

Design Tip #117 Dealing with Data Quality: Don't Just Sit There, Do Something!

By Warren Thornthwaite

Most data quality problems can be traced back to the data capture systems because, historically, they have only been responsible for the level of data quality needed to support transactions. What works for transactions often won't work for analytics. In fact, many of the attributes we need for analytics are not even necessary for the transactions, and therefore capturing them correctly is just extra work. By requiring better data quality as we move forward, we are requiring the data capture system to meet the needs of both transactions and analytics. Changing the data capture systems to get better data quality is a long term organizational change process. This political journey is often paralyzing for those of us who didn't expect to be business process engineers in addition to being data warehouse engineers!

Do not let this discourage you. You can take some small, productive steps in the short term that will get your organization on the road to improving data quality.

Perform Research

The earlier you identify data quality problems, the better. It will take much more time if these problems only surface well into the ETL development task, or worse, after the initial rollout. And it will tarnish the credibility of the DW/BI system (even though it's not your fault).

Your first pass at data quality research should come as part of the requirements definition phase early in the lifecycle. Take a look at the data required to support each major opportunity. Initially, this can be as simple as a few counts and ratios. For example, if the business folks wanted to do geographic targeting, calculating the percentage of rows in the customer table where the postal code is NULL might be revealing. If 20 percent of the rows don't have a postal code, you have a problem. Make sure you include this information in the requirements documentation, both under the description of each opportunity that is impacted by poor data quality, and in a separate data quality section.

The next opportunity for data quality research is during the dimensional modeling process. Defining each attribute in each table requires querying the source systems to identify and verify the attribute's domain (the list of possible values the attribute can have). You should go into more detail at this point, investigating relationships among columns, such as hierarchies, referential integrity with lookup tables, and the definition and enforcement of business rules.

The third major research point in the lifecycle is during the ETL system development. The ETL developer must dig far deeper into the data and often discovers more issues.

A data quality / data profiling tool can be a big help for data quality research. These tools allow you to do a broad survey of your data fairly quickly to help identify questionable areas for more detailed investigation. However, if you don't have a data quality tool in place, don't stop your research until you find the best tool and the funds to purchase it. Simple SQL statements like:

```
SELECT PostalCode, COUNT(*) AS RowCount
FROM Dim_Customer GROUP BY PostalCode ORDER BY 2 DESC;
```

will help you begin to identify anomalies in the data immediately. You can get more sophisticated later, as you generate awareness and concern about data quality.

It's a good idea to include the source systems folks in the research process. If they have a broader sense of responsibility for the data they collect, you may be able to get them to adjust their data collection processes to fix the problems. If they seem amenable to changing their data collection processes, it is a good idea to batch together as many of your concerns as possible while they are in a good mood. Source systems folks often aren't happy at updating and testing their code too frequently. Don't continuously dribble little requests to them!

Share Findings

Once you have an idea of the data quality issues you face, and the analytic problems they will cause, you need to educate the business people. Ultimately, they will need to re-define the data capture requirements for the transaction systems and allocate additional resources to fix them. They won't do this unless they understand the problems and associated costs.

The first major chance to educate on data quality problems is as part of the opportunity prioritization session with senior management. You should show examples of data quality problems, explain how they are created, and demonstrate their impact on analytics and project feasibility. Explain that you will document these in more detail as part of the modeling process, and at that point you can reconvene to determine your data quality strategy. Set the expectation that this is work and will require resources.

The dimensional modeling process is the second major education opportunity. All of the issues you identify during the modeling process should be discussed as part of documenting the model, and an approach to remedying the problem should be agreed upon with key business folks.

At some point, you should have generated enough awareness and concern to establish a small scale data governance effort which will become the primary research and education channel for data quality.

Conclusion

Improving data quality is a long, slow educational process of teaching the organization about what's wrong with the data, the cost in terms of accurate business decision making, and how best to fix it. Don't let it overwhelm you. Just start with your highest value business opportunity and dive into the data.