Data Warehouse

In the 1980's organizations realized the importance of not just using data for operational purposes, but also for deriving intelligence out of it. This intelligence would not only justify past decisions but also help in making decisions for the future. The term **Business Intelligence** became more and more popular and it was during the late 1980's that IBM researchers Barry Devlin and Paul Murphy developed the concept of a **Business data warehouse**. As business intelligence applications emerged, it was quickly realized that data from transactional databases had to first be transformed and stored into other databases with a schema specific for deriving intelligence. This database would be used for archiving, and it would be larger in size than transactional databases, but its design would make it optimal to run reports that would enable large organizations to plan and proactively make decisions. This separate database, typically storing the organization's past and present activity, was termed a ***Data Warehouse***.

Large amount of operational data are routinely collected and stored away in the archives of many organizations. Similar to a real-life warehouse, a Data Warehouse gathers its data from some central source, typically a transactional database and stores and distributes this data in a fashion that enables easy analytics and report generation. The difference between a typical database and a data warehouse not only lies in the volume of data that can be stored but also in the way it is stored. Technically speaking, they use different database designs.

There are several definitions and theories for data warehouse. According to a definition, Data warehousing is a collection of *decision support* technologies, aimed at enabling the *knowledge worker* (executive, manager, analyst) to make better and faster decisions. Data warehousing and on-line analytical processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. According to W. H. Inmon, "A data warehouse is a subject-oriented, integrated, time-variant, and non volatile collection of data in support of management's decision-making process."

Data warehousing technologies have been successfully deployed in many industries: manufacturing (for order shipment and customer support), retail (for user profiling and inventory management), financial services (for claims analysis, risk analysis, credit card analysis, and fraud detection), transportation (for fleet management), telecommunications (for call analysis and fraud detection), utilities (for power usage analysis), and healthcare (for outcomes analysis).

Unlike having multiple decision-support environments operating independently, which often leads to conflicting information, a data warehouse unifies all sources of information. Data is stored in a way that integrity and quality are guaranteed. In addition to a different database design, this is accomplished by using an **Extract, Transform** and **Load** of data process -- also known as **ETL**.

The ETL process for each Data Warehouse System is defined considering a clear objective that serves a specific business purpose. Therefore, the organization's business objective must be well known in advance, as it is essential for the definition of the appropriate transformation of data. The basic goal of the ETL is to filter redundant data not required for analytic reports and to converge data for fast report generation.

Data Warehouse System is subject oriented as it organized around major subjects, such as customer, product, sales. It focuses on the modelling and analysis of data for decision makers and not for daily operations or transaction processing. It provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

It is integrated as it is constructed by integrating multiple, heterogeneous data sources like relational databases, flat files, on-line transaction records. Data cleaning and data integration techniques are applied. When data is moved to the warehouse from databases, it is converted.

The time horizon for the data warehouse is significantly longer than that of operational systems. For operational database the time horizon is current value data whereas for Data warehouse it provide information from a historical perspective (e.g., past 5-10 years). Every key structure in the data warehouse contains an element of time, explicitly or implicitly but the key of operational data may or may not contain "time element"

The non-volatile nature of data warehouse arises from the fact that it is a physically separate store of data transformed from the operational environment. Operational update of data does not occur in the data warehouse environment. It does not require transaction processing, recovery, and concurrency control mechanisms rather requires only two operations in data accessing: *initial loading of data* and *access of data.*

OLTP and OLAP Systems

Online Transaction Processing

Online Transaction Processing (OLTP) refers to workloads that access data randomly, typically to perform a quick search, insert, update or delete. OLTP operations are normally performed concurrently by a large number of users who use the database in their daily work for regular business operations. Typicaly, the data in these systems must be consistent and accurate at all times. The life span of data in an OLTP system is short since its primary usage is in providing the current snapshot of transient data. Hence, OLTP systems need to support *real-time* data insertions, updates and retrievals, and end up having large number of small-size tables.

Consider an online reservation system as an example of an OLTP system. An online user must be presented with accurate data 24 x 7. Reservations must be done in a quick fashion and any updates on the reservation status must be reflected immediately to all other users. In addition to online reservations systems, other examples of OLTP systems, include banking applications, eCommerce, and payroll applications. These systems are characterized by their simplicity and efficiency, which help enable 24 x 7-support to end users.

OLTP systems use simple tables to store data. Data is ***normalized***, that is, redundancy is reduced or eliminated while still ensuring data consistency. Data is stored in its utmost raw form for **each** customer transaction.

Online Analytical Processing

Online Analytical Processing (**OLAP**) refers to workloads where large amounts of historical data are processed to generate reports and perform data analysis. Typically, OLAP databases are fed from OLTP databases, and tuned to manage this type of workload. An OLAP database stores a large volume of the same transactional data as an OLTP database, but this data has been transformed through an ETL process to enable best performance for easy report generation and analytics. OLTP systems are tuned for extremely quick inserts, updated and deletes, while OLAP systems are tuned for quick reads only.

The lifespan of data stored in a Data Warehouse is much longer than in OLTP systems, since this data is used to reflect trends of an organization's business over time and help in decision making. Hence OLAP databases are typically a lot larger than OLTP ones. For instance, while OLTP databases might keep transactions for six months or one year, OLAP databases might keep accumulating the same type of data year over year for 10 years or more.

As compared to OLTP systems, data in an OLAP data warehouse is less normalized than an OLTP system. Usually OLAP data warehouses are in the 2nd Normal Form (2NF). The great advantage of this approach is to make database design more readable and faster to retrieve data. Some examples of OLAP applications are business reporting for sales, marketing reports reporting for management and financial forecasting.

The large size of a data warehouse makes it not economically viable to have a high availability and disaster recovery setup in place. Since OLAP systems are not used for real-time applications, having another exact replica of an existing huge system would not justify neither the costs, nor the business needs.

Differentiating Factor	**OLTP Systems**	**OLAP Systems**
Business Needs and Usage	These systems are data stores for real time transactional applications and are typically the first data entry point for any organization. They are critical for controlling and running the fundamental business tasks of an organization.	These systems are needed by an organization to generate reports, run analytics useful for decision making in a multiple decision support environment. Data in such systems is sourced from various OLTP systems and consolidated in a specific format.
Nature of Data Stored	Data stored in such systems represent the current snapshot of transient data. Data is collected real time from user applications. There is no transformation done to the data before storing it into the system.	Such systems contain historical data that is gathered from operational databases over a period. The data stored reflects the business trends of the organization and helps in forecasting. After transformation (ETL), data is generally loaded into such systems periodically
Database Tuning	Database is tuned for extremely fast inserts, updates and deletes.	Database is tuned only for quick reads.
Data Lifespan	Such systems deal with data of short lifespan.	Such systems deal with data of very large lifespan (historic).
Data Size	Data in OLTP systems is raw and it is stored in numerous	Data in OLAP systems is first transformed and usually stored

	but small-size tables. The data size in such systems is hence not too big.	in the form of fact and dimension tables. The data size of such systems is huge.
Data Structure	Data is stored in the highest normalized for possible. Usually 3NF.	Data is somewhat denormalized to provide better performance. Usually under 2NF (important: this denormalization applies to dimension tables only).
Data Backup and Recovery	One of the main requirements of an OLTP system is its reliability. Since such systems control the basic fundamental tasks of an organization, they must be tuned for high availability and data recovery. Such systems cannot afford to go offline since they often have mission critical applications running on them. Hence, an HADR setup with the primary and secondary (with their respective storage) installed in different geographies is recommended.	Such systems need not require high availability and data may be archived in external storage such as tapes. In case such a system goes down, it would not necessarily have a critical impact on any running business. Data can be reloaded from archives when the system comes up again.
Examples	Banking Applications, Online Reservations systems, ecommerce etc.	Reporting for sales, marketing, management reporting and financial forecasting.