

An Introduction to Data Warehousing

An organization manages information in two dominant forms: operational systems of record and data warehouses. Operational systems are designed to support online transaction processing (OLTP) whereas data warehousing systems are designed to support online analytical processing (OLAP).

Operational systems concentrate on high-volume transaction processing on a day-to-day basis using real-time data. These systems are generally process-oriented and usually focus on specific business tasks such as registering students, updating financial transactions, and managing employee timesheets. They are optimized for simplicity and speed of modification, allowing for efficient and effortless data entry and retrieval. Such systems also track historical and transactional data, but not to the degree required by research queries.

While the operational systems primarily focus on current data management, the data warehouses update and store historical data. They are generally subject-specific and usually carry data from multiple operational systems to support organizational decision-making. Data warehouses can be used to address issues in academic institutions regarding student satisfaction, the effectiveness of new instructional techniques, and the attrition rate. In response to a concern, relevant data can be extracted, or mined, and utilized for data analysis and report generation.

Definition:

According to William H. Inmon, a data warehouse is a “subject-oriented, integrated, time-varying, non-volatile collection of data in support of the management’s decision-making process” (Inmon, 2005, p. 32). A data warehouse is a centralized repository that stores data from multiple

information sources and transforms them into a common, multidimensional data model for efficient querying and analysis.

A data warehouse has the ability to address a wide variety of phenomena. A faculty member may ask, “How can I modify my instruction to help my students learn to write more effective essays?” A manager in the Department of Human Resources may ask, “What kind of training or orientation is necessary for new employees?” Individuals from administration may ask, “Are students who attend classes full-time more likely to succeed academically than those who take classes on a part-time basis?” Each of these questions requires more information about the situation in order to conduct research. This information originates from the data warehouse.

A data warehouse is a storehouse of an organization’s historical data. Information from operational systems is extracted and imported into the data warehouse on a regular basis. As a result, complex inquiries, or queries, can be conducted through the data warehouse with minimal interruptions to the operational systems. The imported data is read-only and only adds to the data existing in the data warehouse. With greater amounts of data, the value of the data warehouse to the user increases, since analyses looking over a longer period of time become possible. When a user query is submitted to the warehouse, all relevant historical data addressing that query is readily available and current to support in the decision-making.

Goals:

The fundamental goal of the data warehouse is to support strategic planning, modeling and forecasting at the organizational level. It must fulfill the need for knowledge for an area of uncertainty or growth in the organization. In order to accomplish this task, it must provide a single, comprehensive and consistent view of the organization.

The data must be easily accessible and understandable to the user. It must be simple yet intuitive, quick, and easy to use. Also, the data warehouse must present the information consistently and securely to its users. When data is collected from the source systems, it must complete various measures of quality assurance to confirm its accuracy. The data must be verified, appropriately labeled, and fully accounted for before it can be made available to the users. Also, the data must be resilient and able to seamlessly adapt to change without discrediting the existing data. Effective data warehousing can help create a meaningful relationship between information technology and business, facilitating enterprise-level strategic planning and growth (Cohen, 2006).

Components:

A data warehouse has four main components: operational systems of record, the data staging area, the data presentation area, and data access tools (Kimball & Ross, 2002, p. 7). Each component of the data warehouse serves a unique function in preparing data for manipulation and examination.

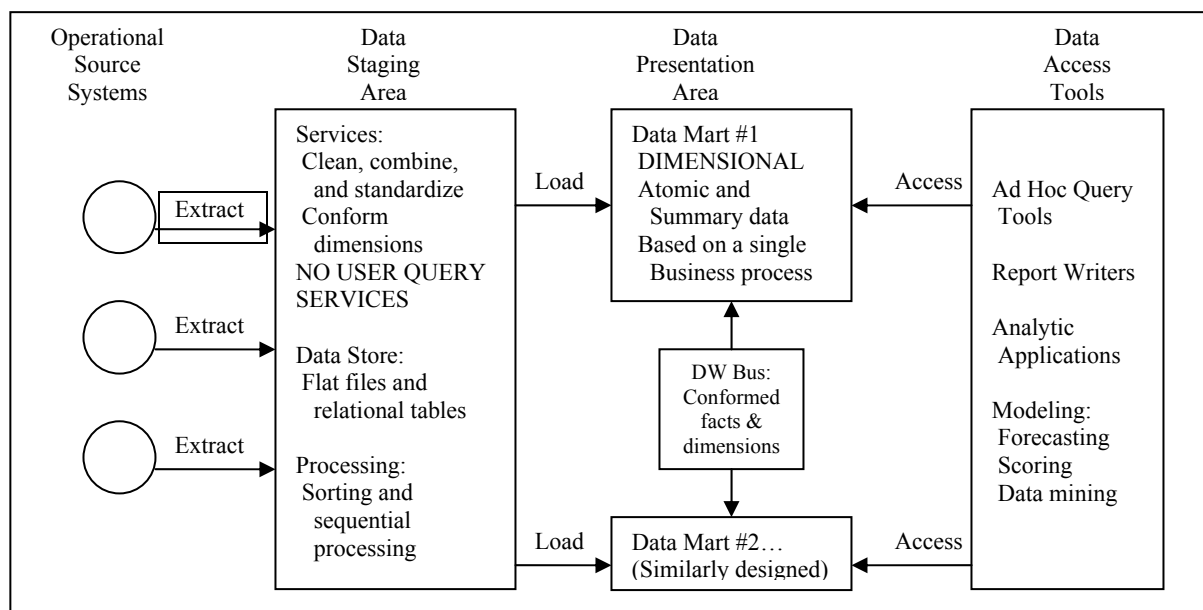


Figure 1. Basic elements of the data warehouse (Kimball & Ross, 2002, p. 7).

As aforementioned, the operational systems of records, or source systems, capture and process the organization's day-to-day transactions. They concentrate heavily on efficient processing performance, since they are dealing with a high volume of transactions. They function in isolation and do not typically share common data with other source systems. The data that is acquired through these systems is uploaded into the data staging area.

The data staging area acts doubly as a storage area for the captured data and as a platform for the set of processes called extract-transformation-load (ETL). This set of processes occurs to standardize the raw data and incorporate it into the data warehouse environment. First, the data is extracted from various source systems and copied into the data staging area. There, the data is combined, cleansed and transformed into a standard format and structure. Missing elements, incorrect labels, duplicate data, misspellings, and other errors are manipulated and corrected in this phase. Once the data is standardized, it is loaded into the data presentation area, where it is finally accessible to users.

The formatted data is organized, located and available for user queries in the data presentation area. The data presentation area is considered to be a set of integrated data marts. A data mart is a subset of the data warehouse and represents select data regarding a specific business function (Inmon, 1999). An organization can have multiple data marts, each one relevant to the department for which it was designed. For example, the English department may have a data mart reflecting historical student data including demographics, placement scores, academic performance, and class schedules. The data contained in the data presentation area must be detailed and logically organized.

Once the data presentation area contains the formatted data, users can utilize various data access tools to perform queries. Some data access tools include ad hoc query tools, data mining

applications and sophisticated forecasting tools. Users can use these tools to customize queries to search specific segments of the data presentation area.

In addition to these components of a data warehouse, it is imperative to discuss the importance of a strong metadata structure. Metadata, or data about the data, contains vital information that guides the process of converting the raw data from the operational systems of record into accessible data in the data presentation area (Kimball & Ross, 2002, p. 14). Due to its value, the metadata resources must be as carefully categorized, protected, and accessible as the data itself.

Analysis:

Data warehouses are effective in the transformation from intuitive information gathering to systematic and objective investigation (Zikmund, 2003, p. 5). They provide users access and control to a wide variety of centralized and formatted data to choose the best course of action and support business decisions. Users can manipulate and customize the data to support specific queries that will enable positive changes at various business levels. Since the various stages increase data accuracy and integrity, complex queries can be conducted with a strong sense of confidence.

Although there are many benefits to data warehousing, there are several challenges and drawbacks as well. Depending on their design, data warehouses can be highly risky since they have complex architectures, long development cycles, poor information quality, and an incapability to adapt as quickly as business conditions change. Furthermore, since the operational source systems provide the data that eventually makes its way into the data presentation area, data warehouses are limited by these source systems. Thus, each organization should focus on

continuous evaluation and improvement of its data warehouse, as well as its source systems, to ensure its effectiveness in conducting research and supporting business decisions.

Conclusion:

Since the primary task of management is effective decision making, the primary task of research, and subsequently data warehouses, is to generate accurate information for use in that decision making. It is imperative that an organization's data warehousing strategies reflect changes in the internal and external business environment in addition to the direction in which the business is traveling. Playing an integral role in the growth, development and success of an organization, data warehouses facilitate meaningful research which facilitates effective management.

References:

- [1] Cohen, Rich (2006). Business Intelligence Strategy: Seven Principles for Enterprise Data Warehouse Design. *DM Review*. Retrieved December 18, 2006, from http://www.dmreview.com/article_sub.cfm?articleId=1045818.
- [2] Inmon, William H., Building the Data Warehouse, 4th Edition, Wiley Publishing, Indianapolis, 2005.
- [3] Inmon, William H. (1999). Data Mart Does Not Equal Data Warehouse. *DM Review*. Retrieved January 2, 2007 from http://www.dmreview.com/article_sub.cfm?articleId=1675.
- [4] Kimball, Ralph; Ross, Margy. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, John Wiley and Sons, Inc., Chichester, 2002.
- [5] Zikmund, William G., Business Research Methods, 7th Edition, Dryden Press, New York, 2003.