The Matrix

By Ralph
Kimball

Over the years, I have found that a matrix depiction of the data warehouse plan is a pretty good planning tool once you have gathered the business requirements and performed a full data audit. This matrix approach has been exceptionally effective for distributed data warehouses without a center. Most of the new Web-oriented, multiple organization warehouses we are trying to build these days have no center, so it is even more urgent that we find a way to plan these beasts.

The matrix is simply a vertical list of data marts and a horizontal list of dimensions. Figure 1 is an example matrix for the enterprise data warehouse of a large telecommunications company. You start the matrix by listing all the *first-level* data marts that you could possibly build over the next three years across the enterprise. A first-level data mart is a collection of related fact tables and dimension tables that is typically:

- *Derived* from a single data source
- *Supported and implemented* by a single department
- *Based on the most atomic data possible* to collect from the source
- *Conformed* to the "data warehouse bus."

First-level data marts should be the smallest and least risky initial implementations of an enterprise data warehouse. They form a foundation on which a larger implementation can be brought to completion in the least amount of time, but that are still guaranteed to contribute to the final result without being incompatible stovepipes.

You should try to reduce the risk of implementation as much as possible by basing the first-level data marts on single production sources. In my experience, the cost and complexity of data warehouse implementation, once the "right" data has been chosen, turns out to be proportional to the number of data sources that must be extracted. Each separate data source can be as much as a six-month programming and testing exercise. You must create a production data pipeline from the legacy source through the data staging area and on to the fact and dimension tables of the presentation part of the data warehouse.

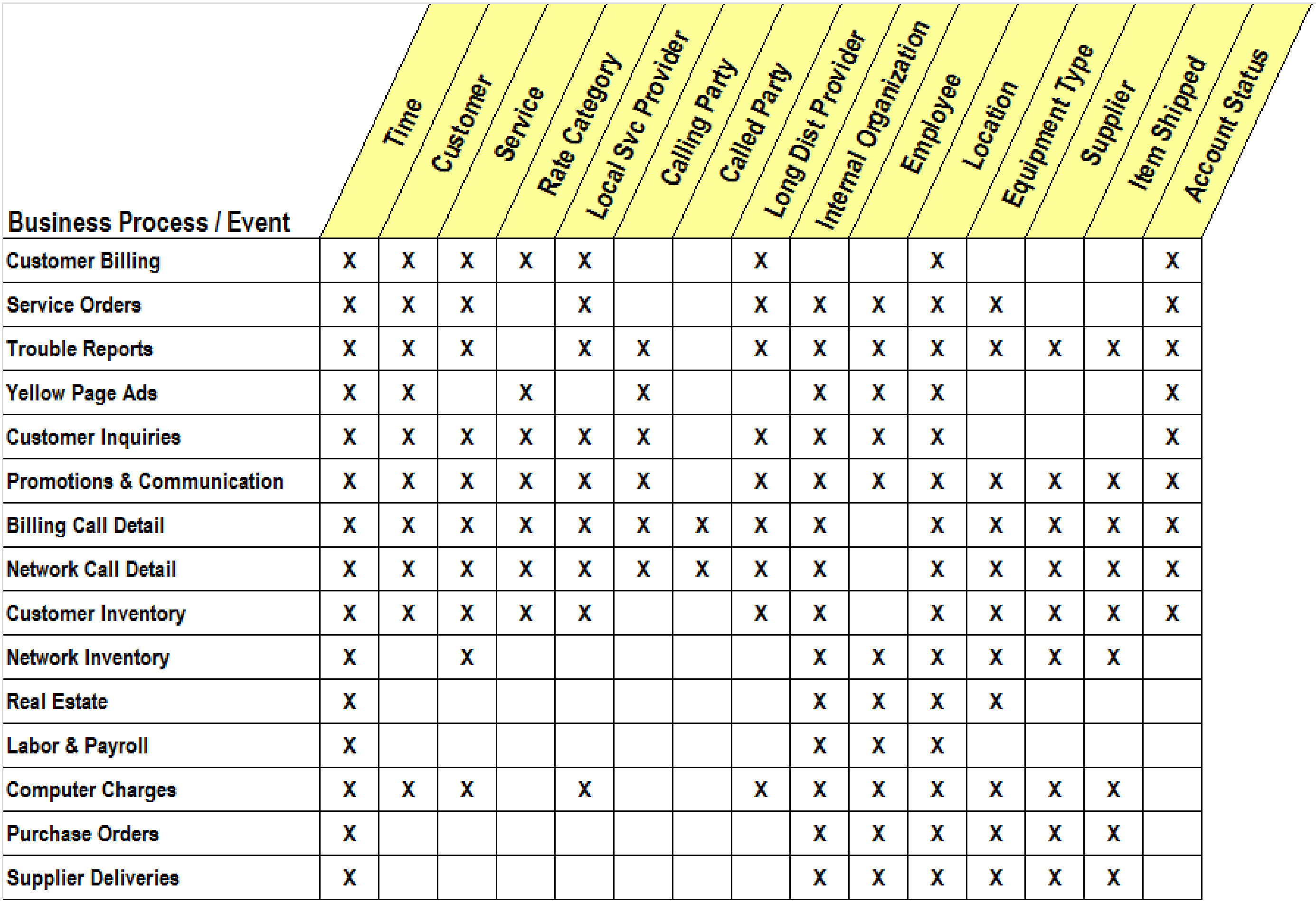
In Figure 1, the first-level data marts for the telecommunications company are many of the major production data sources. An obvious production data source is the customer billing system, listed first. This row in the matrix is meant to represent all the base-level fact tables you expect to build in this data mart. Assume this data mart contains one major base-level fact table, the grain of which is the individual line item on a customer bill. Assume the line item on the bill represents the class of service provided, not the individual telephone call within the class of service. With these assumptions, you can check off the dimensions this fact table needs. For customer bills, you need Time, Customer, Service, Rate Category, Local Service Provider, Long Distance Provider, Location, and Account Status.

Continue to develop the matrix rows by listing all the possible first-level data marts that could be developed in the next three years, based on known, existing data sources. Sometimes I am asked to include a first-level data mart based on a production system that does not yet exist. I usually decline the offer. I try to avoid including "potential" data sources, unless there is a very specific design and implementation plan in place. Another dangerously idealistic data source is the grand corporate data model, which usually takes up a whole wall of the IT department. Most of this data model cannot be used as a data source because it is not real. Ask the corporate data architect to highlight with a red pen the tables on the corporate data model that are currently populated with real data. These red tables are legitimate drivers of data marts in the planning matrix and can be used as sources.

The planning matrix columns indicate all the dimensions a data mart might need. A real enterprise data warehouse contains more dimensions than those in Figure 1. It is often helpful to attempt a comprehensive list of dimensions before filling in the matrix. When you start with a large list of dimensions, it becomes a kind of creative exercise to ask whether a given dimension could possibly be associated with a data mart. This activity could suggest interesting ways to add dimensional data sources to existing fact tables. If you study the details of Figure 1, you may decide that more X's should be filled in, or that some significant dimensions should be added. If so, more power to you! You are using the matrix as it was intended.

Inviting Data Mart Groups to the Conforming Meeting

Looking across the rows of the matrix is revealing. You can see the full dimensionality of each data mart at a glance. Dimensions can be tested for inclusion or exclusion. But the real power of the matrix comes from looking at the columns. A column in the matrix is a map of where the dimension is required.



Business Process / Event	Time	Customer	Service	Rate Category	Local Svc Provider	Calling Party	Called Party	Long Dist Provider	Internal Organization	Employee	Location	Equipment Type	Supplier	Item Shipped	Account Status
Customer Billing	X	X	X	X	X			X			X				X
Service Orders	X	X	X		X			X	X	X	X	X			X
Trouble Reports	X	X	X		X	X		X	X	X	X	X	X	X	X
Yellow Page Ads	X	X		X		X			X	X	X				X
Customer Inquiries	X	X	X	X	X	X		X	X	X	X				X
Promotions & Communication	X	X	X	X	X	X		X	X	X	X	X	X	X	X
Billing Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X
Network Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X
Customer Inventory	X	X	X	X	X			X	X		X	X	X	X	X
Network Inventory	X		X						X	X	X	X	X	X	
Real Estate	X								X	X	X	X			
Labor & Payroll	X								X	X	X				
Computer Charges	X	X	X		X			X	X	X	X	X	X	X	
Purchase Orders	X								X	X	X	X	X	X	
Supplier Deliveries	X								X	X	X	X	X	X	

FIGURE 1 The Matrix Plan for the enterprise data warehouse of a large telecommunications company.

The first dimension, Time, is required in every data mart. Every data mart is a time series. But even the Time dimension requires some thought. When a dimension is used in multiple data marts, *it must be conformed*. Conformed dimensions are the basis for distributed data warehouses, and using conformed dimensions is the way to avoid stovepipe data marts. A dimension is conformed when two copies of the dimensions are either exactly the same (including the values of the keys and all the attributes), or else one dimension is a perfect subset of the other. So using the Time dimension in all the data marts implies that the data mart teams agree on a corporate calendar. All the data mart teams must use this calendar and agree on fiscal periods, holidays, and workdays.

The grain of the conformed Time dimension needs to be consistent as well. An obvious source of stovepipe data marts is the reckless use of incompatible weeks and months across the data marts. Get rid of awkward time spans such as quad weeks or 4-4-5-week quarters.

The second dimension in Figure 1, Customer, is even more interesting than Time. Developing a standard definition for “customer” is one of the most important steps in combining separate sources of data from around the enterprise. The willingness to seek a common definition of the customer is a major litmus test for an organization intending to build an enterprise data warehouse. Roughly speaking, if an organization is unwilling to agree on a common definition of the customer across all data marts, the organization should not attempt to build a data warehouse that spans these data marts. The data marts should remain separate forever.

For these reasons, you can think of the planning matrix columns as the invitation list to the conforming meeting! The planning matrix reveals the interaction between the data marts and the dimensions.

Communicating With the Boss

The planning matrix is a good communication vehicle for senior management. It is simple and direct. Even if the executive does not know much about the technical details of the data warehouse, the planning matrix sends the message that standard definitions of calendars, customers, and products must be defined, or the enterprise won’t be able to use its data.

A meeting to conform a dimension is probably more political than technical. The data warehouse project leader does not need to be the sole force for conforming a dimension such as Customer. A senior manager such as the enterprise CIO should be willing to appear at the conforming meeting and make it clear how important the task of conforming the dimension is. This political support is very important. It gets the data warehouse project manager off the hook and puts the burden of the decision making process on senior management’s shoulders, where it belongs.

Second-Level Data Marts

After you have represented all the major production sources in the enterprise with first-level data marts, you can define one or more second-level marts. A second-level data mart is a combination of two or more first-level marts. In most cases, a second-level mart is more than a simple union of data sets from the first-level marts. For example, a second-level profitability mart may result from a complex allocation process that associates costs from several first-level cost-oriented data marts onto products and customers contained in a first-level revenue mart. I discussed the issues of creating these kinds of profitability data marts in my column, “Not so Fast.”

The matrix planning technique helps you build an enterprise data warehouse, especially when the warehouse is a distributed combination of far-flung data marts. The matrix becomes a resource that is part technical tool, part project management tool, and part communication vehicle to senior management.