http://www.kimballgroup.com/2008/09/slowly-changing-dimensions-part-2/

Slowly Changing Dimensions, Part 2

By Ralph
Kimball

The owner of the data warehouse must decide how to respond to the changes in the descriptions of dimensional entities like Employee, Customer, Product, Supplier, Location and others. In 30 years of studying this issue, I have found that only three different kinds of responses are needed. I call these slowly changing dimension (SCD) Types 1, 2 and 3. In last month's column, I described Type 1, which overwrites the changed information in the dimension. In this column I describe Types 2 and 3.

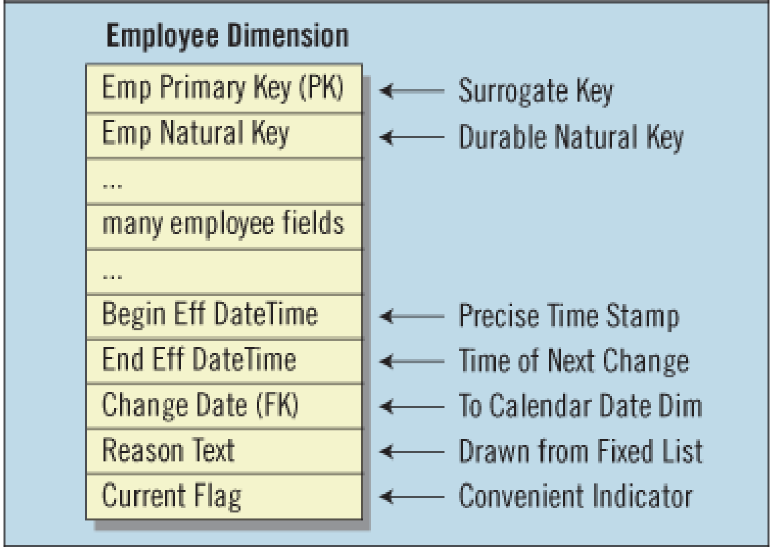
Type 2: Add a New Dimension Record

Let's alter the scenario of the previous column where I overwrote the Home City field in Ralph Kimball's employee record to assume that Ralph Kimball actually moved from Santa Cruz to Boulder Creek on July 18, 2008. Assume our policy is to accurately track the employee home addresses in the data warehouse. This is a classic Type 2 change.

The Type 2 SCD requires that we issue a new employee record for Ralph Kimball effective July 18, 2008. This has many interesting side effects:

1. Type 2 requires that we generalize the primary key of the Employee dimension. If Ralph Kimball's employee natural key is G446, then that natural key will be the "glue" that holds Ralph Kimball's multiple records together. I do not recommend creating a smart primary key for Type 2 SCDs that contains the literal natural key. The problems with smart keys become especially obvious if you are integrating several incompatible HR systems with differently formatted natural keys. Rather, you should create completely artificial primary keys that are simply sequentially assigned integers. We call these keys surrogate keys. You must make a new surrogate primary key whenever you process a Type 2 change in a dimension.
2. In addition to the primary surrogate key, I recommend adding five additional fields to a dimension that is undergoing Type 2 processing. These fields are shown in Figure 1. The datetimes are full time stamps that represent the span of time between when the change became effective and when the next change becomes effective. The end-effective-datetime of a Type 2 dimension record must be exactly equal to the begin-effective-datetime of the next change for that dimension member. The most current dimension record must have an end-effective-datetime equal to a fictitious datetime far in the future. The reason text for the change should be drawn from a preplanned list of reasons for a change, in our example, to the employee attributes. Finally, the current-flag provides a rapid way to isolate exactly the set of dimension members that is in effect at the moment of the query. These five administrative fields allow end users and applications to perform many powerful queries.
3. With a dimension undergoing Type 2 processing, great care must be taken to use the correct contemporary surrogate keys from this dimension in every affected fact table. This assures that the correct dimension profiles are associated with fact table activity. The extract, transform and load (ETL) process for aligning the dimension tables with fact tables at load time is called the surrogate key pipeline and is covered extensively in my articles and books.

Figure 1: Employee Dimension Designed for Type 2 SCD



Employee Dimension

Field	
Emp Primary Key (PK)	← Surrogate Key
Emp Natural Key	← Durable Natural Key
...	
many employee fields	
...	
Begin Eff DateTime	← Precise Time Stamp
End Eff DateTime	← Time of Next Change
Change Date (FK)	← To Calendar Date Dim
Reason Text	← Drawn from Fixed List
Current Flag	← Convenient Indicator

Type 3: Add a New Field

Although the Type 1 and 2 SCDs are the primary workhorse techniques for responding to changes in a dimension, we need a third technique for handling alternate realities. Unlike physical attributes that can only have one value at a point in time, some user-assigned attributes can legitimately have more than one assigned value depending on the observer's point of view. For example, a product category can have more than one interpretation. In a stationery store, a marking pen could be assigned to the household goods category or the art supplies category. End users and applications need to be able to choose at query time which of these alternate realities applies.

The requirement for an alternate reality view of a dimension attribute usually is accompanied by a subtle requirement that separate versions of reality be available at all times in the past and in the future, even though the request to make these realities visible arrived at the data warehouse today.

In the simplest variation, there is only one alternate-reality. In this case, for the product category example, we add a new field in the dimension, perhaps called Alternate Category. If the primary category of our marking pen used to be household goods and now should be art supplies, then in a Type 3 treatment, we push the household goods label into the Alternate Category field and we update the regular Category field with art supplies by overwriting. The overwriting step is similar to a Type 1 SCD and provokes all the same caveats in last month's column.

With Type 3 machinery in place, end users and applications can switch seamlessly between these alternate realities. If the environment requires more than one alternate reality, this approach can be generalized by adding more Alternate fields, although obviously this approach does not scale gracefully beyond a few choices.

The three SCD approaches to handling time variance in dimensions have enormous applicability in the real-world situations encountered by the data warehouse. Type 2, in particular, allows us to make good on the data warehouse pledge to preserve history faithfully.