

# Preparing for Business Intelligence

Ross Morrissey – ross@tantiva.com

International Spectrum – Sept/Oct 2008

Sooner or later, every MultiValue shop will need to rework their data for some analytical reporting project. From the full-blown data warehouse to the humble pivot table, successful analytical reporting projects rely on dimensional modeling. This article will introduce dimensional modeling and describe some of the decision making that go into translating MultiValue transactional data into a basic dimensional model. I've refined this design process with IT and business staff from over one hundred MultiValue shops over the past few years; it's straight forward, reduces risk, and can guide both business and IT users to powerful solutions.

## Dimensional Modeling

Dimensional modeling originated with the design of large databases that aggregated grocery store sales for General Mills in the 1960s. It was popularized by Ralph Kimball in his 1996 classic "The Data Warehouse Toolkit" (now in an excellent 2<sup>nd</sup> edition). Dimensional modeling revolves around facts or measurements and dimensions that identify facts. A typical fact is a count or amount – a quantity or dollar figure. On its own, a number doesn't tell the whole story, e.g. 42. Dimensions - like customer, store, product, and time – provide the context for analysis. This fact-dimension relationship can be captured in a classic data warehousing star

schema or a flat file, but the design decisions are the same. A quick example will show the difference between facts and dimensions.

All the examples in this article use the HS.SALES database that ships with UniVerse. Here are the first two customer records:

```
Customer ID....      1
Contact Name... Mr. Samuel Smith
Company Name... Better Beer, Inc.
Street Address.10 Commercial St.
City..... Concord
State..... NH
State name..... New Hampshire
Zip..... 02131
Telephone..... (603)555-3212
```

Product	Product Description.	Serial#	Date Purchased	Date paid	List price	Price..	Discount
M2000	Moderate duty, monochrome copier	501278	01/07/91	01/28/91	\$4,490	\$4,200	6.5

```
Customer ID....      10
Contact Name... Dr. Andrew McCaig
Company Name... HGT Dental Center
Street Address. 999 Hill Road
City..... Brattleboro
State..... VT
State name..... Vermont
Zip..... 03356
Telephone..... (802)555-6534
```

Product	Product Description.	Serial#	Date Purchased	Date paid	List price	Price..	Discount
M1000	Low cost, entry level, light duty, monochrome copier	203510	01/28/91	02/14/91	\$1,990	\$1,990	0.0
M1000	Low cost, entry level, light duty, monochrome copier	203600	01/29/91	02/28/91	\$1,990	\$1,900	4.5
C2000	Moderate duty, entry level, color copier	600791	01/30/91		\$6,890	\$6,500	5.7

In this case a fact record might include the Price. Obvious Dimensions would be customer, product, and time. A simple Retrieve statement pulls this information out:

```

SORT CUSTOMER WITH PRODID NE "" CUSTID COMPANY STATE BY-EXP PRODID NE "" PRODID BUY_DATE PRICE
ID-SUPP
Customer ID      Company Name..... State      Product      Date Purchased      Price..
10      HGT Dental Center      VT      C2000      01/30/91      $6,500
2      Fast Copy Center      MA      C2000      01/08/91      $6,600
4      Fast Copy Center      MA      C3000      01/09/91      $16,500
10      HGT Dental Center      VT      M1000      01/28/91      $1,990
10      HGT Dental Center      VT      M1000      01/29/91      $1,900
5      Ocean State Fish      RI      M1000      01/14/91      $1,900
      Company
5      Ocean State Fish      RI      M1000      02/15/91      $1,900
      Company
1      Better Beer, Inc.      NH      M2000      01/07/91      $4,200
3      Fast Copy Center      MA      M2000      01/08/91      $4,250
7      Central Hospital      NH      M2000      01/20/91      $4,490
2      Fast Copy Center      MA      M3000      01/08/91      $12,000
8      Copies, Inc.      MA      M3000      01/21/91      $12,000
7      Central Hospital      NH      S2000      01/20/91      $990
2      Fast Copy Center      MA      S3000      01/22/91      $900
8      Copies, Inc.      MA      S3000      01/21/91      $900

```

15 records listed.

Notice that some customer names are repeated when there are customers with multiple transactions. This (de)normalization or flattening of the data is tightly bound to the dimensional model and will be covered shortly. This flattened data lends itself to analysis with pivot tables or for inclusion in many business intelligence solutions. Some of these solutions will handle the extraction or flattening, but not decisions about which fields to use as facts and dimensions. The remainder of this article will focus on these decisions.

**The Dimensional Modeling Process**

There are four steps to this process: formulate a vision statement, identify the transaction granularity, select a primary source file, and choose the dimensions and facts. The order of the steps is important – it is common (and risky) for IT staff to jump straight to choosing a source

file because they make that choice instinctively as a way of thinking about the project; while this is frequently correct, the downside of choosing the wrong file can be substantial.

The HS.SALES account will provide examples for the four steps.

### **Formulate a Vision Statement**

Before any project (not just a BI project) it makes sense to set out a basic vision statement outlining the business process being modeled and the shared vision of the scope and capability of the finished project. This helps in two ways: first, without a clear idea of the scope of the project, it is difficult to know when the project is finished. Second, developers don't need to go back to the client for every micro decision if the vision statement is clear enough. If the vision statement includes the phrase "fast and cheap" little time will be spent choosing the colors for the staff's Aeron chairs. If decisions are outside or contrary to the vision statement this is a red flag to rethink or reset expectations - early.

A vision statement should be crafted in consultation with the intended audience. Understand their goals and how this project will support those goals. It is sometimes helpful to collect questions to be answered by this project. A key question can be used to validate the design: "which products receive the highest discounts?" Reach agreement with all interested parties. This vision statement will be used for the rest of the article: "Provide upper management with information on sales and discounts by customer and product." This simple statement includes the audience, the business process, and specific requirements.

## **Choosing the Transaction Granularity**

Transaction Granularity refers to the atomic level of the data. Each line in a flat file is at this level of granularity. Baseball standings have team-level granularity. Line scores have inning granularity. The example vision statement looks for “product sales”. This leads directly to a “line item” granularity – products or SKUs are on invoice lines. Had the vision statement been restricted to “customer profitability” then “invoice” or “customer” granularity would be sufficient. Three factors come into play: a more summarized granularity will create a smaller, less resource intensive solution; a more detailed granularity will provide more flexibility; and matching the granularity of existing models can potentially provide extended capabilities and help validate the accuracy of the solution.

## **Granularity and Multivalues**

If the granularity of the facts is at a multivalued level, all fact fields must be associated multivalues and any multivalued dimension fields must be associated. Non-multivalued fields must be repeated to flatten the data. For example, if five multivalued line items are associated with five sets of SKUs, quantities, and prices, a single shipping charge cannot be included as a fact. A shipping method for the entire order could be included as a dimension field, but it will need to appear five times in the flattened data. Artificial facts like line counts would appear as five associated multivalues, in this case, each with a value of 1.

## Choosing a Source File

With an understanding of the transaction granularity, the primary source file choice is usually straight forward. There will be supporting files, but any SELECT and extraction will be based on this primary source file. An “invoice line item” granularity will typically only yield a few choices for a typical MultiValue system. The history file may contain all invoices for all time – but only contain summary data. An open orders file may have the appropriate granularity, but not the required history. Another file might contain both current and historical transactions, but be highly volatile and difficult to validate.

If there are multiple candidate files, consider choosing one that is already used in the reporting solution or one that has an existing body of reports. This will make it easier to audit the solution.

The ideal is a file that best meets the vision statement while supplying the requisite transaction granularity. If there is no appropriate source file, do not proceed. Revisit the vision statement with the project stakeholders to see if a compromise with existing data is possible, or revise your application to start capturing the required data going forward. There is little point in trying to build an analytical reporting solution without the data to do the job.

The example vision statement leads to “line item” granularity. This level of granularity is actually stored in a sales history embedded in each CUSTOMER record – and CUSTOMER will be the primary source file. Can there be multiple source files? Yes, but usually only one primary source file matches the required granularity. It may be logically or physically distributed, in which case the files need to be merged as part of the data extraction.

## **Choosing the Dimensions and Facts**

Stepping through the fields in the CUSTOMER record, the vision statement will guide the treatment of each field and determine what part it plays in the dimensional model.

### **Customer ID.... 1**

Include in Customer Dimension – this allows upper management to locate transactions for a particular customer. In a star schema, this will be the key to the customer dimension.

### **Contact Name... Mr. Samuel Smith**

Do not include – upper management will not be interested in this level of detail.

### **Company Name... Better Beer, Inc.**

Include in Customer Dimension – this allows upper management to locate transactions for a particular customer.

### **Street Address. 10 Commercial St.**

Do not include – upper management will not be interested in this level of detail.

### **City..... Concord**

Include in Customer Dimension – this allows upper management to look for geographic patterns.

**State..... NH**

Include in Customer Dimension – this allows upper management to look for geographic patterns.

**State name..... New Hampshire**

Include in Customer Dimension – this allows upper management to look for patterns. Do not include this in a separate “State” table. This is what is known as snow-flaking dimensions and leads to exceptionally slow query speeds.

**Zip..... 02131**

Include in Customer Dimension – this allows upper management to look for geographic patterns.

**Telephone..... (603)555-3212**

Do not include – upper management will not be interested in this level of detail.

**Product..... M2000**

Include in Product Dimension – this allows upper management to locate transactions for a particular product. In a star schema, this would be the key to the product dimension.

**Description.... Moderate duty, monochrome copier**

Include in Product Dimension – this allows upper management to look for patterns.

**Serial#..... 501278**

Do not include – upper management will not be interested in this level of detail.

**Date Purchased. 01/07/91**

This is the time dimension. Every transaction-based dimensional model will have a time dimension.

**Date paid..... 01/28/91**

Do not include –the amount of time it took customers to pay has no bearing on the vision statement. The fact that they paid might.

**List price..... \$4,490**

Include as a fact.

**Price..... \$4,200**

Include as a fact. Combined with List price yields discount percentage at any level of aggregation.

**Discount..... 6.5**

Do not include as a fact – Discount percentage is a semi-additive measure. These percentages cannot be added, but can be calculated from the sums of list and price. Individual order

discount might be used to drive a discount cohort dimension where ranges of discounts are grouped for analysis, e.g. 0%, under 5%, under 10%, over 10%.

**Flat File Solution:**

Customer ID.... 1

Contact Name... Mr. Samuel Smith

Company Name... Better Beer, Inc.

City..... Concord

State..... NH

State name..... New Hampshire

Zip..... 02131

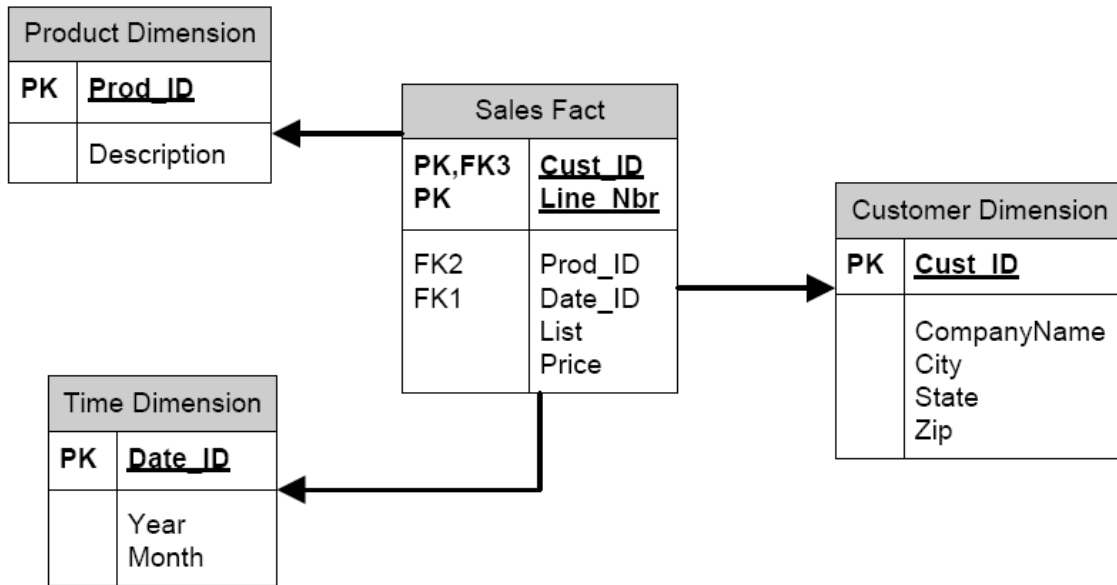
Product..... M2000

Description.... Moderate duty, monochrome copier

Date Purchased. 01/07/91

List price..... \$4,490

Price..... \$4,200



**Star Schema Solution** (figure 1)

The same dimensional model yields a "flat file" representation or a "star schema." Before moving on to extract and transfer the data from MultiValue, this is a good spot to "sanity check" the dimensional model against the vision statement and primary question. This "sign-off on the blueprints" makes an excellent milestone - and can potentially save a tremendous amount of effort. In this case the goal was to calculate discounts by product. With this model this is possible at any level of detail - as long as our reporting tool allows us to do calculations on intermediate totals in columns.

There are other "non-transactional" dimensional models, and additional complexities caused by stretching a transactional model that captures the truth at an instant to fit a reporting model that accurately shows a flow over time. These will be covered in future articles.