

## **Design Tip #86 Creating a Reference Dimension for Infrequently-Accessed Degenerates**

By Bob Becker

In Design Tips #81 and #84, Ralph and I explored several circumstances when it may be advantageous to create a surrogate key for fact tables. This design tip leverages that earlier discussion and introduces a concept called the reference dimension where we stash rarely used fact table elements, such as degenerate dimension reference numbers, in a separate table that's linked to the fact table either through a regular dimension surrogate key or the fact table's surrogate key.

While degenerate dimensions (Design Tip #46) often play an important role in supporting reporting and analytic requirements, some degenerate dimensions are not analytically valuable; they are included in the schema for reference purposes only. They help provide occasional ties back to the operational source systems, support audit, compliance or legal requirements, or are included simply because "we might need it someday." The result can be a fact table that contains a large number of degenerate dimensions with perhaps only two or three that are truly important and interesting. In healthcare, for example, the degenerate dimensions might reference the provider network contract number, fee schedule number, and claim microfiche number. Since fact tables are the largest tables in our schema often containing hundreds of millions or billions of rows, we'd like to keep them as tight as possible. It certainly seems wasteful to populate the fact table with ten or more large alphanumeric degenerate dimensions especially if they are not typically used for reporting or analytics. This is where the reference dimension can be helpful.

The concept of the reference dimension is to break the fact table apart, moving the seldom used degenerate dimension values to a separate reference dimension table with the fact table or reference dimension surrogate key preserving the relationship between the tables. The analytically-valuable degenerate dimensions should not be moved to the reference dimension; they need to be retained in the fact table where they can be used most effectively. It is very important to move only the degenerate dimensions that are not used to support analytic or reporting requirements. While it is generally not considered good dimensional modeling practice to create a dimension with a potential one-to-one relationship with the fact table, in this situation we will accept it as a reasonable tradeoff. The important advantage we gain is dramatically reducing the length of the fact table row resulting in a tighter design that will perform better. This design tradeoff works because the reference dimension should very seldom actually join back to the fact table. If the design assumptions were correct and the degenerate dimensions moved to the reference dimension are not required to support reporting and analytics, most users will never use the reference dimension. There should be just a few occasional investigatory requirements that need to access this information.

In the rare occasion when some component of the reference table is required, it will be necessary to join it to the fact table which may be an expensive and slow running query. User expectations when using the reference dimension need to be carefully managed. Frequent complaints about the performance of the reference dimension likely means some data element needs to migrate back to the fact table as it's important for supporting reporting or analytic needs.

**Caution:** This design tip should not to be construed as granting design teams permission to build large dimensions with potentially one-to-one relationships with the fact table. Rather, the reference dimension is a specific design response to a particular situation that a design team may confront. We would not expect to see reference dimensions in most dimensional designs. They should be the rare exception rather than the rule.