Building a Data Quality Scorecard for Operational Data Governance

A White Paper by David Loshin

WHITE PAPER

Table of Contents

Introduction

There are few businesses today that do not rely on high-quality information to support performance and productivity. In today's organizations, the importance of high-quality data is dictated by the needs of the operational and the analytical applications that will process the data. Data governance is a means for data quality assurance in two contexts:

1. The ability to protect against negative business impacts by identifying data-quality issues before any material impact takes place (such as failure to comply with regulations or allowing fraudulent transactions to occur).

2. Establishing trust in the data and providing confidence that the organization can take advantage of business opportunities as they arise.

Operational data governance is the manifestation of the processes and protocols necessary to ensure that an acceptable level of confidence in the data effectively satisfies the organization's business needs. A data governance program defines the roles, responsibilities and accountabilities associated with managing data quality. Rewarding individuals who are successful at their roles and responsibilities can ensure the success of the data governance program. To measure this, a data quality scorecard provides an effective management tool for monitoring organizational performance with respect to data quality control.

Establishing Business Objectives

In this paper, we look at taking the concepts of data governance into general practice as a byproduct of the processes of inspecting and managing data quality control. By considering how the business is affected by poor data quality – and establishing measurable metrics that correlate data quality to business goals – organizational data quality can be quantified and reported within the context of a scorecard that describes the level of trustworthiness of enterprise data.

Business Drivers

Levels of scrutiny are increasing across the enterprise – industry organizations are dictating expected practices for participation within the community, while municipal, state and federal governments are introducing regulations and policies for both data-quality processes and data quality itself. Successful implementation of automated business processing streams is related to high-quality data as well. The increased use of business intelligence platforms for measuring performance against operational and strategic goals is indicative of a maturing view of what the organization's business drivers are, and how performance is supported by all aspects of quality, including data quality.

Establishing the trust of a unified view of business information and decreasing the need for redundant storage and seemingly never-ending stream of reconciliations helps improve operational efficiency. Reviewing the specific ways that information supports the achievement of business objectives helps analysts clarify the business drivers for data governance and data quality and lays out the parameters of what "acceptable data quality" means within the organization.

For example, business clients making decisions using analytic applications dependent on data warehouse data may have to defer making decisions or, even worse, be at risk for making incorrect decisions when there is no oversight in controlling the quality of the data in the warehouse. The business user would not be able to provide usable insight into which customers to target, which products to promote or where to concentrate efforts to maximize the supply chain. In this scenario, a business driver is to ensure an acceptable level of confidence in the reporting and analysis that satisfies the business needs defined by the use of enterprise information. Similar drivers can be identified in relation to transaction processing, regulatory compliance or conforming to industry standards.

Success Criteria

Identifying the business drivers establishes the operational governance direction by enabling the data governance team to prioritize the information policies in relation to the risk of material impact. Listing the expectations for acceptable data suggests quantifiable measurements, and this allows business analysts or data stewards to specify acceptability thresholds for those emerging metrics. By listing the critical expectations, methods for measurement, and specifying thresholds, the business clients can associate data governance with levels of success in their business activities.

For our analytic application example, the success criteria can be noted in relation to the ways that data quality improvement reduces time spent on diagnosis and correction. Success will mean increasing the speed of delivering information as well as increasing confidence in the decisions. Articulating specific achievements or milestones as success criteria allows managers to gauge individual accountability and reward achievement.

Data Quality Control and Operational Data Governance

A data quality control framework enables the ability to identify and document emerging data issues, then initiate a workflow to remediate these problems. Operational data governance leads to an increase in the level of trust in the data, as the ability to catch an issue is pushed further and further upstream until the point of data acquisition or creation. A data quality control process provides a safety net that eliminates the need for downstream users to monitor for poor-quality data. As long as the controls are transparent and auditable, those downstream users can trust the data that feeds their applications.

Data Quality Inspection and Control

For years, nobody expected that data flaws could directly affect business operations. However, the reality is that errors – especially those that can be described as violations of expectations for completeness, accuracy, timeliness, consistency and other dimensions of data quality – often impede the successful completion of information processing streams and, consequently, their dependent business processes. However, no matter how much effort is expended on data filters or edits, there are always going to be issues requiring attention and remediation.

Operational data governance combines the ability to identify data errors as early as possible with the process of initiating the activities necessary to address those errors to avoid or minimize any downstream impacts. This essentially includes notifying the right individuals to address the issue and determining if the issue can be resolved appropriately within an agreed time frame. Data inspection processes are instituted to measure and monitor compliance with data quality rules, while service level agreements (SLAs) specify the reasonable expectations for response and remediation.

Note that data quality inspection differs from data validation. While the data validation process reviews and measures conformance of data with a set of defined business rules, inspection is an ongoing process to:

- Reduce the number of errors to a reasonable and manageable level.
- Enable the identification of data flaws along with a protocol for interactively making adjustments to enable the completion of the processing stream.
- Institute a mitigation or remediation of the root cause within an agreed time frame.

The value of data quality inspection as part of operational data governance is in establishing trust on behalf of downstream users that any issue likely to cause a significant business impact is caught early enough to avoid any significant impact on operations. Without this inspection process, poor-quality data pervades every system, complicating practically any operational or analytical process.

Data Quality Service Level Agreements

A key component of governing data quality control is an SLA. For each processing stream, we can define a data quality SLA incorporating a number of items:

- Location in the processing stream that is covered by the SLA.
- Data elements covered by the agreement.
- Business effects associated with data flaws.
- Data quality dimensions associated with each data element.
- Expectations for quality for each data element for each of the identified dimensions.
- Methods for measuring against those expectations.

- Acceptability threshold for each measurement.
- The individual to be notified in case the acceptability threshold is not met.
- Times for expected resolution or remediation of the issue.
- Escalation strategy when the resolution times are not met.

Monitoring Performance of Data Governance

While there are practices in place for measuring and monitoring certain aspects of organizational data quality, there is an opportunity to evaluate the relationship between the business impacts of noncompliant data as indicated by the business clients and the defined thresholds for data quality acceptability. The degree of acceptability becomes the standard against which the data is measured, with operational data governance instituted within the context of measuring performance in relation to the data governance procedures.

This measurement essentially covers conformance to the defined standards, as well as monitoring the staff's ability to take specific actions when the data sets do not conform. Given the set of data quality rules, methods for measuring conformance, the acceptability thresholds defined by the business clients, and the SLAs, we can monitor data governance. And we can observe not only compliance of the data to the business rules, but also the compliance of data stewards to observing the processes associated with data risks and failures.

Data Quality Metrics and the Data Quality Scorecard

Putting the processes in place for defining a data quality SLA for operational data governance depends on measuring conformance to business expectations and knowing when the appropriate data stewards need to be notified to remediate an issue. This requires two things: a method for quantifying conformance and the threshold for acceptability.

Since business policies drive the way the organization does business, business policy conformance is related to information policy conformance. Data governance reflects the way that information policies support the business policies and impose data rules that can be monitored throughout the business processing streams. In essence, performance objectives center on maximizing productivity and goodwill while reducing organizational risks and operating costs. In that context, business policies are defined or imposed to constrain or manage the way that business is performed, and each business policy may loosely imply (or even explicitly define) data definitions, information policies, and even data structures and formats.

Therefore, reverse engineering the relationship between business impacts and the associated data rules provides the means for quantifying conformance to expectations. These data quality metrics will roll up into a data quality scorecard. This suggests that a good way to start establishing relevant data quality metrics is to evaluate how data flaws affect the ability of application clients to efficiently achieve their business goals. In other words, evaluate the business impacts of data flaws and determine the dimensions of data quality that can be used to define data quality metrics.

Evaluating Business Impacts and Dimensions of Data Quality

In the context of data governance, we seek ways to effectively measure conformance to the business expectations that are manifested as business rules. Categorizing the impacts associated with poor data quality can help to simplify the process of evaluation – distinguishing monetary impacts (such as increased operating costs or decreased revenues) from risk impacts (such as those associated with regulatory compliance or sunk development costs) or productivity impacts (such as decreased throughput).

Correlating defined business rules, based on fundamental data quality principles, allows one to represent different measurable aspects of data quality, and can be used in characterizing relevance across a set of application domains to support the data governance program. Measurements can be observed to inspect data quality performance at different levels of the operational business hierarchy, enabling monitoring of both line-of-business and enterprise data governance.

At the data element and data value level, intrinsic data quality dimensions focus on rules relating directly to the data values themselves out of a specific data or model context. Some examples of intrinsic dimensions are:

- Accuracy – the degree with which data values agree with an identified source of correct information.
- Lineage – documentation of the ability to identify the originating source of any new or updated data element.
- Structural consistency – characterizing the consistency in the representation of similar attribute values, both within the same data set and across the data models associated with related tables.

Contextual dimensions depend on the ways that business policies are imposed over the systems and processes relating to data instances and data sets. Some sample contextual dimensions are:

- Timeliness – the time expectation for accessibility of information.
- Currency – which information is current with the world that it models.
- Consistency – relationships between values within a single record, or across many records in one or more tables.
- Completeness – the expectation that certain attributes are expected to have assigned values in a data set.

Defining Quantifiable Data Quality Metrics

Having identified the dimensions of data quality that are relevant to the business processes, we can map the information policies and their corresponding business rules to those dimensions. For example, consider a business policy that specifies that personal data collected over the Web may be shared only if the user has not opted out of that sharing process. This business policy defines information policies; the data model must have a data attribute specifying whether a user has opted out of information sharing, and that attribute must be checked before any records may be shared. This also provides us with a measurable metric: the count of shared records for those users who have opted out of sharing.

The same successive refinement can be applied to almost every business policy and its corresponding information policies. As we distill out the information requirements, we also capture assertions about the business user expectations for the result of the operational processes. Many of these assertions can be expressed as rules for determining whether a record does or does not conform to the expectations. The assertion is a quantifiable measurement when it results in a count of nonconforming records, and therefore monitoring data against that assertion provides the necessary data control.

Once we have reviewed methods for inspecting and measuring against those dimensions in a quantifiable manner, the next step is to interview the business users to determine the acceptability thresholds. Scoring below the acceptability threshold indicates that the data does not meet business expectations, and highlights the boundary at which noncompliance with expectations may lead to material impact to the downstream business functions. Integrating these thresholds with the methods for measurement completes the construction of the data quality control. Missing the desired threshold will trigger a data quality event, notifying the data steward and possibly even recommending specific actions for mitigating the discovered issue.

Automating the Scorecard Process

Articulating data quality metrics is a valuable exercise, and in fact may supplement metrics or controls that already are in place in some processing streams. However, despite the existence of these controls for measuring and reporting data validity, frequently there is no framework for automatically measuring, logging, collecting, communicating and presenting the results to those entrusted with data stewardship. Moreover, the objective of data governance is not only to report on the acceptability of data, but also to remediate issues and eliminate their root causes with the reasonable times established within the data quality SLA.

Identifying the metrics is good, but better yet is integrating their measurements and reporting into a process that automatically inspects conformance to data expectations (at any point where data is shared between activities within a processing stream), compares the data against the acceptability thresholds, and initiates events to alert data stewards to take specific actions. It's these processes that truly make governance operational.

Capturing Metrics and Their Measurements

The techniques that exist within the organization for collecting, presenting and validating metrics must be evaluated in preparation for automating selected repeatable processes. Cataloging existing measurements and qualifying their relevance helps to filter out processes that do not provide business value and reduces potential duplication of effort in measuring and monitoring critical data quality metrics. Surviving measurements of relevant metrics are to be collected and presented in a hierarchical manner within a scorecard, reflecting the ways that individual metrics roll up into higher level characterizations of compliance with expectations while allowing for drill-down to isolate the source of specific issues. As is shown in Figure 1, collecting the measurements for a data quality scorecard would incorporate:

1. Standardizing business processes for automatically populating selected metrics into a common repository.

2. Collecting requirements for an appropriate level of design for a data model for capturing data quality metrics.

3. Standardizing a reporting template for reporting and presenting data quality metrics.

4. Automating the extraction of metric data from the repository.

5. Automating the population of the reporting and presentation template.

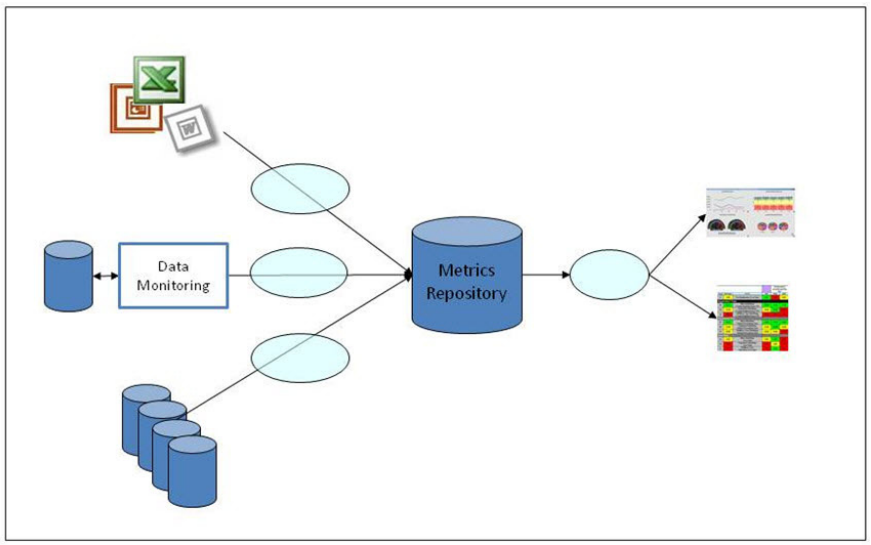

Figure 1: Automating the collection and reporting of data quality metrics

Reporting and Presentation

In Figure 2 we have an example of a high-level data quality scorecard reflecting three aspects of measurements. The first, Data Quality Score, is an accumulated score computed as a function of the underlying data quality metrics. The second, Data Quality Policy, refers to the degree to which the data governance team has identified business impacts and defined corresponding metrics, whether those metrics have processes for measurement, and whether acceptability thresholds and Data Quality Service Level Agreements (DQ SLAs) have been agreed to by business clients from the key business areas. The third, Data Governance, provides an accumulated score reflecting the observance of the DQ SLAs by the team members and functional area data stewards (such as resolving data quality issues within specified time periods).

In this example, scores are qualified as acceptable (green), at risk (yellow), unacceptable (red), or not yet defined (blue). Sample scores for Data Quality Policy might be to assign green if more than 90 percent of the metrics have processes and thresholds; yellow if between 50 percent and 90 percent, and red if less than 50 percent. If governance processes are not yet in place, we can designate a “not yet assigned” score for Data Governance.

	Data Quality Score	Data Quality Policy	Data Governance
Sales			
Marketing			
Human Resources			
Finance			
Fulfillment			
Manufacturing			
Supplier			

	Acceptable
	At Risk
	Unacceptable
	Not yet defined

Figure 2: Sample Data Quality Scorecard

Interestingly, the process described here for monitoring the performance of the data governance activities is not significantly different from the processes used for monitoring any type of performance. Many organizations already have an infrastructure to support the definition of operational performance indicators and the supporting measurements that feed a hierarchical view of productivity within a business intelligence framework. The data quality metrics repository essentially acts as a data mart that feeds front-end reporting and analytics, and the more sophisticated tools may provide visualization widgets and drill-down to support the data stewardship activity.

Summary

In this paper we have described a target state for operational data governance that is managed via a comprehensive data quality scorecard that communicates:

- The qualified oversight of data quality along business lines.
- The degree of levels of trust in the data in use across the application infrastructure.
- The ability for data stewards to drill down to identify the area of measurement that contributes most to missed expectations.

Processes can be put in place to facilitate the definition of data quality service level agreements and the metrics that support those SLAs. The collection of statistics associated with data governance and the presentation of the resulting scores to the stakeholders will demonstrate that with respect to data, the business processes are in control and that the data is of a predictable level of acceptable quality. Providing a data quality scorecard provides transparency to the data governance process by summarizing the usability of the data as defined by the business users. The data governance team will work with the business users to integrate the hierarchies of data quality expectations and rules into the metrics collection and reporting framework and enable drill-through to track down specific issues that affect organizational data. The processes for instituting data quality business rules and data validation can then be used to demonstrate an auditable process for governing the quality of organizational data.

About the Author

David Loshin, President of Knowledge Integrity Inc., is a recognized thought leader and expert consultant in the areas of data quality, master data management and business intelligence. Loshin is a prolific author regarding data management best practices and has written numerous books, white papers and Web seminars on a variety of data management best practices. His book *Business Intelligence: The Savvy Manager's Guide* has been hailed as a resource allowing readers to "gain an understanding of business intelligence, business management disciplines, data warehousing and how all of the pieces work together." His book *Master Data Management* has been endorsed by data management industry leaders, and his valuable MDM insights can be reviewed at mdmbook.com. Loshin is also the author of the recent book *The Practitioner's Guide to Data Quality Improvement.* He can be reached at loshin@knowledge-integrity.com.

About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976, SAS has been giving customers around the world THE POWER TO KNOW®.

SAS Institute Inc. World Headquarters +1 919 677 8000
To contact your local SAS office, please visit: **sas.com/offices**