

Data Warehousing: A Technology Review and Update

Vernon Hoffner, Ph.D., CCP
EntreSoft Resources, Inc.

Abstract

Data warehousing has been around for over a decade. Therefore, when you read the articles in professional magazines and journals, most authors assume that you know and understand the terminology and acronyms. This can be a problem when you are now trying to understand this area of the technology. This paper reviews the technology and terminology (buzzwords) and provides an integrated foundation for understanding the benefits and potential hazards for data warehousing. The paper begins with a list of definitions and their implications for understanding data warehousing. Next, is a discussion of how data warehousing fits within the IT infrastructure, its benefits and application. Then data warehousing is reviewed from a business and organizational perspective, with consideration of organizational impact and risks.

Introduction

Data warehousing has become one of the significantly growing areas of information technology during the past decade. Our ability to store increasingly large volumes of data in organizational databases is no longer a technical problem, but potentially an organizational asset that needs to be utilized to business advantage. The growth in interest in data warehousing applications is the result of a convergence of technological capability and business needs. The technological capabilities are the growth in processor speed, the increase in data storage capacity and an increased sophistication and ease of use of data warehousing and data analytical software. This enables information systems staff and the end users to efficiently and effectively run software against volumes of business data to obtain information to improve business activities. Usable results are now available in a relatively short time frame.

In the past, and currently, in many organizations, access to information to support tactical and strategic business decisions was a difficult and time-consuming process. (Past tense will be used for consistency, but this still refers to current situations in many companies.) In some cases it was impossible to obtain the relevant information on a timely basis. The problem was not that the data was not available, but that the data resided in more than one legacy system. This required the services of a programmer or systems analyst to write programs to extract the needed data from the various sources, and there was a backlog of requests for the programmer to perform. There could be a significant time delay before the programmer could start the request. Accessing multiple data sources would also be a problem. Knowledge of each data source would need to be acquired, security permissions would have to be obtained for access, and programs written to extract the required data from each source. If the data were not extracted from the operational source, then in many cases, the load on the operational systems would be greatly increased resulting in degradation of operational efficiency in processing normal business transactions. Then the data would have to be organized into a format to perform the required analysis and reporting. Figure 1, illustrates a typical situation with volumes of data to be extracted, sifted, and organized prior to it being usable to provide information to the decision making process. This process could take an extended amount of time, several weeks or longer depending on the programmer's knowledge of the data sources, programming proficiency, and communication skill in determining and understanding the user's request. There were times when the request could not be fulfilled within the time frame required by the end user. This was not a good situation for business decision makers. In addition, there was always the problem when different users made similar requests that different programmers would respond

to those requests, and the results would be different. An example of this was the difficulty of obtaining a count of employees at an automotive assembly plant for one of the automotive companies at which I had worked. Depending on the organizational source of the response, a wide variety on numbers could be obtained. All of these counts would be accurate, depending on the definition of “employee” and the point in time of the count.

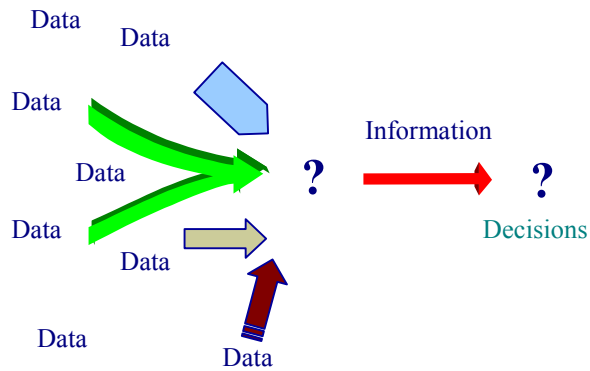


Figure 1. Current, existing situation for extracting data for decision support in many organizations.

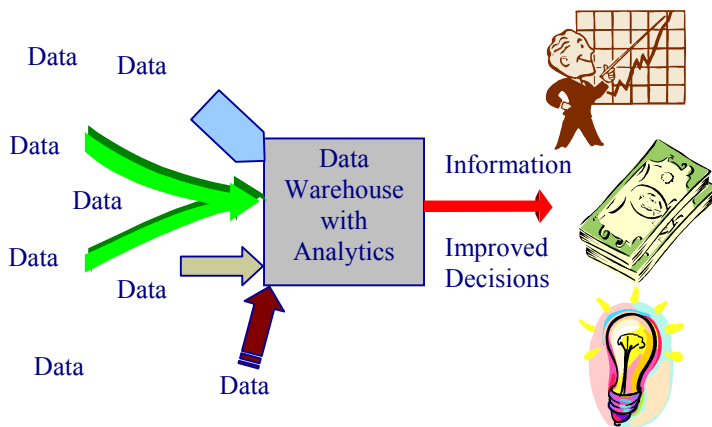


Figure 2. The desired situation, with integrated, consistent and accurate data, leading to improved decisions.

The primary reason for a company to implement a data warehouse is to resolve the problem illustrated in the preceding paragraph. Data is extracted from several sources into a single repository, providing a single source for information to be used in decision-making. There are several inherent advantages of the data warehouse solution, as illustrated in Figure 2. The data can be quality assured so that the information derived from the data warehouse is reliable and there will be only a single number for any specific query against the data. This eliminates the problem of management trying to reconcile a variety of numbers from different departmental perspectives. The data in the data warehouse will be structured for efficient data analysis and ease of use by the business analysts. The tools that are available today, e.g. SAS, provide the end users with ability to access the data without requiring data gathering and programming support from programmers in the Information Systems department. Results of queries will be available within minutes or hours, not the days and sometimes weeks that were required when the data was not available in a data warehouse. In addition, if the data warehouse resides on its own system, then the analytical processing will not add additional processing load to the operational/transaction processing systems leading to poor response time for transaction processing.

The resulting data warehouse system provides the foundation for a decision support system or an enterprise information system. Depending on the design and structure of the data warehouse, it will provide data that is described in standard business terms. The data names and structures are designed for non-technical users. The data is also preprocessed during the loading process with standard business rules of the organization. In addition, the focus of the data will be on business entities, such as, customer, product, and marketing channel. The user will have access to summary data and also have the ability to drill down to the detail data when needed for their analysis. We now have a system that has the potential to support tactical and strategic decision-making. This is not usually possible with ERP and operational legacy systems.

Data Warehouse Defined

In the preceding paragraphs the rationale for and general description of a data warehouse have been presented. Next are presented several definitions of a data warehouse. More than one definition is presented to provide different views and perspectives on the structure and use of a data warehouse. The first is a classic definition from one of the early writers about data warehousing. A data warehouse is “a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management decisions.” (Inmon, p. 33) This definition provides the focus of the data in the data warehouse.

The data is subject oriented. Operational (legacy) systems are organized around business processes (applications), Order Entry, Accounts Payable, Customer Billing which each application supports. But data for tactical and strategic decision-making is focused around categories or subject areas, such as, Customers, Products, and Vendors. The data warehouse is can be the focused source for data a major subject area for a business.

The data is integrated. Data is integrated by enterprise consistency in the measurement of variables, naming conventions, and physical data definitions. The inconsistencies that have developed over the years in the legacy systems are reduced or eliminated as the data is included in the data warehouse. The process of extracting, transforming, and loading the data into the data warehouse should result in obtaining data from multiple sources, combining all of the source data for a single, and as complete as possible, view of subject area.

The data is nonvolatile. The data in the data warehouse is a snapshot of the enterprise status at specific points in time. Therefore, the data is a series of historical data points, and as such should not be updated,

changed or modified. If the data is incorrect, then the data in the operational systems is incorrect. In this situation the operational systems should be corrected. In very unusual cases, it may be necessary to correct the contents of the data warehouse. The data in the data warehouse is usually unchanging.

The data is time variant. The data, as a series of data snapshots, provides a historical time series of the status of the enterprise at regular points in time. This provides a time dimension to the data that is not possible in the operational (legacy) systems. One of the key components of the data structure of the data warehouse is time. Operational systems contain current data and very little historical data. A 30 to 90 day storage of data would be normal for most operational systems. A data warehouse may contain 5 to 10 years of data, as appropriate for the subject area.

The second definition provides more of a focus on the use of a data warehouse by business analysts and managers. A data warehouse is “a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context.” (Devlin, p. 20) This definition also takes into account the quality of the data with the consideration of a complete and consistent store of data. However, the data is only useful if it can be accessed easily by end users. The data is structured to be used by end users in the process of making business decisions in and across the business functional areas. This implies the development of an enterprise data model that fits the organization’s business environment and decision making requirements needed to support strategic and tactical planning and decision making processes. This really means that the development of a data warehouse is a business development task, not a technical development task. The implication is that the data warehouse should be a single, complete source of data to support decision-making. As a single source, it should be easily accessible through several tools that are easy for business analysts to use. The data and supporting information are also structured for ease of supporting decision-making and data exploration.

A third definition considers data warehousing as a two part process that identifies the collection of the data into a repository as the first part of the process, and the second part as the process that supports decision making within the organization. A data warehouse is “the extraction, cleansing, and transformation of operational data for the purpose of Decision Support” (Welbrock, p. 9), and “a process of fulfilling Decision Support enterprise needs through the availability of information.” (Welbrock, p. 10) This definition adds a component that was absent from the previous definitions – the activity of cleansing the data prior to loading it into the data warehouse. Another term that has also been applied to this activity is “data scrubbing.” The data cleansing is the process of removing any inconsistencies or inaccuracies from the data prior to loading the data into the data warehouse. The definition of the requirements for the extraction, cleansing, and transformation of the source data is the most difficult part of building the data warehouse. In most data warehouse projects it is the most time consuming step and its scope is frequently underestimated. The first part of this definition concentrates on the technological aspects of the data warehouse. Without this emphasis the data warehouse cannot be successful. However, the end purpose of the data warehouse is to support organizational decision-making. If improved decision-making is not the result of the data warehousing process, then the data warehouse application is a failure.

Many authors have different definitions to support their specific focus or approach to building and/or using a data warehouse. Some focus on the technological aspects, some focus on the project management activities, but the ultimate purpose of the data warehouse is to lead to improved organizational decision making. Ralph Kimball (Kimball, p. xxiii-xxv) has succinctly stated the goals or requirements for a data warehouse as:

1. The data warehouse provides *access* to corporate or organizational data.
2. The data in a data warehouse is *consistent*.
3. The data in a data warehouse can be separated and combined by means of every possible measure in the business (the classic **slice and dice** requirement.)

4. The warehouse is not just data, but also a set of tools to *query, analyze, and present* information.
5. The data warehouse is the place where we *publish used data*.
6. The quality of the data in the data warehouse is a *driver of business reengineering*.

These requirements lead to a number of capabilities and results that can be obtained from the data warehouse. First, the data is easily accessible by managers and analysts in the organization. The access is immediate, on demand, and performance with respect to response time performance is fast. Access is not made through another individual, but the business analyst or manager can access the data from their desktop. No intermediary is required. Second, the same question from different people will result in the same answer. The single, consistent source of the data means that any question about an organizational characteristic will result in the same single number. Third, this implies a multidimensional data storage approach. Although this is technological consideration, the focus is on capability desired by the end users and their need to analyze the data. The “slice and dice” means that they can view the data along any dimension that is important to their analysis requirements, for example; product, geography, time, customer, etc. Fourth, the real value of the data warehouse comes with the set of tools the support the extraction of business intelligence from the data. These tools should include data access and query capability, report writers/generators, decision support tools including statistical functions and on-line analytical processing (OLAP), and potentially the use of an Executive Information System interface. Fifth, the data is collected, merged, integrated, cleansed, quality assured, transformed, and then placed into the data warehouse. Only complete, reliable data is released to the data warehouse. The primary sources of most of the data for the data warehouse are the various operational systems within the organization. In most instances a significant amount of processing must be performed on the source data in order to insure that quality data is loaded into the data warehouse. Sixth, the better the quality of the data, the better we can understand our business environment and our business processes, which can lead to better direction for changes in our business. Inaccurate and incorrect data can lead to a misunderstanding of our business environment and business processes. This misunderstanding will lead to poor decisions.

The Business Need

The business need is for more business intelligence, not smarter people, but better information about the business environment and its impact on the organization. There is enough data and statistics to bury the workforce, business analysts, and managers. The need is to identify and disseminate usable data that helps people make better decisions. Not only does business need to collect the right data, but they also need to be able to analyze it correctly. Knowing the sales trend over the last several months may help predict the sales next month, but analyzing buyer behavior tells them who is buying. Analyzing product sales volumes provides insight into which products are currently in demand. Knowing the level, function, industry, and how much was purchased provides the information needed to derive conclusions about product development, marketing strategies and focused sales efforts that will be more effective and efficient in the marketplace. This is the kind of business intelligence that is obtainable from a data warehousing system.

The IT Infrastructure for Data Warehousing

The traditional operational systems were designed to process the typical business transactions of the organization. In the case of on-line transaction processing (OLTP) many systems process high volumes of transactions requiring rapid processing and response. Much of the data created by these systems are used to support the daily activities for business functional areas. These operational systems are designed for and are very good at putting data into databases quickly, accurately, and efficiently, one transaction at a time. However, they are not designed for decision support activities. Decision support, or on-line analytical processing (OLAP) is designed to provide data for analyzing a problem or situation. This is accomplished by analyzing patterns or trends in the data. This usually requires the processing of a large number of business records, the results of many of the individual transactions processed by the OLTP applications. In order to avoid seriously

impacting the performance of the OLTP systems, the data warehouse needs to reside on a separate computer system. An example of the structure for a data warehousing system is illustrated in Figure 3.

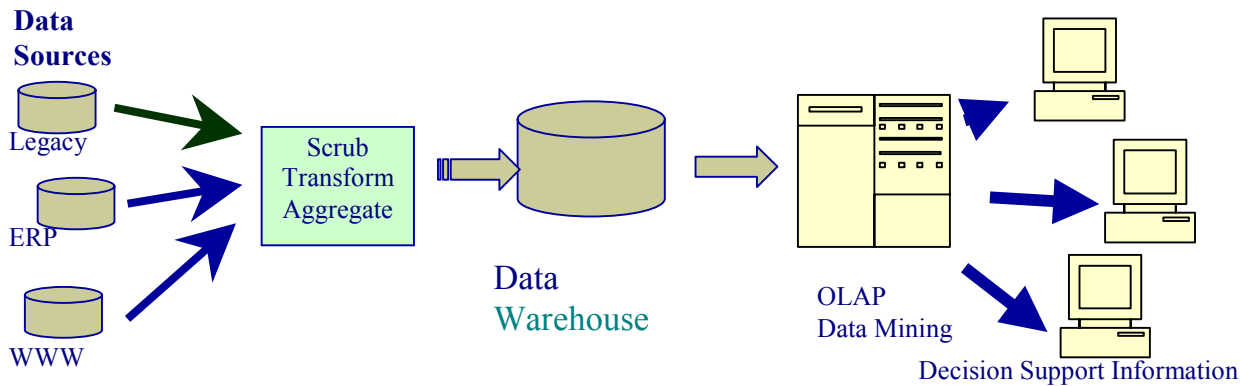


Figure 3. Example data warehousing system for decision support.

The process illustrated in Figure 3, shows the complete process for an operational data warehousing system. The left side of the diagram pictures the extract, transform and load (ETL) process. The first step is the extraction or collection of data from a variety of sources, most of which will be internal to the organization. The sources could include older legacy systems, enterprise resource planning systems, and any existing Internet e-business applications. The next step is to process the data into a form so that it can be added to the data warehouse. This processing could include any required consolidation, merging, cleansing, transforming, summarizing, and quality assurance activities to insure that only specific quality data is added to the data warehouse. Now the data is ready to be loaded (added) into the data warehouse. The ETL process can occur as frequently as required by the organization. Typically this will occur on a regular monthly, weekly, or daily cycle.

The center of the diagram is the data storage for the system. This could be a relational database management system. It could be a multidimensional database. It could be a SAS database. The size of this data storage could be only a few gigabytes up to tens of terabytes. The data warehouse should also contain the metadata for the system. This is data that describes the content and processes of the data warehouse in fine detail. Complete documentation of the ETL processes should be included in order to track back the source and transformation of every data element, when that becomes necessary. The metadata should also provide complete documentation of every data element in business terms for ease of use by the business analyst or manager.

The right side of the diagram is the data access and analytical tools used to access the data and generate the information needed to support understanding of the business environment and support the organizational decision-making. Some of the analytical processing may occur on the server on which the database resides and the results transmitted to the desktop of the end user. An alternative may be that some data is extracted from the data warehouse, downloaded to the desktop and then the analysis performed with desktop tools. This architecture provides the end user with easy access to the data and the analytical tools to support the required analysis.

An Example

A good example of the consolidating of customer order data into a data warehouse is Eddie Bauer. (Hess, Feb. 2001, SAS Communications, No. 1 & 2, 2000) Customers can purchase items at their stores, through their

catalogs, and online via the Internet. They have been accumulating sales from 2 web sites, 600 stores, and 110,000,000 catalogs for 15 to 20 million customers for the past 5 years. The analytical applications involve predictive modeling. Which customers are most profitable? Eddie Bauer has learned that tri-channel customers are the most profitable. However, the relationship is not linear but exponential. They have utilized this knowledge to improve their methods of marketing to their special customers. Another question they have addressed is which of the 44 different catalogs, with a circulation of 110,000,000 copies, should be sent to which customers in order to maximize revenue and profit from a fixed advertising budget? This is an example of a customer relationship management (CRM) application. CRM is one of the most frequently talked about applications of data warehousing, particularly for retailing, telecommunications, and financial companies.

Summary

This paper has reviewed the concepts relating to data warehousing. It is an interesting application of information technology. New hardware and software support the analysis of large amounts of data in relatively short periods of time. The process of creating a data warehouse provides the potential of increased availability of information to support management decision-making. The challenge is integration of this new technology into the decision-making activities of the organization. As indicated in the example, the potential benefit of this is increased profitability for the organization and enhanced and improved relations between the company and the customer.

References

“data mining is a perfect fit for retailer for Eddie Bauer,” SAS Communications, No. 1 & 2, 2000.

Barry Devlin. *Data Warehouse from Architecture to Implementation*. Addison-Wesley, Reading, Mass., 1997.

Ed Hess, “The ABCs of CRM,” Integrated Solutions, February, 2000.

W.H. Inmon. *Building the Data Warehouse*, 2nd. John Wiley & Sons, New York, NY, 1996

Ralph Kimball. *The Data Warehouse Toolkit*. John Wiley & Sons, New York, NY, 1996.

Peter R. Welbrock. *Strategic Data Warehousing Principles Using SAS Software*. SAS Institute, Cary, NC, 1998.

Vernon Hoffner, Ph.D., CCP
Chief Technology Officer
EntreSoft Resources, Inc.
A SAS Alliance Partner
27600 Northwestern Highway
Suite 280
Southfield, MI 48034

Mobil 248-514-2826
Office 248-350-1350 ext. *13

M W S U G

Information Visualization



