IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4719

Data Push versus Metric Pull: Competing Paradigms for Data Warehouse Design And Their Implications

John M. Artz
Department of Management Science
The George Washington University
Washington, DC 20052
(202) 994-4931
jartz@gwu.edu

INTRODUCTION

Although data warehousing theory and technology have been around for well over ten years, it may well be the next really hot technology. How can it be that a technology sleeps for a decade and then begins to move rapidly to the foreground? There can be several answers to this question. It could be that the technology had not yet caught up to the theory. It could be that computer technology ten years ago did not have the capacity to delivery what the theory promised. Or it could be that the ideas and the products were just ahead of their time. All of 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 shift and that paradigmatic shift will reposition data warehousing to meet demands of the future.

This past summer I taught a course in data warehousing. Since it is a new course, and I only get to try out new courses in the summer, I have only been able to teach this course three times so far. Nonetheless, 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 I would be teaching. "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."

This issue, I believe, that she 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 will call the "data push" view of data warehouse design, begins with existing organizational data. This data has more than likely been produced by existing transaction processing systems. It is cleansed and summarized and is used to gain greater insight into the functioning of the organization. The analysis that can be done is a function of the data that was 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 will call the "metric pull" 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 is 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 needs to be acquired somehow. The metric push 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 DATA PUSH VIEW OF DATA WAREHOUSE 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 on top of" 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 or data push 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.

There are several advantages to this data driven model. First, it is much more concrete. The data in the data warehouse is 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, since 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 represents 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'. This also, unfortunately, undermines data mining. Data mining helps us find hidden relationships in the data. But if the data does 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 push view. Current research in data warehousing focuses on: 1) data

extraction and integration; 2) data aggregation and production of summary sets; 3) query optimization; and 4) update propagation. All of this issues address the production of summary data based on operational data stores.

A POVERTY OF EPISTEMOLOGY

The primary flaw in the data push model is that it is based on an impoverish epistemology. That is to say, when you derive information from a data warehouse based on the data push model, 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 thirty for their age, then summed those ages and divided by thirty, I should have the average age of the class assuming that everyone reported their age accurately. If I were to generate a list of thirty random numbers between twenty and forty and take the average, that average would be the average of the numbers in that data set and would have nothing to with the average age of the class. In between those two extremes there 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 measurement of a specific phenomenon and those measurements are likely to represent that phenomenon faithfully. The epistemological problem in data push data warehousing 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.

THE METRIC PULL VIEW OF DATA WAREHOUSE DESIGN

The metric pull view of data warehousing 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. Once these key business processes are defined, they are modeled in a dimensional data model and the 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 it is the metrics that are the driver, not the availability of data from operational data stores.

A relational database models the entities or things of interest to an organization. These things of interest may include customers, products, employees and the like. The entity model represents these things and the relationships between them. As occurrences of these entities enter or leave the organization, that addition or deletion is reflected in the database. As these entities change in state, somehow, those state changes are also reflected in the database. So, theoretically, at any point in time, the database faithfully represents the state of the organization. Queries can be submitted to the database and the answers to those queries should, indeed, be the answers to those questions if they were asked and answered with respect to the organization.

A data warehouse, on the other hand, models the business processes in an organization to measure and track those processes over time. Processes may include sales, productivity, the effectiveness of promotions and the like. The dimensional model contains facts that represent measurements over time of a key business process. It also contains dimensions that are attributes of these facts. The fact table can be thought of as the dependent variable in a statistical model and the dimensions can be thought of as the independent variables. So the data warehouse becomes a longitudinal dataset tracking key business processes.

A PARALLEL WITH PRE RELATIONAL DAYS

We can see certain parallels between the state of data warehousing and the state of database prior to the relational model. The relational model was introduced in 1970 but was not realized in a commercial product until the early 1980's. At that time there were a large number of non-relational database management systems. All of these products handle data in different ways because they were software products developed to handle the problem of storing and retrieving data. They were not developed as implementations of a theoretical model of data. When the first relational product came out, the world of databases changed, almost overnight. Every non-relational product attempted, unsuccessfully, to claim that is was really a relational product. But the claims were not believed and the non-relational products lost their market share almost immediately.

Surprisingly, the relational model held its ground over the next couple of decades and is probably one of the few, if not the only, body of knowledge, in information systems, that has not changed much during that time. Certainly, relational database theory has advanced over the past two decades and the products have become more sophisticated. But the basic tenets of relational database theory have not changed, nor are they likely to. They are a theory of data and the nature of data is unlikely to change.

Similarly, there are a wide variety of data warehousing products on the market today. Some are based on the dimensional model and some are not. The dimensional model provides a basis for an underlying theory of data which tracks processes over time rather than the current state of entities. Admittedly, this model of data needs quite a bit of work, but the relational model did not come into dominance until it was coupled with entity theory, so the parallel still holds. We may never have an announcement, in data warehousing, as dramatic as Codd's paper in relational theory. It is more likely that a theory of temporal dimensional data will accumulate over time. However, in order for data warehousing to become a major force in the world of databases an underlying theory of data is needed and it will eventually be developed.

THE IMPLICATIONS FOR RESEARCH

The implications for research in data warehousing are rather profound. Current research focuses on issues such as data extraction and integration, data aggregation and summary sets, query optimization and update propagation. All of these problems are applied problems in software development and do not advance our understanding of the theory of data.

But a metric driven approach to data warehouse design introduces some problems, whose resolution can make a lasting contribution to the theory of data. Research problems in metric pull data warehousing include: 1) how do we identify key business processes; 2) how do we construct appropriate measures for these processes; 3) how do we know those measures are valid; 4) how do we know that a dimensional model has accurately captured the independent variables; 5) can we develop an abstract theory of aggregation so that the data aggregation problem can be understood and advanced theoretically; and, finally, 6) can we develop an abstract data language so that aggregations can be expressed mathematically by the user and realized by the machine?

PUSHING METRIC PULL

Initially, data push and metric pull appear to be legitimate competing paradigms for data warehousing. The epistemological flaw is a little difficult to grasp and the distinction - that information derived from a data push model is information about the dataset while information derived from a metric pull model is information about the organization – may also be a bit elusive. However, the implications are enormous. The data push model has little future, in that it is founded on a model of data exploitation rather than a model of data. The metric pull model, on the other hand, is likely have some major impacts and implications. A few of those are provided here.

THE IMPACT ON WHITE COLLAR WORK

The data push view of data warehousing limits the future of data warehousing to the possibilities inherent in summarizing large collections of old data without a specific purpose in mind. The metric pull view of data warehousing opens up vast new possibilities for improving

the efficiency and productivity of an organization by tracking the performance of key business processes. The introduction of quality management procedures in manufacturing a few decades ago dramatically improved the efficiency and productivity of manufacturing processes, but such improvements have not occurred in white-collar work.

The reason that we have not seen such an improvement in white-collar work is that we have not had metrics to track the productivity of white-collar workers. And even if we did have the metrics we did not have a reasonable way to collect them and track them over time. The identification of measurable key business processes and the modeling of those processes in a data warehouse provides the opportunity to perform quality management and process improvement on white-collar work.

Subjecting white-collar work to the same rigorous definition as blue-collar work may seem daunting, and indeed that level of definition and specification will not come easily. So what would motivate a business to do this? The answer is simple - businesses will have to do this when the competitors in their industry do it. Whoever does this first will achieve such productivity gains that competitors will have to follow suit in order to compete. In the early 1970's corporations were not revamping their internal procedures because computerized accounting systems were fun. They were revamping their internal procedures because they could not protect themselves from their competitors with out the information for decision making and organizational control provided by their accounting information systems. A similar phenomenon is likely to drive data warehousing.

DIMENSIONAL ALGEBRAS

The relational model introduced Structured Query Language, an entirely new data language that allowed non-technical people to access data in a database. SQL also provided a means of thinking about record selection and limited aggregation.

Research in data warehousing will likely yield some sort of a dimensional algebra that will provide, at the same time, a mathematical means of describing data aggregation and correlation, and a set of concepts for thinking about aggregation and correlation. To see how this could happen, think about how the relational model led us to think about the organization as a collection of entity types, or how statistical software made the concepts of correlation and regression much more concrete.

A UNIFIED THEORY OF DATA

In the organization today, the database administrator and the statistician seem worlds apart. Of course, the statistician may have to extract some data from a relational database in order to do his or her analysis. And the statistician may engage in limited data modeling in designing a data set for analysis using a statistical tool. The database administrator on the other hand will spend most of his or her time in designing, populating and maintain a database. A limited amount of time may be devoted to statistical thinking when counts, sums or averages are derived from the database. But largely these two individuals will view themselves as participating in greatly differing disciplines.

But with dimensional modeling the gap between database theory and statistics begins to close. In dimensional modeling we have to begin thinking in terms of construct validity and temporal data. We need to think about correlations between dependent and independent variables. We begin to realize that the choice of data types (e.g. interval or ratio) will affect the types of analysis we can do on the data and hence potentially limit the queries. So the database designer has to address concerns that have traditionally been the domain of the statistician. Similarly, the statistician cannot afford the luxury of constructing a data set for a single purpose or a single type of analysis. The data set must be rich enough to allow the statistician to find relationships that may not have been considered when the data set was being constructed. Variables must be included that may potentially have impact, or may have impact at some times but not others, or may have impact in conjunction with other variables. So the statistician has to address concerns that have traditionally been the domain of the database designer.

What this points to is the fact that database design and statistical exploitation are just different ends of the same problem. Once these two ends have been connected by data warehouse technology, a single theory of data must be developed to address the entire problem. This unified theory of data would include entity theory and measurement theory at one end, and statistical exploitation at the other. The middle ground of this theory will show how decisions made in database design will affect the potential exploitations so that intelligent design decisions can be made that will allow full exploitation of the data to serve the organization's needs to model itself in data.

CONCLUSIONS

Data warehousing is undergoing a theoretical shift from a data push model to a metric pull model. The metric pull model rests on a much firmer epistemological foundation and promises a much richer and more productive future for data warehousing.

REFERENCES

Inmon, W et al. (2000) **Exploration Warehousing : Turning Business Information into Business Opportunity**. John Wiley & Sons.

Inmon, W. (2002) **Building the Data Warehouse**. John Wiley & Sons.

Jarke, M et al. (2000) **Fundamentals of Data Warehouses.** Springer Verlag.

Kimball, R. (1996) **The Data Warehouse Toolkit.** John Wiley & Sons.

Kimball, R et al. (1998) **The Data Warehouse Lifecycle Toolkit.** John Wiley & Sons.

Thomsen, E. (2002) **OLAP Solutions.** John Wiley & Sons.