http://www.kimballgroup.com/2007/07/keep-to-the-grain-in-dimensional-modeling/

Keep to the Grain in Dimensional Modeling

By Ralph
Kimball

When developing fact tables, aggregated data is NOT the place to start. To avoid “mixed granularity” woes including bad and overlapping data, stick to rich, expressive, atomic-level data that’s closely connected to the original source and collection process.

The power of a dimensional model comes from a careful adherence to “the grain.” A clear definition of the grain of a fact table makes the logical and physical design possible; a muddled or imprecise definition of the grain poses a threat to all aspects of the design, from the ETL processes that fetch the data all the way to the reports that try to use the data.

What, exactly, is the grain? The grain of a fact table is the business definition of the measurement event that creates a fact record. The grain is exclusively determined by the physical realities of the source of the data.

All grain definitions should start at the lowest, most atomic grain and should describe the physical process that collects the data. Thus, in our dimensional modeling classes, when we start with the familiar example of retail sales, I ask the students “what is the grain?” After listening to a number of careful replies listing various retail dimensions such as product, customer, store and time, I stop and ask the students to visualize the physical process. The salesperson or checkout clerk scans the retail item and the register goes “BEEP.” The grain of the fact table is BEEP!

Once the grain of the fact table is established with such clarity, the next steps of the design process can proceed smoothly. Continuing with our retail example, we can immediately include or exclude possible dimensions from the logical design of the retail fact table. Benefiting from the very atomic definition (BEEP), we can propose a large number of dimensions including date, time, customer, product, employee (perhaps both checkout clerk and supervisor), store, cash register, sales promotion, competitive factor, method of payment, regional demographics, the weather, and possibly others. Our humble little BEEP turns into a powerful measurement event with more than a dozen dimensions!

Of course, in a given practical application, the design team may not have all these dimensions available. But the power of keeping to the grain arises from the clarity that the grain definition supplies to the design process. Once the logical design has been proposed, the design team can systematically investigate whether the data sources are rich enough to “decorate” the BEEP event with all these dimensions in the physical implementation.

The BEEP grain illustrates why the atomic data is the place to start all designs. The atomic data is the most expressive data because it is the most precisely defined. Aggregated data, for example, store sales by product by month, can easily be derived from the atomic data, but necessarily must truncate or delete most of the dimensions of the atomic data. Aggregated data is NOT the place to start a design!

This design tip was motivated in part by an article that appeared recently in a trade magazine discussing the design of dimensional (star) schemas. The author, who otherwise writes pretty clearly on this subject, stated “a star needs to be defined by a set of business questions and [metrics are assigned] to stars based on common reporting and queries.” This is terrible advice! A dimensional model that is designed around business questions and reports has no clear grain. It has lost its connection to the original data source and is hostage to the “report of the day” mentality that causes such a database to be tweaked and altered until no one can explain why a record is in the table or not.

When a dimensional design has lost its connection to the grain, it becomes vulnerable to a subtle problem, called mixed granularity, that is nearly impossible to fix. In this case, records in the same fact table may represent different physical measurement events that are not comparable or may even overlap. A simple example deviating from the BEEP grain would be a fact table containing “combo-pack” sales records in addition to the sales records of the individual items comprising the combo-pack. This is dangerous because without a careful constraint in the query tool or report, the sales of these items would be double counted. A corollary to keeping to the grain is “don’t require the end user tools to correct problems with the grain.”

Keeping to the grain means building physical fact tables around each atomic measurement event. These tables are the least difficult to implement and they provide the most durable and flexible foundation for addressing business questions and reports-of-the day.