Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2001 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2001

Recent Developments in Data Warehousing

Hugh Watson
University of Georgia

RECENT DEVELOPMENTS IN DATA WAREHOUSING

Hugh J. Watson
University of Georgia
Hwatson@terry.uga.edu

Abstract

Data warehousing is a critical enabler of strategic initiatives such as B2C and B2B e-commerce, customer relationship management, and Balanced Scorecards. It requires the extraction of data from source systems, the use of data cleansing and transformation processes, and the loading of the data to a data mart or warehouse. When the data is in place, users access the data to support decision making. Harrah's Entertainment provides an example of how a data warehouse can be used to build better, more profitable relationships with customers, and Owens&Minor illustrates how it can integrate the activities of trading partners along the supply chain.

Introduction

In less than 15 years, data warehousing has joined the mainstream of organizational computing. Once practiced by only a few firms, today virtually all large and many medium and small size firms have a data warehouse in place. Data warehousing is critical to important initiatives such as B2C and B2B e-commerce, customer relationship management, and Balanced Scorecards. The market for data warehousing continues to grow. The Palo Alto Management Group predicts that the data warehousing market will increase to $113.5 billion in 2002, including the sales of systems, software, services, and in-house expenditures (Eckerson 1998).

This paper provides a tutorial on data warehousing. It covers its definition; history; alternative development strategies and architectures; data extraction, loading, cleansing, and loading processes; data storage and representation alternatives; and organizational uses. Experiences at Harrah's Entertainment and Owens&Minor are used to illustrate current, leading-edge practices.

Data Warehousing Defined

Most fundamentally, a data warehouse is a specially created repository of data designed to support decision making (Gray and Watson 1998). The concept is not new. In the mid 1970s, conceptual articles on decision support systems called for the creation of a separate decision support data base (Sprague and Watson 1975). What is new is the terminology, the enabling technology, and the huge amounts of data that can be captured and analyzed.

Bill Inmon (1992), considered to be "the father of data warehousing," provides the most commonly used definition for data warehousing. According to Inmon, a data warehouse is a:

- subject oriented (data are organized around areas of interest, such as customers)
- integrated (data from multiple sources are combined around a common key)
- time-variant (historical data are maintained)
- nonvolatile (users do not update the data)
- collection of data in support of management decision processes.

While the data warehouse is the repository for the data, the entire process, that begins with placing data into the warehouse and then ultimately using the data for decision-support purposes, is called data warehousing. With this view, data warehousing is a more inclusive term.

The Emergence of Data Warehousing

The first data warehouses were built in the mid 1980s (Kelly 1997). Later in the decade, Bill Inmon coined the data warehouse term. At about the same time, database machines (early parallel processing computers with built-in databases) became popular for handling large volumes of data. Teradata, which was later acquired by NCR, is an example of this kind of technology. Data warehousing received a boost when IBM began promoting the "information warehouse" in the early 1990s. In the late 1990s, as companies focused on customer relationship management (CRM), the need to store and analyze terabytes of data emerged. Data warehousing capabilities have continued to grow in importance as the various forms of e-commerce have emerged.

Competing Strategies: Marts versus Warehouses

Even though data warehouses are widespread, there is no common agreement about the best development methodology to use. In fact, there are two competing strategies or approaches. The first is associated with Bill Inmon (1992) who recommends that companies use a top-down, enterprise data warehouse approach. With this approach, the initial design focuses on a warehouse that contains multiple subject areas and integrates data from multiple sources. Its successful execution requires considerable senior level support, enterprise-wide governance, and ample resources. The second approach is associated with Ralph Kimball [1992], another highly respected consultant and writer on data warehousing, who recommends that companies use a bottom-up, data mart approach. This is a "start small, think big" approach and typically begins with a specific business need for data, often in the sales or marketing areas (Watson et al. 2001). The initial data mart contains data for only a single or a limited number of subject areas and draws data from a small number of source systems. Because of its limited scope, a data mart can be developed quickly, at a relatively low cost, and provide a fast return on investment. If the data mart is successful (thus providing a "proof of concept" for data warehousing), the project team expands the data mart by adding more subject areas, users, and applications. When properly executed, the data mart strategy ultimately leads to an enterprise-wide data warehouse.

Data Extraction, Cleansing, Transformation, and Loading

It is not unusual for a data warehouse to source data from over 100 internal systems. External data may also be sourced, especially for CRM applications. Understanding the data and extracting it from these systems is a major undertaking. While there is special-purpose software available to assist with the work, most of it is "pick and shovel" labor, with custom-written COBOL routines commonly used (Watson and Haley 1997). It is common to store the data in a staging area before it is loaded into the warehouse. This approach allows the data to be cleaned, integrated, and transformed. While some extraction, transformation, and loading (ETL) software has rudimentary data cleansing capabilities, it is common to use specialized cleansing software. Common keys for data integration are not always readily available because of differences in source systems. Mergers and acquisitions also complicate integrating the data. The data also must be transformed. Transformation includes putting data in the same format (e.g., M for male, F for female) and applying business rules (e.g., how aggregations are performed). The data is then loaded into the warehouse, using either a complete or incremental refresh strategy. The emerging approach is to "trickle" data to the warehouse; the warehouse is updated incrementally, with very short refresh cycles. NCR Teradata calls this the "active" warehouse. This approach is becoming more important as data warehouses are used to support operational systems where current data are needed.

Data Stores

Warehouse data is most often stored in a relational database (e.g., Oracle 9i, DB2). To provide a simple view of the data and to improve the response times for users, dependent data marts may be used. With this approach, data are sourced from the warehouse and placed on a data mart. By sourcing the data from the warehouse, a "single version of the truth" is maintained. Everybody accesses the same, accepted, accurate data. When relational data base technology is used for the data mart, it is units common to use a star schema as the underlying data model, with the data maintained in fact and dimension tables (Kimball 1992). The fact tables contain the data to be analyzed (e.g., the number of sold) and the dimensions are the ways the data can be "sliced and diced" (e.g., by time period, sales person). Multidimensional databases (e.g., Essbase) can also be used to improve response times.

In addition to data warehouses and marts, many companies maintain operational data stores (ODS). ODSs are similar to data warehouses in that they contain clean, integrated, subject oriented data. They differ in that they contain only a limited amount of historical data (e.g., 30 days) and the data are used by operational systems. For example, a telecommunications company might

create an ODS to integrate its long distance, broadband, and wireless customer databases. A customer service representative would then use the ODS to gain a complete understanding of a customer's use of the company's products.

Use of a Data Warehouse

The end users of a warehouse range from analysts to senior executives to business trading partners. Typically, analysts access warehouse data either through SQL queries, a managed query environment (e.g., Business Objects, Brio), or a data mining tool (e.g., SAS, Teraminer). On-line Analytical Processing (OLAP) is the term commonly used to describe the multidimensional analysis of data by users (Gray and Watson 1998). At the other end of the spectrum, executives might be provided with a set of reports generated from the warehouse or have access to an executive information system that relies on warehouse data.

The Web has had a significant impact on how warehouse data are accessed, and most warehouses are now Web enabled. The Web has also facilitated giving partners along the supply chain access to warehouse data.

Data warehouses have moved from being a useful source of decision support data to being a necessity for many business initiatives. It is virtually impossible to have significant CRM, B2B, and B2C activities without the use of data warehouse.

The following brief descriptions of the data warehousing activities at Harrah's Entertainment and Owens&Minor illustrate what leading-edge companies are doing. These companies won the prestigious Leadership Award from The Data Warehousing Institute in 2000 and 1999, respectively.

Data Warehousing at Harrah's Entertainment

Harrah's Entertainment (Watson and Volonino 2001) is assuming a leadership role in the gaming industry through a CRM business strategy, enabled by data warehousing, that focuses on knowing their customers well, giving them great service, and rewarding their loyalty so that they seek out a Harrah's casino whenever and wherever they play.

In 1993, changing gaming laws allowed Harrah's to expand into new markets through the building of new properties, and the company has followed an acquisition strategy with the purchase of Showboat casinos, the Rio All-Suite Casino, and Players International. As management thought about how it could create the greatest value for its shareholders, it was decided that a brand approach should be taken. With this approach, the various casinos would operate in an integrated manner rather than as separate properties. Critical to their strategy was the need to understand and manage relationships with their customers.

Harrah's had to understand where their customers gamed, how often and what games they played, how much they gambled, their profitability, and what offers would entice them to visit a Harrah's casino. Armed with this information, Harrah's could better identify specific target customer segments, respond to customers' preferences, and maximize profitability across the various casinos.

To execute their business strategy, Harrah's made a substantial investment in information technology to create WINet. WINet sources data from casino, hotel, and event systems. The data is then integrated into a patron database that serves as an operational data store. The data store is used to support operations, such as facilitating the check-in of customers at Harrah's hotels, and in generating offers to customers to visit Harrah's casinos. It is also used with Total Rewards, Harrah's customer loyalty program. Customers accumulate Reward Credits based on their gaming and other activities at any of Harrah's properties. These Reward Credits can be redeemed for cash or comps on hotel accommodations, meals, and shows. Data from the patron database is then loaded in the marketing workbench, which serves as Harrah's data warehouse. The marketing workbench supports analytical applications such as customer segmentation and profiling, and identifying customers to send offers to. NCR's Teradata database software is used to provide fast response times to analysts' queries. WINet also provides the foundation for "closed loop marketing." With this data-driven approach, campaigns are designed, tested, and the results retained for future use. Harrah's creative marketing, innovative uses of information technology, and operational excellence resulted in many benefits, including: a doubling in the response rate to offers to customers; consistent guest rewards and recognition across properties; a brand identity for Harrah's casinos; an increase in customer retention worth several million dollars; a 72 percent increase in the number of customers who play at more than one Harrah's property, increasing profitability by more than $50 million; and a 62 percent internal rate on return on the information technology investments.

Data Warehousing at Owens&Minor

Owens&Minor (Wixom et al. 2001) is the nation's leading distributor of name-brand medical and surgical supplies, and over the past four years, the company transformed its business model by integrating supply chain management, e-business, data warehousing, and Internet technologies. Especially important is WISDOM (WebIntelligence Supporting Decisions from Owens&Minor), a Web-based decision support system that provides information to Owens&Minor's employees, and perhaps more importantly, to OM's suppliers and customers as a supply chain extranet. WISDOM is instrumental in creating new business, expanding relationships with existing customers and suppliers, and retaining customers and suppliers.

The development of a data warehouse was an important step in turning information into knowledge into profit. Initially, the warehouse focused on providing internal personnel with better information. However, the most exciting initiative was to give customers and suppliers access to OM's data warehouse to strengthen relationships with trading partners, increase sales with existing customers, generate business with new customers, and extend OM's data warehouse as a value-added service over the entire supply chain.

Through a Web browser interface, customers and suppliers use WISDOM in a variety of ways. For example, customers can check their purchases of items on and off contract; items purchased off contract are more expensive than those bought on contract. They can investigate whether they might save money through product standardization by buying items from a smaller number of suppliers. Suppliers can investigation their sales penetration into different hospitals. This information can help suppliers increase sales and hospitals to decrease costs. The information that trading partners can access from WISDOM allows them to save money and provide better service.

Because the information in WISDOM is so valuable to trading partners, OM turned it into a product. Customers and suppliers pay to access their own data because they cannot or do not chose to generate the information from their own internal systems.

WISDOM created significant business benefits, including: winning over $60M in new business; expanding existing business by over $100M; generating $125,000 from customers and suppliers' use of WISDOM in 1999; over $500,000 in 2000; and an estimated $2 million run rate projected for 2001; and creating a perception in the industry that OM values its customer relationships.

OM recognizes that it is difficult to sustain a competitive advantage using information technology. To stay ahead of the competition, on-going IT initiatives are required. Therefore, OM is creating the second generation of WISDOM, which will store all of their trading partners data, not just that created by business transactions with OM.

References

Eckerson, W. W., "Post-Chasm Warehousing," Journal of Data Warehousing (3:3), Fall 1998, pp. 38-45.
Gray, P., and Watson, H.J. Decision Support in the Data Warehouse, Prentice-Hall, 1998.
Inmon, W.H. Building the Data Warehouse, New York: Wiley, 1992.
Kelly, S. Data Warehousing in Action, Chichester, Wiley, 1997.
Kimball, R. The Data Warehousing Toolkit, New York, Wiley, 1992.
Sprague, R.H., Jr., and Watson, H.J. "MIS Concepts," Journal of Systems Management (26:1), January 1975, pp. 34-37.
Watson, H.J., and Haley, B.A. "Data Warehousing: A Framework and Survey of Current Practices," Journal of Data Warehousing (2:1), January 1997, pp. 10-17.
Watson, H.J., and Volonino, L. "Harrah's High Payoff from Customer Information," http://www.terry.uga.edu/~hwatson/harrah's), 2001.
Watson, H.J., Annino, D.A., Wixom, H.W., Avery, L.A., and Rutherford, M. "Current Practices in Data Warehousing," Information Systems Management (18:1), Winter 2001, pp.47-55.
Wixom, B.H., Watson, H.J., and Stoller, D. "Adding WISDOM to Customer and Supplier Relationships," (http://terry.uga.edu/~hwatso/ owens&minor), 2001.