KIMBALL GROUP KIMBALL UNIVERSITY

www.kimballgroup.com Number 57, August 2, 2004

Kimball Design Tip #57: Early Arriving Facts

By Ralph Kimball

Data warehouses are usually built around the ideal normative assumption that measured activity (the fact records) arrive in the data warehouse at the same time as the context of the activity (the dimension records). When we have both the fact records and the correct contemporary dimension records, we have the luxury of bookkeeping the dimension keys first, and then using these up-to-date keys in the accompanying fact records.

Basically three things can happen when we bookkeep the dimension records.

1) If the dimension entity (say Customer) is a new member of the dimension, we assign a fresh new surrogate dimension key.
2) If the dimension entity is a REVISED version of a Customer, we use the Type 2 slowly changing dimension technique of assigning a new surrogate key and storing the revised Customer description as a new dimension record.
3) Finally if the Customer is a familiar, unchanged member of the dimension, we just use the dimension key we already have for that Customer.

For several years, we have been aware of special modifications to these procedures to deal with Late Arriving Facts, namely fact records that come into the warehouse very much delayed. This is a messy situation because we have to search back in history within the data warehouse to decide how to assign the right dimension keys that were in effect when the activity occurred at the right point in the past. See the Intelligent Enterprise article (*Backward in Time*) on this subject at www.intelligententerprise.com/000929/webhouse.jhtml.

If we have Late Arriving Facts, is it possible to have Early Arriving Facts? How can this happen? Are there situations where this is important?

An early arriving fact takes place when the activity measurement arrives at the data warehouse without its full context. In other words, the statuses of the dimensions attached to the activity measurement are ambiguous or unknown for some period of time. If we are living in the conventional batch update cycle of one or more days latency, we can usually just wait for the dimensions to be reported to us. For example, the identification of the new customer may come in a separate feed delayed by several hours. We may just be able to wait until the dependency is resolved.

But if we are in a real time data warehousing situation in which the fact record must be made visible NOW, and we don't know when the dimensional context will arrive, we have some interesting choices. Our real-time bookkeeping needs to be revised, again using Customer as the problem dimension.

1) If the natural Customer key on the incoming fact record can be recognized, then we provisionally attach the surrogate key for the existing most recent version of that Customer to the fact table, but we also hold open the possibility that we will get a revised version of this Customer reported to us at a later time. At that point, we 2) add the revised Customer record to the dimension with a new surrogate key and then go in and destructively modify the fact record's foreign key to the Customer table. 3) Finally, if we believe that the Customer is new, we assign a new Customer surrogate key with a set of dummy attribute values in a new Customer dimension record. We then return to this

© Copyright Kimball Group, 2004. All rights reserved.

dummy dimension record at a later time and make Type 1 (overwrite) changes to its attributes when we get more complete information on the new Customer. At least this step avoids destructively changing any fact table keys.

There is no way to avoid a brief provisional period where the dimensions are “not quite right.” But these bookkeeping steps try to minimize the impact of the unavoidable updates to keys and other fields. If these early arriving records are all housed in a “hot partition” pinned in memory, then aggregate fact table records should not be necessary. Only when the hot partition is conventionally loaded into the static data warehouse tables at the end of the day (and when the dimensions have caught up with the facts) do you need to build the aggregates.

© Copyright Kimball Group, 2004. All rights reserved.