Data Driven vs. Metric Driven Data Warehouse Design

John M. Artz
The George Washington University, USA

INTRODUCTION

Although data warehousing theory and technology have been around for well over a decade, they may well be the next hot technologies. How can it be that a technology sleeps for so long and then begins to move rapidly to the foreground? This question can have several answers. Perhaps the technology had not yet caught up to the theory or that computer technology 10 years ago did not have the capacity to delivery what the theory promised. Perhaps the ideas and the products were just ahead of their time. All these answers are true to some extent. But the real answer, I believe, is that data warehousing is in the process of undergoing a radical theoretical and paradigmatic shift, and that shift will reposition data warehousing to meet future demands.

BACKGROUND

Just recently I started teaching a new course in data warehousing. I have only taught it a few times so far, but I have already noticed that there are two distinct and largely incompatible views of the nature of a data warehouse. A prospective student, who had several years of industry experience in data warehousing but little theoretical insight, came by my office one day to find out more about the course. "Are you an Inmonite or a Kimballite?" she inquired, reducing the possibilities to the core issues. "Well, I suppose if you put it that way," I replied, "I would have to classify myself as a Kimballite." William Inmon (2000, 2002) and Ralph Kimball (1996, 1998, 2000) are the two most widely recognized authors in data warehousing and represent two competing positions on the nature of a data warehouse.

The issue that this student was trying to get at was whether or not I viewed the dimensional data model as the core concept in data warehousing. I do, of course, but there is, I believe, a lot more to the emerging competition between these alternative views of data warehouse design. One of these views, which I call the data-driven view of data warehouse design, begins with existing organizational data. These data have more than likely been produced by existing transaction processing systems. They are cleansed and summarized and are used to gain greater insight into the functioning of the organization. The analysis that can be done is a function of the data that were collected in the transaction processing systems. This was, perhaps, the original view of data warehousing and, as will be shown, much of the current research in data warehousing assumes this view.

The competing view, which I call the metric-driven view of data warehouse design, begins by identifying key business processes that need to be measured and tracked over time in order for the organization to function more efficiently. A dimensional model is designed to facilitate that measurement over time, and data are collected to populate that dimensional model. If existing organizational data can be used to populate that dimensional model, so much the better. But if not, the data need to be acquired somehow. The metric-driven view of data warehouse design, as will be shown, is superior both theoretically and philosophically. In addition, it dramatically changes the research program in data warehousing. The metric-driven and data-driven approaches to data warehouse design have also been referred to, respectively, as metric pull versus data push (Artz, 2003).

MAIN THRUST

Data-Driven Design

The classic view of data warehousing sees the data warehouse as an extension of decision support systems. Again, in a classic view, decision support systems sit atop management information systems and use data extracted from management information and transaction processing systems to support decisions within

the organization. This view can be thought of as a data-driven view of data warehousing, because the exploitations that can be done in the data warehouse are driven by the data made available in the underlying operational information systems.

This data-driven model has several advantages. First, it is much more concrete. The data in the data warehouse are defined as an extension of existing data. Second, it is evolutionary. The data warehouse can be populated and exploited as new uses are found for existing data. Finally, there is no question that summary data can be derived, because the summaries are based upon existing data. However, it is not without flaws. First, the integration of multiple data sources may be difficult. These operational data sources may have been developed independently, and the semantics may not agree. It is difficult to resolve these conflicting semantics without a known end state to aim for. But the more damaging problem is epistemological. The summary data derived from the operational systems represent something, but the exact nature of that something may not be clear. Consequently, the meaning of the information that describes that something may also be unclear. This is related to the semantic disintegrity problem in relational databases. A user asks a question of the database and gets an answer, but it is not the answer to the question that the user asked. When the somethings that are represented in the database are not fully understood, then answers derived from the data warehouse are likely to be applied incorrectly to known somethings. Unfortunately, this also undermines data mining. Data mining helps people find hidden relationships in the data. But if the data do not represent something of interest in the world, then those relationships do not represent anything interesting, either.

Research problems in data warehousing currently reflect this data-driven view. Current research in data warehousing focuses on a) data extraction and integration, b) data aggregation and production of summary sets, c) query optimization, and d) update propagation (Jarke, Lenzerini, Vassiliou, & Vassiliadis, 2000). All these issues address the production of summary data based on operational data stores.

A Poverty of Epistemology

The primary flaw in data-driven data warehouse design is that it is based on an impoverish epistemology. *Epistemology* is that branch of philosophy concerned with theories of knowledge and the criteria for valid knowledge (Fetzer & Almeder, 1993; Palmer, 2001). That is to say, when you derive information from a data warehouse based on the data-driven approach, what does that information mean? How does it relate to the work of the organization? To see this issue, consider the following example. If I asked each student in a class of 30 for their ages, then summed those ages and divided by 30, I should have the average age of the class, assuming that everyone reported their age accurately. If I were to generate a list of 30 random numbers between 20 and 40 and took the average, that average would be the average of the numbers in that data set and would have nothing to do with the average age of the class. In between those two extremes are any number of options. I could guess the ages of students based on their looks. I could ask members of the class to guess the ages of other members. I could rank the students by age and then use the ranking number instead of age. The point is that each of these attempts is somewhere between the two extremes, and the validity of my data improves as I move closer to the first extreme. That is, I have measurements of a specific phenomenon, and those measurements are likely to represent that phenomenon faithfully. The epistemological problem in data-driven data warehouse design is that data is collected for one purpose and then used for another purpose. The strongest validity claim that can be made is that any information derived from this data is true about the data set, but its connection to the organization is tenuous. This not only creates problems with the data warehouse, but all subsequent data-mining discoveries are suspect also.

METRIC-DRIVEN DESIGN

The metric-driven approach to data warehouse design begins by defining key business processes that need to be measured and tracked in order to maintain or improve the efficiency and productivity of the organization. After these key business processes are defined, they are modeled in a dimensional data model and then further analysis is done to determine how the dimensional model will be populated. Hopefully, much of the data can be derived from operational data stores, but the metrics are the driver, not the availability of data from operational data stores.