A Systems Approach to Dimensional Modeling in Data Marts

By

Joseph M. Firestone, Ph.D.

White Paper No. One

March 12, 1997

OLAP's Purposes And Dimensional Data Modeling

Dimensional Data Modeling (DDM) for Data Marts needs to be viewed from the standpoint of OLAP requirements. Let's take Nigel Pendse's and Richard Creeth's ***F***ast ***A***nalysis of ***S***hared ***M***ultidimensional ***I***nformation (**FASMI**) definition of OLAP [1] as the basis of discussion, since it seems to reflect a greater industry consensus than earlier definitions.

In terms of FASMI, DDM needs to support:

- delivery of most responses to queries in (F) five seconds, with simple queries coming back in less than one second, and all but the very few most difficult queries taking no longer than 20 seconds;
- access to records one needs to perform a variety of analyses (A), including
  - database segmenting (or subsetting according to the criteria specified in a query, also known as "dicing"),
  - rotating (also known as "data slicing,") to examine a different view of the multidimensional data being queried without having to reassemble the view from more basic data,
  - aggregating or disaggregating multidimensional data to display higher or lower levels in an analytic hierarchy such as time periodicity, geography, or business/social/ government organizational hierarchy (known as "rolling up" or drilling down).
  - predictive modeling,
  - time series analysis,
  - measurement modeling combining database attributes (to develop good measures of important abstractions such as corporate or government performance, customer satisfaction, strength of customer bonding, and many other properties not adequately measured by a single database attribute or variable),

- nonlinear, even fuzzy, causal and structural modeling combining measurement and causal models (to develop impact modeling and further refine predictive modeling),
- short- and long-term forecasting,
- automated exploratory data analysis (data mining) to aid Knowledge Discovery in Databases (KDD),
- validation analysis (see my White Paper, "Data Mining and KDD" [2] for an explanation of why validation of data mining is necessary) of patterns discovered through data mining;

- a multidimensional (M) conceptual view of the data in the application;
- a comprehensive organization of all the data (I) that may be needed to achieve KDD.

While current approaches are pretty effective for most of this list, they're not enough for all of it. In particular, not for supporting measurement, causal, and structural modeling, or long-term forecasting; and not for supporting a comprehensive-enough specification and organization of data for KDD. So, I'm moved to propose the following systems approach to dimensional data modeling in data marts. I use the term "systems approach" because it's designed to support development of business area models of system dynamics (in the broadest sense of this term, not to be confused with Jay Forrester's name for his well-known modeling technique, popular in the late 'sixties and the 'seventies). Also, while the statement of this approach is more explicit than his, and is heavier on the prior conceptualization and data inventory side of DDM development, much of it is implicit in Ralph Kimball's brilliant published work. [3]

A Systems Approach to DDM

STEP ONE: Develop a conceptual framework dividing the subject matter to be covered by the data mart into conceptual domains. Each domain should be a category composed of individual concepts, each having two or more possible values. The concepts should be high level abstractions, not data attributes.

Arrive at the domains and their concepts by thinking in terms of the following categories.

(I) Conditions, states, or things that users of the data mart value as goals, or objectives, or criteria related to goals and objectives, of their enterprise or division. These are endogenous to the system being tracked by the data mart, and should be viewed as the effects of certain causes. Favorite examples are Amount of Sales, Unit Costs, Quantity Sold, and Discounts.

When arriving at these concepts don't forget to include possible concepts relating to possible "side effects" only indirectly related to enterprise goals and objectives. Side effects are important because they can lead to feedback effects on key goals and objectives even though they initially may seem unrelated to them.

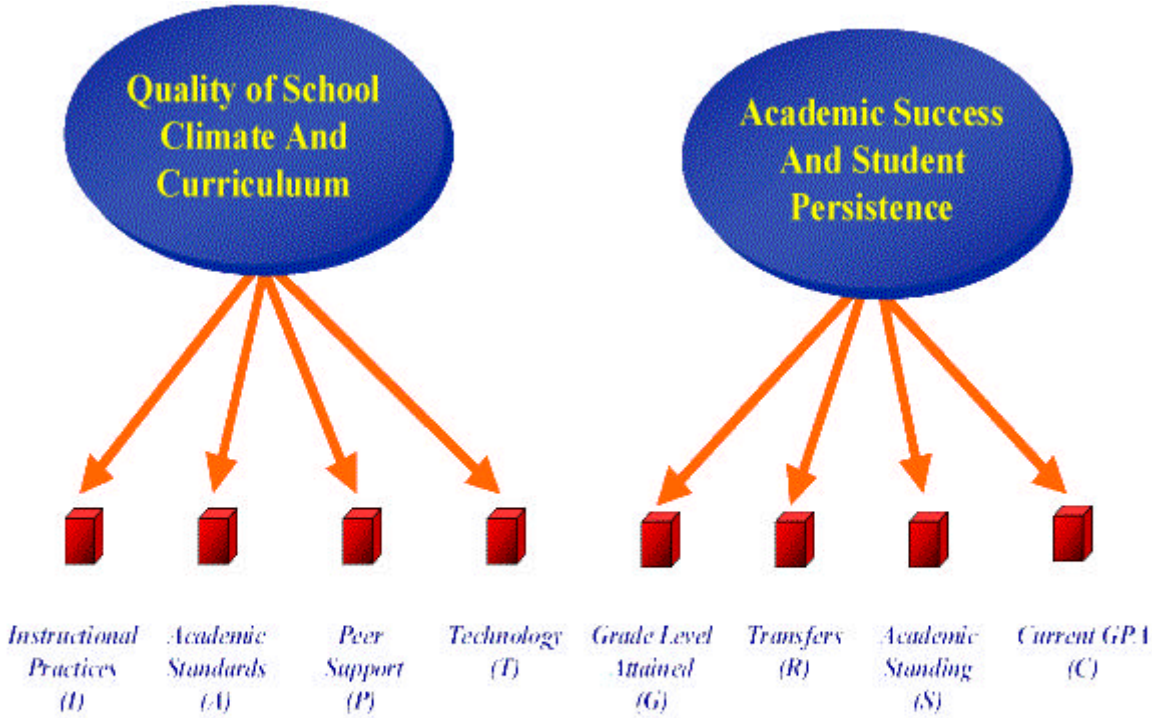

Figure One -- Two Conceptual Domains and Some Associated Concepts

(II) Conditions, states, or things that may conceivably have an impact on the goals or objectives of the enterprise or division developing the data mart. In other words, the possible causes or exogeneous factors such as: advertising or direct mail promotions, interest rates, demographic background of customers, and appearance of technologically sophisticated competitors.

(III) Conditions, states, or things representing descriptive properties that provide a framework for segmenting effects into subtypes useful for understanding the details of effects viewed as business process outcomes. Some examples are product type, and product department. The distinction between Type II causal concepts and Type III descriptive concepts is not hard and fast. It depends on what you think about causality. But in every data mart you'll think of some concepts as causal, and others as primarily descriptive. And your views on these matters may change as the data mart is used and more analysis and data mining is done.

(IV) Components of analytic hierarchies defining levels of description and analysis that provide a framework for either globalizing or localizing descriptions through geographical and social space and/or time. Examples are the time, geographic, organizational, political, product hierarchies whose components are levels of analysis characterized by parent/child or global/local relations.

STEP TWO: Develop the conceptual framework further by specifying abstract cause-effect relations between category (II) causal concepts, and category (I) effects concepts, and by distinguishing the analytic hierarchies for globalizing or localizing measurement. To specify

cause-effect relations you can use a set of formless functional equations that lay out possible or conceivable relationships, but make no comitment to particular linear or non-linear forms. You can go further and formulate more specific causal relations if you have good reasons for doing this, but the data mart DDM need only *provide a framework* for formulating alternative causal models and theories of system dynamics. Specific models can be formulated during subsequent analysis and modeling activities.

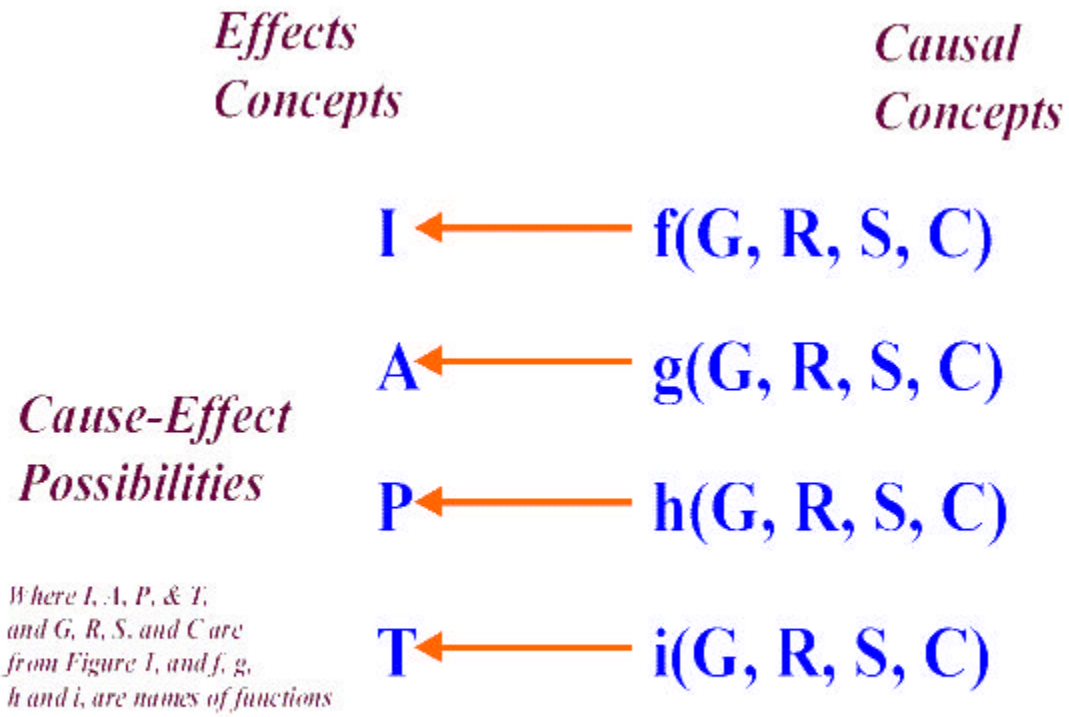

Figure Two -- Specifying Abstract Cause-Effect Relations

STEP THREE: From the initial framework resulting from STEP TWO, extract the conceptual domains and associated concepts you will want to use for the current data mart. Conceptual domains will eventually map to candidate relational tables, though not in general one-to-one, and their associated concepts will guide you to data variables that, in turn, will provide measures of these concepts. Find these data variables by doing a data inventory of all data sources accessible to the data mart team, both internal and external to the sponsoring enterprise. Then map the data variables to the concepts and conceptual domains in order to define candidate attributes for the eventual DDM tables.

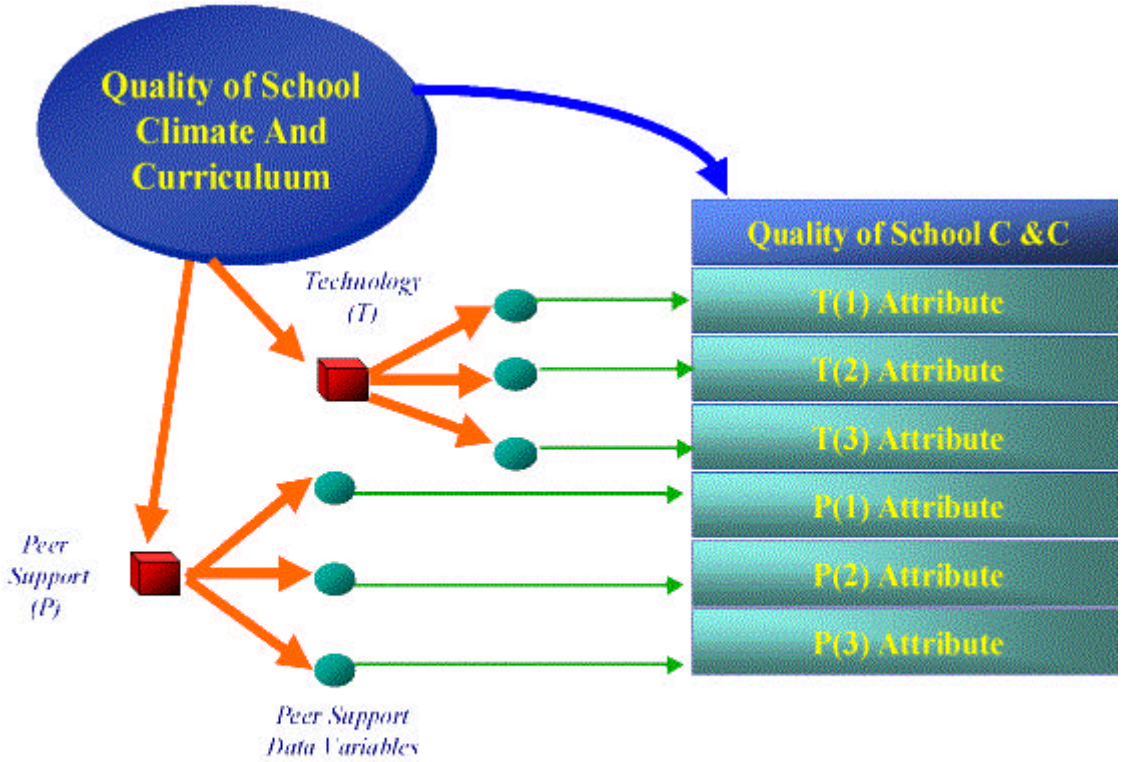

Figure Three -- Mapping a Conceptual Domain to a Table and Data Variables to Attributes

This step provides a foundation in the data mart for later measurement modeling. If the earlier specification of concepts and the data inventory of this step are sufficiently comprehensive, the data mart will have the conceptual and data attribute raw material necessary to provide semantic content to measurement models. While these measurement models will eventually become part of the data mart, they are not part of the dimensional data model. It specifies only the Tables and attributes for the measurement models to operate on.

STEP FOUR: Find the conceptual domains (candidate tables) that contain the candidate attributes measuring the presence, absence or closeness of approach to goals and objectives. Pick these domains as sources of attributes for candidate Fact Tables. Think about the attributes you need to model in this data mart and decide on the meaning of the lowest level record in the data mart containing these attribute(s). Following Kimball, this is called deciding on the grain of the Fact Table.

Now, identify the attribute(s) that will serve as the grain attribute(s) for the Fact Table(s) of the data mart. The grain attributes of the Fact Tables provide data useful in measuring the effects of causal attributes on the goals and objectives of the enterprise. It is this data, the changes in it, and the measures derived from it, that are an important focus of tracking in the data mart.

STEP FIVE: Identify the conceptual domains whose attributes may have a causal impact on the Fact Table grain attribute(s). Map these attributes to relational tables bearing the names of the conceptual domains the attributes are intended to measure or describe. Specify primary keys for

these tables.

STEP SIX: Identify the conceptual domains whose attributes can provide more detailed descriptions of the Fact Table grain attribute(s). Map these attributes to relational tables bearing the names of the conceptual domains they are intended to measure or describe. Specify primary keys for these tables.

Sometimes causal and descriptive attributes will share a conceptual domain and will map to the same relational tables. Sometimes mapping will result in tables that have only descriptive or causal attributes in them apart from keys. Depending on results it may be possible to clearly distinguish certain dimensions as causal, others as descriptive, and others as combinations of both.

Kimball's promotion dimension is distinctly causal. All the attributes in it are attributes measuring exogenous manipulative variables, under control of the enterprise, that can impact sales. His treatment suggests that when a dimension represents enterprise policies, strategies, or tactics, it can always be considered a causal dimension. There are some purely causal dimensions though, that the enterprise may not be able to manupulate. For example, dimensions representing external economic conditions. Or dimensions representing stock market cycles. In addition, though other dimensions such as product or personnel dimensions may not be purely causal, there will be some attributes measuring causal factors in these dimensions.

STEP SEVEN: Identify the conceptual domains that lay out analytic hierarchies for localizing or generalizing the description provided by the Fact Table grain, including description and measurement of change over time. Decide whether to incorporate these domains as separate Hierarchy Tables, or as fields in other previously defined tables. Specify their primary keys.

A time dimension table is almost always included in data marts. But frequently, other hierarchies such as geography, sales organization, product type, etc., are included in the same tables as non-hierarchical factors. How the tables are formulated is guided by performance considerations.

STEP EIGHT: Having specified the keys in all dimensional tables, go back to the candidate Fact Tables and add the primary keys of Causal, Descriptive, Hierarchy, and combined dimensions as foreign keys. Consider, next, whether the Fact Tables need to be supplemented with Factless Event or Coverage Tables. In general, Factless Tables will be needed to measure the impact of hypothesized causes on Fact Table effects, because these tables measure the absence of effects. Create the Factless tables by deciding whether event, coverage, or both types of tables are necessary in your design. Include the same foreign keys in the Factless tables as are in the corresponding Fact Tables.

STEP NINE: Use the hierarchical and combined dimensions of the DDM to define logic for computing aggregative tables from the Fact, Factless, Causal, Descriptive, and combined dimensions.

DDM Benefits

This systems approach (Figure 4) requires highly explicit, top-down conceptualization and data inventory steps in order to sketch out the broadest possible view of the outlines of the range of measurement and cause-effect relations underlying the data mart. The conceptualization and data inventory activities establish the semantic content of the DDM. They provide the foundation necessary for future efforts at KDD to use the concepts and attributes underlying the data mart to successfully conclude a wide range of measurement, causal, structural, time-series, forecasting, and dynamic analyses including explicit modeling, data mining and data validation.

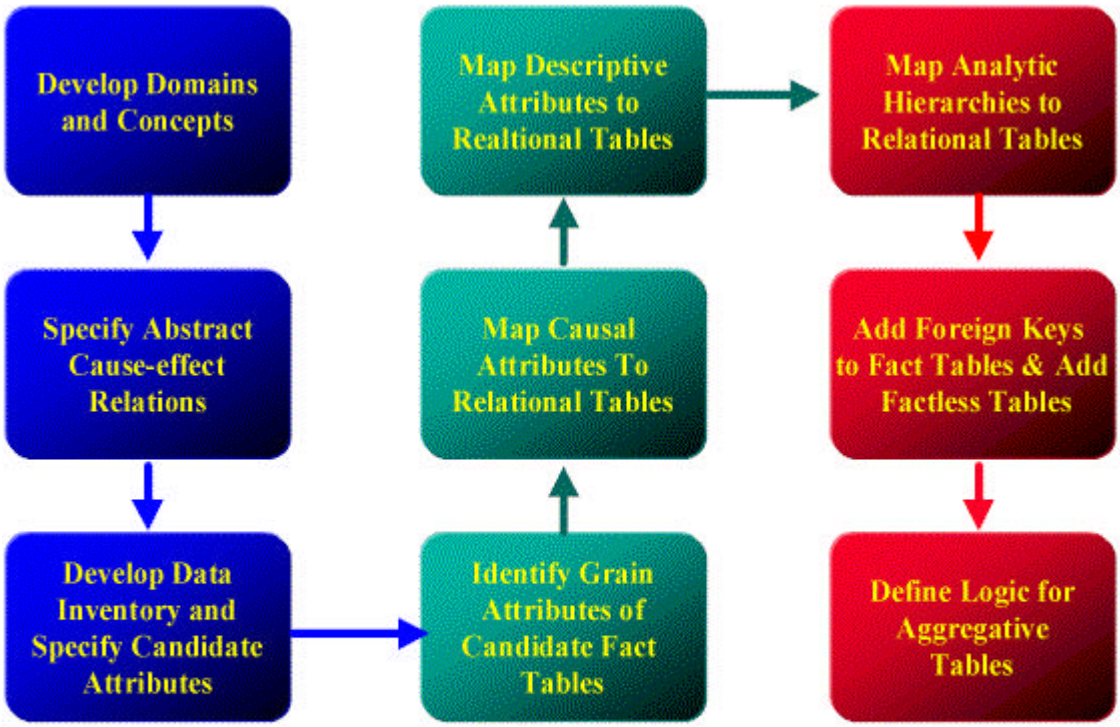

Figure 4 -- The Nine Steps of the Systems Approach to DDM

In following this approach the data modeler also justifies selection of tables and attributes because of their actual and possible connections to the business goals and objectives of the enterprise, and the low level criteria represented by the Fact Table attributes related to these goals and objectives. In addition, the approach follows Kimball's work closely enough to ensure that all the other aspects of OLAP as defined in the FASMI definition are also supported.

References

[1] Nigel Pendse and Richard Creeth, "Synopsis of the OLAP Report," Business Intelligence, Inc., Norwalk, CT, 1997 (available at http://www.busintel.com/synopsis.htm).

[2] Joseph M. Firestone, "Data Mining and KDD: A Shifting Mosaic," White Paper No.2, Executive Information Systems, Inc. Wilmington, DE, March 12, 1997 (available from the author).

[3] Ralph Kimball, The Data Warehouse Toolkit. New York: N. Y. (John Wiley and Sons) 1996.

Biography

Joseph M. Firestone is an independent Information Technology consultant working in the areas of Decision Support (especially Data Marts and Data Mining), Business Process Reengineering and Database Marketing. He formulated and is developing the idea of Market Systems Reengineering (MSR). In addition, he is developing an integrated data mining approach incorporating a fair comparison methodology for evaluating data mining results. You can e-mail Joe at eisai@moon.jic.com.

[Up] [Data Warehouses and Data Marts: New Definitions and New Conceptions]
[Is Data Staging Relational: A Comment]
[DKMA and The Data Warehouse Bus Architecture]
[The Corporate Information Factory or the Corporate Knowledge Factory]
[Architectural Evolution in Data Warehousing]
[Dimensional Modeling and E-R Modeling in the Data Warehouse]
[Dimensional Object Modeling] [Evaluating OLAP Alternatives]
[Data Mining and KDD: A Shifting Mosaic]
[Data Warehouses and Data Marts: A Dynamic View]
[A Systems Approach to Dimensional Modeling in Data Marts]
[Object-Oriented Data Warehousing]