

DMMReview

Information Is Your Business

June 2007/Volume 17, Number 6

www.dmreview.com

PROOF 5-29-07

Data Quality Assessment

On Hunting Mammoths and Measuring Data Quality

By Arkady Maydanchik



On Hunting Mammoths and Measuring Data Quality

By Arkady Maydanchik



Illustrated by Arnold Roth

Imagine we could travel back in time 20,000 years. We would find ourselves in the middle of a stone-age landscape. The reason we call it stone age is because stone was the most important resource for the early humans. If we could walk into a typical hunter's hut (or caveman's cave), we would certainly find spears with sharpened stone points. These spears were used by our early ancestors to hunt many large animals, including mammoths. Of course, hunting mammoths was a dangerous business, and in many cases, the spear would not penetrate mammoth's skin.

Over time, though, a different breed of hunters developed. These hunters would stretch and hang the skin of a previously killed mammoth on the wall and engage in strange cabalistic rituals. They danced, chanted incantations and threw the spears at the wall. All the time, they observed which spear, thrown from which angle and distance, and accompanied by which dance and incantation, penetrated the mammoth's skin the best. What they learned was incredibly useful to make better spears; it also helped develop hunting strategies that best accommodated deficiencies of the existing weapons. These rituals were the earliest examples of quality assessment.

Now, let's come back to our time. We live in an information age - the term emphasizes that information is our most precious resource. Modern humans hunt their mammoths with information. The success of corporations and government institutions largely depends on the efficiency with which they can collect, organize and utilize data about products, customers, competitors and employees.

Yet, we tolerate enormous inaccuracies in databases. Data errors are the cancer of information systems, spreading

from place to place, wreaking operational and financial havoc. Even more disheartening is that the magnitude of the data quality problems is continuously growing, fostered by exponential increases in the size of the databases and further proliferation of information systems.

Over the years, the onus of data quality improvement was placed on modern database technologies and better information systems. I remember well an HR executive confidently telling me over lunch that now that his company has implemented a modern payroll system, it would no longer have any problems.

"PeopleSoft does not have data quality problems," he said. "Maybe PeopleSoft does not have data quality problems," I replied, "but your company certainly does, and they are not going away. In fact, they are about to get much bigger." He did not believe me; people do not like those who predict rain on a sunny day. A few years later, the recently appointed new HR executive of that company called me for advice about their data quality strategy.

In reality, if we do nothing, data quality will continuously deteriorate to the point where the data will become a huge liability. The only way to address the data quality challenge is with a systematic, ongoing program, which would assess and improve existing data quality levels as well as continuously monitor data quality and prevent its future deterioration as much as possible.

A major obstacle on the path to higher data quality is that most organizations, though aware of its importance, have no idea about the extent of the problems with their data. Typically, data quality is either grossly underestimated or grossly overestimated. The impact of data quality levels on business processes is also rarely understood. This causes the failure of many data-driven projects (such as new system implementations). Data quality improvement initiatives also typically fail when no method for measuring data quality improvements is provided.

To improve data quality, we must first learn to measure it. And while dancing and incantations at numerous meetings are of some value, the problem requires a more drastic solution – systematic data quality assessment.

The question, "Who shall be responsible for data quality assessment?" is most frequently asked. Part of the reason for the uncertainty is that the data quality profession is still in its infancy. Even companies that form data quality departments often have them staffed by handpicked employees with general IT and data expertise but no specific data quality knowledge. As a result, data quality assessment projects

possible to determine whether or not each data element in the database is correct and, if incorrect, where it came from and what had caused the error. In practice, this proves a formidable challenge. Indeed, the obvious way to ensure that a piece of data is correct is to compare it with some "trusted" source, that is, a source which is correct 100 percent of the time. Such a source may not exist or at least may not

Rule fine-tuning involves much analysis, and it usually takes many iterations to make error reports reasonably accurate.

tend to follow one of two polar scenarios.

In the first scenario, projects fall into the laps of technical people within the IT group. An attendee of one of my classes, a database administrator, was asked by her boss to outline a data quality assessment strategy. Why her? Because, according to her boss, data quality assessment involves writing queries, manipulating data and understanding databases – all parts of her resume. This, of course, makes as much sense as asking me to be a reporter for the sports section of the *Chicago Tribune* because I can type, have published some articles and watch sports from the comfort of my living-room couch.

In the second scenario, data quality assessment is performed inside business units by the data users. This appears to make some sense, as the data users can tell good data from bad and are mostly in need of quality data. Of course, the problem is that business users lack technical expertise, which is why I keep getting this question at almost every class, "Is there a tool that can do data quality assessment without any custom coding or querying?" My answer does not make those who ask this question happy. Data quality assessment is an IT discipline and requires IT expertise.

In reality, it takes two to tango, so a data quality assessment team must include both IT specialists and business users. In addition, it needs seasoned data quality practitioners – the new breed of information age mammoth hunters.

The next question is how to identify data errors. In an ideal world, it would be

be readily available. Even if a trusted source is available, this approach is expensive and impractical for large databases.

Luckily, there is a better way. Modern databases have two important characteristics that distinguish data from all other resources. First, databases allow users to access and process data with dramatic speeds. Secondly, myriad data elements stored in them are tied by equally huge numbers of data dependencies and business rules. The combination of these two factors allows validating the data en masse by data quality rules – constraints that validate data accuracy and consistency and can be implemented in computer programs. Miraculously, a well-designed and fine-tuned collection of rules will identify a majority of data errors in a fraction of the time compared to manual validation. What is even better, the same setup can be reused over and over again to reassess data quality periodically with minimal effort.

Using data quality rules brings comprehensive data quality assessment from fantasy world to reality. However, it is by no means simple, and it takes a skillful skipper to navigate through the powerful currents and maelstroms along the way. Considering the volume and structural complexity of a typical database, designing a comprehensive set of data quality rules is a daunting task. The number of rules will often reach hundreds or even thousands. When some rules are missing, the results of the data quality assessment can be completely jeopardized. Thus, it is important to consider all rule types, rule

sources and rule design strategies. Data quality rules fall into five broad categories:

1. Attribute domain constraints restrict allowed values of individual data attributes. They are the most basic of all data quality rules because they apply to the most atomic data elements.
2. Relational integrity rules are derived from the relational data models. They are more complex, apply to multiple records at a time and enforce identity and referential integrity of the data.
3. Rules for historical data include timeline constraints and value patterns for time-dependent value stacks and event histories. Because time-dependent data is the most common database type (and also most error prone), these rules typically are a key part of data quality assessment.
4. Rules for state-dependent objects constrain the lifecycle of objects described by so-called state-transition models (e.g., insurance claims or job applications). Data for such objects is most important and can only be validated by a special class of data quality rules.
5. General attribute dependency rules describe complex attribute relationships, including constraints on redundant, derived, partially dependent and correlated attributes.

A serious challenge in rule implementation is how to manage error reports. A comprehensive error catalog must support the following functionality:

- ▮ Aggregate, filter and sort errors across various dimensions.
- ▮ Identify overlaps and correlations between errors for different rules.

- ▮ Identify data records affected by a particular error or a group of errors.
- ▮ Identify all errors for a particular data record or set of records.


It is very hard to design perfect data quality rules. The ones we come up with will often fail to spot some erroneous records and falsely accuse others. They may not tell you which data element is erroneous, even when the error is identified. They may identify the same error in many different ways. Error reports produced by such rules tend to suffer from the same malady as the data itself – poor quality. This imperfection, if not understood and controlled, will overrun and doom any data quality assessment effort. Rule fine-tuning involves much analysis, and it usually takes many iterations to make error reports reasonably accurate.

Eventually, we get to the point where the error reports can be trusted. Now comes the time to analyze the results; however, making conclusions based on the error reports is overwhelming. Indeed, data quality rules produce endless listings of data errors. Each error applies to one or several data elements from one or several tables. The sight of a 500-page printout or even of an electronic listing with 20,000 lines of error messages will make most data quality professionals run for cover. To obtain value from a data quality assessment initiative, you must aggregate the error reports into meaningful summaries and create a data quality scorecard.

Aggregate scores provide high-level measures of data quality. Each score aggregates errors identified by the data quality rules into a single number – a per-

centage of good data records among all target data records. By selecting different groups of target data records, you can create many aggregate scores for a single database. Well-designed scores are goal driven and allow you to make better decisions and take action. They can measure 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.

The ROI of data quality assessment grows exponentially when you organize the error reports into a dimensional data quality scorecard, which allows both aggregated analysis and detailed drill downs. At the top level of the scorecard are aggregate quality scores. At the bottom level is information about data quality of individual data records. In the middle are various score decompositions and error reports, allowing you to analyze and summarize data quality across various dimensions and for different objectives. Building a comprehensive data quality scorecard is the final step of data quality assessment.

Measuring data quality is a daunting task once you roll up your sleeves and dive into the technical details. So was hunting the mammoths. The fact that humans are still around and mammoths are not, while somewhat sad, makes me hopeful about the future of the data quality. 

*Arkady Maydanchik (arkadym@dataqualitygroup.com)
is co-founder of Data Quality Group LLC.*