Data Warehouse

By Anupama Kota
Computer Science Dept

Introduction

2.1 Introduction

The data warehouse is a new approach to enterprise-wide computing at the strategic or architectural level. A data warehouse can provide a central repository for large amounts of diverse and valuable information. By filing the data into a central point of storage, the data warehouse provides an integrated representation of the multiple sources of information dispatch across the enterprise. It ensures the consistency of management rules and conventions applied to the data. It also provides the appropriate tools to extract specific data, convert it into business information, and monitor for changes and, hence, it is possible to use this information to make insightful decisions. A data warehouse is a competitive tool that gives every end user the ability to access quality enterprise-wide data. A data ware house ware house supports business analysis and decision making by creating an enterprise wide integrated data base of summarized historical information. It integrates data from multiple incompatible sources (non relative).

Operation Vs Analytical Applications

Databases are created to store data, but the way they are designed depends on your business objectives. Most business applications store data in an OLTP (On-Line Transaction Processing) database, which is accessed by numerous users to perform fast, simple queries. When you go to the supermarket, the Point-of-Sale system at the cash register uses an OLTP database. Another example is that of a bank, with tellers storing data for each transaction in an OLTP system. Even when you send a text from your smartphone to someone, an OLTP system is running on the backend.

Operation Vs Analytical Applications

OLTP is designed to store day-to-day business transactions and is well-suited for querying specific records, for instance, the email address of customer ABC. Thousands of such queries can be run simultaneously on an OLTP database, but when you need a strategic view of business data, queries start to get increasingly complex and require aggregations among numerous tables. For instance, a query for compiling year-over-year profits is best suited for an OLAP (On-Line Analytical Processing) database, which provides a multi-dimensional view of enterprise data rather than a transaction-level view.

Together, OLTP and OLAP form the two sides of the data warehousing coin. OLTP systems are the original, disparate data sources across the enterprise. On the other hand, OLAP systems integrate data from these transactional sources and present a multi-dimensional view for reporting and analytics.

Operational vs. Analytical

Operational databases

- Used to track and assist in daily "business" activities
- Data typically changes frequently over time
- Examples
  - Human resources
  - Mailing lists
  - Inventory management
  - Accounting systems
  - Point of sale systems (cash registers)

Analytical databases

- Tend to be more static
- Historical data is analyzed for patterns or trends
- Often support the strategic activities of an organization
- Goals may include
  - Predicting the future
  - Summarizing historical data
  - Prove historical assumptions

Data Warehouse Definition

2.2 Definition

W H Inmon (1993) offers definition of a data warehouse as "subject oriented, integrated, time varying, non-volatile collection of data in support of the management's decision making process"

Subject Oriented: A data ware house is organized around major subjects such as customer, product, sales etc. Data are organized according to the subject instead of application so that the information necessary for the decision support processing is available easily. For example an insurance company keeps the details of the customers, the premiums paid and the claims made by the customer.

Non – Volatile: A data warehouse is always a physically separate store of data. The data are not updated or changed in any way once they enter data ware house but are only loaded, refreshed and accessed for queries.

Time Variant: Data is stored in a data warehouse to provide a historical perspective. Every key structure (attribute) in the data warehouse contains implicitly or explicitly an element of time. The data warehouse contains a place for sorting data that are 5 to 10 years old or older for comparisons.

Integrated: A data warehouse is constructed by integrating multiple heterogeneous sources such as relational databases, flat files and OLTP files (On Line Transaction Files). When data resides in separate applications, in operational environment the encoding is inconsistent. Appropriate data cleaning and integration techniques are used to make the data consistent.

	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

Advantages and Disadvantages

Advantages

- Clean data.
- Query processing: multiple options.
- Security: data and access.

Disadvantages

- Long initial implementation time and associated high cost.
- Adding new data sources takes time and associated high cost.
- Typically, data is static and dated.