Getting started with a data quality program

The data quality challenge

Organizations depend on quality data to inform business decisions, support customers, develop plans and manage other essential tasks. But the proliferation of data sources and exponential growth in data volumes can make it difficult to maintain high-quality data. These shortcomings sometimes result in humorous anecdotes, such as the case of a 50-year-old man receiving a letter from his healthcare provider to confirm his upcoming ultrasound appointment to check on his pregnancy. However, poor data quality can also have serious repercussions. In one incident, a young mother in the United Kingdom died of breast cancer because her cancer diagnosis and subsequent treatment were delayed. The cause of the delay? An error in the hospital's patient record, which listed her house number as "16" instead of "1b." As a result, the hospital's letters never reached the patient, and she missed crucial medical appointments.[1]

By addressing data quality issues, organizations can avoid unfortunate consequences and improve business outcomes. Besides human error, most data quality problems arise from a lack of enterprise-wide information standards on how data is stored and uniquely identified. Inconsistency across sources hinders the understanding of relationships between critical business entities, such as parties and products. In many cases, no reliable, persistent key exists to retrieve all information across the enterprise that is associated with a single party or product.

High-quality data enables strategic systems to integrate all related data to provide a complete view of the organization and the interrelationships within it. Without high-quality data across the enterprise, organizations cannot count on a return on the investments they have made in critical business applications such as data warehouses, business intelligence tools and master data systems. By implementing a data quality program, organizations can enhance data integrity to get the most out of their information assets.

A variety of tools go into a data quality solution: data standardization software, data matching engines, metadata workbenches, IT/business terminology consoles and data integration software, to name a few. Integrating these tools into a complete data quality solution all at once can be overwhelming for many IT organizations. In addition, organizations may have conflicting data quality priorities. One department may simply want to scrub addresses in a master customer file, while another department is merging finance systems from a recently acquired company and still another is under pressure from management over the suspect reliability of monthly sales reports.

The key to addressing these challenges is flexibility. A common data integration platform that can be applied to many data quality issues can play a major role in an organization's information governance strategy. Such a platform may be a better investment than a series of non-integrated point solutions purchased to solve particular problems as they arise.

Effective information governance across the information supply chain

A typical organization may host hundreds, or even thousands, of different systems. Information can originate in many places (transaction systems, operational systems, document repositories, external information sources) and in many formats (data, content, streaming). Meaningful relationships often exist between the various sources and types of data. This information supply chain flows throughout an organization and beyond its boundaries (see Figure 1). Unlike entities in a traditional supply chain, those in an information supply chain have a many-to-many relationship. The same data about a person—who may be a customer, an employee and a partner, for example—can come from many sources, and that information ends up in many reports and applications. Disparate systems may define the information differently as well.

Given this complexity, integrating information, ensuring its quality and interpreting it correctly are crucial steps for supporting sound decisions. Information must be turned into a trusted asset and governed to maintain quality over its life cycle. The underlying systems should be cost-effective and easy to maintain, and perform well for the assigned workloads even as the volume of information they handle grows at an astronomical rate.

Effective information governance enhances the quality, availability and integrity of an enterprise's data by fostering cross-organizational collaboration and structured policy making. It balances departmental silos with organizational interest to help increase data confidence—which can directly affect key business concerns such as increasing revenue, lowering costs and reducing risks. The effects of poor data quality include failed business processes, decreased productivity and wasted materials. Lost, inaccurate or incomplete information also can generate high costs and extra work, such as hunting down or reconciling information.

Characteristics of high-quality data

- **Completeness:** All relevant data is linked. 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 have been cleaned up.
- **Availability:** Required data has already been discovered and is available on demand; users do not need to manually search for the information.
- **Timeliness:** How much value does a sales report have if it's missing the most recent month?

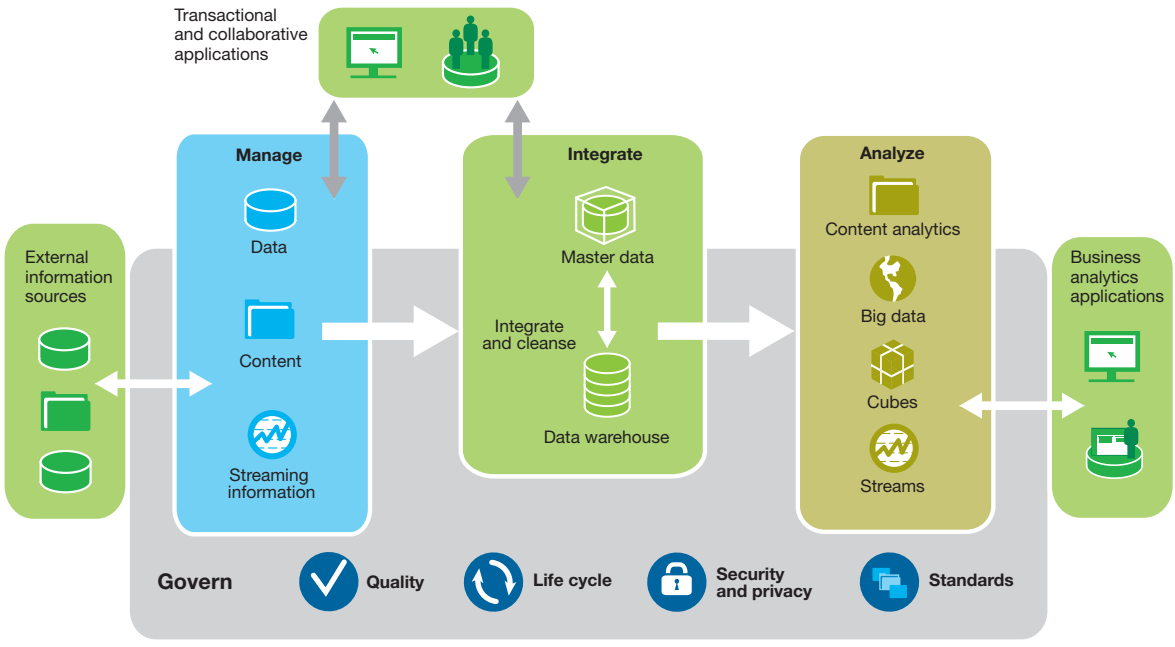

Figure 1: The information supply chain.

Excellent data quality is essential for success; for example, it helps provide a clear understanding of customers, partners and suppliers, which can make the difference between growing a business and failing to compete.

Establishing a data quality program: Getting started

To achieve proper information governance, organizations should establish a data quality program that is based on business objectives and priorities. Not all data quality issues have the same impact on business results; trying to address all of an organization's data quality challenges can be overwhelming and inefficient. Consider the following questions to establish return on investment (ROI) for each potential data quality initiative in the organization:

- What are the most critical business processes that rely on information?
- What information is most important to those processes?
- What is the cost of poor information to the effectiveness of those processes?
- What is the cost of maintaining high-quality information for those processes?
- What is the net benefit to the organization for maintaining data quality for those processes?

Whatever path is chosen, data quality needs will likely change over time, so it is important to invest in technologies that can scale and be leveraged across the enterprise. A point solution that solves today's problems—such as an address cleansing solution to improve accuracy and consistency—may be ill equipped to support tomorrow's data quality requirements.

Besides planning for future needs, a data quality program should address two fundamental questions. First, the organization must agree on the definition of quality. What is "good" data? Is it data that produces an error rate of less than 1 percent? Or can the organization tolerate an error rate of 10 percent?

Take the case of a government agency that needs highly accurate information at border checkpoints. Data errors here can have catastrophic consequences. However, in the address database of a clothing retailer, errors are less likely to have such dire results. Understand the goal (this is key to allocating resources and managing program costs) and do not assume perfection is necessary or even desirable in every case.

The next question involves reporting: Given a definition of quality, what metrics must be tracked to ensure the quality threshold is being maintained? The data quality program should include a systematic, business-driven approach for capturing and reporting these metrics, and the organization must articulate priorities to be formally documented and tracked over time.

The initial focus of data quality efforts depends on the answers that an organization provides to these questions.

The first entry point to data quality: The data quality assessment

Whether embarking on new initiatives or addressing data governance and risk mitigation issues, many organizations find that a good starting point is a data quality assessment, which establishes a baseline: How good is your data, and where are the greatest opportunities for improvement?

Many organizations find it challenging to obtain a consistent understanding of their data across the enterprise, particularly as systems and applications are adapted to changing requirements and as mergers and acquisitions expand existing data sets. Business units and IT use different semantics and applications record data in multiple ways—with different identifiers (such as customer and account IDs), different formats (such as dates stored in a date format in a database, but a string format in a file) or different values (such as gender recorded as "M" or "F" in one system but "0" or "1" in another). The organization's approach to knowledge management can further complicate the situation.

Often, knowledge about data is simply tacit, stored in people's heads based on their specific work. Or it may be recorded in documentation that has not been kept up to date with the latest business processes or system changes. When individuals move within or out of an organization, much of that knowledge is lost or fragmented.

A data quality assessment is intended to provide insight into this complicated picture, establishing a foundational practice for subsequent work and a knowledge base within a shared metadata repository that can be used and reused across multiple projects and initiatives. A data quality assessment focuses on and illustrates a business problem based on the underlying data. Such an effort helps bring data quality issues to light, but a data or business analyst must still review the results and draw conclusions, particularly to define the business impact.

As shown in Figure 2, the data analyst is at the center of a data quality assessment. Analysts must understand the big picture: the scope, objectives and deliverables of the assessment. If the analyst does not know what the organization is trying to achieve, how can he or she identify a true anomaly or problem? The sheer number of tables and attributes may be overwhelming. Simply selecting many columns to analyze and running an analysis job is not the end of the process—the subsequent review is critical.

IBM® InfoSphere® Information Server provides capabilities that help analysts learn and apply data analysis techniques. Training in the use of these capabilities—available from IBM—focuses on core analysis steps and best practices:

- Identification of and approach to the data sources in question
- Advantages of automated data content-driven functions
- Usage of data classifications to focus analysis
- Validation of data formats and domains
- Reporting and delivery of findings and results
- Retention of analysis results over time

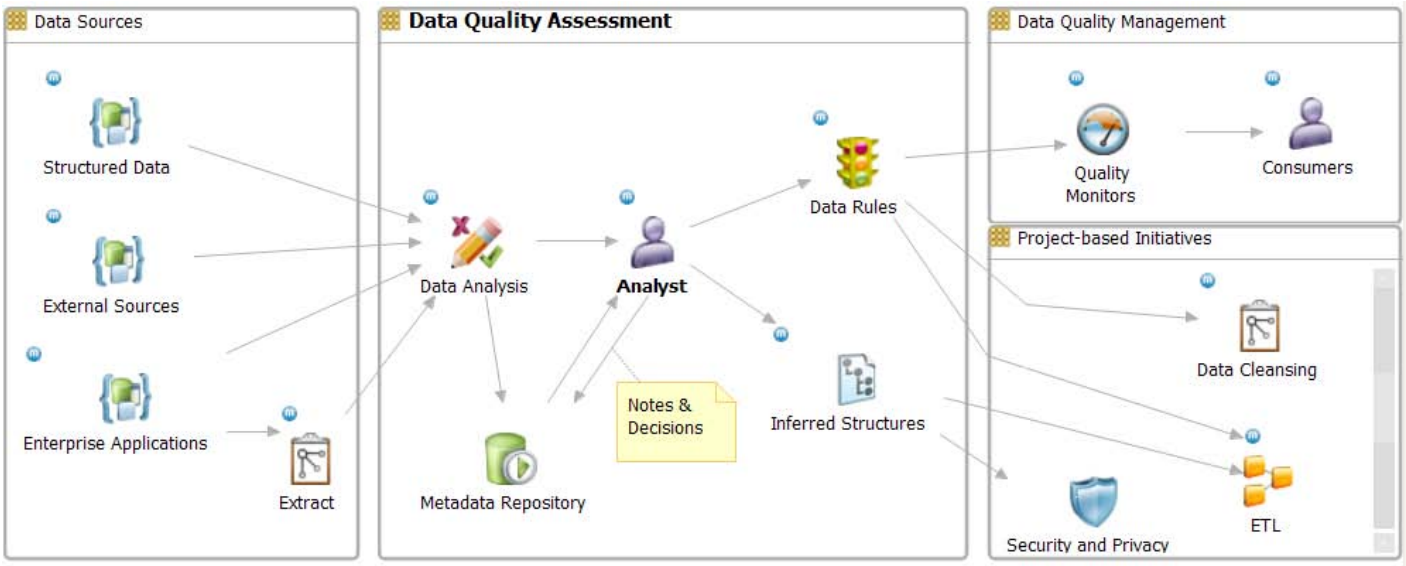

Figure 2: Landscape for a data quality assessment

The second step: A comprehensive platform for assessing data quality

InfoSphere Information Server supports data analysis and data quality monitoring capabilities that enable the creation of a data quality assessment. Using a multi-tier architecture, common services, shared metadata and a parallel processing engine, it provides a common platform capable of analyzing a broad range of data sources, processing high volumes of data, storing extensive results and capturing analyst insights in a secured, project-based environment (see Figure 3).

Using InfoSphere Information Server to perform a core analysis, the analyst may discover a broad range of issues in many domains, such as defaulted data, missing values and non-unique or duplicated keys. The analyst can use additional techniques to focus on certain types of conditions, which can typically be expressed as data validation rules. These rules can be used to test for valid value combinations, correct formulas and aggregations or complex format requirements, as well as to obtain a comprehensive assessment of entire records or tables. The additional tests can be reported, retained and trended over time in InfoSphere Information Server.

The shared metadata from the core analysis performed in InfoSphere Information Server is directly available to users of other capabilities within the platform. Data modelers and database administrators can use the inferred structures and identified classifications to establish staging areas with the correct structure or to refine privacy and governance policies. Developers focused on data transformation or data cleansing can utilize the statistics and annotations to help ensure that appropriate cleansing routines are applied to the data, and they can incorporate reference tables generated from the analysis. Further, developers can plug the data validation rules from the analysis directly into data cleansing and extract, transform and load (ETL) processes through an integrated stage that applies the rules in flight. This capability helps ensure that problematic and invalid data conditions are addressed before the data is loaded into the target environments.

InfoSphere Information Server enables an enterprise to focus first on one source system and then continuously expand the basic data quality assessment in initiatives across the enterprise, including data cleansing, information integration and data governance projects. Analysis processes, data validation rules and

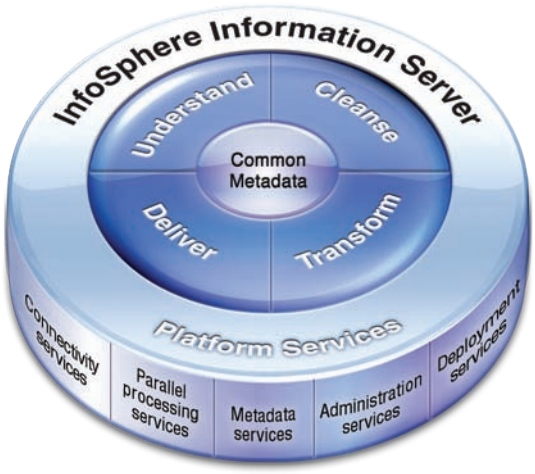

Figure 3: InfoSphere Information Server is built on a foundation of shared metadata, parallel processing and other services.

reports can be scheduled to run on a regular basis to provide ongoing data quality monitoring. The insight into an enterprise's data domains provided by InfoSphere Information Server supports and addresses the challenges inherent in the continuous expansion and acquisition of data, systems and applications that are the foundation of every organization's business.

Other data quality entry points supported by InfoSphere Information Server

In addition to data quality assessment and information profiling and analysis, InfoSphere Information Server supports other entry points to a full-scale data quality program that may be appropriate depending on an organization's priorities.

Define a common business language

Difficulties in understanding and interpreting data, determining what data is important and managing the data can create 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.

Business glossary functionality within InfoSphere Information Server helps organizations create, manage and share an enterprise-wide controlled vocabulary. Creating this common language between business and IT is a critical step in aligning technology with business goals. In addition to a controlled vocabulary, the hierarchy and classification systems provide additional business context.

Understand data and data relationships

Before implementing an information governance program or information-centric project, organizations must gain a complete picture of their data: 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: a time-consuming, error-prone process that slows time to value, establishes doubt about the accuracy of data within the new system, and creates the possibility that the new system will never become operational.

InfoSphere Information Server 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, prototyping and testing for data consolidation and automated transformation discovery. It also uses heuristics and sophisticated algorithms that automate analysis to help organizations realize time and cost savings compared to performing the same tasks manually using a profiling solution.

Cleanse, standardize and match information

To ensure quality and consistency in tasks such as address cleansing and record deduplication, organizations need tools that include reliable and easy-to-use standardization and matching as well as data integration, especially if multiple sources and/or multiple targets are involved. InfoSphere Information Server enables enterprises to create and maintain an accurate view of master data entities such as customers, vendors, locations and products. It also provides a development environment with a powerful and flexible set of capabilities:

- Provide 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
- Match data using probabilistic algorithms designed to ensure that the information needed to run the enterprise is accurate, complete and trustworthy
- Process global data on a massively scalable, parallel platform for optimal performance in demanding environments
- Facilitate the creation and maintenance of high-quality master data to drive benefits across a variety of critical enterprise initiatives, including master data management and data governance
- Bring data quality capabilities to data integration situations through seamless data flow integration
- Employ an intuitive, "design-as-you-think" user interface

Maintain data lineage

InfoSphere Information Server is designed to integrate and enrich information across disparate source systems. By leveraging an active and shared metadata repository layer, it supports 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.

The InfoSphere Information Server platform 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. The platform delivers a centralized, holistic view across the entire landscape of data integration processes, with visibility into data transformations that operate inside and outside of InfoSphere Information Server. This visibility enables organizations to track information back to original sources, establishing trust and confidence in the information received—particularly critical in audit or legal discovery situations.

A flexible, scalable data quality solution

Business decisions are increasingly informed by customer, partner and operational information. For successful outcomes, these decisions must be based on high-quality data. Establishing clear business priorities up front, supported by a comprehensive data quality program, enables organizations to focus on investment planning. IBM InfoSphere Information Server provides the flexibility to address today's high-priority data quality issues while readily scaling to support future requirements. For organizations just getting started on data quality and information governance initiatives, it offers total flexibility in a comprehensive, common data integration platform.

For more information

To learn more about data quality and its role as part of your information governance strategy, please contact your IBM sales representative or IBM Business Partner, or 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
March 2012

[1] "Mother with young son dies of cancer at 38 after hospital typing error sent urgent letters to the wrong address," The Daily Mail, March 14, 2011. www.dailymail.co.uk/news/article-1366056/Mistyped-address-leaves-mother-dead-cancer-son-8-orphan.html

Please Recycle

IMW14623-USEN-00