Availability:** Data must be available upon demand.
- **Timeliness:** Current data must be available.

IBM InfoSphere Information Server supports successful information governance

The success of an information governance program hinges upon robust data quality. IBM InfoSphere Information Server offers end-to-end data quality capabilities that help organizations to:

- Define a common business language to reduce miscommunication between business and IT.
- Understand data and their relationships to gain a complete picture before beginning a project.
- Analyze and monitor data quality continuously to reduce the proliferation of incorrect or inconsistent data.
- Cleanse, standardize and match data to assure its quality and consistency and to provide a single version of the truth.
- Maintain data lineage so end users can trace data back to original sources, establishing trust and confidence in the information received.

The data quality capabilities of InfoSphere Information Server use a parallel processing infrastructure that provides leverage and automation across the platform, offers connectivity to almost any data or content source and can deliver information through a variety of mechanisms.

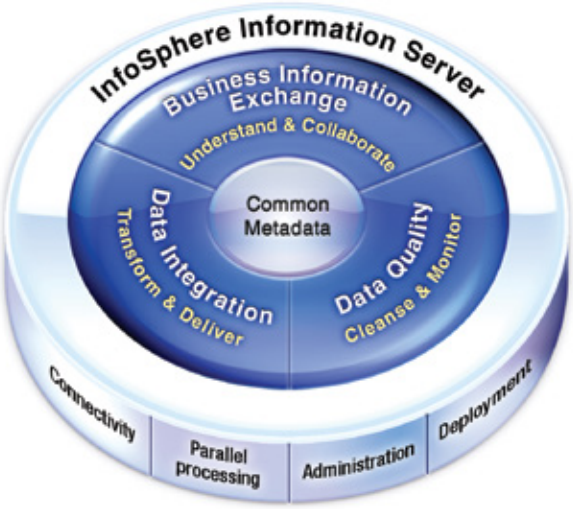

Figure 2: IBM InfoSphere Information Server is built on a foundation of parallel processing and other services

Define a common business language

Roadblocks arise when people have difficulty understanding and interpreting data, determining what data is important or managing that information, affecting collaboration and impacting both IT and businesspeople. These inconsistencies in business definition across enterprise environments are often attributed to the absence of an enterprise-wide data dictionary and stewardship program.

Having a common business language is critical in aligning technology with business goals. In addition to a controlled vocabulary, the hierarchy and classification systems provide additional business context. IBM InfoSphere Information Server helps organizations create, manage and share an enterprise-wide controlled vocabulary that acts as the common language between business and IT.

With InfoSphere Information Server metadata services, data stewards are able to link business terms to technical artifacts shared between IBM InfoSphere Data Architect, InfoSphere Information Server or a third-party data integration solution. The result is a common set of semantic tags used by data modelers, data analysts, business analysts, governance stewards, data architects, developers and end users. To help ensure high quality and tight security, only authorized data stewards can use the administrative functions within InfoSphere Information Server to create and manage the glossary.

The glossary also serves as a history of records to help ensure compliance with regulatory rules, such as the Sarbanes-Oxley Act and Basel II. Business terminology is always subject to change—what defines a high-value customer today may be different tomorrow as business requirements evolve. The ability to see the history of what changed, why it changed and who changed it is as important as the change itself. Such history is critical to data governance protocols because it increases the trust and understanding of the information.

Understand data and data relationships

Before implementing an information governance program or information-centric project, organizations must know what data they have, where it is located and how it relates between systems. For most organizations, the data discovery process is manual, requiring months of human involvement to discover business objects, sensitive data, cross-source data relationships and transformation logic. The result is a time-consuming, error-prone process that can slow time-to-value, establish doubt about the accuracy of the data within the new system and create the possibility that the new system will never become operational.

IBM InfoSphere Information Server provides a full range of capabilities to automate the data discovery process. It addresses single-source profiling, analysis of cross-source data overlap, discovery of matching keys and automated transformation, and prototyping and testing for data consolidation. InfoSphere Information Server also uses heuristics and sophisticated algorithms that automate analysis to help organizations realize 10 times more cost reduction and time savings compared to manually performing the same tasks using a profiling solution.[1]

This automated data discovery function includes several key capabilities:

- **Data profiling:** Gain advanced data profiling with results that are fit for purpose, including column analysis, automated primary-foreign key discovery and simultaneous cross-source column overlap analysis of multiple data sources. These sources can be as simple as text files on a PC or as complex as virtual storage access method (VSAM) on the IBM System z® mainframe—or both at the same time.
- **Unified Schema Builder:** The Unified Schema Builder component takes the output of overlap analysis and uses it as input into a process for helping a data analyst determine the rules by which data will be consolidated for data migration, master data management (MDM) or a data warehouse, to name a few possibilities. It includes automation software with an embedded workflow to help organizations complete consolidation projects on time and within budget.
- **Transformation Analyzer:** The Transformation Analyzer component automates the discovery of complex cross-source transformations and business rules by analyzing data values and patterns across two data sources.

The Transformation Analyzer is used when two data sources are related, but the relationship cannot be described by simple overlaps in data values and requires determining how data is transformed between the two sources. Data migration, application retirement, data warehousing and MDM almost always require the mapping and discovery of complex transformation logic between two or more data sources. The Transformation Analyzer helps accelerate this process by automating much of the analysis involved, replacing tedious manual work.

The InfoSphere Information Server data discovery analysis process establishes an understanding of data sources and how they relate to each other. It generates actionable output that can be immediately consumed by a wide range of information projects, including archiving, test data management, data privacy, data integration, MDM and data consolidation.

Analyze and monitor data quality

IBM InfoSphere Information Server helps scope data quality projects and develop measurements, rules and metrics to form a complete picture of data quality. It provides a dashboard that helps organizations continuously monitor data health and quickly identify issues through a graphical overview. Data owners can use the delivered artifacts to focus on detecting and responding to critical data quality issues and to deliver trusted data to the enterprise.

The risk of proliferating incorrect or inaccurate data can be reduced by using rules-driven analysis. Creating and reusing rules across multiple data sources increases time-to-value and achieves highly consistent, correct data.

Rules analysis is a key data assessment capability that extends the ability to compare, evaluate, analyze and monitor expected data quality. It consists of rules that evaluate data through focused and targeted testing of that data against user-defined conditions. The combination of multiple rules provides a broad, holistic assessment of records and data sources, allowing rules analysis at multiple levels.

InfoSphere Information Server includes several data quality assessment features:

- **Comprehensive data analysis:** A comprehensive set of metrics, based on data profiling, offers a holistic picture of data from many angles and enables analysts to immediately document all discovered data anomalies, including structural integrity, format consistency and data duplication, as well as identify incomplete and invalid values.
- **Drill-down capabilities**: End users can view individual records from data profiling results in real time. For example, if an invalid value in a column is discovered, an analyst can easily drill down to the actual record for further investigation.
- **Integrated rules analysis:** This robust capability provides development, deployment and evaluation of critical data rules on an ongoing basis. It features holistic, multilevel rule assessment at the rule, record and source levels for great insight into potential quality issues. Rules can be built free-form or through a structured builder, tested and reviewed, which helps the end user readily compose standard data conditions.
- **Reusable deployments:** As rules are defined logically, they can be developed once and applied repeatedly and consistently to any number of data sources. The resulting data rules can be run in ad hoc or scheduled modes, or deployed into production environments for ongoing data quality monitoring.
- **Application of data quality rules against data at rest or in flight:** The same rule that can be deployed against multiple data sources can also be applied as part of an extract, transform and load (ETL) or data cleansing job. This capability can help proactively detect, and possibly resolve, data quality issues automatically before the data is further distributed or loaded into trusted repositories, such as a warehouse or an MDM system.
- **Validation of rules across sources:** Certain data validation rules require that data across different databases is compared — for example, that the profit figure stored in a data warehouse equals the revenue data from source A minus the cost data from source B. This capability allows analysts to specify such rules, monitor them and track corresponding exceptions.
- **Ongoing quality monitoring:** Results of rules, or comprehensive rule sets, can be measured and monitored against established benchmarks or thresholds. Additional metrics can also be applied against the generated statistics to create key performance indicators or to establish costs or weights to errors. Any of these measures can be tracked and trended over time.

InfoSphere Information Server assesses data quality up front, and also establishes rigorous and relevant data rules based on business needs. Consequently, organizations are able to continuously assess and monitor trends in information quality that provide confidence in information delivered, and they have the means to proactively target quality improvement as part of an information integration and data governance initiative.

Cleanse, standardize and match information

With IBM InfoSphere Information Server, enterprises can create and maintain an accurate view of master data entities, such as customers, vendors, locations and products. InfoSphere Information Server is designed to provide a development environment with a powerful and flexible set of capabilities.

IBM InfoSphere Information Server:

- Includes the Standardization Rule Designer, a graphical, drag-and-drop interface that makes creating, editing and changing rules as simple as a few mouse clicks
- Provides a single set of standardization, cleansing, matching and survivorship rules for core business entities — executed in batch, real time or as a web service
- Matches data using probabilistic algorithms designed to ensure that the information needed to run an enterprise is accurate, complete and trustworthy
- Processes global data on a massively scalable parallel platform for optimal performance in demanding environments
- Makes creation and maintenance of high-quality master data a reality, to drive benefits across a variety of critical enterprise initiatives, including MDM and data governance
- Monitors data against user-defined quality thresholds to help assure the health of data over time
- Brings data quality capabilities to data integration situations through seamless data flow integration
- Employs an intuitive, design-as-you-think user interface

InfoSphere QualityStage provides a comprehensive process to manage and maintain data quality. Its core functions include:

- **Investigation:** enables an understanding of the nature and extent of data anomalies, as well as effective cleansing and matching
- **Standardization:** creates a standardized view of customer, partner or product data; facilitates global address cleansing, geolocation, and validation and certification for significant postal discounts in select localities
- **Probabilistic matching:** provides an industry-leading matching engine to help ensure the best match results possible; built on a platform enabled for high connectivity and scalability
- **Survivorship:** helps ensure the optimum consolidation, householding or linked view of record information; enables a consolidated and accurate view of customers, partners, products and more

The probabilistic matching capability and dynamic weighting strategies of InfoSphere Information Server help organizations create high-quality, accurate data. Business users can consistently identify core business information, such as customer, location and product, throughout the enterprise. It standardizes and matches any type of information. By helping ensure data quality, InfoSphere Information Server can reduce the time and cost to implement customer relationship management, enterprise resource planning, business intelligence and other strategic customer-related IT initiatives.

Maintain data lineage

InfoSphere Information Server is designed to be a complete platform for integrating and enriching information across disparate source systems. By leveraging an active and shared metadata repository layer, InfoSphere Information Server can support a full range of integration activities and user roles with collaboration and reuse principles. These artifacts include technical metadata about the various sources of information; business metadata that describes the business meaning and usage of information; and operational metadata that describes what happens within the integration process.

InfoSphere Information Server provides a powerful metadata management interface that supports InfoSphere Information Server metadata and other key 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 organizations with critical information that can lead to sound decisions.

IBM InfoSphere Information Server includes several key metadata features:

- Web-based navigation of information assets through an interactive and powerful interface provides an easy way for business and IT users to access critical information.
- Visual cross-tool and cross-platform data lineage enables an understanding of the information lineage — including where the data came from and what happened to it as it moved across data integration processes — with extended visibility into enterprise data flows outside of InfoSphere Information Server.
- Visual cross-tool impact analysis provides a thorough understanding of a change's impact before the change is made, even when the impact extends beyond a single tool.
- Reporting on information assets, through simple and advanced search with save, repeat and publish capabilities, helps business and IT users quickly understand complex environments.
- Automated linkages to InfoSphere Information Server metadata services help organizations reduce their overall IT costs while accelerating productivity.

A comprehensive data quality platform for information governance

The InfoSphere Information Server data quality suite is a fully integrated software platform that helps users understand, maintain and cleanse information. It enables collaboration to develop and support an information governance strategy that helps firms derive value from the complex, heterogeneous information spread across multiple sources. With InfoSphere Information Server facilitates, companies achieve operational efficiency and reduced business risk.

For more information

To learn more about information quality and its role as part of your information governance strategy, please visit:

ibm.com/software/data/integration/capabilities/cleanse.html

ibm.com/software/data/db2imstools/solutions/data-governance.html

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
October 2012

[1] Time and cost savings based on reports from IBM client engagements.

Please Recycle

IMW14537-USEN-02