www.kimballgroup.com Number 48, August 7, 2003

Kimball Design Tip #48: De-Clutter With Junk (Dimensions)

By Margy Ross

When developing a dimensional model, we often encounter miscellaneous indicators and flags that don't logically belong to the core dimension tables. These unattached attributes are usually too valuable to ignore or exclude. Designers sometimes want to treat them as facts (supposed textual facts) or clutter the design with numerous small dimensional tables. A third, less obvious but preferable, solution is to incorporate a junk dimension as a holding place for these flags and indicators.

A junk dimension is a convenient grouping of flags and indicators. It's helpful, but not absolutely required, if there's a positive correlation among the values. The benefits of a junk dimension include:

- Provide a recognizable, user-intuitive location for related codes, indicators and their descriptors in a dimensional framework.
- Clean up a cluttered design that already has too many dimensions. There might be five or more indicators that could be collapsed into a single 4-byte integer surrogate key in the fact table.
- Provide a smaller, quicker point of entry for queries compared to performance from constraining directly on these attributes in the fact table. If your database supports bit-mapped indices, this potential benefit may be irrelevant, although the others are still valid.

An interesting use for a junk dimension is to capture the context of a specific transaction. While our common, conformed dimensions contain the key dimensional attributes of interest, there are likely attributes about the transaction that are not known until the transaction is processed.

For example, a healthcare insurance provide may need to capture the context surrounding their claims transactions. The grain for this key business process is one row for each line item on a claim. Due to the complexities of the healthcare industry, similar claims may be handled quite differently. They may design separate junk dimensions to capture the context of how the claim was processed, how it was paid, and the contractual relationship between the healthcare providers at the time of the claim.

There are two approaches for creating junk dimensions. The first is to create the junk dimension table in advance. Each possible, unique combination generates a row in the junk dimension table. The second approach is to create the rows in the junk dimension on the fly during the extract, transformation, and load (ETL) process. As new unique combinations are encountered, a new row with its surrogate key is created and loaded into the junk dimension table.

If the total number of possible rows in the junk dimension is relatively small, it is probably best to create the rows in advance. On the other hand, if the total number of possible rows in the junk dimension is large, it may be more advantageous to create the junk dimension as unique rows are encountered. One of the junk dimensions encountered in the recent healthcare design had over 1 trillion theoretical rows, while the actual number of observed rows was tens of thousands. Obviously, it did not make sense to create all the theoretically possible rows in advance. If the number of rows in the junk dimension approaches or exceeds the number of rows in the fact table, the design should be clearly re-evaluated.

Since a junk dimension includes all valid combinations of attributes, it will automatically track any changes in dimension attributes. Therefore slowly changing dimension strategies do not need to be considered for junk dimensions.

For more information, see Ralph's Intelligent Enterprise article at http://www.intelligententerprise.com/000320/webhouse.shtml. Junk dimensions are referred to as mystery dimensions in this article.