Populating a Data Quality Scorecard with Relevant Metrics
Table of Contents

Introduction ................................................................. 1
Useful vs. So-What Metrics .............................................. 2
   The So-What Metric .................................................. 2
   Defining Relevant Metrics ......................................... 3
Relating Business Objectives and Quality Data ...................... 3
   Business Impact Categories ........................................ 4
Relating Issues to Effects ................................................. 4
Dimensions of Data Quality .............................................. 5
   Accuracy ................................................................. 6
   Lineage. ................................................................. 6
   Completeness ........................................................ 6
   Currency ............................................................... 7
   Reasonability ........................................................ 7
   Structural Consistency ............................................... 7
   Identifiability ........................................................ 7
   Additional Dimensions .............................................. 8
Reporting the Scorecard ................................................... 8
   The Data Quality Issues View .................................. 8
   The Business Process View .................................... 8
   The Business Impact View .................................... 9
   Managing Scorecard Views .................................... 9
Summary ........................................................................ 10

Contributor:

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 lothin@knowledge-integrity.com.
Introduction

Once an organization has decided to institute a data quality scorecard, its first step is determining the types of metrics to use. Too often, data governance teams rely on existing measurements for a data quality scorecard. But without establishing a relationship between these measurements and the business’s success criteria, it can be difficult to react effectively to emerging data quality issues and to determine whether fixing these problems has any measurable business value. When it comes to data governance, differentiating between “so-what” measurements and relevant metrics is a success factor in managing expectations for data quality.

This paper explores ways to qualify data control and measures to support the governance program. It will examine how data management practitioners can define metrics that are relevant. Organizations must look at the characteristics of relevant data quality metrics and provide a process for understanding how specific data quality issues affect their business. The next step is to provide a framework for defining measurement processes to quantify the business value of high-quality data.

Processes for computing raw data quality scores for base-level metrics can then feed different levels of metrics using different views to address the scorecard needs of various groups across the organization. Ultimately, this drives the description, definition and management of base-level and complex data quality metrics so that:

- Scorecard relevancy is based on a hierarchical rollup of metrics.
- The definition of the metrics is separated from its context, thereby allowing the same measurement to be used in different contexts with different validity thresholds and weights.
- Appropriate reporting can be generated based on the level of detail expected for the data consumer’s specific role and accountability.
Useful vs. So-What Metrics

Famous physicist and inventor Lord Kelvin said about measurement – “If you cannot measure it, then you cannot improve it.” This is the rallying cry for the data quality community. The need to measure has driven quality analysts to evaluate ways to define metrics and their corresponding processes of measurement. Unfortunately, in their zeal to identify and use different kinds of metrics, many have inadvertently flipped the concept, thinking that “If you can measure it, you can improve it,” or even less helpful: “If it is measured, you can report it.”

The So-What Metric

The fascination with metrics for data quality management is based on the idea that continuously monitoring data can ensure that enterprise information meets business process owner requirements, and that data stewards are notified to take action when an issue is identified. The desire to have a framework for reporting data quality metrics often supersedes the need to define the relevant metrics. Analysts assemble data quality scorecards and then search the organization to find existing measurements from auditing or oversight activities to populate the scorecard.

A challenge emerges when it becomes evident that existing measurements taken from auditing or oversight processes may be insufficient for data quality management. For example, in one organization a number of data measurements performed as part of a Sarbanes-Oxley (SOX) compliance initiative were folded into an enterprise data quality scorecard. Because the compliance efforts resulted in completeness and reasonability of data, the thought was that at least some should be reusable. A number of these measures were selected for inclusion. At one of the information-risk team’s meetings, a senior manager questioned one randomly selected data element’s measure of more than 97 percent completion: Was that an acceptable percentage or not? No one on the team was able to answer. The inability to respond reveals some insights:

- It is important to report metric scores that are relevant to the intended audience. The score itself must be meaningful and convey the true value of the quality of the monitored data. Reusing existing metrics (established for different intentions) may not satisfy the needs of those staff members who are the customers of the data quality scorecard.

- Analysts who rely on existing measurements (performed for alternate purposes) lose control over the definition of those metrics. The resulting scorecard is dependent on measurements that may change at any moment for any reason.

- The inability to respond to the question not only reveals a lack of understanding of the value of the metric for data quality management, but it also calls into question the use of the metric for its original purpose. Essentially, measurements that quantify an aspect of data without qualified relevance (so-what metrics) are of little use for data quality scorecarding.
Defining Relevant Metrics

Good data quality metrics exhibit certain characteristics. Defining metrics that share these characteristics will lead to a data quality scorecard that is meaningful:

- **Business Relevance.** The metric must be defined within a business context that explains how its score relates to improved business performance.

- **Measurable.** The metric must have a process that quantifies it within a discrete range.

- **Controllable.** The metric must reflect a controllable aspect of the business process so that when the measurement is not in a desirable range, some action to improve the data should be triggered.

- **Reportable.** The metric’s definition should provide the right level of information to the data steward when the measured value is not acceptable.

- **Traceable.** Documenting a time series of reported measurements must provide insight into the result of improvement efforts over time as well as support statistical process control.

In addition, recognizing that reporting a metric summarizes its value, the scorecarding environment should also provide the ability to reveal underlying data that contributed to a particular metric score. Reviewing data quality measurements and evaluating the data instances that contributed to any unsatisfactory scores suggest the need to be able to drill down into the performance metric. This provides a better understanding of any existing or emergent patterns contributing to a poor score, an assessment of the impact, and help in root-cause analysis.

Relating Business Objectives and Quality Data

If metrics must be defined within a business context that relates the score to improved business performance, then data quality metrics should reflect that business processes (and applications) depend on reliable data. Tying data quality metrics to business activities means aligning rule conformance with the achievement of business objectives.
Business Impact Categories

There are many potential areas that may be affected by poor data quality, but computing their exact cost is often a challenge. Classifying those effects helps to identify discrete issues and relate business value to high-quality data. Evaluating the effect of any type of recurring data issue is easier when there is a process for classification that depends on a taxonomy of impact categories and subcategories. Effects can be assessed within each subcategory, and the measurement and reporting of the value of high-quality data can be a combination of separate measures associated with how specific data flaws prevent the achievement of business goals.

An organization can focus on four high-level areas of performance: financial, productivity, risk and compliance. Within each of these areas there are subcategories, as shown in Figure 1, that further refine how poor data quality can affect the organization.

![Figure 1](image)

Relating Issues to Effects

Every organization has some framework for recording data quality issues, and each issue is reported because somebody in the organization felt that it affected one or more aspects of achieving business objectives. By reviewing the list of issues, the analyst can determine what types of impacts were incurred through a comparison with the impact categories. Each impact can be reviewed so that the economic value can be estimated in terms of the relevance to the business data consumers.

For example, missing customer telephone and address information is a data quality issue that may be associated with a number of effects:

- The absence of a telephone number may affect the telephone sales team’s ability to make a sale (missed revenue).
- Missing address data may affect the ability to deliver a product (increased delivery charges) or payment collection (cash flow).
- Complete customer records may be required for credit checking (credit risk) and government reporting (compliance).
Once a collection of issues and their corresponding effects are reviewed, the cumulative impact can be aggregated. Issues and their remediation tasks can then be prioritized in relation to their business costs. But this is only part of the process. Once issues are identified, reviewed, assessed and prioritized, two aspects of data quality control must be addressed. First, any critical outstanding issues should be subjected to remediation. But more importantly, the existence of any issues that conflict with business-user expectations may require conformance monitoring to provide metrics that have business relevance.

The next step is to define metrics that meet the standard described earlier in this paper: relevant, measurable, controllable, reportable and traceable. The analysis process determines the relevance and control, leading to the deployment of a specification scheme based on dimensions of data quality to ensure measurement, reporting and tracking.

**Dimensions of Data Quality**

Data quality dimensions frame the performance goals of a data production process, and guide the definition of discrete metrics that reflect business expectations. Dimensions of data quality are often categorized according to the metrics associated with business processes, such as the quality of data associated with data values, data models, data presentation and conformance with governance policies. Dimensions associated with data values and data presentation are often automatable and are nicely suited to continuous data quality monitoring. The main purposes for identifying data quality dimensions for data sets include:

- Establishing universal metrics for assessing data quality.
- Determining the suitability of data for its intended targets.
- Defining minimum thresholds for meeting business expectations.
- Monitoring whether measured levels of quality meet business expectations.
- Reporting a qualified score for the quality of organizational data.

Providing a means for organizing data quality rules within defined data quality dimensions simplifies the processes for defining, measuring and reporting levels of data quality as well as supporting data stewardship. When a data quality measurement does not meet the acceptability threshold, the data steward can use the data quality scorecard to help determine the root causes.
There are many potential classifications for data quality; but for practical purposes, the set of dimensions can be limited to ones that lend themselves to straightforward measurement and reporting. In this paper, we focus on a subset of the many dimensions that could be measured: accuracy, lineage, completeness, consistency, reasonability, structural consistency and identification. It is the data analyst's task to evaluate each issue, determine which dimensions are being addressed, identify a specific characteristic that can be measured and then define a metric. That metric describes a measureable aspect of the data with respect to the selected dimension, a measurement process and an acceptability threshold.

**Accuracy**

Data accuracy refers to the degree to which data values correctly reflect attributes of the “real-life” entities they are intended to model. In many cases, accuracy is measured by how the values agree with an identified source of correct information. There are different sources of correct information: a database of record, a corroborative set of similar data values from another table, dynamically computed values or perhaps results from a manual process.

Examples of measurable characteristics of accuracy include:

- **Value precision.** Does each value conform to the defined level of precision?
- **Value acceptance.** Does each value belong to the allowed set of values for the observed attribute?
- **Value accuracy.** Is each data value correct when assessed against a system of record?

**Lineage**

An important measure of trustworthiness is tracing the originating sources of any new or updated data element. Documenting data flow and sources of modification supports root-cause analysis, which also supports the value of measuring the historical sources of data as part of the overall assessment. Lineage is made traceable by ensuring that all data elements include time-stamp and location attributes (including creation or initial introduction) and that audit trails of all modifications can be reconstructed.

**Completeness**

An expectation of completeness indicates that certain attributes should always be assigned values in a data set. Completeness rules can be assigned to a data set in two levels:

- Mandatory attributes that require a value.
- Optional attributes, which may have a value based on some set of conditions.
Note that inapplicable attributes (such as maiden name for a single male) may not have an assigned value.

Completeness can be prescribed or can be dependent on the values of other attributes within a record. Completeness may be relevant to a single attribute across all data instances or within a single data instance. Some aspects that can be measured include the frequency of missing values within an attribute and conformance to optional null value rules.

**Currency**

Currency refers to the degree to which information is up to date. Data currency may be measured as a function of the frequency rate at which data elements are expected to be refreshed, as well as verifying that newly created or updated data is sent to dependent applications within a specified time. Another aspect includes temporal consistency rules that measure whether dependent variables are consistent (e.g., that the start date is earlier than the end date).

**Reasonability**

This dimension includes general statements regarding expectations of consistency or the reasonableness of values either in the context of existing data or over time, for example, multivalue consistency rules where the value of one set of attributes is consistent with the values of another set of attributes, or temporal reasonability rules in which new values are consistent with expectations based on previous values (e.g., today’s transaction count should be within 5 percent of the average daily transaction count for the past 30 days).

**Structural Consistency**

Structural consistency refers to the consistency in the representation of similar attribute values, both within the same data set and across related tables. One aspect of structural consistency involves reviewing whether data elements that share the same value set have the same size and are the same data type. You also can measure the degree of consistency between stored values and the data types and sizes used for information exchange. Conformance with syntactic definitions for data elements reliant on defined patterns (such as addresses or telephone numbers) can also be measured.

**Identifiability**

Identifiability refers to the unique naming and representation of core conceptual objects as well as the ability to link data instances based on identifying attribute values. One measurement is entity uniqueness, which ensures that a new record is not created if there is an existing record. Uniqueness of the entities within a data set implies that no entity exists more than once within the data set and that there is a key that can be used to access each entity (and only that entity) within the data set.
Additional Dimensions

There is a temptation to rely only on an existing list of dimensions or categories for assessing data quality; but in fact, every industry is different, every organization is different and every group within an organization is different. The types of effects and issues you might encounter in one organization may be completely different than another. It is reasonable to use these data quality dimensions as a starting point for evaluation, but then look at other categories that could be used for measuring and qualifying the quality of information.

Reporting the Scorecard

The degree of reportability and controllability may differ depending on your role within the organization, and correspondingly, so will the level of detail reported in a data quality scorecard. Data stewards may focus on continuous monitoring in order to resolve issues using defined service level agreements, while senior managers may be interested in observing the degree to which poor data quality introduces risk.

The need to present higher-level data quality scores makes a distinction between two types of metrics. The types discussed in this paper so far can be referred to as base-level metrics. They quantify acceptable levels of defined data quality rules. A higher-level measurement would be the complex metric representing a rolled-up score computed as a function (such as a sum) of assigning specific weights to a collection of existing metrics. The complex metric provides a qualitative overview of how data quality affects the organization in different ways, because the scorecard can be populated with metrics rolled up across different dimensions depending on the audience. Complex data quality metrics can be accumulated for reporting in a scorecard in one of three different views: by issue, by business process or by business impact.

The Data Quality Issues View

Evaluating the effects of a specific data quality issue across multiple business processes highlights the negative organizationwide effects caused by specific data flaws. The scorecard approach, which is suited to data analysts attempting to prioritize tasks for diagnosis and remediation, provides a comprehensive view of the effects created by each data issue. Analyzing the scorecard sheds light on the root causes of poor data quality, as well as identifying “rogue processes” that require greater attention when instituting monitoring and control processes.

The Business Process View

Operational managers overseeing business processes may be interested in a scorecard view based on business process. In this view, an operational manager can examine the risks and failures that are preventing the achievement of the expected results. This scorecard approach consists of complex metrics representing the effects associated with each issue. This view can be used for isolating the source of data issues at specific stages of the business process, as well as for diagnosis and remediation.
The Business Impact View

This reporting approach displays the aggregation of business impacts from the different issues across different process flows. For example, one scorecard could report aggregated metrics of the credit risk, compliance with privacy protection, and decreased sales. Analysis of the metrics will point to the business processes from which the issues originate, and a more thorough review will point to the specific issues within each of the business processes. This view is for more senior managers seeking a high-level view of the risks associated with data quality issues and how that risk is introduced across the enterprise.

Managing Scorecard Views

Each of these views requires the construction and management of a hierarchy of metrics based on various levels of accountability. But no matter which approach is employed, each is supported by describing, defining and managing base-level and complex metrics so that:

- Scorecards for business relevance are driven by a hierarchical rollup of metrics.
- The definition of metrics is separated from their contextual use, thereby allowing the same measurement to be used in different contexts with different acceptability thresholds and weights.
- The appropriate level of presentation can be materialized based on the level of detail expected for the data consumer's specific data governance role and accountability.
Summary

Scorecards are effective management tools that can summarize important organizational knowledge as well as alert the appropriate staff members when diagnostic or remedial actions need to be taken. Crafting a data quality scorecard to support an organizational data governance program requires defining metrics that correlate a score with acceptable levels of business performance. This means that the data quality rules being observed and monitored as part of the governance program are aligned with the achievement of business goals.

The processes explored in this paper simplify the approach to evaluating business effects associated with poor data quality and how you can define metrics that capture data quality expectations and acceptability thresholds. The impact taxonomy enables the necessary level of precision in describing the business effects, while the dimensions of data quality guide the analyst in defining quantifiable measures. Applying these processes will result in a set of metrics that can be combined into different scorecard approaches that effectively address senior-level manager, operational manager and data steward responsibilities to support organizational data governance.
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 60,000 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®.