Delivering information you can trust

The benefits of leveraging IBM capabilities to deliver quality data

Contents

Executive summary

A decade ago, 5 or 10 million records would have been considered a large volume of data. Today, the volume of data stored by enterprises is often in the terabyte or even exabyte range. The data explosion is not limited to structured data: in fact, most of the new volume flows from unstructured sources, such as email, images and documents. How can timely decisions be made when there is so much data? How can busy executives have any confidence that the reports they see reflect accurate information culled from reliable data?

Imagine two companies, Brute Force, Inc., and Easy Corp. At Brute Force, a senior executive asks for a report that shows growth in key customer accounts over a three-year period. The Brute Force business analysts ask their IT department for customer sales data. The challenge here is that customer data and sales data are stored in different systems. IT can produce raw data for sales on a monthly or quarterly basis, and IT can produce customer data that shows what products different customers have ordered in the past. The business analysts understand that linking this data will take a long time unless a lot of help is available. Knowing that reconciling data to produce reports is challenging, the company hires many analysts to sort through the data. After a week's worth of work, a team of 10 people produces the one-page report the senior executive wanted, with one catch: current month sales are in a different system, and so are not in the report.

When the analysts at Easy Corp. receive a similar executive request, they simply design a report using their business intelligence (BI) tool that queries an underlying data warehouse. The data in the warehouse is compiled from numerous source systems, with all data refreshed every night with any adds, updates and deletes. After a few moments, an analyst generates a draft report, which is then emailed to the analyst's manager for review. The manager may suggest a couple of changes, which the analyst handles (again, using the BI tool) to produce a final report for the requesting executive. In approximately one hour, one person creates a report like the one that took Brute Force 10 people and one week to produce.

Which company more closely resembles your organization?

IBM® InfoSphere® Information Server provides a data quality suite that can make a big difference in helping an organization move toward an Easy Corp. profile. It is the foundation of many successful data quality initiatives, helping organizations derive optimal value from the complex, heterogeneous information spread across their systems. InfoSphere Information Server provides a resilient, reliable, high-performance platform for mission-critical data.

The importance of effective information governance

An organization typically has hundreds or even thousands of different systems. Information can come from many places—such as transaction systems, operational systems, document repositories and external information sources—and in many formats, including data, content and streaming. There are often meaningful relationships between the data, wherever it originates. Organizations must be able to manage all this information, integrate it to build warehouses and analyze it to make business decisions.

This supply chain of information flows throughout an organization (see Figure 1). Unlike a traditional supply chain, an information supply chain has a many-to-many relationship. The same data about a 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 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 turned into a trusted asset and governed to maintain quality over its life cycle. The underlying systems must be cost-effective and easy to maintain and must perform well for the workloads they need to handle, even as information continues to grow at astronomical rates.

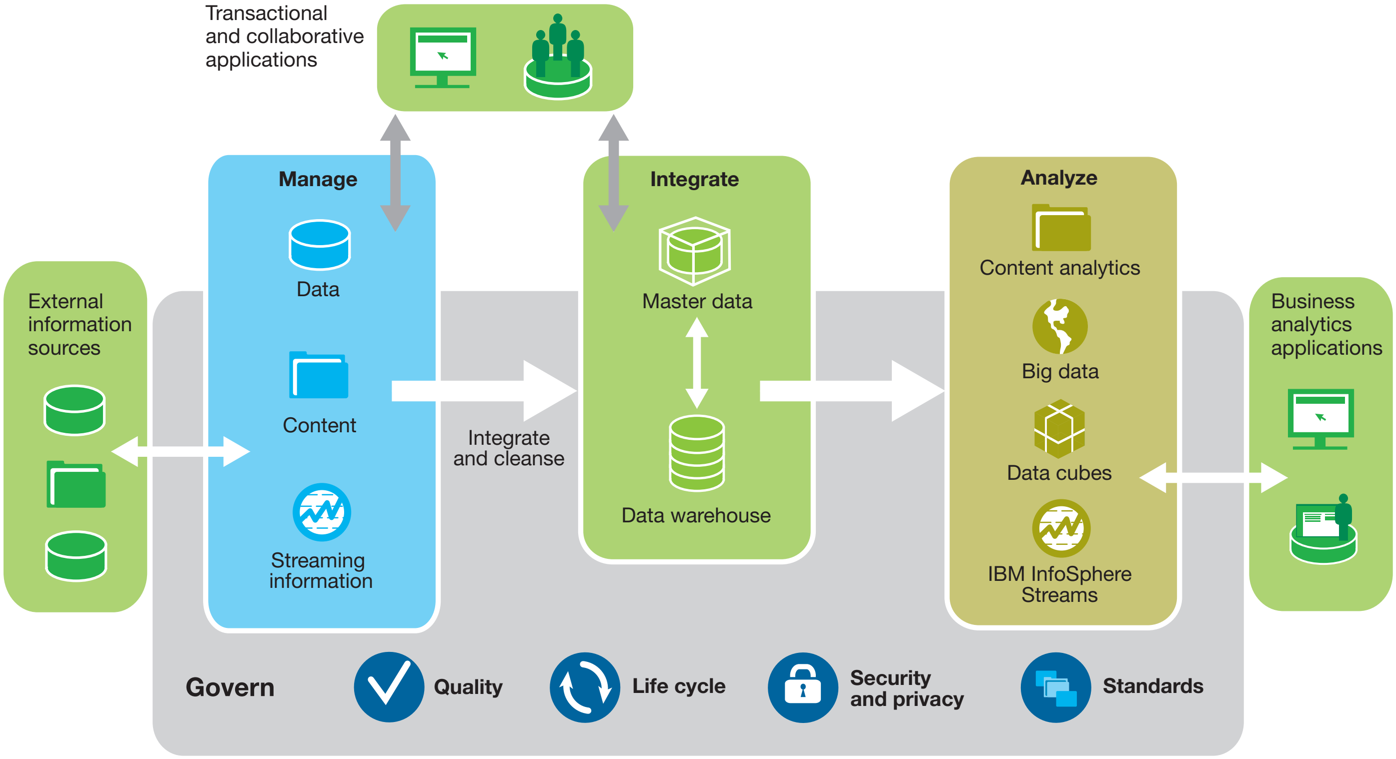


Figure 1: The information supply chain

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 factional silos with organizational interest, directly affecting four factors that are critical to an organization: increasing revenue, lowering costs, reducing risks and increasing data confidence.

A clear understanding of customers, partners and suppliers can mean the difference between growing a business and failing to compete. Excellent data quality, which is essential for success, has the following attributes:

- **Completeness:** To qualify as complete, all relevant data should be linked together. For example, a complete customer record may include all accounts, addresses and relationships that the company has for that customer.
- **Accuracy:** Common data problems like misspellings, typos, random abbreviations and the like must be cleaned up.
- **Availability:** Quality data must be available on demand; data that must be searched manually is not quality data.
- **Timeliness:** How much value does a sales report have if it's missing the most recent month?

Effects of poor data quality include failed business processes, low productivity and wasted materials. Lost, inaccurate or incomplete information also can generate high costs and extra work, such as hunting down information or reconciling data.

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 accomplish the following tasks:

- Define a common business language to reduce miscommunication between business and IT
- Understand data and data relationships to gain a complete picture of data before beginning a project
- Analyze and monitor data quality continuously to reduce the proliferation of incorrect or inconsistent data
- Cleanse, standardize and match information 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 (see Figure 2). InfoSphere Information Server also offers connectivity to almost any data or content source and can deliver information through a variety of mechanisms.

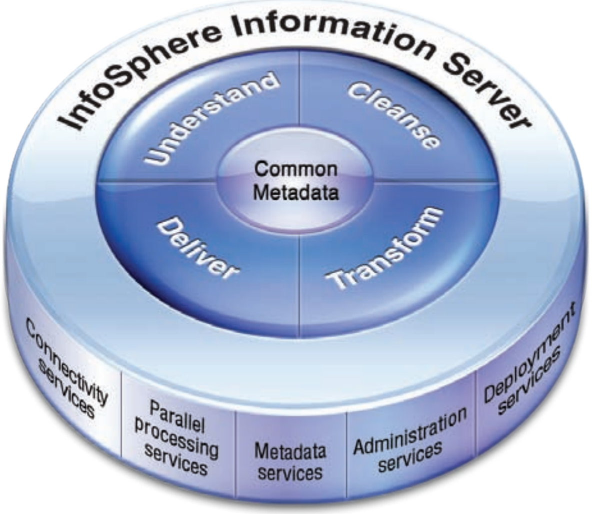


Figure 2: IBM InfoSphere Information Server is built on a foundation of parallel processing and other services

Define a common business language

Difficulty understanding and interpreting data, determining what data is important and then managing that information creates roadblocks as business and IT users attempt to collaborate for effective information integration. The problem of business definition inconsistency across enterprise environments is often attributed to the absence of an enterprise-wide data dictionary and stewardship program.

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

Actively connected to InfoSphere Information Server metadata services, InfoSphere Business Glossary enables data stewards 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 Business Glossary 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. Being able to see the history of what changed, why it changed and who changed it is as important as the change itself. Such a 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 Discovery provides a full range of capabilities to automate the data discovery process. It addresses single-source profiling, cross-source data overlap analysis, matching key discovery, automated transformation discovery and prototyping and testing for data consolidation. InfoSphere Discovery also uses heuristics and sophisticated algorithms that automate analysis to help organizations realize 10 times more time and cost savings compared to performing the same tasks manually using a profiling solution.[1]

InfoSphere Discovery includes several key capabilities:

- **Data profiling:** InfoSphere Discovery provides 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. Unified Schema Builder delivers automation software with an embedded workflow to help organizations complete consolidation projects on time and within budget.
- **Transformation Analyzer:** The Transformation Analyzer component is designed to automate discovery of complex cross-source transformations and business rules by analyzing data values and patterns across two data sources.

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. Transformation Analyzer helps accelerate this process by automating much of the analysis involved and replacing tedious manual work.

The InfoSphere 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

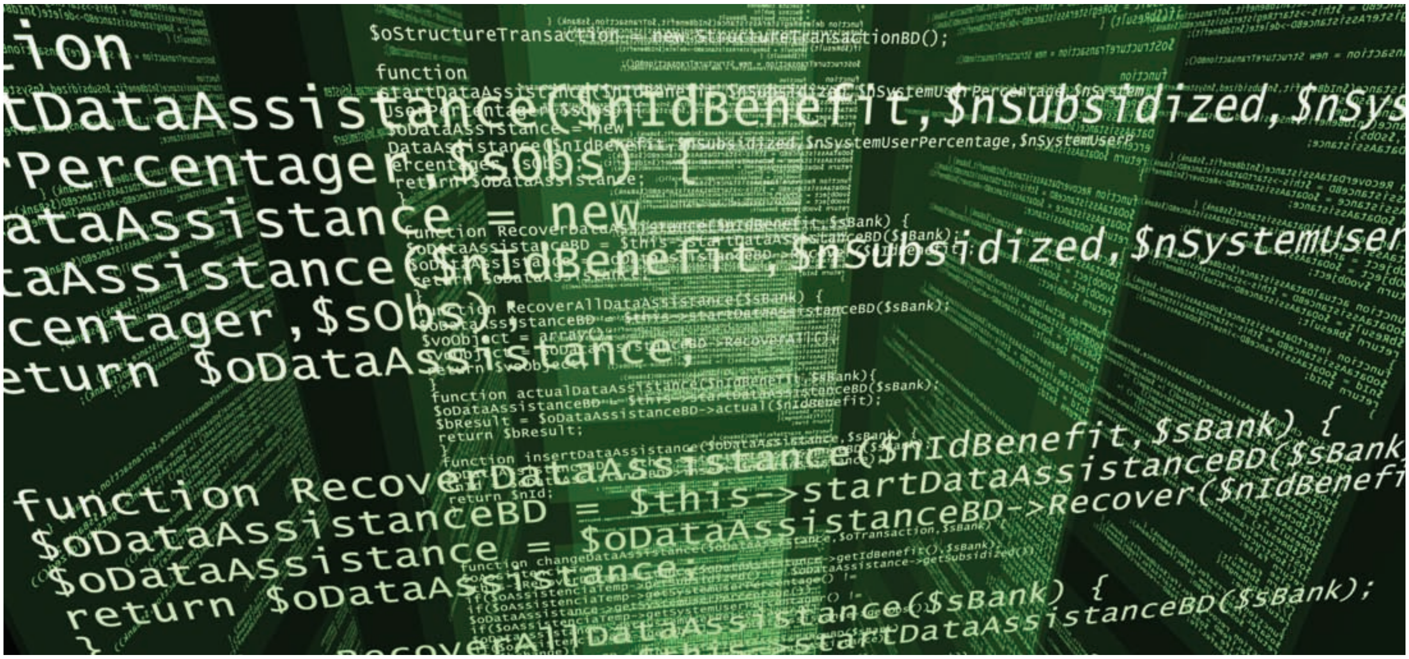
IBM InfoSphere Information Analyzer helps scope data quality projects and develop measurements, rules and metrics to form a complete picture of data quality. It provides a dashboard that helps organizations continuously monitor data health and quickly identify issues through a graphical overview. The artifacts delivered by Information Analyzer enable data owners 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 rules analysis. Creating and reusing rules across multiple data sources enables increased time-to-value and 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 Analyzer 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 identifying 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 freeform 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 stored in a warehouse equals the revenue data from source A minus the cost data from source B. The Information Analyzer Exception Management 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 Analyzer not only assesses data quality up-front, but also establishes rigorous and relevant data rules based on business needs. Consequently, InfoSphere Information Analyzer enables organizations to continuously assess and monitor trends in information quality that provide confidence in information delivered and delivers the means to proactively target quality improvement as part of an information integration and data governance initiative.

Cleanse, standardize and match information

IBM InfoSphere QualityStage® software enables enterprises to create and maintain an accurate view of master data entities, such as customers, vendors, locations and products. InfoSphere QualityStage is designed to provide a development environment with a powerful and flexible set of capabilities:

- Provides a single set of standardization, cleansing, matching and survivorship rules for core business entities—executed in batch, in 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
- Brings data quality capabilities to data integration situations through seamless data flow integration
- Employs an intuitive, design-as-you-think user interface

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

- **Investigation:** Enables 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 consolidated and accurate view of customers, partners, products and more

The probabilistic matching capability and dynamic weighting strategies of InfoSphere QualityStage help organizations create high-quality, accurate data. With InfoSphere QualityStage, 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 QualityStage can reduce the time and cost to implement customer relationship management (CRM), enterprise resource planning (ERP), BI 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.

IBM InfoSphere Metadata Workbench, a product module of InfoSphere Information Server, provides a powerful metadata management interface that supports not only InfoSphere Information Server metadata but also 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.

InfoSphere Metadata Workbench includes several key 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** allows 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 and accelerate productivity.
- **Collaboration and shared metadata** with InfoSphere Business Glossary promote data stewardship, business and IT alignment and better understanding of information assets.

Make your information work harder for you

The InfoSphere Information Server data quality suite is a fully integrated software platform that helps you understand, maintain and cleanse information. It enables collaboration to develop and support an information governance strategy that helps you derive value from the complex, heterogeneous information spread across source systems. InfoSphere Information Server facilitates novel ways of using information to support innovation, 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
January 2012

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation. Statements regarding IBM’s future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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

♲ Please Recycle