: What is a data warehouse?

by Dan Power

Editor, DSSResources.com

A data warehouse is a database designed to support decision making in organizations. It stores historical operating data used for reporting and for generating business intelligence. IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse" concept in the late 1980s. In 1990, Bill Inmon published his influential book Building the Data Warehouse. A data warehouse is a data management strategy organizations use to collect, store, and analyze large amounts of historical business data. There are two major approaches to storing data in a data warehouse — the dimensional approach and the normalized approach. A hybrid or combination approach is sometimes used.

The dimensional approach of Ralph Kimball advocates modelling a data warehouse using a Dimensional Model/star schema and the normalized approach also called the 3NF model suggested by Bill Inmon argues a data warehouse should be modeled using an E-R model/normalized model.

In a dimensional approach, transaction data are partitioned into either "facts", which are generally numeric transaction data, or "dimensions", which are the reference information that gives context to the facts. For example, a sales transaction can be broken up into facts such as the number of products ordered and the price paid for the products, and into dimensions such as order date, customer name, product number, order ship-to and bill-to locations, and salesperson responsible for receiving the order.

Bill Inmon, one of the first authors on the subject of data warehousing, has defined a data warehouse as a centralized repository for the entire enterprise. Inmon is one of the leading proponents of the top-down approach to data warehouse design, in which the data warehouse is designed using a normalized enterprise data model. Inmon defines a data warehouse as subject-oriented, non-volatile, integrated, and time-variant.

In 1994, Inmon and Kelley created a list of 12 rules defining a data warehouse:

- The data warehouse and operational environments are seperated.

- The data warehouse data are integrated.

- The data warehouse contains historical data over a long time horizon.

- The data warehouse data are snapshot data captured at a given point in time.

- The data warehouse data are subject-oriented.

- The data warehouse data are mainly read-only with periodic batch updates from operational data. No online updates are allowed.

- The data warehouse development life cycle differs from classical systems development. The data warehouse development is data-driven; the classical approach is procee-driven.

- The data warehouse contains data with several levels of detail: current detail data, old detail data, lightly summarized , and highly summarized data.

- The data warehouse environment is characterized by read-only transactions to very large data sets. The operational environment is characterized by numerous update transactions to a few data entities as a time.

- The data warehouse environment has a system that traces data sources, transformations, and storage.

- The data warehouse's metadata are a critical component of the environment. The metadata identify and define all data elements. The metadataprovide the source, transformation, integration, storage, usage, relationships, and history of each data element.

- The data warehouse contains a chargeback mechanism for resource usage that enforces optimal use of the data by end users.

Inmon recommends starting with building a large centralized enterprise-wide data warehouse (EDW) with smaller departmental databases to serve the analytical needs of departments (later known as "data marts"). Hence, his approach is called "top down".

Kimball's philosophy recommends starting with building several data marts that serve the analytical needs of departments, followed by "virtually" integrating these data marts for consistency through an Information Bus. So no EDW and a "bottom up" incremental approach.

: What is a data warehouse?

A data warehouse is used as the data store of a data-driven decision support information system.

References

Data warehouse, from Wikipedia, the free encyclopedia.
http://en.wikipedia.org/wiki/Data_warehouse

Devlin, Barry A. and Paul T. Murphy: An Architecture for a Business and Information System. *IBM Systems Journal* vol. 27(1), 1988, pp. 60-80.

Inmon, W. H., Building the Data Warehouse, Boston: QED Information Sciences, Inc., 1990.

Inmon, W. H., Using Oracle to Build Decision Support Systems, Boston: QED Information Sciences, Inc., 1990.

Inmon, W. H., Database Machines and Decision Support Systems: Third Wave Processing, Boston: QED Information Sciences, Inc., 1991.

Inmon, W.H. and C. Kelley, "The Twelve Rules of Data Warehouse for a Client/Server World." *Data Management Review*, vol. 4(5), May 1994, pp. 6-16.

Rob, P. and C. Coronel, Database Systems, Cambridge, MA: Course Technology, 2000.

Kimball, Ralph and Ross, Margy. The Data Warehouse Toolkit Second Edition, John Wiley and Sons, Inc., 2002.

W.H. Inmon, Using Oracle to Build Decision Support Systems (1990)

This book covers the issues of ORACLE in the end-user environment. These issues include: the credibility of data, the accessibility of data, the ability to create views of data spontaneously, the ability to manage large amounts of data effectively, and the ability to manage data efficiently in a read-only mode. Following these guidelines allows construction of an effective DSS/ORACLE environment that addresses these issues. The author looks at the long-term consequences of building DSS systems across an organisation and gives solid reasons why organisations should resist the temptation to build haphazard DSS systems quickly. This book provides the first discussion of how to use the data accumulated by operational systems and shows how to apply ORACLE in the real world. IT managers, developers, database administartors and information centre professionals will find this book of great use.

W. H. Inmon, Third Wave Processing: Database Machines and Decision Support Systems

Identifies obstacles to effective DSS processing, demonstrating how database machines provide efficient and potent paths around these obstructions. Using Teradata as an example, it explains the benefit of database devices, how they support atomic level data, manage large amounts of data, expand economically as atomic data level grows and how they access such diverse sys tems as DB2 and LANs.

Author: Daniel Power
Last update: 2010-11-28 12:28