Data warehousing

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process." (W. H. Inmon 1992).

This definition presents the major features of a data warehouse:

- Subject-oriented
- Integrated
- Time-variant
- Nonvolatile

These four features distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems. Let's take a closer look at each of these key features.

Subject-oriented: A data warehouse is organized around major subjects, such as customer, vendor, product, and sales etc. rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated: A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, files, and on-line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.

Time-variant: Data are stored to provide information from a historical perspective (e.g., the past 5-10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

Nonvolatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

Data warehousing is the process of constructing and using data warehouses.

OLTP: The major task of on-line operational database systems is to perform on-line transaction and query processing. These systems are called on-line transaction processing (OLTP) systems. They cover most of the day-to-day operations of an organization, such as, purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

OLAP: Data warehouse systems, on the other hand, serve users or "knowledge workers" in the role of data analysis and decision making. Such systems can organize and present data in various

formats in order to accommodate the diverse needs of the different users. These systems are known as on-line analytical processing (OLAP) systems.

The major distinguishing features between OLTP and OLAP are summarized as follows:

1. **Users and system orientation**: An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.

2. **Data contents**: An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier for use in informed decision making.

3. **Database design**: An OLTP system usually adopts an entity-relationship (ER) data model and an application oriented database design. An OLAP system typically adopts either a star or a snowflake model, and a subject-oriented database design.

4. **View**: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

5. **Access patterns**: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly read-only operations (since most data warehouses store historical rather than up-to-date information), although many could be complex queries.

Differences are presented here in a tabular format

Feature	OLTP	OLAP
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long term informational requirements, decision support
DB design	E-R based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated

Feature	OLTP	OLAP
View	detailed, at relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
# of records accessed	tens	millions
# of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Data cube: A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

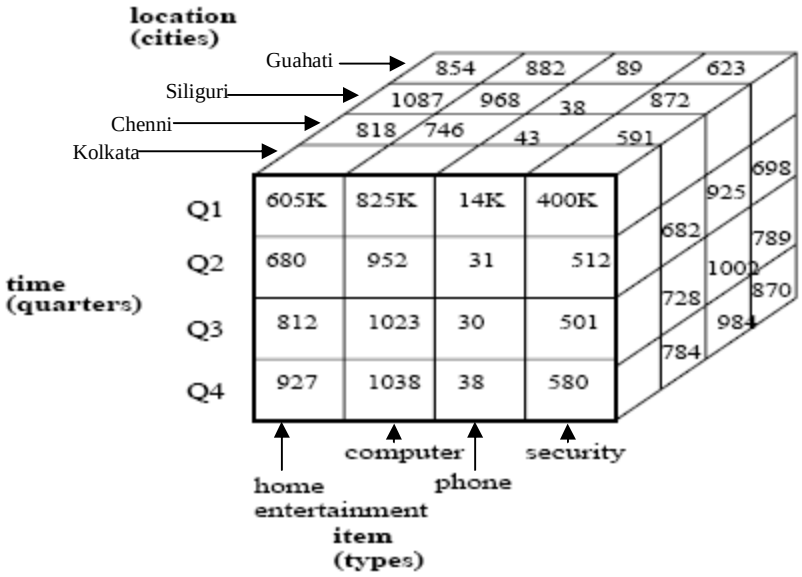
In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records. For example, ***AllElectronics*** may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch, and location. These dimensions allow the store to keep track of things like monthly sales of items, and the branches and locations at which the items were sold. Each dimension may have a table associated with it, called a **dimension table**, which further describes the dimension. For example, a dimension table for ***item*** may contain the attributes ***item_name***, ***brand***, and ***type***. Dimension tables can be specified by users or experts, or automatically generated and adjusted based on data distributions.

A 2-D view of sales data for AllElectronics: Dimensions are.
Sales for all locations in Kolkata

Time (Quarter)	Locations = Kolkata			
	Item (Type)			
	Home Entertainment	Computer	Phone	Security
Q1	605K	825K	14K	400K
Q2	680K	952K	31K	512K
Q3	812K	1023K	30K	501K
Q4	927K	1038K	38K	580K

A 3-D view of sales data for AllElectronics: Dimensions are ***Time***, ***Item*** and ***location***.
Sales for all locations in Kolkata, Chhenni, Siliguri and Guahati

Time	Locations = Kolkata				Locations = Chenni				Location = Siliguri				Location = Guahati			
	HmEnt	Comp	Ph	Sec	HmEnt	Comp	Ph	Sec	HmEnt	Comp	Ph	Sec	HmEnt	Comp	Ph	Sec
Q1	605K	825K	14K	400K	818K	746K	43K	591K	1087K	968K	38K	872K	854K	882K	89K	623K
Q2	680K	952K	31K	512K	849K	769K	52K	682K	1130K	1024K	41K	925K	943K	890K	64K	698K
Q3	812K	1023K	30K	501K	940K	795K	58K	728K	1034K	1048K	45K	1002K	1032K	924K	59K	789K
Q4	927K	1038K	38K	580K	978K	864K	59K	784K	1142K	1091K	54K	984K	1129K	992K	63K	870K



A 3-D data cube representation of the data in the above table, with the dimensions ***time***, ***item***, and ***location***.