



Data Quality and the Single Version of the Truth

Tom Breur
November 2009

Introduction

Data quality *emerges* as users create value from working with data. It implies value *to* someone; it is not a property that is intrinsic in the data itself. When nobody uses data, it has zero value. In order to execute corporate strategy you need to know what's going on. To make data usable, it is eminently important to construct some uniform and consistent structure that houses the data. A data warehouse (DWH).

Traditionally, in data warehousing we have taken source data, and applied extract, transform, load (ETL) to clean, scrub, and move data to our data warehouse (DWH) star schemas. We don't want "bad" data going into the DWH, because that means it would show up in corporate reports (an architectural exception would be a Data Vault). This might trigger business users to question the accuracy of the DWH, which is supposed to be *the* trusted source for integrated corporate data.

When you derive information from disparate, unconnected source systems, there is a fair chance the numbers don't align. As businesses grow more complex and ever more digitized, we've seen an overwhelming proliferation of data streams. This becomes too much to handle, even for die-hard spreadsheet warriors. Consistent information becomes commensurately more difficult to produce.

This dynamic has driven the quest for our holy grail: one single version of the truth. Either front-end systems are connected through enterprise application integration (EAI) and enterprise information integration (EII) solutions, or data are integrated on the back end. Or both. But without tight application coupling, the DWH can't "magically" make numbers match that are fundamentally out of kilter.

In many organizations, confusion reigns. Spreadsheets and isolated databases are often the norm. When you ask what "the truth" is, you are likely to get as many different answers as the number of people you asked. That doesn't sound very singular. What does it take, then, to arrive at a "*single* version of the truth?"

A single version of the truth?

In all fairness, "the truth" isn't always obvious, not even when it's staring you in the face. For many years, the hole in the ozone layer had been there, and data on it were recorded. However, data that were collected seemed so out of line with "normal" expectations that these measurements were habitually discarded (as errors). Sometimes the blatantly obvious can remain invisible, simply because we don't "want" to see it.

This ozone example illustrates another point: what we consider to be "the truth" is an amalgam of facts and their interpretation. And the same holds for data warehousing. We record facts, but our interpretation of the truth only arises after we apply our model of "the world" to those observations.

In data warehousing we sometimes need to force fit fundamentally irreconcilable data streams (filtering out "errors"), to arrive at "the truth." We should never forget that the business rules we apply to change "facts" into "reality" can always be subject to change as our model evolves of what is real and what is not. Or at least what we *consider* to be "real."

A DWH records facts, and as such can become the central (historical) repository. Our interpretation of the facts may well change over time. So we could choose to "rewrite" history as our interpretation of *past* facts changes. And it does, just like in the example of "facts" about the hole in the ozone layer that turned out to exist much earlier than we thought. This is what I mean when I say that a DWH is the single source of the *facts*. And, thus, principally *not* of the truth.

What *is* the single version of the truth, then?

As we just discussed, a DWH does not contain any "truths", only facts. If you think about it, "truth" really seems more of a philosophical or religious concept. Even mathematics has no "truth", but instead axiomatic derivations from prior statements. Empirical sciences like physics have no "truth", either. Merely hypotheses that last until someone proves them wrong.

For BI professionals, a single version of the truth seems to equate to multiple colleagues all agreeing on a common interpretation of facts as recorded in a (central) database. When no one feels the urge anymore to corroborate database facts by triangulating with real world observations (or other database systems), we jointly buy into "the truth." One could say that a common version of the truth equates to sharing the same point of view and "speaking the same language."

How would this look in real life? If you asked *either* Finance, *or* Logistics, *or* Sales, *or* Marketing what the turnover was in Q1, they'd all give you the same response. That implies "Q1" refers to the same time period

(fiscal, reporting, sales), all agree on which sales should be attributed to this period, and they use the same definition for "turnover." But wait. If your definition of "turnover" changes later, they might still all agree but now report a *different* number. So our "single version of the truth" means that the entire organization responds in the same way, to the same question, *at any given point in time*. But the answer they give to this question, might be subject to change over time.

Conclusion

Having good quality, readily accessible data is a tremendous asset. Disgruntled data analysts who leave out of frustration with a poor data infrastructure are a costly "brain drain" for any organization. Usability deserves attention. Here too.

The data warehouse team is often faced with irreconcilable data streams from disparate data silos. They can't help it either that data don't "match", but they have to deal with it nonetheless. With proper senior sponsorship and a data governance program in place, they will do their best to arrive at a mutually agreeable view on "reality."

The fact that data models in your warehouse may be simpler than those in supplying source systems, can "strain" their ability to represent "the facts" well. But since a wholesale copy of your source system data models to the warehouse is (usually) not an option, they are simply asked to do the best modeling job they can.

After everybody comes to trust and rely on your DWH, a common reality emerges. This includes both facts, as well as an agreed upon *interpretation* of these facts. Together they comprise our often sought "single version of the truth."