Effecting Data Quality Improvement through Data Virtualization

Prepared for Composite Software by:

David Loshin
Knowledge Integrity, Inc.
June, 2010

Introduction

The application infrastructure in most businesses has grown organically. There are some organizations that have adopted approaches of enterprise architecture that attempt to impose a high-level structure over both the business applications and their underlying data systems. However, in most environments, there are clearly-defined lines of demarcation between any set of operational business applications, as well as clear boundaries separating operational systems from analytical systems such as the reporting and analytics supported by siloed or enterprise data warehouses. Yet as the interest in consolidating data for the purposes of reporting has grown, the dependence on (and frequent failure of) enterprise-wide business applications such as Enterprise Resource Planning (ERP) or Business Intelligence and analytics only demonstrates that the traditional attempts for ensuring high quality data are not living up to enterprise expectations.

Traditionally, the approach to data quality assurance has presumed that data inconsistencies, errors, or incompleteness are unavoidable, and that the data sets can only be subjected to data quality improvement once they have been subsumed within the data warehouse. Yet rampant data extraction, and transformation without taking source metadata and definitions into account mean that the data consolidation process often introduces new issues when the information is being repurposed for downstream consumers.

For example, similar data concepts may be represented in different ways using different data types, lengths, and formats. Data validations that are sufficient for operational use may not support the data quality expectations for reporting and analysis. Decentralized data use may drive multiple business data consumers to apply the same transformations and cleansing techniques, reducing inefficiency and potentially increasing operational costs. And every time data is extracted and copied, it presents an opportunity for introducing new data flaws.

Although there are some data quality services that rely on access to replicated data sets, many kinds of data quality issues can be addressed using data virtualization. Data virtualization is a method of introducing layers of abstraction over a variety of native data source to provide reusable relational views and data services without requiring that data be extracted from its source. The data abstraction layers typically deployed within a data virtualization environment standardize a logical representation of enterprise data concepts, enabling multiple business consumers to see a structurally and semantically consistent view.

In this paper, we consider some significant data quality challenges that become magnified in relation to the growing need for information sharing. The paper then reviews different methods that are applied to improve data quality, and suggests that data virtualization provides a fertile environment for embedding these methods. We then look at how incorporating data quality methods within the abstracted relational views and data services provided by data virtualization addresses some of our critical data issues.

The ability to integrate data quality capabilities as part of a data virtualization infrastructure leads to improved data requirements analysis, metadata management, and data standards. By forcing organizations to improve their data governance practices, data virtualization has the potential to revolutionize ways of data quality assurance and improvement.

Data Quality Challenges in an "Enterprise World"

When business applications within an organization use data solely for operational purposes, business users expect sufficient levels of data quality to meet each application's needs. Yet the increasing need for line-of-business and organizational reporting, the expanding interest in value-added analytics, combined with the growing volumes of information drives an insatiable need for sharing large amounts of data among numerous producers and consumer across the enterprise.

This "enterprise world" is frequently characterized with a variety of downstream consumers with widely different information requirements reusing multiple data sets through repeated copying and replication. Yet inconsistency in reporting leads to an ongoing need for reconciliation of generated reports and analyses. Variance in use of commonly-accepted reference data concepts leads to inconsistencies and inaccuracies as well. In essence, the repeated pattern of extracting and transforming data for a particular use exposes incrementally introduced inconsistencies, ultimately leading to demand for greater levels of trust in the quality of the data, among other, more complex data quality challenges, specifically regarding these areas are shown in Figure 1.

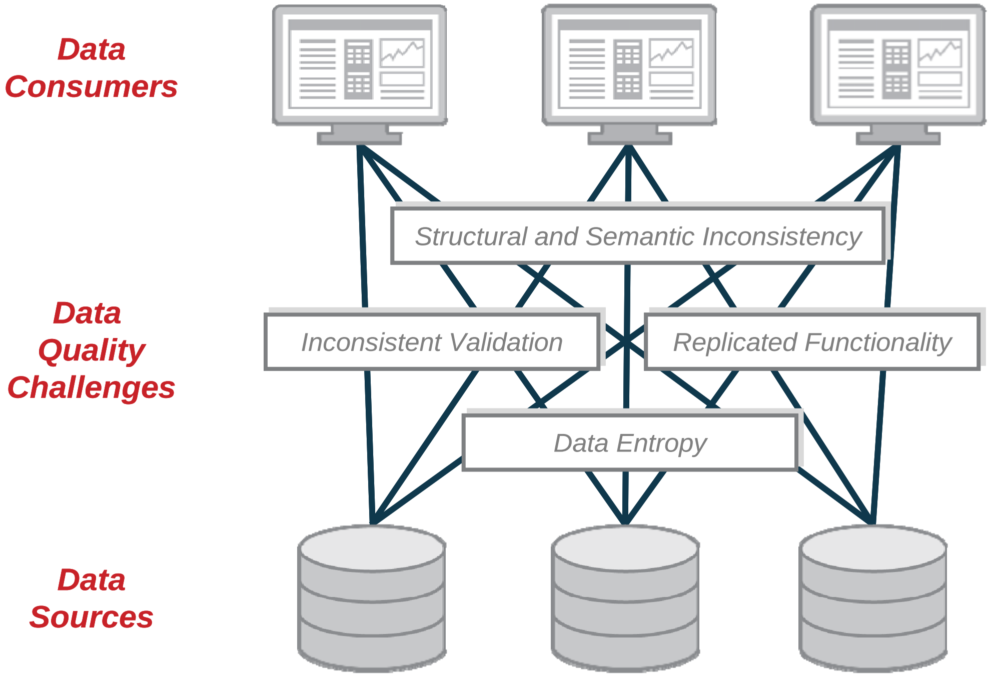

Figure 1: Data quality challenges

These challenges include:

1. Structural and Semantic inconsistency: Differences in formats, structures, and semantics presumed by downstream data consumers may confuse conclusions drawn from similar analyses;
2. Inconsistent validations: Data validation is inconsistently applied at various points in the business processes, with variant impacts downstream;
3. Replicated functionality: Repeatedly applying the same (or similar) data cleansing and identity resolution applications to data multiple times increases costs but does not ensure consistency;
4. Data entropy: Multiple copies of the same data lead to more data silos in which the quality of the data continues to degrade, especially when levels of service for consistency and synchronization are not defined or not met.

Bandages vs. Cures: Eliminating Root Causes

The lion's share of these data quality issues emerge as a result of the continuous extraction, copying, and transformation of data in different ways in preparation for various secondary uses, especially when one data consumer's data quality expectations differ from the expectations of others. Data consumers may attempt to rectify apparent data flaws through commonly applied data quality management techniques, such as:

- Data validation, in which data instances are subjected to inspection of conformance with defined data quality rules, such as ensuring that required data fields are properly populated, or that data attribute values are drawn from the appropriate data domain (such as US States);
- Data parsing and standardization, where data values are scanned and potentially reformatted (based on predefined rules) and converted into a standardized (or commonly accepted) representation, such as individual names or street addresses;
- Data cleansing, applied to sets of data instances to identify potential record matches, resolve duplication, and automatically correct known errors; and
- Data enrichment, in which a data instance is improved through the addition of value-added content, such as adding geographic location data (geocoding), or additional demographic information.

These approaches can be sound when applied judiciously. But when they are applied multiple times in different ways by multiple consumers at the point of consumption, not only is there a risk of introducing new errors and inconsistencies, the repetitive application of similar functionality in variant ways may not be cost effective as it allows for different tools and techniques to be used, as well as potentially degrades performance. Instead of helping to reduce the business impacts of data inconsistencies, repeated application of data quality techniques may actually have an inverse effect by amplifying inconsistencies.

Data Quality Improvement and Data Virtualization

In other words, copying primary data sources into multiple secondary repositories in preparation for reuse is actually a root cause of a number of data quality issues. If so, then eliminating that root cause by changing the approach to sharing data should alleviate many of those issues! An alternative approach to data sharing is *data virtualization*, which allows the data to remain in its primary data source until required for specific downstream needs.

Data virtualization provides layers of abstraction between the consuming applications and the primary data sources, and these abstraction layers present data taken from its original structures using canonical representations that simplify data reuse while enabling common data services to meet downstream consumer performance and data quality expectations. From a business application perspective, the abstraction provided through virtualization reduces complexity by standardizing the representations; by incorporating our data quality techniques directly into these layers of abstraction we can also address most of our data quality challenges as well. Embedding data quality management as part of data virtualization integrates data management best practices *directly into the application infrastructure*, and identifies data issues early in the information production process, and enables cleansing or other remediation processes to be performed before material impacts can be incurred.

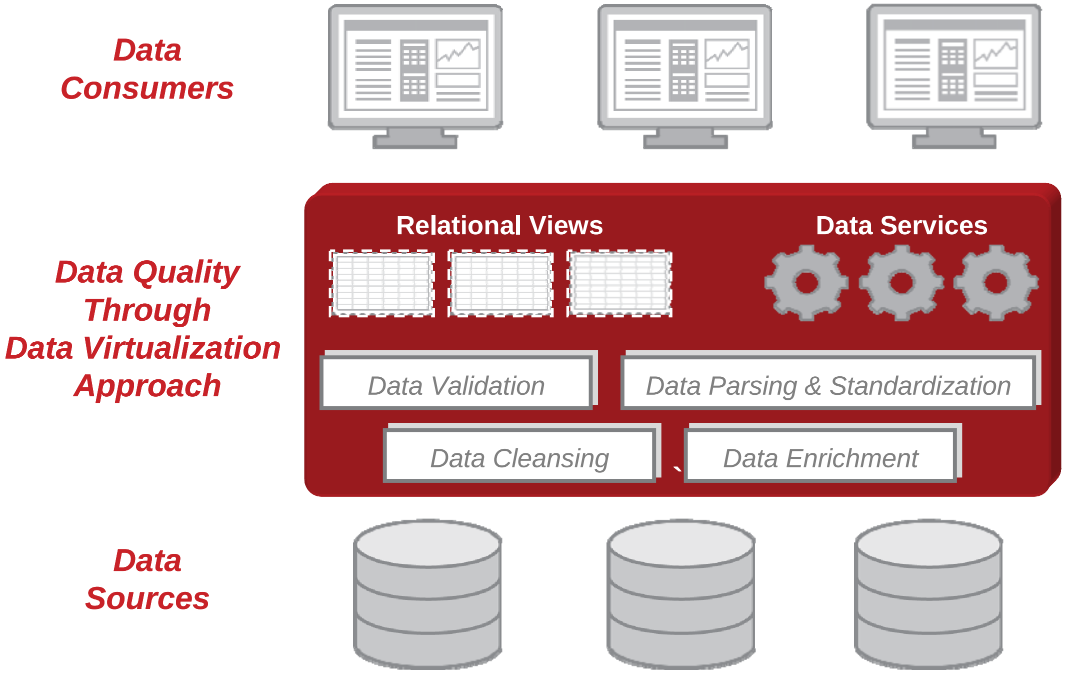

Figure 2: Integrating data quality within the data abstraction provided by data virtualization.

Resolving Structural and Semantic Inconsistency

The first of our data quality challenges is structural and semantic inconsistency. Structural inconsistency occurs when the same data concepts are represented using different structural and format representations. As an example, one customer data set may allocate 20 characters for the customer's last name, while another might allocate 25 characters.

Semantic inconsistency happens when one business process's understanding of a data concept differs from other business process's understandings. As an example, the sales business process might consider a *customer* to be "a party that to whom the company has provided a product or service in return for payment or transfer of value," while the customer service process might consider a *customer* to be "a party with whom the organization has an agreement for providing service." In either situation, differences in presumed structures, meanings, and definitions for similar data concepts leads to confusion.

One common root issue is inconsistent or absent enterprise metadata, caused by absent metadata and data standards, which can lead to variance and misinterpretation of semantics, especially when data sets are copied from various sources. When there are few standard representations or definitions, each business application must reinterpret the data to meet its needs, sometimes in ways that slightly, or even violently disagree.

Fortunately, resolving structural and semantic differences is a byproduct of the abstraction provided by a virtualized data environment. There is a need to transform a collection of physical models into a canonical representation of the critical data concepts employed by downstream data consumers. This drives the data management practitioners to improve the metadata management processes that drive standards for data definitions, semantics, and structures, which leads to a harmonized approach to data sharing. This canonical model presents a conformed representation derived from the numerous entities and attributes from the candidate data sources, with the data attributes from the source data sets grouped together into semantically similar entities as a way of assembling a common data dictionary across the business.

Enterprise Requirements and Interoperable Data Validation

Our next data quality challenge is the inconsistent application of extraction, transformation, and cleansing rules due to different requirements by the downstream consumers. The traditional approach to data quality improvement involves the application of data cleansing rules at the point of use. But even if the data sets are accumulated within the same warehouse or operational store, each downstream business process will have its own data expectations.

Typically there is no coordination among the data consumers, and they do not necessarily apply the same transformation or cleansing rules. And even if their rules are the same or similar, in the absence of any enterprise-wide coordination it is unlikely that the rules would be executed in the same order, or that the thresholds for acceptability would be the same. As a result, even though the same sources are being used, the results of the extraction and cleansing will vary, as will the corresponding results of the business applications.

Uncontrolled application of data extraction and cleansing rules reflect gaps in enterprise data governance, especially with respect to soliciting, collating, and resolving data requirements from the pool of downstream data consumers. When there are well-defined processes for enterprise data requirements analysis, data quality expectations can be collected from all the data users. Assessing enterprise data requirements is a core aspect of a data governance program, and will generally provide better quality data for all consumers.

Through the abstracted views and services provided via data virtualization, data consumers are presented with a canonical representation, and there is a reduced ability to modify copies of the original sources in uncontrolled ways. Therefore, the introduction of virtualization provides an opportunity to engage the downstream consumers to solicit their data quality requirements. In turn, most straightforward, data attribute-based validations can be embedded directly within one of the layers of abstraction, allowing the data management team to layer data validation constraints at specific points in the information provisioning process, which reduces the risk of inconsistency. Consolidating data quality requirements and implementing data validations as early as possible in the information production flow helps to not only ensure that the data quality is sufficient to meet all downstream data consumers, but also that any potential issues can be identified and remediated as early as possible.

Standardization of Data Quality Functionality

In a typical organization, when data sets are extracted and used by downstream business processes, the corresponding records and data attributes are subjected to transformations, standardization, and potentially data cleansing services. Yet in a distributed environment, the owners of each consuming business process may have selected their own approaches to implementing these services. Without any level of enterprise coordination, this introduces a potential for replicated functionality, our third data quality challenge.

Although each data consumer has its own set of tools and processes to qualify data, the types of transformations and corrections that are applied are probably similar, if not identical. These transformations are introduced to meet specific needs, and can be implemented in a variety of ways, embedded within different end-user computing tools such as the data warehouse dimensions, OLAP products, or even the report delivery and presentation tools. As a result, even when the intention is to apply the same standardizations and corrections, this inadvertent replication of functionality, applied in different ways, increases data inconsistency, both between the end product and the original sources, and among the collection of data consumers.

This replicated (and often inconsistent) functionality may have other business impacts. For example, one organization may hold more than one license to the same data integration and cleansing products, may have licenses for competing products, and may assign multiple staff personnel for implementation and support of the same functionality.

Implementing data virtualization forces the data consumers to reconsider the approaches used for data transformation and cleansing. Instead of duplicating the same functionality and incurring increased license and maintenance costs, data transformations and cleansing can be embedded within one of the layers of abstraction. This allows the organization to consolidate their cleansing and transformation resources, standardize the tools and rules, and reduce the overhead in maintaining quality by applying those approaches in a consistent manner *to the data derived from the original sources*. This approach finesses the issues associated with siloed data ownership and stewardship. Consistent application of data quality techniques to source data reduces downstream confusion, but does not impose additional requirements on the source data owners!

Reducing Unnecessary Data Set Replication

Our last significant data quality challenge involves increased opportunities for introducing new errors as data sets are repeatedly copied. In general, data systems are engineered in concert with the operational business application they are intended to support. Using that data for additional purposes such as reporting and analytics is typically only an afterthought. And with business applications having twenty year (or longer!) lifetimes, data reuse is bound to be characterized by large-scale data extraction and subsequent replication. This is particularly true with operational data stores, data warehouses, and front-end analytical tools, where data sets are often replicated multiple times before being presented to the end-consumer.

Of course, each time the data is copied, it is also subjected to any number of data transformations, each of which provides an opportunity for the introduction of data errors. Each subsequent copy resembles the original source less. Copying data can only lead to entropy and inconsistency.

It is true that there are situations requiring data sets to be copied. For example in some types of data analytics applications, some data replication is necessary in preparation for any consolidation necessary for realignment of data warehouse dimensions. However, situations like that are independent of data quality procedures. Data virtualization reduces the severity of data entropy as an issue. With today's hardware, software, and network capabilities, the high performance query algorithms and caching provided by data virtualization can reduce, or even eliminate access latency. With data virtualization, instead of repeatedly copying and modifying data, one can allow the source data to remain in its original location, and access and enrich/cleanse the data when it is accessed.

Integrating Data Quality into the Infrastructure

It is true that there are some data quality activities that are more effectively performed using replicated data. For example, identity resolution and duplicate entity analysis from multiple data sources require that the data sets are all available for scanning and analysis. There are some situations where certain cross-table validations (such as ensuring conformance to cardinality constraints, consistency/reasonableness expectations along multiple layers of aggregation, or ensuring multiple table referential integrity) can be performed in a more efficient manner when there is rapid access to entire data sets.

But all of the issues discussed in this paper manifest themselves more frequently in environments that allow for ungoverned data extraction and replication. And since data flaws that are introduced through any of the information production flows may lead to multiple impacts downstream, reducing data replication and consolidating data quality services will likewise reduce negative business impacts related to data errors.

When the data quality team members have a comprehensive approach to soliciting data quality expectations from data consumers, they can consolidate the requirements for data validation and institute data controls as a “firewall” to prevent errors from entering the enterprise in the first place. By leveraging data virtualization, the data quality team can integrate data validation close to the point of data acquisition and can consistently apply the same transformations, standardizations, and corrections in support of all downstream data consumers.

Incorporating data inspection, monitoring, and notification in the event of a data failure early in the data management flow will prevent process failures from occurring, as well as reduce the need for cleansing. In addition, consolidating data quality expectations and standardizing the approaches for validation, inspection, and monitoring will lead to predictable and consistent levels of quality across the spectrum of downstream consumers, which reduces the need for report reconciliations. In other words, a repeatable process of data requirements assessment, review and consolidation of data quality expectations, and development of high-quality views and data services will support the integration of a large degree of data quality management within the abstraction layer provided by data virtualization.

Considerations –Data Governance Strategy and Data Virtualization

Because of the historically organic application development, it is not surprising that the quality of the typical data source is sufficient to meet the needs of the original transactional or operational business application's intent. However, as data sources are tapped for multiple additional uses such as business intelligence and reporting by downstream consumers, the expectations for quality change as the data sets are copied, replicated, and used in different ways. Yet when the downstream consumers do not have any administrative control over the original sources, and cannot influence changes by the original application owners, the ability to meet data quality expectations diminishes as the data sets are put to their alternate purposes.

The approaches typically employed today to enable the data to meet data consumer expectations have their own consequences, especially as each consumer applies similar transformations and corrections in different ways and execution order. Each downstream business process is forced to interpret the data to meet its own specific needs, leading to structural inconsistency, semantic inconsistency, inconsistently applied validations, and increased needs for reconciliations.

In light of the existing constraints where the data cannot be fixed upstream and it does not make sense to cleanse it in multiple different ways downstream, data virtualization offers a fresh approach, in effect enabling consistent, high quality data for multiple uses "mid-stream." By providing reusable relational views and data services that combine common data validation, standardization, and transformations, data virtualization can bring clarity to the source *without forcing any changes to the source.* But standardizing the clarification enables a consistent view that is presented to all downstream consumers, providing the benefits of integrated data quality without any of the implied organizational consequences.

Enterprise data quality can be viewed as a Gordian knot and data virtualization provides a fundamental approach to slice through this barrier to consistency and finesse the organizational challenges associated with data quality improvement. Any data quality team faced with the challenges discussed in this paper would benefit from considering data virtualization as an approach to implementing data quality control and consistency.

About the Author

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data quality, master data management, and business intelligence. David is a prolific author regarding BI best practices, via the expert channel at www.b-eye-network.com and numerous books and papers on BI and data quality. His book, “Master Data Management,” (August 2009) has been endorsed by data management industry leaders, and his valuable MDM insights can be reviewed at www.mdmbook.com. Learn about data quality improvement and data governance in “The Practitioner’s Guide to Data Quality Improvement,” (October 2010).

David can be reached at loshin@knowledge-integrity.com.

About the Sponsor

COMPOSITE
SOFTWARE

Composite Software, Inc. is the data virtualization gold standard at ten of the top 20 banks, six of the top ten pharmaceutical companies, four of the top five energy firms, major media and technology organizations; and multiple government agencies. These are among the hundreds of global organizations with disparate, complex information environments that count on the Composite to increase their data agility, cut costs and reduce risk.

Backed by nearly a decade of pioneering R&D, Composite is the data virtualization performance leader, scaling from project to enterprise for data federation, data warehouse extension, enterprise data sharing, real-time and cloud computing data integration. Composite Software is a privately held, Silicon Valley-based corporation. For more information, please visit www.compositesw.com.