



## **The Data Modeling Problem**

### **EXECUTIVE SUMMARY**

The enterprise data warehouse is a repository of the enterprise's data that is extracted from the operational databases. This repository is organized in a manner that facilitates querying, reporting and analysis. Since the data warehouse's primary function is to support end-user analysis and facilitate decision making, the method in which the data warehouse is organized becomes critical to its success. Two methods of modeling the data warehouse are explored.

The first method is the entity relationship model that was developed to design high-performance relational transaction processing systems (TPS). While the E-R model is a sound approach to TPS, it creates a complex layout of data in to a multitude of tables across to represent the enterprise. The E-R model is oriented around the data and its relationships to other data. This creates an unacceptable level of complexity for the end-user that focuses on business processes.

The second approach examined is the dimensional model. Dimensional modeling is a method built around a business process and it's associated data elements. The dimensional model presents the end-user with an easy to conceptualize view of the data warehouse but has technical issues when the size of the data warehouse is very large.

A review of the literature reveals ways of trying to work around the issues of each model. Of the different methods discussed, the most intriguing was the concept of hybrid models. Hybrid models are an effort to combine two or more models in to a "hybrid model" that maximizes the strengths of other models to compensate for the weakness in the primary model.

The analysis of the information presented in the literature summarizes that using hybrid models can be a way to solve the issues associated with a pure modeling concept. The use of hybrid models may limit the organization to a smaller set of available tools and limiting the robustness of the overall data warehouse.

This study concludes recommending a combination of both models not as a hybrid but as a process for developing a data warehouse. The dimensional model is recommended for the front-end of the data warehouse providing ease of use and access to the end-user. The E-R model is recommended as a back-end process to support the extract of date from the production system and the loading of the data warehouse.

### **INTRODUCTION**

The continued drive for businesses to gain a competitive advantage and the rapid growth in information technology has spurred the expansion of data warehousing as a key technological approach to business analysis and decision support. Data warehousing is being promoted as a new and improved decision support system (DSS, MIS and EIS) that can provide organizations a method to leverage their information resources to operate more efficiently and effectively. The concept of data warehousing is not new, but only recently have the tools, techniques, methodologies, processor capacity and disk storage advanced enough to make it possible to develop a highly effective data warehouse.

Data warehousing is the process of creating a repository of enterprise wide data that is off-loaded from production databases and is stored in a separate database that can be queried, reported and analyzed [14]. The goal of the data warehouse is to create an integrated view of the enterprise's data from all activities. The data warehouse brings together a wide variety of internal and external data that stored in a manner that allows easy access by the end-user and integrates with analytical software to support end-user analysis and decision making.

Data marts are a scaled down version of the enterprise wide data warehouse. Data marts are typically subject or department oriented and have a more narrow scope than the data warehouse. Though the scope of a data mart may be limited, its scale (size) is only limited by the technology behind it. Data marts can be integrated across the enterprise to create a data warehouse.

Though data warehouses and data marts are distinctly different from each other their function, they do share the same technologies and modeling techniques. For the purposes of this discussion, data

warehouse can be interchanged with the data mart as the logical data model for data warehousing is explored.

## **THE DATA MODELING PROBLEM**

The fundamental issue in data warehousing is database design. The database design will determine the capabilities of the data warehouse and in most cases, the software that will be needed to run it. The key to the database design issue is which data modeling technique to use. The two predominate models are the E-R model and the dimensional model.

### ***Database Design***

The development of a data warehouse can be a daunting task at best. The complexity of the potential queries the end-user may try to execute against the data warehouse requires a flexible design that has the ability to expand and change with the demands of the user.

Database design is typically divided into a four-stage process. The four stages are requirements collection, conceptual design, logical design and physical design [10]. The requirement collection stage gathers the end-users information needs and combines them into a preliminary specification of requirements for the organization. The conceptual model is then developed incorporating the data gathered from the requirements phase. The logical modeling of the database is the key focal point of the database design process since it translates the conceptual model into a specific data model used in the physical design of the database. The physical design is the stage where the logical model is mapped into a specific database management system.

Of the four processes, the logical design is the most critical to the design of a database that will be used for data warehousing. The models used in this phase will determine the database management system and consequently the subset of tools that will be available to the end-user for query and analysis.

### ***Entity-Relationship Model***

The majority of modern enterprise information systems are built using the entity relationship model (E-R model). The growth of relational database management systems in the last decade has made the E-R model one of the best known and frequently used models in database design [10]. The E-R model focuses on removing redundancy of data elements in the database. This model was utilized in relational database systems to increase the speed and accuracy of the relational databases in transaction processing systems. The E-R model is used to demonstrate the detailed relationships between the data elements. Once the relationships were determined, all data redundancy is removed. This process is known as normalization. Once normalization is taken to its highest level, the transaction processing system can perform at high levels of speed and accuracy since the transactions are reduced to very simple deterministic processes [4]. The normalization process also drives a multitude of simple tables in the database that are related to each other through keys that facilitate the relationships between the table and the referential integrity of the database.

The E-R model presents some issues for data warehousing such as query complexity and performance. Since the end-user needs to query the data warehouse, he/she must understand the model of the database in order to understand the results of the queries. This presents an interesting situation since the E-R Model for an enterprise may have thousands of tables for a single functional process within the business [4, 9, 10]. Even if the end-user is able to conquer the mental task of understanding the E-R model, the user must mentally still map the E-R model to the business process(es) being analyzed. The fundamental purpose of the data warehouse is to support decision making. The analytical and decision making process uses a combination of structured and unstructured data. The E-R Model strives to completely structure data. This contrast in purpose and need makes the pure E-R model ill-suited for higher level systems such as MIS and EIS without major advances in end-user software.

### ***Dimensional Model***

Dimensional modeling is technique used to conceptualize business processes as a standard set of measures that describe the ordinary facets of the business [9]. The dimensional model seeks to provide the user with an intuitive framework to operate within and also provide for high performance access to the data within the data warehouse [4].

The Dimensional model adheres to a discipline that incorporates the relational model with restrictions. The dimensional model is composed of a table called a fact table that has multi-part keys and a set of smaller tables that are called dimension tables. The dimension tables have a single-part primary key that relates to only one of the components of the multi-part keys in the fact table. This structure is known as the "star join" and dates back to the earliest days of relational databases [4, 13].

Dimensional models are oriented toward a specific business process or "subject". This approach keeps the

model instinctive for the end-user and keeps the data elements associated with the business process only one join away. The E-R Model modeled the whole enterprise while the dimensional approach models one business process per model. The dimensional model is then extended over the enterprise process by process. The dimensional model is illustrated in Figure 1 [4]. The fact tables can share the dimension tables as long as the primary key rule is not violated. The enterprise is represented in the data warehouse as groups of fact tables models connected by their related dimension tables. The enterprise model that brings the individual dimensional models together is called a Multi-dimensional Model [9].

Source: Kimball, Ralph. "A Dimensional Modeling Manifesto", DBMS. 10(9). 1997 Aug.

IT professionals are sometimes challenged with the dimensional model since they are trained to think in terms of modeling the data and programming the business processes into the system. The key to understanding the dimensional model is to understand the E-R model can be broken into multiple dimensional models. To convert from an E-R model diagram to a dimensional model diagram, separates each business process out of the E-R Model. Once the business processes are separated the many-to-many tables in the E-R model and convert them to dimensional model fact tables. Then final step is then to take the remaining tables in the E-R model and "de-normalize" them into dimension tables for the dimensional model [4, 5]. This process of "de-normalization" is troubling to IT professionals that have only used the E-R model and believe that this makes the dimensional model violate the relational model develop in the 1980's. The "de-normalization" does not violate the relational model but in fact represents some of the earliest relational models that were not adequate for high-speed transaction processing prior to the embracement of the E-R model.

The dimensional model is simple and easy to understand by end-users since they think in terms of the business processes naturally. This level of understanding makes querying the dimensional model by the end-user very intuitive. The simplicity of the tables also promotes very high performance by the database management system.

### **LITERATURE REVIEW & DISCUSSION**

The literature that discusses the E-R model and the dimensional model in data warehousing demonstrates the variety of views taken both in industry and research.

Ralph Kimball, founder and former CEO of Red Brick Systems, is a staunch supporter of the dimensional model. In his dimensional modeling manifesto, he goes as far to say that the E-R model should be avoided for end-user delivery of data warehouse information [4]. Kimball does suggest that the E-R model is the proper design to use for transaction processing and even the data extract and load process for data warehousing [5]. Kimball's work is cited and has been synthesized into the primary method for developing data warehousing models [3, 9, 12, 13, and 14]. Though Kimball's methods have established a wide following, he does not claim to be the inventor of the dimensional model. In fact, Kimball freely admits the initial development of the dimensional model was done by IT professionals that were trying to simplify E-R modeled databases for use by end-users.

Francett [2] dimensional modeling has a difficult downside to it in the area of integration of other products and technologies. Francett's also states that when the dimensional model is also the physical model of the data, the file structures become very proprietary. The formation of the OLAP council by several database vendors is in response to this issue.

Raden [8], Scheffy [11], and van den Hoven [13] all use the dimensional model in their descriptions of designing data warehousing and data marts. These three authors discuss the dimensional model as a necessary design for on-line analytical processing (OLAP) and data warehousing. Raden [8] also points out that the dimensional model is not tied to a physical representation of data and that the dimensional model can be implemented on hierarchical, relational and object-oriented database management systems [8].

Supporters of the E-R model are not necessarily critical of the dimensional model [1, 3, 6, 7, and 15]. Rather, these authors tend to discuss dimensional modeling as an option for the designer and then they promote their view of database modeling.

Francett [2], Ram [10], Papazoglou [8] and Wilson [15] extend the E-R model by addressing the shortcomings of the E-R model with technological advances or by integrating other modeling techniques such as object oriented modeling.

Francett supports the E-R model if the analyst can determine if the user "knows what he wants to know" [2]. If the user is not sure what information is needed, Francett throws her support to an "enhanced" dimensional model. This position recognizes the inherent strengths and weaknesses of each model discussed earlier.

University of Arizona professor, Sudha Ram, explores tools to assist in the interface between conceptual

and logical database design. These tools derive the functional dependencies, which are the logical associations that link the attributes of the E-R model to its components. The functional dependencies are used in both decomposition and synthesis in an E-R model [10]. If the relational database management system and the end-user software can make use of the functional dependencies, the complexity of the E-R model can be reduced to a level that the end-user should be able to operate within the data warehouse.

Wilson [15] discussed the use of object oriented extensions with the E-R model to create a hybrid model. The object model is introduced into the E-R model to standardize and re-use object code. Though there are major implementations of this strategy (Oracle, IBM, SABRE Technologies), this hybrid has not caught on in the data warehousing world [15]. Even though the hybrid models are relatively new, they do lend once approach to work around the technical issues in the E-R model and the dimensional model.

Monk [7] explores a totally different technology for databases. Monk proposes the object oriented model and the concept of "dynamic instance conversion" to express views of the database rather than the results of a query. This method is better suited for object oriented databases and may help break this emerging technology into the forefront of data warehousing.

Papazoglou [8] attacks the complexity of the E-R model by proposing a method that combines object-oriented semantics to the conceptual database schema. This approach develops semantic relationship between objects. The objects in this model are equivalent to the entities in an E-R model. This proposed methodology abstracts the precise structural nature of the E-R model into concepts that the end-user can understand [8]. Papazoglou's proposal is intriguing but it still requires a large amount of predetermined relationship knowledge of the information requirements of the end-user, albeit much less than the traditional E-R models.

Linthicum's discourse on mixed object oriented and E-R models asks the question "what is the business drivers?" [6]. This is a key question in a technology where implementation costs can be very high. The advantage to the object relational hybrid is the ability to include complex objects such as images and documents [6]. Linthicum drives to the point that even though hybrid object/relation models are the latest rage, the technology and tools are not there to support the mainstream IT shops.

## **ANALYSIS**

It is clear that the issues with both the E-R model and the dimensional model are enough to keep either model from being the ultimate solution to data warehousing.

The literature suggests that some combination of these models and other technologies might address the shortcomings of each model. These hybrid solutions will need to be supported by products and tools in the marketplace. A growth in tools and products for hybrid models will perpetuate the growth of data warehousing techniques that are flexible enough to meet the user's no matter how structured the decision making may or may not be.

It is also evident that the two models can be combined into a separate model that utilizes both technologies in the data warehousing process. Kimball supports the idea of the E-R model being a "back room" solution for the data warehouse and the dimensional model as the front-end solution for querying the database. This idea is extended into the table depicted in Figure 2:

By defining each step in the data warehousing process, each model can then be exploited for its strengths and minimize the impact of its weaknesses.

The idea of using object oriented concepts combined with E-R model and dimensional modeling will need to be explored more fully once the technology advances enough to assist the designer in developing a successful data warehouse model.

## **CONCLUSION**

E-R model and dimensional modeling have been explored as the primary approaches to data warehouse modeling. While both models have issues that can be addressed for specific implementations, the technology needs to be expanded to develop a more robust model that will support a broad general base of implementations.

To accomplish the more general solution to modeling the data warehouse, both the E-R model and the dimensional model will need to be extended into the next level of modeling. This next level will need to combine the simplicity of the dimensional model and the efficiency of the E-R model. This next level of modeling should also combine the concept encapsulation of the object-oriented model to support recent trends in distributed computing.

For the designer that needs to implement a solution today, it is recommended that a combination of the E-R model and the dimensional model be used. The E-R model should be used as the back-end data

loading model since the production database is most likely relational or hierarchical. The dimensional model should be used on the front-end to facilitate user understanding and acceptance. The multiple dimensional models can then be related to one another across the enterprise using the concepts in the dimensional model. The E-R model can also be modified to support the linking of the dimensional models if the fact tables are treated as entities, the dimensions are treated as relationships and normalization is not induced to the tables.

## REFERENCES

1. Date, C. J. "A Fruitful Union", *Computerworld*. 27(24): 130. 1994 Jun 14.  

Francett, Barbara. "Database Technologies vie for Data Warehouse Occupancy", *Software Magazine*. 15(4): 70-78. 1995 Apr.  

Gray, Paul. "Mining for Data Warehousing Gems", *Information Systems Management*. 82-86. Winter 1997.
4. Kimball, Ralph. "A Dimensional Modeling Manifesto", *DBMS*. 10(9). 1997 Aug.
5. Kimball, Ralph. The Data Warehouse Toolkit. New York: John Wiley & Sons. 1996.  

Linthicum, David S. "Mixing Tuples and Objects: Object relational databases are all the rage, but do they really fulfill a need?", *DBMS*, 10(13): 45-49, 1997 Dec.  

Monk, Sr. R. "View Definition in an Object-oriented Database", *Information & Software Technology*. 36(9): 549-554. 1994 Sep.
8. Papazoglou, M. P. "Unraveling the semantics of Conceptual Schemas", *Communications of the ACM*. 38(9): 80-94. 1995 Sep.  

Raden, Neil. "Modeling the Data Warehouse", Manuscript of an article by Neil Raden that was excerpted in the January 29, 1996 issue of *Information Week*, [http://members.aol.com/nraden/iw0196\\_1.htm](http://members.aol.com/nraden/iw0196_1.htm).  

Ram, Sudha. "Deriving Functional Dependencies from the Entity-Relationship Model", *Communications of the ACM*. 38(9): 95-107. 1995 Sep.
11. Rudin, Ken. "What's New in Data Warehousing", *DBMS*. 9(9): 54-63, 1996 Aug.  

Scheffy, Hugh. "Cube the power of your spreadsheets with OLAP", *Management Accounting*. 79(8): 50-54. 1998 Feb.
13. "Star Schemas and STARjoin? Technology", A Red Brick Systems White Paper.  
<http://www.redbrick.com>  

Van den Hoven, John. "Data Warehousing: Bringing it all Together", *Information Systems Management*. Spring 1998.
15. Wilson, Linda. "Hybrid Databases Enter Warehouse", *Computerworld*. 32(3): 71-72. 1998