Data-driven DSS: What do I need to know about Data Warehousing/OLAP?

The answer to this question depends upon who is asking. Managers need to be familiar with some DW/OLAP terminology (the basic "what is" questions) and they need to have an idea of the benefits and limitations of these decision support components (the why questions). More technical people in Information Systems need to know how and when to develop systems using these components. This short DW/OLAP FAQ consolidates answers from some recent email questions and from a number of questions previously answered at DSSResources.COM. The bias in this FAQ is definitely towards what managers need to know. Some more technical question related to DW/OLAP were answered in the Ask Dan! of February 17, 2002. That Ask Dan! answered the following questions: Is a Data Warehouse a DSS? What is a star schema? How does a snowflake schema differ from a star schema? Also, for people who want definitions for technical terms like derived data, hypercube, pivot and slice and dice the OLAP Council glossary (1995) is online at http://dssresources.com/glossary/olaptrms.html.

Q. What is a Data Warehouse?

A. A data warehouse is a database designed to support a broad range of decision tasks in a specific organization. It is usually batch updated and structured for rapid online queries and managerial summaries. Data warehouses contain large amounts of historical data. The term data warehousing is often used to describe the process of creating, managing and using a data warehouse.

Q. What is On-line Analytical Processing (OLAP)?

A. OLAP is software for manipulating multidimensional data from a variety of sources. The data is often stored in a data warehouse. OLAP software helps a user create queries, views, representations and reports. OLAP tools can provide a "front-end" for a data-driven DSS.

Q. What is the difference between data warehousing and OLAP?

A. The terms data warehousing and OLAP are often used interchangeably. As the definitions suggest, warehousing refers to the organization and storage of data from a variety of sources so that it can be analyzed and retrieved easily. OLAP deals with the software and the process of analyzing data, managing aggregations, and partitioning information into cubes for in-depth analysis, retrieval and visualization. Some vendors are replacing the term OLAP with the terms analytical software and business intelligence.

Q. When should a company consider implementing a data warehouse?

A. Data warehouses or a more focused database called a data mart should be considered when a significant number of potential users are requesting access to a large amount of related historical information for analysis and reporting purposes. So-called active or real-time data warehouses can provide advanced decision support capabilities.

Q. What data is stored in a data warehouse?

A. In general, organized data about business transactions and business operations is stored in a data warehouse. But, any data used to manage a business or any type of data that has value to a business should be evaluated for storage in the warehouse. Some static data may be compiled for

initial loading into the warehouse. Any data that comes from mainframe, client/server, or web-based systems can then be periodically loaded into the warehouse. The idea behind a data warehouse is to capture and maintain useful data in a central location. Once data is organized, managers and analysts can use software tools like OLAP to link different types of data together and potentially turn that data into valuable information that can be used for a variety of business decision support needs, including analysis, discovery, reporting and planning.

Q. Database administrators (DBAs) have always said that having non-normalized or de-normalized data is bad. Why is de-normalized data now okay when it's used for Decision Support?

A. Normalization of a relational database for transaction processing avoids processing anomalies and results in the most efficient use of database storage. A data warehouse for Decision Support is not intended to achieve these same goals. For Data-driven Decision Support, the main concern is to provide information to the user as fast as possible. Because of this, storing data in a de-normalized fashion, including storing redundant data and pre-summarizing data, provides the best retrieval results. Also, data warehouse data is usually static so anomalies will not occur from operations like add, delete and update of a record or field.

Q. How often should data be loaded into a data warehouse from transaction processing and other source systems?

A. It all depends on the needs of the users, how fast data changes and the volume of information that is to be loaded into the data warehouse. It is common to schedule daily, weekly or monthly dumps from operational data stores during periods of low activity (for example, at night or on weekends). The longer the gap between loads, the longer the processing times for the load when it does run. A technical IS/IT staffer should make some calculations and consult with potential users to develop a schedule to load new data.

Q. What are the benefits of data warehousing?

A. Some of the potential benefits of putting data into a data warehouse include: 1. improving turnaround time for data access and reporting; 2. standardizing data across the organization so there will be one view of the "truth"; 3. merging data from various source systems to create a more comprehensive information source; 4. lowering costs to create and distribute information and reports; 5. sharing data and allowing others to access and analyze the data; and 6. encouraging and improving fact-based decision making.

Q. What are the limitations of data warehousing?

A. The major limitations associated with data warehousing are related to user expectations, lack of data and poor data quality. Building a data warehouse creates some unrealistic expectations that need to be managed. A data warehouse doesn't meet all decision support needs. If needed data is not currently collected, transaction systems need to be altered to collect the data. If data quality is a problem, the problem should be corrected in the source system before the data warehouse is built. Software can provide only limited support for cleaning and transforming data. Missing and inaccurate data can not be "fixed" using software. Historical data can be collected manually, coded and "fixed", but at some point source systems need to provide quality data that can be loaded into the data warehouse without manual clerical intervention.

Data-driven DSS: What do I need to know about Data Warehousing/OLAP?

Q. How does my company get started with data warehousing?

A. Build one! The easiest way to get started with data warehousing is to analyze some existing transaction processing systems and see what type of historical trends and comparisons might be interesting to examine to support decision making. See if there is a "real" user need for integrating the data. If there is, then IS/IT staff can develop a data model for a new schema and load it with some current data and start creating a decision support data store using a database management system (DBMS). Find some software for query and reporting and build a decision support interface that's easy to use. Although the initial data warehouse/data-driven DSS may seem to meet only limited needs, it is a "first step". Start small and build more sophisticated systems based upon experience and successes.

Some of the questions and answers in this DW/OLAP FAQ are based on material in "Microsoft SQL Server 7.0 DBA Survival Guide" by SAMS.

The above response is from Power, D., What do I need to know about Data Warehousing/OLAP? DSS News, Vol. 4, No. 5, March 02, 2003.

Author: Daniel Power
Last update: 2005-08-06 21:59