

Data Integration Alternatives – Managing Value and Quality

Using a Governed Approach to Incorporating Data Quality Services
Within the Data Integration Process

WHITE PAPER:

DATA QUALITY & DATA INTEGRATION

David Loshin • President
Knowledge-Integrity, Inc.



Data Integration Alternatives – Managing Value and Quality

Using a Governed Approach to Incorporating Data Quality Services Within the Data Integration Process

ABSTRACT

WITH THE RATE OF DATA VOLUME GROWTH INCREASING AT A BREAKNECK PACE, SMART ORGANIZATIONS ARE INCREASINGLY RELYING ON REPORTING AND ANALYTICS TO NOT JUST RUN, BUT IMPROVE THE WAY THAT BUSINESS IS DONE. YET WHILE DECADES OF INVESTMENTS IN TRANSACTIONAL AND OPERATIONAL BUSINESS APPLICATIONS HAVE LED TO VIRTUAL “ISLANDS OF DATA,” THERE IS A GROWING LIST OF ENTERPRISE APPLICATIONS SUCH AS BUSINESS INTELLIGENCE AND DATA WAREHOUSING, CUSTOMER RELATIONSHIP MANAGEMENT, ENTERPRISE RESOURCE PLANNING, AND EVEN MORE COMPLEX APPROACHES TO BUSINESS ANALYTICS THAT REQUIRE ACCESS TO DATA SETS FROM A VARIETY OF SOURCES.

KEY DATA GOVERNANCE PROCESSES FACILITATE THE COLLECTION AND STANDARDIZATION OF ENTERPRISE DATA QUALITY REQUIREMENTS

Data Integration is Everywhere

With the rate of data volume growth increasing at a breakneck pace, smart organizations are increasingly relying on reporting and analytics to not just run, but improve the way that business is done. Data centralization becomes key to deploying strategic enterprise applications. Operational data stores, data warehouses, data marts, mash ups, federated operational systems, self-service reporting, data exchanges, and other analytic and operational applications require a greater degree of data sharing than ever before. Satisfying the information demands of these secondary-use business applications becomes a primary objective, and that means moving data from the original sources to the target business data systems. In other words, those needs must be satisfied using *data integration*.

Data Integration Alternatives

As the needs of downstream consumers have become more sophisticated, different approaches to data integration have evolved. The more traditional “*Extract, Transform, Load*” (ETL) approach which takes the data from its sources to a staging area in which data sets are manipulated and transformed into a target representation. An alternate approach is *data virtualization*, in which the data remains stored at the source and a conceptual view is materialized on demand.

Traditional ETL

The most common approach to data integration relies on some variation of the ETL paradigm. Because data sources that are used to populate downstream and secondary-use business applications often live in many different formats, file and/or table structures, and sometimes even using different underlying character encodings, there is a predisposition to normalize data set representations before attempting to merge them into a target downstream system. The extraction component implies specially-engineered routines employed to fetch data from the sources, which

will also require specially-designed transformations that apply a series of functions to normalize, cleanse, standardize, derive, translate, and other functions necessary to massage the data into a format that is suitably consistent with other transformed data sources in preparation for the target data systems. At that point, the data is ready to be propagated and loaded into the target destination, either overwriting the existing data or periodically augmenting the existing target data set.

Data Virtualization

As opposed to the traditional approach of extracting data from multiple sources and temporarily storing those data sets at a staging area, a different approach, called *data federation* or *data virtualization* allows the source data sets to remain in their original locations. Data virtualization introduces abstraction layers over a variety of native data sources and, as a byproduct, provides relational views without requiring that data be extracted from its source. This approach to abstraction enables the creation of reusable data services, and the data abstraction layers typically deployed within a data virtualization environment allow for the presentation of a standardized logical representation of enterprise data concepts, thereby allowing many different downstream data consumers to see a view of the data that is both structurally and semantically consistent.

Data Challenges: Completeness, Consistency, Reasonableness

With an increased interest in developing business applications that repurpose primary data sources for secondary uses, downstream consumers may have widely different expectations of the data, especially in terms of data quality. But when data sets are used for purposes for which they were not originally intended, those data users are often forced to redefine and reinterpret the meaning of the original data sets.

Data Integration Alternatives – Managing Value and Quality

Using a Governed Approach to Incorporating Data Quality Services Within the Data Integration Process

4

Extracting and transforming the data multiple times in different ways for different applications may lead to variant results and continual need for reconciliations, leading to mistrust, wasted time, rework, and questionable results. Some typical issues include:

- Missing data element values that skew counts and other aggregations;
- Variance in use of commonly-accepted reference data introduce inconsistencies and inaccuracies;
- Differences in formats, structures, and semantics presumed by downstream business applications may lead to drawing different conclusions from similar analyses;
- Inconsistency in reporting that leads to an ongoing need for reconciliation of generated reports;
- Different implied semantics leads to misinterpretations and missed reasonableness expectations.

Essentially, the absence of standards for structure, formats, and definitions of repurposed data leads to issues that emerge as a result of ungoverned data integration processes – incomplete data, inconsistent data, and data that does not meet reasonableness expectations.

Adding Value Using Governed Data Quality Services

If the challenges are introduced as a byproduct of the need for data integration, one approach to solving it is by recognizing that the issues exist, figuring out what the issues are, and then retooling the data integration process so that it incorporates ways to remediate data flaws before they have material impact. Fortunately, a governed approach to incorporating data quality services within the data integration process can alleviate a number of the issues that can emerge as a byproduct of ungoverned

data consolidation. The first step is to institute some key data governance processes to facilitate the collection and standardization of enterprise data quality requirements. The second step is embedding data quality management techniques within the data integration strategy.

Data Governance Practices

While an enterprise data governance program will encompass a wide variety of data management processes, the practical demands of data integration suggest focusing on a subset of those practices that directly support data integration, such as:

- Data requirements analysis – Typical business application development considers the collection of data requirements as subsidiary to the functional requirements analysis process. But because enterprise projects such as data warehousing and customer relationship management cross line-of-business boundaries, there is a need for a well-defined process for soliciting, documenting, and synthesizing the collected information expectations that all downstream users will expect, and then translate those expectations into data requirements to be imposed on all candidate data sources. Not only does this impose a radical change in requirements gathering, it also requires the kind of oversight provided by a data governance infrastructure.
- Data standards review – Defining data standards can address the challenge of inconsistency, especially aligning data element definitions and semantics. When key stakeholders from across the enterprise participate in a review and approval process for proposed data standards, there is a degree of confidence that the standards will be defined so that the collected downstream data consumer requirements will be observed.

HAVING DATA GOVERNANCE PRACTICES IN PLACE SIMPLIFIES THE INCORPORATION OF DATA QUALITY TECHNIQUES

- Metadata management – These include processes for documenting the approved standard structures and definitions for reference data domains and data exchange and providing a means for communicating those standards.

Data Quality Services

Having data governance practices in place simplifies the incorporation of data quality techniques such as these:

- Parsing and Standardization – Parsing is a process that relies on defined formats, patterns, and structures to determine when data values conform to a common set of expectations. Parsing is used in concert with a set of standardization rules triggered to transform the input data into a form that can be more effectively used, either to standardize the representation (presuming a valid representation) or to correct the values (should known errors be identified). Parsing and standardization can employ a library of data domains and rules to split data values into multiple components and rearrange the components into a normalized format. Standardization can also change full words to abbreviations, or abbreviations to full words, transform nicknames into a standard name form, translate across languages (e.g., Spanish to English), correct common misspellings, and reduce value variance to improve record linkage for deduplication and data cleansing.
- Data cleansing – When data values are recognized as being inconsistent or incorrect, and the data flaws cannot be corrected at the point of origin, an alternative is to apply transformation rules to impute data values, correct names or addresses, eliminate extraneous and/or meaningless data, and even merge duplicate records. Cleansing the data directly ensures that the data that meets some level of suitability; incorporating the same approach to parsing, standardization, and cleansing as part of the data integration process standardizes the

transformations so that a consistent view of the data is provided to all downstream data consumers.

- Data validation – When there is no coordination among the data consumers they might not necessarily apply the same data validations in the same way. Even if their rules are the same or similar, when validations are applied at the point of use, 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 validation may vary as well. By incorporating a standard set of data validations within the data integration process, the constraints can be tested at specific points in the information flow process, thereby reducing the risk of inconsistency. Soliciting data quality requirements from the collection of downstream data consumers allows you to define data quality rules; implementing validation of compliance to these rules early in the process can help ensure that the quality of the data is sufficient to meet the business needs and allow any potential issues to be identified and remediated early and consistently.

Considerations

The approach used for implementing data integration should not interfere with the desire to improve the quality of the data in a coherent and consistent manner. Most traditional ETL tools vendors have recognized the need for incorporating data quality techniques, and many have either forged partner relationships with, or have completely acquired data quality tools vendors. Today, it is rare to find an end-to-end extraction, transformation, and loading tool that does not encourage the definition and implementation of embedded parsing, standardization, cleansing, and validation.

Alternatively, data virtualization tools are also increasingly tethered to data quality tools and technology. The

Data Integration Alternatives – Managing Value and Quality

Using a Governed Approach to Incorporating Data Quality Services Within the Data Integration Process

6

abstraction provided via data virtualization provides an opportunity for the data management team to engage the downstream consumers, solicit their data quality requirements, and directly embed data attribute-based validations within one of the layers of abstraction. Consolidating data quality requirements and implementing data validation constraints at specific points in the information production flow reduces the risk of inconsistency, helps to ensure that the data quality is sufficient to meet the downstream data consumer needs, and alerts the data stewards to any potential issues that can be identified and remediated early in the data integration process.

Here is the bottom line: Data integration is becoming pervasive across the organization. By introducing governed process that simplify information reuse in a consistent manner, trust in reporting and analytics will increase, benefitting all stakeholders across the organization!

FOR MORE INFORMATION ON DATA QUALITY AND DATA INTEGRATION SOLUTIONS, CALL PITNEY BOWES BUSINESS INSIGHT TODAY OR VISIT OUR WEBSITES.

David Loshin is the President of Knowledge Integrity, Inc., a consulting and development company focusing on customized information management solutions including information quality solutions consulting, information quality training and business rules solutions. Loshin is the author of Master Data Management, Enterprise Knowledge Management – The Data Quality Approach and Business Intelligence – The Savvy Manager’s Guide and is a frequent speaker on maximizing the value of information. David can be reached at loshin@knowledge-integrity.com or at (301) 754-6350.

UNITED STATES

One Global View
Troy, NY 12180
1.800.327.8627
pbbi.sales@pb.com
www.pbinsight.com

CANADA

26 Wellington Street East
Suite 500
Toronto, ON M5E 1S2
1.800.268.3282
pbbi.canada.sales@pb.com
www.pbinsight.ca

EMEA HEADQUARTERS

Minton Place
Victoria Street
Windsor, Berkshire SL4 1EG
+44.800.840.0001
pbbi.europe@pb.com
www.pbinsight.co.uk

ASIA PACIFIC HEADQUARTERS

Level 7, 1 Elizabeth Plaza
North Sydney NSW 2060
+61.2.9437.6255
pbbi.australia@pb.com
pbbi.singapore@pb.com
pbbi.china@pb.com
www.pbinsight.com.au