Drill Down to Ask Why, Part 1

By Ralph
Kimball

Boiled down to its essence, the real purpose of a data warehouse is to be the perfect platform for decision-making. Most data warehouse (DW) and business intelligence (BI) architects accept this view, but how many stop and think carefully about what is decision-making, exactly? Every DW/BI architect can describe his or her technical architecture, but how many can describe the architecture of decision-making? If there even is an architecture of decision-making, how does the DW/BI system interact with its components, and what specific demands does decision-making place on the DW/BI system?

In 2002, Bill Schmarzo, a former member of the Kimball Group, proposed a very useful architecture for decision-making, which he called the *analytic application* process. According to Bill, an analytic application consists of five stages:

1. **Publish reports.** Provide standard operational and managerial "report cards" on the current state of a business.

2. **Identify exceptions.** Reveal the exceptional performance situations to focus attention

3. **Determine causal factors.** Seek to understand the "why" or root causes behind the identified exceptions.

4. **Model alternatives.** Provide a backdrop to evaluate different decision alternatives.

5. **Track actions.** Evaluate the effectiveness of the recommended actions and feed the decisions back to both the operational systems and DW, against which stage one reporting will be conducted, thereby closing the loop.

I have found these analytic application stages to be very useful when I think about the architecture of a DW/BI system. Publishing reports (stage one) is the traditional legacy view of the data warehouse. We pump out reports and we stack them on the end users' desks. There isn't a lot of interactive BI in stage one! We also have been identifying exceptions (stage two) with thresholds, alerts and red/green blinking graphics for many years. At least in stage two, the choice of which alerts and thresholds we want on our desktops implies some judgment and involvement by the end user.

But it is in stage three, where we determine the causal factors behind the exceptions, that life really gets interesting. A good DW/BI system should let the decision-maker bring his or her full intellectual capital to bear on understanding what the system is bringing to our attention. This stage can be summarized by one all-important word:*why.*

Suppose that you work for an airline as a fare planner. In this role, a critical key performance indicator (KPI) is the yield, which according to Wikipedia is "the revenue or profits from a fixed, perishable resource such as airline seats or hotel room reservations. The challenge is to sell the right resources to the right customer at the right time for the right price."

This morning, in your job as a fare planner, the DW/BI system produces a yield report (stage one) and highlights a number of airline routes for which the yield has dropped significantly (stage two). So, how does the DW/BI system

support the all-important stage three? How does the DW/BI system support the fare planner when you ask, "Why are my yields down?"

Imagine five ways in which the fare planner might ask why. I'll arrange these in order of increasing breadth and complexity:

1. **Give me more detail.** Run the same yield report, but break down the high-level routes by dates, time of day, aircraft type, fare class and other attributes of the original yield calculation.

2. **Give me a comparison.** Run the same yield report, but this time compare to a previous time period or to competitive yield data if it is available.

3. **Let me search for other factors.** Jump to nonyield databases, such as a weather database, a holiday/special events database, a marketing promotions database or a competitive pricing database to see if any of these exogenous factors could have played a role.

4. **Tell me what explains the variance.** Perform a data mining analysis, perhaps using decision trees, examining hundreds of marketplace conditions to see which of these conditions correlates most strongly with the drop in yield (*explaining the variance* in data mining terminology).

5. **Search the Web for information about the problem.** Google or Yahoo! the Web for "airline yield 2008 versus 2007."

Twenty years ago, when we drilled down in a data warehouse to ask why, we rarely provided more than the first capability above. I like to think of the longer list of all five capabilities as 2008's definition of drilling down to ask why. Next month, I'll explore how to implement these capabilities in a real DW/BI system.