

Building and Using Data Quality Scorecard

Data quality scorecard is the centerpiece of any data quality management program. It provides comprehensive information about quality of data in a database, and allows both aggregated analysis and detailed drill-downs. A well-designed data quality scorecard is the key to understanding how well the data supports various reports, analytical and operational processes, and data-driven projects. It is also critical for making good decisions about data quality improvement initiatives.

Data Quality Scorecard Defined

It is a common misconception that the objective of a data quality assessment project is to produce error reports. Such a view significantly diminishes the ROI of assessment initiatives. Project teams spend months designing, implementing, and fine-tuning data quality rules; they build neat rule catalogues and produce extensive error reports. But, without a data quality scorecard, all they have are raw materials and no value-added product to justify further investment into data quality management. Indeed, no amount of firewood will make you warm in the winter unless you can make a decent fire. The main product of data quality assessment is the data quality scorecard!

The picture below represents the data quality scorecard as an information pyramid. At the top level are aggregate scores which are high-level measures of the data quality. Well-designed aggregate scores are goal driven and allow us to evaluate data fitness for various purposes and indicate quality of various data collection processes. From the perspective of understanding the data quality and its impact on the business, aggregate scores are the key piece of data quality metadata. At the bottom level of the data quality scorecard is information about data quality of individual data records. In the middle are various score decompositions and error reports allowing us to analyze and summarize data quality across various dimensions and for different objectives. Let's consider these components more in detail.



Aggregate Scores

On the surface, the data quality scorecard is a collection of aggregate scores. Each score aggregates errors identified by the data quality rules into a single number – a percentage of good data records among all target data records. Aggregate scores help make sense out of the numerous error reports produced in the course of data quality assessment. Without aggregate scores, error reports often discourage rather than enable data quality improvement.

You have to be careful when choosing which aggregate scores to measure. The scores that are not tied with a meaningful business objective are useless. For instance, a

simple aggregate score for the entire database is usually rather meaningless. Suppose, we know that 6.3% of all records in the database have some errors. So what? This number does not help me at all if I cannot say whether it is good or bad, and I cannot make any decisions based on this information.

On the other hand, consider an HR database that is used, among other things, to calculate employee retirement benefits. Now, if you can build an aggregate score that says 6.3% of all calculations are incorrect because of data quality problems, such a score is extremely valuable. You can use it to measure the annual cost of data quality to the business through its impact to a specific business process. You can further use it to decide whether or not to initiate a data-cleansing project by estimating its ROI.

The bottom line is that good aggregate scores are goal driven and allow us to make better decisions and take actions. Poorly designed aggregate scores are just meaningless numbers.

Of course, it is possible and desirable to build many different aggregate scores by selecting different groups of target data records. The most valuable scores measure data fitness for various business uses. These scores allow us to estimate the cost of bad data to the business, to evaluate potential ROI of data quality initiatives, and to set correct expectations for data-driven projects. In fact, if you define the objective of a data quality assessment project as calculating one or several of such scores, you will have much easier time finding sponsors for your initiative.

Other important aggregate scores measure quality of various data collection procedures. For example, scores based on the data origin provide estimates of the quality of the data obtained from a particular data source or through a particular data interface. A similar concept involves measuring the quality of the data collected during a specific period of time. Indeed, it is usually important to know if the data errors are mostly historic or were introduced recently. The presence of recent errors indicates a greater need for data collection improvement initiatives. Such measurement can be accomplished by an aggregate score with constraints on the timestamps of the relevant records.

To conclude, analysis of the aggregate scores answers key data quality questions:

- What is the impact of the errors in your database on business processes?
- What are the sources and causes of the errors in your database?
- Where in the database can most of the errors be found?

Score Decompositions

Next layer in the data quality scorecard is composed of various score decompositions, which show contributions of different components to the data quality. Score decompositions can be built along many dimensions, including data elements, data quality rules, subject populations, and record subsets.

For instance, in the above example we may find that 6.3% of all calculations are incorrect. Decomposition may indicate that in 80% of cases it is caused by the problem with the employee compensation data; in 15% of cases the reason is missing or incorrect employment history; and in 5% of cases the culprit is invalid date of birth. This can be

used to prioritize a data cleansing initiative. Another score decomposition may indicate that over 70% of errors are for employees from a specific subsidiary. This may suggest a need to improve data collection procedures in that subsidiary.

The level of detail obtained through score decompositions is enough to understand where most data quality problems come from. However, if we want to investigate data quality further, more drill-downs are necessary. The next step would be to produce various reports of individual errors that contribute to the score (or sub-score) tabulation. These reports can be filtered and sorted in various ways to better understand the causes, nature, and magnitude of the data problems.

Finally, at the very bottom of the data quality scorecard pyramid are reports showing the quality of individual records or subjects. These atomic level reports identify records and subjects affected by errors and could even estimate the probability that each data element is erroneous.

Summary

Data quality scorecard is a valuable analytical tool that allows to measure the cost of bad data for the business and to estimate ROI of data quality improvement initiatives. Building and maintaining a dimensional time-dependent data quality scorecard must be one of the first priorities in any data quality management initiative.