Artigo Clássico:

Data Mart Does Not Equal Data Warehouse (Bill Inmon)

The single most important issue facing the information technology manager this year is whether to build the data warehouse first or the data mart first.

The data mart vendors have said that data warehouses are difficult and expensive to build, take a long time to design and develop, require thought and investment, and mandate that the corporation face difficult issues such as integration of legacy data, managing massive volumes of data and cost justifying the entire DSS/data warehouse effort to the management committee. The picture painted by the data mart advocates for building the data warehouse is gloomy. It is also self-serving and incorrect.

The data mart vendors look upon the data warehouse as an obstacle between themselves and the revenue that comes from making sales. Of course, they want to shun the data warehouse. The data warehouse lengthens their sales cycle, regardless of the long-term effect of building a bunch of data marts with no data warehouse. The data mart vendors are selling a very short-term perspective at the expense of long-term architectural success.

The data mart advocates suggest that there may be alternate, much easier paths to DSS success than building a data warehouse. One of those paths is to build several data marts and when they grow big enough, call them a data warehouse, rather than build an actual data warehouse. The data mart advocates argue that the data mart can be built much more quickly and cheaply than a warehouse. When you build the data mart there is no need for a great amount of organizational hassle or discipline and no concern for the long-term architecture that is created by the data marts.

Unfortunately, by avoiding the visceral organizational and design issues of warehousing, the data mart advocates miss much of the point of warehousing. By building an architecture consisting entirely of data marts, the data mart advocates lead the organization into an even larger mess. Instead of messy legacy operational systems, now we have messy legacy operational systems AND messy data marts. Stovepipe data marts and stovepipe DSS applications are what result from building nothing but data marts. There is no integration when all that you build is data marts. And a DSS environment without integration is like a man without a skeletal system--hardly a useful, viable entity.

A Change of Approaches

In the early days of the data warehouse marketplace, the data mart vendors tried to jump on the warehouse gravy train by proclaiming that a data warehouse was the same thing as a data mart. In trade show after trade show, the data mart vendors confused people with what a data warehouse is and what a data mart is. The data mart vendors spread half truths and misinformation about data warehousing. The result was confusion.

The obfuscation sowed by the data mart vendors caused a few confused customers to build data marts with no actual warehouse. After about the third data mart, the customer discovered something was rotten in
Denmark. The architectural deficiency of building nothing but data marts was unmasked. The customer discovered that when you don't build a data warehouse, there is:

- massive redundancy of detailed and historical data from one data mart to another,
- inconsistent and irreconcilable results from one data mart to the next,
- an unmanageable interface between the data marts and the legacy application environment, etc.

In short order, the world discovered that a DSS environment without a data warehouse was an extremely unsatisfactory thing.

Now that the world has found that building data marts is not the proper way to proceed in DSS, the data mart vendors and their spokesmen are back again and are sowing a different brand of confusion. This time they have altered their original words a little and have promised a new and improved path to easy success. In a slight twist of concept from the first time around, the notion now being spread is that a data warehouse is merely a collection of integrated data marts (whatever that is). The notion that multiple data marts can be integrated is oxymoronic. The whole essence of data marts is that mart users do their own thing so that they don't have to integrate with other marts.

Simply stated, for a variety of very powerful reasons, you cannot build data marts, watch them grow and magically turn them a data warehouse when they reach a certain size. And by the same token, integrating data across data marts is equally unthinkable because each department that owns its own data mart has its own unique specifications.

In order to understand why one or more data marts cannot be transformed into a data warehouse, you must first understand what a data mart is and a data warehouse is.

**Different Architectural Structures**

A data mart and a data warehouse are essentially different architectural structures, even though when viewed from afar and superficially, they look to be very similar.

**What is a Data Mart?**

A data mart is a collection of subject areas organized for decision support based on the needs of a given department. Finance has their data mart, marketing has theirs, sales has theirs and so on. And the data mart for marketing only faintly resembles anyone else's data mart.

Perhaps most importantly, the individual departments OWN the hardware, software, data and programs that constitute the data mart. The rights of ownership allow the departments to bypass any means of control or discipline that might coordinate the data found in the different departments.

Each department has its own interpretation of what a data mart should look like and each department's data mart is peculiar to and specific to its own needs. Typically, the database design for a data mart is built around a star-join structure that is optimal for the needs of the users found in the department. In order to shape the star join, the requirements of the users for the department must be gathered. The data mart
contains only a modicum of historical information and is granular only to the point that it suits the needs of the department. The data mart is typically housed in multidimensional technology which is great for flexibility of analysis but is not optimal for large amounts of data. Data found in data marts is highly indexed.

There are two kinds of data marts--dependent and independent. A dependent data mart is one whose source is a data warehouse. An independent data mart is one whose source is the legacy applications environment. All dependent data marts are fed by the same source--the data warehouse. Each independent data mart is fed uniquely and separately by the legacy applications environment. Dependent data marts are architecturally and structurally sound. Independent data marts are unstable and architecturally unsound, at least for the long haul. The problem with independent data marts is that their deficiencies do not make themselves manifest until the organization has built multiple independent data marts.

**What is a Data Warehouse?**

Data warehouses are significantly different from data marts. Data warehouses are arranged around the corporate subject areas found in the corporate data model. Usually the data warehouse is built and owned by centrally coordinated organizations, such as the classic IT organization. The data warehouse represents a truly corporate effort.

There may or may not be a relationship between any department's subject areas and the corporation's subject areas. The data warehouse contains the most granular data the corporation has. Data mart data is usually much less granular than data warehouse data (i.e., data warehouses contain more detail information while most data marts contain more summarized or aggregated data). The data warehouse data structure is an essentially normalized structure. The structure and the content of the data in the data warehouse do not reflect the bias of any particular department, but represent the corporation's needs for data. The volume of data found in the data warehouse is significantly different from the data found in the data mart. Because of the volume of data found in the data warehouse, the data warehouse is indexed very lightly. The data warehouse contains a robust amount of historical data. The technology housing the data warehouse is optimized on handling an industrial strength amount of data. The data warehouse data is integrated from the many legacy sources.

In short, there are very significant differences between the structure and content of data that resides in a data warehouse and the structure and content of data that resides in a data mart.
Figure 1: shows some of the differences between a data mart and a data warehouse.

Because data that is granular, integrated and historical resides in a data warehouse, the data warehouse attracts a significant volume of data. Because the warehouse attracts a significant amount of data, it is advisable to build the warehouse iteratively. If you don't build the warehouse iteratively, you will spend years building the warehouse. From the very first literature that was ever written on data warehousing, it has been recognized that there was a need to get concrete, tangible results in front of the end user as quickly as possible. The best advice of the writers and consultants of the data warehousing industry has consistently been to build the warehouse quickly and to avoid large, lengthy efforts.

Interestingly, the data mart advocates and their spokesmen claim that data warehouses take a long time to build. It is only in the hype issued by the data mart advocates that it has ever been suggested that the warehouse be built in "galactic" proportions.
Figure 2: shows the recommended construction path for data warehouses.

The most recent theory of the data mart advocates is that you can build one or more data marts, integrate them (although no one is very clear as to what that means) and then when they grow to a certain size, they can be (magically!) turned into a warehouse. This suggestion is sadly mistaken for a variety of reasons:

* The data mart is designed to suit the needs of a department. Many departments with very different objectives must be satisfied. That is why there are many different data marts in the corporation, each with its own distinctive look and feel. The data warehouse is designed to suit the collective needs of the entire corporation. A given design can be optimal for a single department or the corporation, but not both. The design objectives for the corporation are very different from the design objectives for a given department.

* The granularity of data in the data mart is very different from the granularity of data in the data warehouse. The data mart contains aggregated or summarized data. The data warehouse contains the most detailed data that is found in the corporation. Since the data mart granularity is much higher than that found in the data warehouse, you cannot easily decompose the data mart granularity into data warehouse granularity. But you can always go the other direction and summarize detailed units of data into summarizations.

  • The structure of the data in the data mart (commonly a star join structure) is only faintly compatible with the structure of the data in the warehouse (a normalized structure).
  • The amount of historical data found in the data mart is very different from the history of the data found in the warehouse. Data warehouses contain robust amounts of history. Data marts contain only modest amounts of history.
  • The subject areas found in the data mart are only faintly related to the subject areas found in the data warehouse.
  • The relationships found in the data mart are not those relationships that are found in the data warehouse.
• The types of queries satisfied in the data mart are quite different from those queries found in the data warehouse.

• The kind of users (farmers) that are found in the marts are quite different from the type of users (explorers) that are found in the data warehouse.

• The key structures found in the data mart are significantly different from those key structures found in the data warehouse, and so forth.

Reality

There are simply MAJOR, MAJOR significant differences between the data mart and the data warehouse environment. The assertion that a data mart can be turned into a data warehouse when it reaches a certain size or that data marts can be integrated together is no more valid than saying that when a tumbleweed grows large enough that it can be turned into an oak tree. Reality and genetics being what they are, it is true that a tumbleweed and an oak tree are, at one point in their life, green living organisms planted in the soil and are approximately the same size. But just because those two plants share a few basic characteristics at some moment in time does not mean that a tumbleweed can be turned into an oak tree. Only a misinformed person would mistake a tumbleweed for an oak tree at any stage in the life of the plants.