The Enterprise Data Warehouse - From Data to Decision Making

Richard McCarthy, Central Connecticut State University, mccarthyr@ccsu.edu
George Claffey, Central Connecticut State University, claffey@ccsu.edu

ABSTRACT

An application centric data warehouse is built by focusing on addressing individual application reporting needs into a consolidated and comprehensive data warehouse environment. Application centric warehouses focus on the development of individual data marts as a primary step in the development process. Data centric warehouses unite business experts to define the data needs of the organization prior to the development of the warehouse environment. They are predicated on the development of a comprehensive data model. This paper defines the research framework for an investigation of the relationship between the types of methodology used to define a data warehouse ands the quality of the resulting warehouse. It also poses the question, does model type affect the strength of the relationship between business rules and user evaluations of data warehouse environments.

INTRODUCTION

New tools that allow for the drill down into many layers of data have enabled complex analysis of information. In order to effectively make use of these tools, there must first be an infrastructure in place that supports an integrated set of data that can effectively serve as the basis for supplying a single source of information. Implementation of a data warehouse enables that single source. An enterprise data warehouse is a centralized store of summary and detail information from all relevant sources that is used to analyze a business by allowing for drill down analysis and ad hoc discovery from multiple user groups [8]. A data warehouse contains four characteristics [1] [11], they are (1) subject oriented, (2) non-volatile, (3) time variant, and (4) integrated. Subject orientation enables users to determine not only how their business is performing, but also why. A data warehouse differs from an operational database in that, most operational databases have a product orientation and are tuned to handle transactions that update the database. There is a temporal and granularity mismatch in comparing On-line analytical processing (OLAP) applications driven from a data warehouse to an on-line transaction processing system caused by the amount of detail that is focused on by each application type [5]. The non-volatility of a data warehouse means that the data does not change between updates. This allows the data warehouse to be tuned for improving the performance of accessing information, since issues such as allowance for free space (for data growth) can be ignored. A time variant data warehouse means that it presents data as of a single point in time. All relevant data stores that are utilized are synced up as of a single point in time. An integrated data warehouse means that all of the data needed to manage the business is consolidated in one single location. Relevant data sources may include external data as well as internal operational data. A data warehouse should be designed to challenge people's thinking, not reinforce it [10]. It should lead to asking further questions to analyze information.

Data Warehouse Project Goals

Most data warehouses have six fundamental goals [8]. (1) The data warehouse provides access to organizational data, immediately, and with high performance. (2) The data is consistent. The

organization should have a 'single version of the truth'. (3) The data in the data warehouse can be separated and combined by every organizational measure. (4) The data warehouse provides a set of tools for query and analysis. (5) Only reliable and complete data is published within the data warehouse. (6) The quality of the data in the warehouse can be used to drive business process reengineering.

Data warehouse projects are either data centered or application centered [13]. A data centric warehouse is based upon a data model that is independent of any application. It is designed to support a variety of user needs and a number of applications. The methodological approach to designing a data centric warehouse involves data modeling with a group of business experts who are familiar with the different information views that are needed to support that business. This consists of a top down approach in producing specifications of information needs so as to not leave data behind [11]. A mapping approach should be used to provide a structured approach to classification of data. Data centric warehouses should support flexibility because executive information needs change constantly based upon changes in the underlying business. An application centric warehouse is one that is initially designed to support a single initiative or small set of initiatives. In addition to flexibility, a data warehouse needs to support scalability. The main issues within scalability are the amount of data within the warehouse, the number of concurrent users, and the complexities of user queries [4]. A data warehouse scales both horizontally and vertically. The warehouse will grow as a function of data growth and the need to expand the warehouse to support new business functionality. Scalability can be defined using four dimensions [3]. The first, environmental complexity involves supporting complex data models and queries. The second, user concurrency refers to the number and types of queries that can be supported at the same time. Not all queries are of the same priority or complexity. The third dimension, support for the environment must be planned as the data warehouse grows. The fourth dimension, data volume, refers to how will the current model and subsequent physical implementation support future growth. If the database is very large, it can affect the backup strategy. In many client server applications the use of RAID (redundant array of disks) technology allows the database to be mirrored, so that in the event of a hardware problem, the mirrored copy may be used to keep the database operational. The warehouse design team will have to determine the cost benefit of implementing this approach for a potentially large data warehouse. In addition, the timing of the backups will have to be determined because the amount of data involved may require special processing.

Data Mart Strategies

Some companies begin their move towards a data warehouse strategy through the development of a data mart. A data mart can be either independent or dependent. An independent data mart is a smaller subject area data repository that is not directly connected to the enterprise data warehouse [13]. An independent data mart is usually quicker to construct, and as such can serve as a proof of concept before a full-scale investment is made. This approach can also generate a quicker return on investment by realizing benefits sooner. This approach is viewed as an advantage by departments that are highly focused on control of their data. One of the major disadvantages to developing independent data marts is that with an enterprise data warehouse there is a single integrated data store, which reduces the potential for data integrity issues that arise when data is stored in multiple locations redundantly. An independent data mart strategy can lead to the proliferation of multiple independent silos of information, which can be a

deterrent to developing a single integrated strategy. Further, it can cause to prevent users from analyzing data using similar views that are made possible by the enterprise data warehouse.

A dependent data mart is a subset of the data warehouse organized by subject area (e.g. the marketing data mart). It provides the advantages of using a consistent data model and providing quality data. Dependent data marts support the concept of a single enterprise wide data model, however they require that the data warehouse be constructed first.

For a data mart implementation to be successful, it is important to develop a scope up front and build it into the plan [2]. Scoping should determine how long the requirements definition phase should take. The scope should be narrowly focused on a solid business case. The scoping process should be a high level evaluation designed to answer the significant questions that affect the development of the data mart. A common error in scoping is to combine the requirements gathering phase with the design phase. This can lead to prematurely modeling before the requirements are defined. One of the principle scoping questions therefore is who will translate the business requirements into the physical database design. There are several other issues to consider as well, such as the need for performance tuning, and how often will the data be refreshed.

The Role of Metadata

There are three major classes of software for a data warehouse. The first is the database management system, which will manage the storage and retrieval of data. The second class of software to be considered is the reporting or OLAP (on-line analytical processing) tool(s) to be used that will support the users need to access and analyze information. The third class of software is metadata management tools. Metadata is defined as data about data [4]. Metadata includes definitions of all objects that relate to the data warehouse. These include such things as tables, views, and attributes. It is important for the user in a warehouse environment to have some knowledge of the source and timing of data that comes from multiple sources. Metadata tools can assist the user by keeping track of the source of the data, when it was last updated and if there are any transformation rules that have been applied to the data prior to it being stored in the warehouse. The use of metadata by users requires a greater knowledge of the data than many users have had access to using in previous legacy systems. Additionally it allows them to spend more time as data analysts instead of merely being data reporters. Metadata is used to determine the quality of the content of the data. Metadata may track when data was last updated. It is also used to resolve differences in definition of data elements across the enterprise data warehouse. There are two types of metadata, structural and access [6]. Access metadata sets up drill down rules. It must be maintained so that users can quickly understand attribute names and aliases. Structural metadata tracks the relationships between data to increase the understanding of usage patterns.

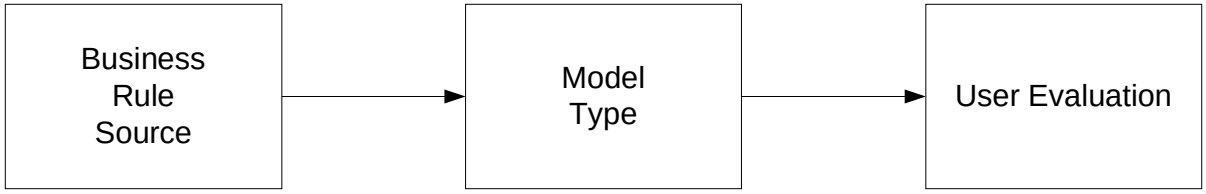
Data Mining Enablement

A data warehouse provides foundation architecture that data mining applications can be built from. Data mining provides for discoveries of patterns and correlations of information to be discovered from the volumes of data that many organizations retain within their information systems. Data mining applications are classified as either discovery driven or hypothesis driven [12]. A hypothesis driven approach is one in which the user seeks to validate a theory based

upon available data. Discovery driven data mining applications make use of the data warehouse by uncovering previously unknown patterns or classifications within the data. Utilizing object views, such as global views, base views and the local interface; the data warehouse can be used as a platform to enable the mining of large amounts of unstructured data [9]. This can enable new knowledge discoveries within an organization.

RESEARCH QUESTIONS

This study will focus on the relationship between the quality of the business rules that are developed for a data warehouse and the user evaluation of the data warehouse. This study will analyze the differences in the quality of the data model that is developed for data centric versus application driven data warehouse projects. The theoretical framework for this study purports that model type (data centric, application driven) moderates the strength of the relationship between the user evaluation of a data warehouse and the quality of the business.



The questions that will serve as the focus of this study are:

H1: User evaluation will be greater for data warehouses that have documented business rules in a centralized repository.
H2: Model type affects the strength of the relationship between user evaluation and business rules source.

METHODOLOGY

Data warehousing projects that are beyond their initial implementation phase will be used as the basis for this study. A survey instrument will be developed to test the model type, business rule source and the end user evaluation of the data warehouse implementation. Questions pertaining to the model type variable will determine if the data warehouse was originally designed to be data centric or application centric. The business rule source will investigate how integrated the data dictionary is with the data warehouse, and if the end-users have access to the data dictionary. User evaluation will measure the extent to which the user feels the data warehouse achieved its intended goals. The sample study will be performed using a population drawn from the same business type (property-casualty insurance companies that are domiciled in the U.S.) to eliminate any potential disparities that may arise to due significant differences in type of business that the data warehouse supports.

CONCLUDING REMARKS

In order to remain competitive, many organizations need to be able to quickly analyze the characteristics of their customers and their markets as an integrated view across

their enterprise. Data warehousing has been used as a means to standardize, centralize and distribute data to achieve this purpose [7]. A data warehouse provides a strategic infrastructure that enables the organization to more effectively analyze their operation, and hence make better decisions. The data warehouse architecture supports the enablement of data mining applications. A critical decision early in a data warehouse project is to decide if a data centric or application driven approach will be used. This study may serve as a guide to managers in the future to determine which approach is more likely to yield a satisfactory result. In either case, the data warehouse provides a foundation to improve the quality and access to information throughout an organization.

REFERENCES

[1] Corey, M., Abbey, M., Abramson, I., Barnes, L. Taub, B., and Venkitachalam, R, (1999), *SQL Server 7 Data Warehousing*, McGraw Hill, Berkely, CA.

[2] Dyche, J. (August 1998), Scoping Your Data Mart Implementation, DBMS, 11(9), 43-50.

[3] Fryer, R. (June 1998). Data-Warehouse Scalability, Byte, 63-64.

[4] Gardner, S. (September 1998). Building the Data Warehouse, Communications of the ACM, 41(9), 52-60.

[5] Jarke, M., Jeusfeld, M., Quix, C., and Vassiliadis, P. (January 1999). Architecture and Quality in Data Warehouses: An Extended Repository Approach, Information Systems, 24(3), .229-253.

[6] Katic, N, Quirchmayr, J. Schiefer, M., Stolba, A., and Tjoa, M. (1998), A Prototype Model for Data Warehouse Security Based on Metadata, IEEE Transactions on Engineering Management, 300-308.

[7] Kim, Y., Yoon, K., Kim, Y., (1996), Success Factors for Data Warehouse Introduction, IT Management and Organizational Innovations, 265- 270.

[8] Kimball, R. (1996). *The Data Warehouse Toolkit*, John Wiley & Sons, New York, NY.

[9] Miller, L., Honavar, V. and Barta, T. (August 1998). Warehousing Structured and Unstructured Data for Data Mining, IEEE Transactions on Engineering Management, 215-224.

[10] Raden, N. (November 1997). Push back in push technology, DBMS, 10, 63-68.

[11] Subramanian, A., Smith, L., Douglas, L., Nelson, A., Campbell, J. & Bird, D., (July 1997). Strategic Planning for Data Warehousing, Information & Management, 99-113.

[12] Turban, E., Aronson, J. (1998), *Decision Support and Intelligent Systems*, Prentice-Hall, Upper Saddle River, NJ, 5th edition.

[13] Watson, H. and Haley, B. (September 1998). Managerial Considerations, Communications of the ACM 41(9), 32-37.