

Design Tip #116 Add Uncertainty to Your Fact Table

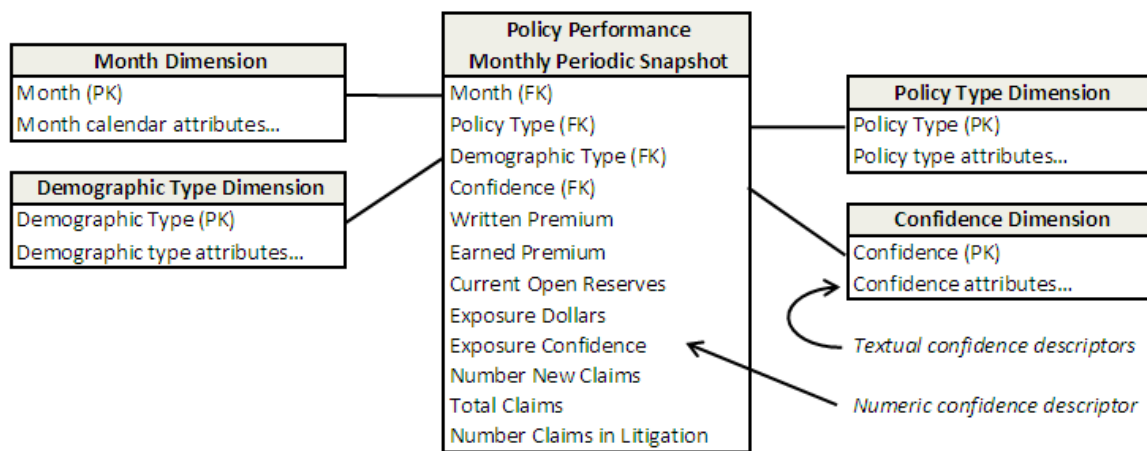
By Ralph Kimball

We always want our business users to have confidence in the data we deliver through our data warehouses. Thus it goes against our instincts to talk about problems encountered in the ETL back room, or known inaccuracies in the source systems. But this reluctance to expose our uncertainties can ultimately hurt our credibility and lessen the business users' confidence in the data if a data quality problem is revealed and we haven't said anything about it.

Almost thirty years ago when I was first exposed to A.C. Nielsen's pioneering data warehouse solution, the Inf*Act data reporting service for grocery store scanner data, I was surprised to see critical key performance indicators in their reports occasionally marked with asterisks indicating a lack of confidence in the data. In this case, the asterisk meant "not-applicable data encountered in the computation of this metric." The key performance indicator nevertheless appeared in the report, but the asterisk warned the business user not to overly trust the value. When I asked Nielsen about these asterisks, they told me the business users appreciated the warning not to make a big decision based on a specific value. I liked this implied partnership between the data provider and business user because it promoted an atmosphere of trust. And, of course, the reminders of data quality issues motivated the data provider to improve the process so as to reduce the number of asterisks.

Today I rarely see such warnings of uncertainty in the final BI layer of our data warehouses. But our world is far more wired to data than it was in 1980. I think it is time we reintroduced "uncertainty" into our fact tables. Here are two places we can add uncertainty into any fact table without changing the grain or invalidating existing applications.

Using property and casualty insurance as an example, one of the fact table's key performance indicators is the exposure dollars for a group of policies with certain demographics. This is an estimate of total liability for the known claims against the chosen set of policies. Certainly the insurance company management will pay attention to this number!



In the above figure, we accompany the exposure dollar value with an exposure confidence metric, whose value ranges between 0 and 1. An exposure confidence value of 0 indicates no confidence in the reported exposure dollars and a value of 1 indicates complete confidence. We assign the exposure confidence value in the back room ETL processes by examining the status of each claim that contributes to the aggregated record shown in the above figure. Presumably a claim associated with a very large exposure but whose claim status is “preliminary estimate” or “unverified claim” or “claim disputed” would lower the overall exposure confidence on the summary record in this fact table. If the doubtful claim’s individual reserve value is given a weight of zero, then the overall reserve confidence metric could be a weighted average of the individual reserve values.

The above figure also shows a confidence dimension containing textual attributes describing the confidence in one or more of the values in the fact table. A textual attribute for the exposure confidence would be correlated with the exposure confidence metric. An exposure confidence between 0.95 and 1 might correspond to “Certain.” An exposure confidence between 0.7 and 0.94 might correspond to “Less Certain,” and an exposure confidence of less than 0.7 might correspond to “Unreliable.” The combination of numeric and textual confidence information in our example allows BI tools to display numeric values in various ways (e.g., using italics for data less than Certain), and allows the BI tool to constrain and group on ranges of confidence.

This example should be a plausible template for providing confidence indicators on almost any fact table. I have found it to be a useful exercise just to imagine what the confidence is for various delivered metrics. And there is no question that the business users will find that exercise useful too.