A Framework for Data Warehousing Research

Lei-Da Chen
The University of Memphis

Mark Gillenson
The University of Memphis

Follow this and additional works at: http://aisel.aisnet.org/amcis1999

A Framework for Data Warehousing Research

Lei-da Chen, The University of Memphis, ldchen@memphis.edu
Mark L. Gillenson, The University of Memphis, mgillnsn@memphis.edu

Abstract

This paper proposes a framework for data warehousing research. The framework includes environmental, organizational, personal, data and technological variables, which influence the development and use of a data warehouse, and dependent variables, which are useful in evaluating system success. The framework attempts to provide directions for future data warehousing research and development.

Introduction

A data warehouse is "a subject-oriented, integrated, nonvolatile, time-variant collection of data organized to support management needs" (Inmon and Hackthorn, 1994). The purpose of a data warehouse, as stated in the definition, is to facilitate managerial decision making. Data warehousing provides end users with decision making tools that create a competitive advantage. Data warehouses differ from operational databases, which mainly support the daily business transactions (BT), and management information systems (MIS), which supply information to managers in the form of simple periodic reports. A data warehouse collects information from multiple systems and stores it in a fashion that allows end-users to have faster, easier, and more flexible access to key information (Teresko, 1996). It is important to establish a meaningful and comprehensive framework to identify major issues in the development, use, and evaluation of a data warehouse. A general framework for data warehousing research is presented in Figure 1. This framework is based on the literature on both data mining and IS success. The independent variables (input), process, and dependent variables (output) are discussed in the forthcoming section.

Dependent Variables

Due to its complexity, the success of a data warehouse should be evaluated in a number of aspects. DeLone and McLean (1992) identified six major dimensions of IS success. These six dimensions can be translated into the data warehousing situation as follows:

1. System quality – An evaluation of a data warehouse itself (e.g. response time and code efficiency).
2. Use – The usage level of a data warehouse (e.g. frequency of use).
3. Quality of Information – Accuracy, meaningfulness, and timeliness of the information generated.
4. User satisfaction – End-users' satisfaction with a data warehouse.
5. Individual impact – The effect of the information on the success of an individual's decision making.
6. Organizational impact – The effect of the information on an organization's performance and structure.

While studying all six aspects would be difficult, it is recommended that a researcher choose one or more from the list as the indicator of success. For example, user satisfaction is often used as the single indicator for system success due to its high face validity and its availability as a valid measure (e.g. Doll & Torkzadeh, 1988).

Process: Data Warehouse Development and Use Process

The process can be further divided into two parts: development process and use process. The development process entails activities performed by the IS developer during the planning, analysis, designing, and implementation of the data warehouse, whereas the use process involves a multidimensional analysis of the data using AI and other statistical and mathematical techniques. Systems developers and end users are the two major types of personnel involved in data warehousing, and some of the independent variables discussed below apply to only one or the other of them.

Independent Variables

A wide range of independent variables influence the dependent variables discussed above. A comprehensive review of the IS literature have yielded four distinct categories of independent variables that affect the success of data warehousing. They are environmental, organizational, individual, and data and technological variables.

1. Environmental Variables:
a) Social responsibility
Today, companies cannot afford negative publicity. Loss of confidence in an organization is the single greatest cost of unethical behavior. Since the 1980s, public awareness of ethics in business has increased significantly, and studies have shown that firms are generally penalized for alleged unethical business practices. (Gunthorpe, 1997; Primeaux & Stieber, 1997). Companies implementing data warehouses need to

collect as much detailed data about individuals as they can. Companies should collect this data through legal and ethical means without invading consumers' privacy. Companies that hold the private information of their customers (e.g. medical records) should also honor customers' trust by not using it for profits unethically.

b) Government support and reporting requirements
Studies have shown that government policies such as R&D funding and investment tax credits have a strong positive influence on technology development strategy at the corporate level (Mowery and Rosenberg, 1979). Government support encourages organizations to invest in new technology to improve their competitiveness. Governmental organizations require companies to submit detailed reports disclosing their business activities on a regular basis. The requirements change over time depending on the funding of government programs and legislation (Robertson, 1997). This largely affects the kind of data companies will collect. Understanding the government reporting requirements is imperative in developing an IS, and such tasks become more challenging when a company is operating internationally (Waples & Norris, 1992).

2. Organizational Variables:
a) Industry type
The need for data differs from industry to industry. Some industries are more data intensive than others. This difference determines the value and complexity of a data warehouse. Industries such as retailing and financial services have long been dependent on data for their survival; therefore, a data warehouse will certainly serve as a competitive advantage for them. At the same time, the algorithms and technology employed in their data mining systems are usually more complex than those of the industries with less data intensity. The industry type determines the extent to which an organization is expected to benefit from a data warehouse.

b) Task characteristics
The types of decisions each company faces are highly different. Decisions can generally be divided into structured and unstructured decisions (Mason & Mitroff, 1973). A structured decision is "clear, well-defined, distinct and unambiguous". There are established and agreed procedures for making them, and the possible outcomes are predictable. On the contrary, an unstructured decision is "ill-defined, fuzzy and difficult to tackle". There are no known rules or procedures for making them, and the outcomes are unpredictable (Kaye, 1995). The unique characteristics of a data warehouse determine that it is most suitable for managerial decision making (Inmon & Hackthon, 1994); therefore, the benefits will be more visible where a great deal of unstructured decisions have to be made.

c) Organizational decision making structure
Organizational decision making structure determines where a data warehouse fits in an organization. Malone (1997) has identified 3 stages of decision making: In the first stage, when the communication costs are high, decentralized decision making was preferred for economical reasons. As communication costs fall, it becomes more desirable to bring remote information together and make decisions centrally. When the communication costs fall still further, decentralized decision making turns out to be more effective since decision makers combine remote information with local knowledge to make creative decisions. Fiedler et al. (1996) have divided information system structure into four clusters: centralized, centralized cooperative, decentralized, and distributed cooperative computing. Matching information system structure with organizational structure is one of the most critical decisions of a corporation (Straub & Wetherbe 1989). Different decision structures will affect the use of a data warehouse. A centralized decision structure limits the data warehousing activities to a small number of managers while a decentralized decision structure extends the use of the technology to a much larger user body. At the same time, implementing a data warehouse may in turn affect the decision structure of an organization. Once the tools for decision making are made widely available in the organization, the structure of the organization can be flattened to enable lower level managers to participate in more decision making.

f) Attitude towards new technology and top management support
Studies have shown that an organization's attitude towards a new technology will affect the outcomes of the new information systems (Berger, 1988; Goleniewski, 1988). The average attitudes toward management information systems are positively related to user behavior: a more positive attitude toward new technology results in a higher usage level (Ferguson & Nevell, 1996). Top management support has been consistently found to relate to IS success (Lederer & Mendelow, 1988). The top management's commitment to the technology will clear the barricades in the development and ensure that adequate resources will be devoted to the data warehousing project.

3. Personal Variables (focusing on both end-users and developers):
a) User background
Whether the users of data mining are top management, middle management, or lower management reflect the decision structure of the organization. Due to the different educational backgrounds, learning skills and experience of different users, the outcomes of the system tend to differ. Studies have shown that skilled users can flexibly and meaningfully change critical parts of

decision processes to respond to change in the external decision environment (Dutta et al. 1997). Although much of the decision making has been automated, it is still valuable to have skilled end-users make unstructured decisions in a dynamic decision environment.

b) Understanding of the organizational goals and objectives
A data warehouse is a tool to support unstructured decision making. A great deal of the decisions made using it will affect the business strategy and performance. Therefore, an individual's accurate grasp of organizational goals and objectives will better guide their use of the data warehouse. Lack of it will result in a waste of resources and inferior strategies. On the other hand, formal understanding of the organizational goals is necessary to design effective IS (Goodman, 1994), so developers must also be equipped with an accurate perception of the goals and objectives to ensure the validity and reliability of the data warehouse.

4. Data and Technological Variables:
a) Quality of the data
The quality of the information produced by a data

The delivery of business-critical information in great speed, volumes, and precision depends on the use of robust tools. Most techniques used in data warehousing are computational intensive; therefore, the sophistication of an organization's hardware, software, and network greatly affect the results from the system. A large body of literature has been devoted to the hardware, software and network components of data warehousing (e.g. Jeffery, 1995; Foley, 1996; Krivda, 1996).

Conclusion and Future Research

This paper proposes a framework for data warehousing research. The framework includes environmental, organizational, personal, data and technological variables, which influences the development and use of a data warehouse, and dependent variables, which are useful in evaluating the system success. The framework is based on an extensive literature review, and a natural extension of this study would be validating the framework empirically. The framework attempts to provide directions for future data warehousing research and to assist practitioners in their development and management of data warehouses.
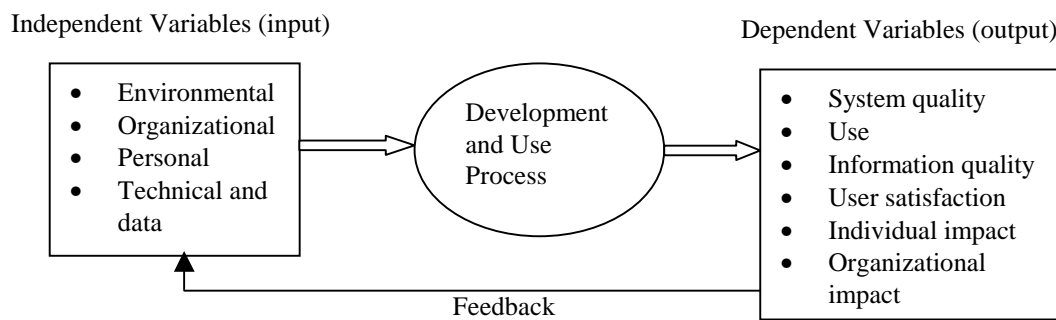
A Framework for Data Warehousing Research
Figure 1

warehouse depends on the quality of the data collected. Quality data is data that is representative of the information needed, well pre-processed for use with the systems, and well organized for easy access. Inmon (1996) identified four elements of data in data warehouses: 1. Integrated data (well structured and consistent data make the mining easier). 2. Detailed and summarized data (detailed data are necessary for uncovering patterns and trends, while summarized data allow end-users to build on the findings of the others to avoid repetitive work). 3. Historical data (historical data are crucial for businesses to understand their seasonality and business cycles). And 4. Metadata (metadata provides the context of data and serves as a road map). The quality of all of this data is important to the success of the system.

b) Degree of the data warehousing tools support

References

Reference available upon request from first author.