Organizational Data Warehousing

John Wang
Montclair State University, USA

Xiaohua Hu
Drexel University, USA

Dan Zhu
Iowa State University, USA

INTRODUCTION

A data warehouse (or smaller-scale data mart) is a specially prepared repository of data created to support decision making. Data are extracted from source systems, cleaned/scrubbed, transformed, and placed in data stores (Gorla, 2003). A data warehouse has data suppliers who are responsible for delivering data to the ultimate end users of the warehouse, such as analysts, operational personnel, and managers. The data suppliers make data available to end users either through structured query language (SQL) queries or custom-built decision-support applications, including decision support systems (DSS) and executive information systems (EIS).

During the mid-to-late 1990s, data warehousing became one of the most important developments in the information systems field. It has been estimated that about 95% of the Fortune 1000 companies either have a data warehouse in place or are planning to develop one (Wixon & Watson, 2001). Data warehousing is a product of business need and technological advances. The business environment has become more global, competitive, complex, and volatile. Customer relationship management (CRM) and e-commerce initiatives are creating requirements for large, integrated data repositories and advanced analytical capabilities. More data are captured by organizational systems or can be purchased from the third party.

Even though there are many success stories, a data warehousing project is an expensive, risky undertaking (Beitler & Leary, 1997). Organizations are spending millions each year on data warehouse development, but the majority of all initial data warehousing efforts fail (Chenoweth, Corral, & Demirkan, 2006). The most common reasons for failure include weak sponsorship and management support, insufficient funding, inadequate user involvement, and organizational politics (Watson, Gerard, Gonzalez, Haywood, & Fenton, 1999).

Conventional wisdom holds that having a management champion with a tightly focused (data mart) design and restrictive tools will lead to success. However, Chenoweth et al. (2006) found that the reverse situation can be just as successful. If the users see the potential of the data warehouse to deliver value to the organization, they can be the champions and convince management to adopt the technology. Furthermore, if users understand both the technology and the organization's business processes, a single data repository may actually be more satisfying for them.

BACKGROUND

In today's business environment, every business owner dreams of having the ability to know what is happening in all aspects of his or her operation and of being able to use that information to better the market and increase profit. In order for an organization to achieve competitive advantage, voluminous data need to be managed, analyzed, and fed into the decision-making process. The introduction of data warehouses, which provide decision support to organizations with the help of analytical databases and analytical applications like online analytical processing (OLAP), answers this need (Gorla, 2003). The technical definition of a data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data that supports managerial decision making (Inmon, 2002). Typically, a data warehouse is housed on an enterprise's mainframe server, but it can reside with a storage service provider. Data in an OLAP data warehouse are extracted and loaded

from various online transaction processing (OLTP) applications and other sources using extract, transfer and load (ETL) tools. Analytical applications such as online analytical processing (OLAP) tools, data mining, statistical modeling, geographical information systems (GIS), DSS, and other user queries are then applied to the repository (Jones, 2001).

There are five major elements of data warehousing including: data acquisition, data modeling and schema, metadata, data management, and data analysis (Inmon, 2002; Jones, 2001). Data acquisition involves identifying, capturing, and transforming data in operational systems so that the data can be loaded into a data warehouse or data mart. Data acquisition is a complex, time-consuming, and costly phase of building and managing a data warehouse, but if this phase is not correctly carried through, the data warehouse will not be effective. During data acquisition, data are extracted, transformed, transported, and loaded. Data modeling is the analysis of data objects used in a business or other context and the identification of the relationships among these data objects. A data model consists of objects (for example, a product, a product price, or a product sale) and expressions of the relationships between each of these objects. The activity of data modeling leads to a schema, which is the organization or structure for a database. Metadata is a definition or description of data, and it is the glue that holds together all components and views of a data warehouse. Data management includes the access and storage mechanisms that support the data warehouse. This is usually a relational, multidimensional, or other specialized database designed to facilitate complex queries. Data analysis applications enable end users to access and analyze data stored in data warehouses or data marts. There are many variants of data analysis software. The main types of data analysis software include data mining tools, OLAP tools, enterprise business intelligence suites, and DSS.

There has been little empirical research on the implementation success of data warehousing projects. Wixon and Watson's (2001) empirical investigation suggest that management support and resources help to address organizational issues that arise during warehouse implementations. Furthermore, resources, user participation, and highly-skilled project team members increase the likelihood that warehousing projects will finish on-time, on-budget, with the right functionality. The implementation's success with organizational and project issues, in turn, influence the system quality of the data warehouse.

There are several issues of interest in the data warehousing literature. In this article, we focus on three issues. The first is data warehousing methodologies that organization may choose, the second is the management of the data warehouse through its operational life cycle, and finally the security of the data warehouse. The later issue is of importance because we believe that organizations must protect their valuable information asset.

MAIN FOCUS

When a small business decides to install a network for the first time, it must choose the operating systems, hardware, network, and software components. The same applies when an organization decides to build a data warehouse because there are several methodologies to choose from. Data integration technologies have experienced explosive growth and a large number of data warehousing methodologies and tools are available to support market growth. Furthermore as the business environment changes rapidly, management of the data warehouse through its operational life, and securing the data, become important because of the costs involved. In this section we present our thoughts on the choice of data warehousing methodologies, managing the data warehousing through time, and data warehouse security.

Choosing Data Warehouse Methodology

With so many data warehousing methodologies to choose from, a major problem for many firms is which one the company should utilize for its situation. We believe that when a technology is in its growth stage there is going to be a variety of methodologies and very little standardization. In such a case, we expect that an organization would use different criteria to evaluate the different options and select the one that meets their need. To confirm our thoughts, we researched recent empirical studies on data warehousing methodologies (Sen & Sinha, 2005). In general, data warehousing methodologies can be classified into three broad categories as follows (Sen & Sinha, 2005): (i) Core-technology vendors are those companies that