Slowly Changing Dimensions

By Ralph
Kimball

The notion of time pervades every corner of the data warehouse. Most of the fundamental measurements we store in our fact tables are time series, which we carefully annotate with time stamps and foreign keys connecting to calendar date dimensions. But the effects of time are not isolated just to these activity-based time stamps. All of the other dimensions that connect to fact tables, including fundamental entities such as Customer, Product, Service, Terms, Location and Employee, are also affected by the passage of time. As data warehouse managers, we are routinely confronted with revised descriptions of these entities. Sometimes the revised description merely corrects an error in the data. But many times the revised description represents a true change at a point in time of the description of a particular dimension member, such as Customer or Product. Because these changes arrive unexpectedly, sporadically and far less frequently than fact table measurements, we call this topic slowly changing dimensions (SCDs).

The Three Types

In more than 30 years of studying the time variance of dimensions, amazingly I have found that the data warehouse only needs three basic responses when confronted with a revised or updated description of a dimension member. I call these, appropriately, Types 1, 2 and 3. I'll start with Type 1 this month, and I will use the Employee dimension to keep the discussion from being too abstract.

Type 1: Overwrite

Suppose we are notified that the Home City field for Ralph Kimball in the Employee dimension has changed from Santa Cruz to Boulder Creek as of today. Furthermore, we are advised that this is an error correction, not an actual change of location. In this case, we may decide to overwrite the Home City field in the Employee dimension with the new value. This is a classic Type 1 change. Type 1 changes are appropriate for correcting errors and for situations where a conscious choice is made not to track history. And of course, most data warehouses start out with Type 1 as the default.

While the Type 1 SCD is the simplest and seemingly cleanest change, there are a number of fine points to think about:

1. Type 1 destroys the history of a particular field. In our example, reports that constrain or group on the Home City field will change. End users will need to be aware that this can happen. The data warehouse needs an explicit, visible policy for Type 1 fields that says, "We will correct errors" and/or "We do not maintain history on this field even if it changes."
2. Precomputed aggregates (including materialized views and automatic summary tables) that depend on the Home City field must be taken offline at the moment of the overwrite and must be recomputed before being brought back online. Aggregates that do not depend on the Home City field are unaffected.
3. In financial reporting environments with month end close processes and in any environment subject to regulatory or legal compliance, Type 1 changes may be outlawed. In these cases, the Type 2 technique must be used.

4. Overwriting a single dimension field in a relational environment has a pretty small impact but can be disastrous in an online analytical processing (OLAP) environment if the overwrite causes the cube to be rebuilt. Carefully study your OLAP system reference manual to see how to avoid unexpected cube rebuilds.
5. All distributed copies of the Employee dimension, as well as aggregates, must be updated simultaneously across the enterprise when Type 1 changes occur, or else the logic of drilling across will be corrupted. In a distributed environment, Type 1 (and Type 3) changes should force the dimension version number to be updated, and all drill across applications must include the dimension version number in their queries. This process was described in detail in my columns on the architecture of the integrated enterprise data warehouse.
6. In a pure Type 1 dimension where all fields in the dimension are subject to overwriting, a Type 1 change like the Home City change for Ralph Kimball will typically affect only one record (the record for Ralph Kimball). But in a more typical complex environment, where some fields are Type 1 and other fields are Type 2, the act of overwriting the Home City field must overwrite all the records for Ralph Kimball. In other words, Type 1 affects all history, not just the current perspective.

In next month's column, instead of responding to a change by overwriting, I'll carefully keep track of the change by issuing a new dimension record. This is the classic Type 2 SCD. And finally, I'll show how to handle a requested change that establishes an alternate reality that coexists with the current truth. This is the Type 3 SCD.