IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB12696

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour © 2006, Idea Group Inc.

Designing a Balanced Data Quality Scorecard

John R. Talburt, University of Arkansas at Little Rock, 2801 South University Ave., Little Rock, AR 72204, jrtalburt@ualr.edu

Traci Campbell, Acxiom Corporation, 601 East 3rd St., Little Rock, AR 72201, trcamp@acxiom.com

ABSTRACT

As organizations embrace strategic data quality management, they seek to implement data quality scorecards that display data quality metric results for their information systems. However, defining the metrics that populate these scorecards can be uncharted territory for the organization's data management team. Negotiations with users over the design of data quality scorecards and scorecard metrics can be facilitated by establishing a common terminology and understanding of dimensions and metric types. This paper presents a classification scheme for metrics and some of the trade-offs related to balancing the composition of a scorecard by types of metrics.

BACKGROUND

With the success of the Balanced Scorecard, Key Performance Indicators (KPI), and other metric-based business management techniques, many organizations are seeking to apply the same principles to data quality management in the form of data quality scorecards and dashboards. Outside of vendor tools for building data quality scorecards, there is little guidance available regarding the design and selection of the underlying metrics.

To clarify terminology in this paper the term *data quality scorecard* means a mechanism by which data quality metric values are displayed in a meaningful and easily understood format. Data quality scorecard and *data quality dashboard* are used interchangeably.

Although the term data quality metric is sometimes used in a generic sense to denote a feature or attribute of a data quality dimension (such as in ISO/IEC (2001) and Naumann, F. (2002)), we use the term *data quality metric* to denote a function that maps the state of an information system (e.g., database tables, flat files, XML) into a numeric value conveying positive or negative covariance with the perceived data quality.

A positive covariance means that increasing metric values indicate increased data quality along the dimension being measured. For example, a metric for completeness defined as the percentage of non-null values for a particular column would exhibit positive covariance — increasing percentages of non-null values indicates higher quality of completeness. Conversely, an access metric for timeliness counting the numbers of hours to successfully complete a daily database build would exhibit negative covariance — increasing hours would indicate lower quality of timeliness.

Lastly, the expression *"displayed in a meaningful and easily understood format"* has broad meaning, pertaining to both delivery medium— e.g., HTML screen display versus printed report— and representation, ranging from a simple table of metric values to a graphic display of color coded icons and charts. Presentation and usability are important aspects of scorecard design; however, the focus of this paper is on the metric design task of selecting and balancing the overall palette of metrics that comprise the scorecard. Whereas presentation represents form and message, the metrics represent content.

DATA QUALITY MANAGEMENT MATURITY

Most organizations begin their data quality journey in the area of tactical data quality management. Tactical data quality management is analogous to management by exception – quality is only addressed when there is an adverse quality event, i.e., an exception or failure. The problem with this approach is that the data quality failures are often detected by the information consumer rather than the information producer.

Just as the capability and maturity of software development corresponds to increasing levels of process repeatability and process improvement, the maturity of data quality management can be related to an organization's adoption of strategic data quality management – the willingness to measure and implement improvements through its own internal processes.

Data quality measurement is a key component of Total Data Quality Management (TDQM), Six Sigma, and other data quality management processes. Therefore, it is not surprising to see the increased use of scorecards in organizations attempting to move up on the data quality management maturity scale.

METRIC STRUCTURE AND DESIGN

Loshin (2005) and others have done a great job of describing the characteristics that a good metric should possess. Loshin's list includes

- Clarity of Definition
- Measurability
- Business relevance
- Controllability
- Representation
- Reportability
- Trackability
- Drill-down Capability

However, this list presupposes that one already has or at least has in mind a metric to evaluate. In the authors' experience, many people—even those with extensive experience in data management—are often at a loss as to the first step in defining a useful metric. No matter how usable the presentation, a scorecard is only as good as the metrics that comprise it.

METRIC TYPES

In order to facilitate data quality metric discussions with users, the authors adopted a classification scheme for metrics based on three independent attributes:

- The primary Data Quality Dimension being measured
- Whether it is an Improvability Metric versus a Process Control Metric
- Whether it is a Composite (Roll-up) Metric versus a Simple Metric

DATA QUALITY DIMENSIONS

The fact that data quality is multi-dimensional is well established, although it is much less clear exactly what those dimensions are. There are as many dimension and category classifications as there are experts in the field. In the spirit that "simpler is better," the authors modified the four-dimensional framework set out very early by Ballou and Pazer (1985). We added a fifth dimension associated with the grouping or integration of data, and broadened the interpretation of timeliness as an aspect of access.

The authors' five dimensions are

- Accuracy. Degree to which facts are recorded correctly. Requires that a "benchmark" be identified that represents a standard for measurement of what is and is not correct.
- Completeness. Degree to which relevant and available information is present. Typically expressed as a percentage of collected values to the total population of possible values.
- Consistency. Uniformity or similarity of data outcomes relative to other results, or in absolute terms as compared to allowable values or pre-defined constraints (validation).
- Access. Covers a wide range of contextual and representational quality issues including – timeliness, relevance, formatting, ease of access, understandability, and interpretability.
- Grouping. Refers to the accuracy with which separate records for the same entity (account, consumer, household, etc.) are brought together within a data integration process.

When implementing a scorecard for the first time, users often fail to consider the broader set of data quality dimensions, instead focusing on one or two, typically completeness and consistency. Balancing a scorecard to include metrics from every dimension may require challenging user thinking about data quality.

IMPROVABILITY VERSUS CONTROL METRICS

Improvability and control represent two often desired, but sometimes incompatible, goals for metrics. Control metrics produce a small set of discrete values, such as 1/0/-1 to represent acceptable/marginal/unacceptable. Control metrics alert the user when some aspect of data quality becomes unacceptable, but does not give any indication of the degree. Perhaps because many data quality champions have quality control backgrounds, there is a tendency to overpopulate scorecards with control metrics.

An important aspect of strategic data quality management is to realize improvement in quality through a repeated process of measurement, analysis, and improvement. This is very much the spirit of the Total Data Quality Management (TDQM) methodology (Huang, 1999). The values of improvability metrics fall into a continuous interval of values, such as zero to one hundred, along with an initial baseline and a goal. Through analysis and improvement, the intent is that over time the measured values will trend from the baseline value and reach or surpass the goal. Improvability metrics can also be interpreted as control metrics if a user establishes threshold values (failure points) that translate the continuous value into a discrete value displayed as a color (red/green) or alert icon on the scorecard.

COMPOSITE VERSUS SIMPLE METRICS

The choice between composite and simple metrics reflects the tension between the scorecard user's desire for detail and the need for summarization. For any information system, there are an unlimited number of metrics that could be defined and included on a scorecard. One of the hardest choices in the design of any scorecard is deciding how many metrics to include.

A simple metric is one that translates a single measurement into a metric value. An example would be a completeness metric that converts the count of non-null values for a particular column of a database table to a percentage of the total number of rows in the table.

If, however, there are many columns of interest amongst several tables, then including each one as a simple metric could create an unusable scorecard displaying hundreds of metrics. One solution to this problem is to use *composite metrics* that combine several measurements into one metric value. There are many ways to create composite metrics, and the actual design will ultimately depend on the user's requirements.

In this example, there are three columns of interest (x, y, and z) in two tables (Table-A and Table-B). Column-x and Column-y are in Table-A, and Column-z is in Table-B. In this scenario, composite metric CM1 could be defined as the average completeness of non-null values for all three columns.

$$CM1 = 100*(nonNull(x)+nonNull(y)+nonNull(z))/(2*rowsOf(A)+rowsOf(B))$$

The trade-off for reducing the number of metrics in this way is the loss of detail. In this example, if composite metric CM1 has a value of 85%, the user only knows that the average completeness of the three columns is 85%, but the actual completeness of each column is not evident. This presents a problem if the two tables were of very different sizes. If Table-A had 900 columns but Table-B only had 100 columns, and column-x and column-y were 100% complete but column-z was 0% complete, the composite metric CM1 could result in a 90% value..

For this reason, composite metrics sometimes work better when combining control metrics. Using the same example, suppose that the user wants to be sure that each of the three columns is at least 80% populated. In this case, a composite metric CM2 could be defined as

$$CM2 = 100*(true(x,80)+true(y,80)+true(z,80))/3$$

Where

true(x,80) = 1 if Column-x is at least 80% populated, otherwise 0

CM2 is the percentage of columns that have acceptable completeness and has only four discrete values, 100%, 67%, 33%, or 0%. While CM2 helps the user understand how many columns have unacceptable levels of completeness, it still does not indicate which one failed.

METRIC REQUIREMENTS WORKSHEET

As a metric for a scorecard is proposed and considered, it is often helpful to create a worksheet for it including:

- Metric Label (6-8 characters)
- Short Working Name
- System and Touch Point
- Primary Dimension and Type
- Tool(s) Used for Generating Statistics
- Measurement Interval (Daily, Weekly, etc)
- Tables/Files and Columns/Elements Involved
- Complete Definition (Algorithm)
- Maximum/Minimum Value, Goal, and Failure Point

CONCLUSION

The process of designing a complete and balanced data quality scorecard can be facilitated by establishing a common understanding of data quality metric classifications with the user. Users may have a relatively narrow view of data quality issues and metrics, and presenting them with a broad overview often leads to more balance and completeness in scorecard design.

REFERENCES

Ballou, D.P. and Pazer, H.L. (1985) "Modeling Data and Process Quality in Multi-input, Multi-output Information Systems," *Management Science* 31, 2 (1985), 150-162

Campbell, T. and Wilhoit, Z. (2003) "How's Your Data Quality? A Case Study in Corporate Data Quality Strategy." *Proceedings: International Conference on Information Quality*, MIT, 2003.

Huang, K., Lee, Y.W., and Wang, R.Y. (1999). *Quality Information and Knowledge*, 1999, Prentice Hall.

ISO/IEC. (2001). *Information technology - software product quality* (No. ISO/IEC 9126-1:2001)

Loshin, D. (2005). "Developing Information Quality Metrics" DM Review, Vol. 15, No. 5, May 2005, pp. 24-27.

Mattison, R. (1996). *Data Warehousing Strategies, Technologies and Techniques*, McGraw-Hill, 1996

Naumann, F. (2002). *Quality-driven query answering for integrated information systems*: Springer-Verlag.

Wang, R.Y., Ziad, M., and Lee, Y.W. (2001) *Data Quality*, 2001, Kluwer Academic Publishers.