Introduction to data warehouse and Data mining

1. **Concepts of Data Warehouse:**
   The term "Data Warehouse" was first coined by Bill Inmon in 1990.

- ❖ **Defn:** A data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data.

- ❖ **Defn:** Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

- ❖ **Facts:**

a. A data warehouse is a database, which is kept separate from the organization's operational database.
b. There is no frequent updating done in a data warehouse.
c. It possesses consolidated historical data, which helps the organization to analyze its business.
d. A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.

- ❖ **Data Warehouse Features**

The key features of a data warehouse are

1) **Subject Oriented** − A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations; rather it focuses on modeling and analysis of data for decision making.
2) **Integrated** − A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.
3) **Time Variant** − the data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.
4) **Non-volatile** − Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore a frequent change in operational database is not reflected in the data warehouse.

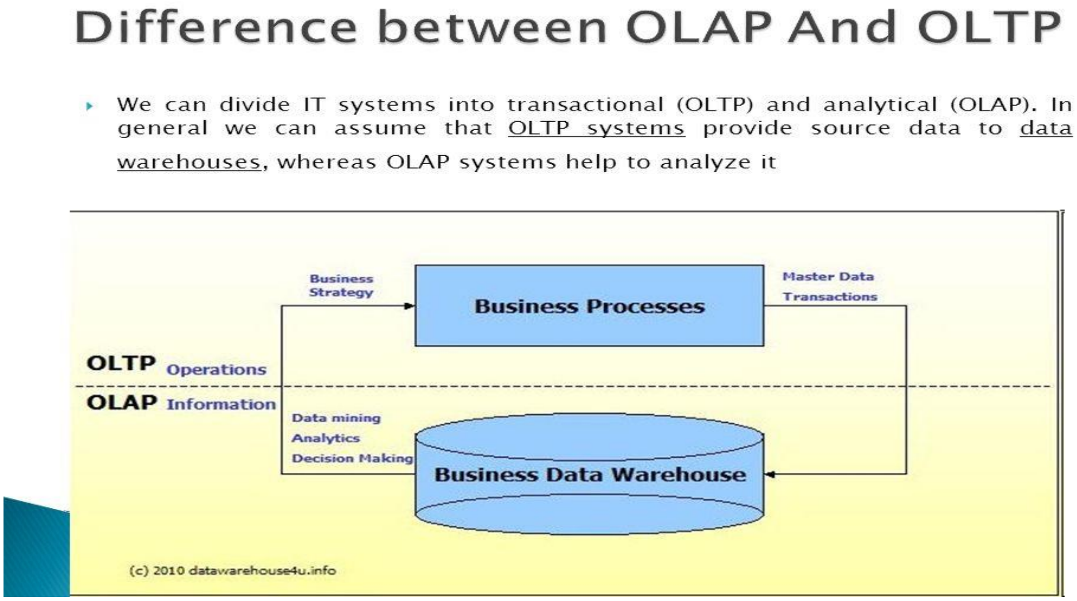
Data warehouses have the following distinctive characteristics.

• Multidimensional conceptual view. • Generic dimensionality. • Unrestricted cross-dimensional operations.
• Dynamic sparse matrix handling. • Client-server architecture. • Multi-user support.
• Accessibility. • Transparency. • Consistent reporting performance. • Flexible reporting

❖ Why a Data Warehouse is Separated from Operational Databases

1) An operational database is constructed for well-known tasks and workloads such as searching particular records, indexing, etc. In contract, data warehouse queries are often complex and they present a general form of data.
2) Operational databases support concurrent processing of multiple transactions. Concurrency control and recovery mechanisms are required for operational databases to ensure robustness and consistency of the database.
3) An operational database query allows reading and modifying operations, while an OLAP query needs only read only access of stored data.
4) An operational database maintains current data. On the other hand, a data warehouse maintains historical data.

Sr.No.	Data Warehouse (OLAP)	Operational Database(OLTP)
1	It involves historical processing of information.	It involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	It is used to analyze the business.	It is used to run the business.
4	It focuses on Information out.	It focuses on Data in.
5	It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
6	It focuses on Information out.	It is application oriented.
7	It contains historical data.	It contains current data.
8	It provides summarized and consolidated data.	It provides primitive and highly detailed data.
9	It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
10	The number of users is in hundreds.	The number of users is in thousands.
11	The number of records accessed is in millions.	The number of records accessed is in tens.
12	The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
13	These are highly flexible.	It provides high performance.



Difference between OLAP And OLTP

- We can divide IT systems into transactional (OLTP) and analytical (OLAP). In general we can assume that OLTP systems provide source data to data warehouses, whereas OLAP systems help to analyze it

❖ Types of Data Warehouse

Information processing, analytical processing and data mining are the three types of data warehouse applications that are discussed below −

- **Information Processing** − A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.
- **Analytical Processing** − A data warehouse supports analytical processing of the information stored in it. The data can be analyzed by means of basic OLAP operations, including slice-and-dice, drill down, drill up, and pivoting.
- **Data Mining** − Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using the visualization tools.

❖ Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities −

- **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** − Involves finding and correcting the errors in data.
- **Data Transformation** − Involves converting the data from legacy format to warehouse format.
- **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** − Involves updating from data sources to warehouse.

- **The term Knowledge Discovery in Databases, or KDD** for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.
  - The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.
  - Here is the list of steps involved in the knowledge discovery process –

    I. Data Cleaning − In this step, the noise and inconsistent data is removed.

    - Data cleaning is defined as removal of noisy and irrelevant data from collection.
    - Cleaning in case of ***Missing values***.
    - Cleaning ***noisy*** data, where noise is a random or variance error.
    - Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.

    II. Data Integration − In this step, multiple data sources are combined.

    - Data integration is defined as heterogeneous data from multiple sources combined in a common source (DataWarehouse).
    - Data integration using ***Data Migration tools***.
    - Data integration using ***Data Synchronization tools***.
    - Data integration using ***ETL***(Extract-Load-Transformation) process.

    III. Data Selection − In this step, data relevant to the analysis task are retrieved from the database.

    - Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
    - Data selection using ***Neural network***.
    - Data selection using ***Decision Trees***.
    - Data selection using ***Naive bayes***.
    - Data selection using ***Clustering***, ***Regression***, etc

    IV. Data Transformation − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

    - Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

    Data Transformation is a two step process:

    - ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
    - ***Code generation***: Creation of the actual transformation program.

    V. Data Mining − In this step, intelligent methods are applied in order to extract data patterns.

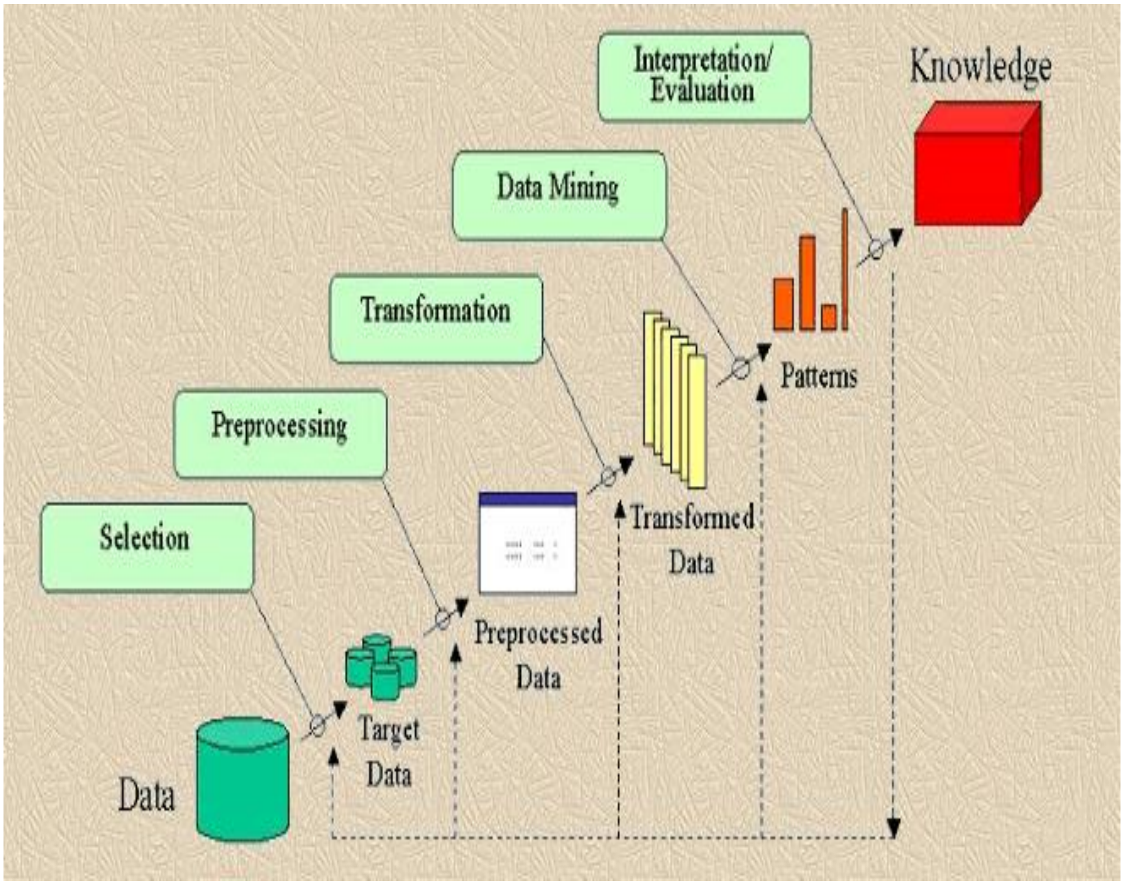
- Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
- Transforms task relevant data into ***patterns***.
- Decides purpose of model using ***classification*** or ***characterization***.

VI. Pattern Evaluation − In this step, data patterns are evaluated.

- Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.
- Find ***interestingness score*** of each pattern.
- Uses ***summarization*** and ***Visualization*** to make data understandable by user.

VII. Knowledge Presentation − In this step, knowledge is represented.

- Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
- Generate ***reports***.
- Generate ***tables***.
- Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.



- **Issues in Data Warehouse**

1. **Data Quality** – In a data warehouse, data is coming from many disparate sources from all facets of an organization. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges.
2. **Understanding Analytics -**When building a data warehouse, analytics and reporting will have to be taken into design considerations. In order to do this, the business user will need to know exactly what analysis will be performed**.**
3. **Quality Assurance –** The end user of a data warehouse is using Big Data reporting and analytics to make the best decisions possible. Consequently, the data must be 100 percent accurate.
4. **Performance-**A data warehouse must also be carefully designed to meet overall performance requirements.
5. **Designing the Data Warehouse-**People generally don't want to "waste" their time defining the requirements necessary for proper data warehouse design. This results in miscommunication between the business users and the technicians building the data warehouse.
6. **User Acceptance** -There are many challenges to overcome to make a data warehouse that is quickly adopted by an organization. Having a comprehensive user training program can ease this hesitation but will require planning and additional resources.
7. **Cost-**there are a multitude of hidden problems in building data warehouses.

- **Data Mining** :

Defn: Data mining refers to extracting or mining knowledge from large amounts of data.
Data mining is also known as data discovery and knowledge discovery.

- **Issues in Data Mining**

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place.

1. Mining Methodology and User Interaction

a) **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.

b) **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

c) **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used.

d) **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

e) **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.

f) **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities.

g) **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

2. **Performance Issues**

a) **Efficiency and scalability of data mining algorithms**− In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

b) **Parallel, distributed, and incremental mining algorithms-** These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions are merged. The incremental algorithms, update databases without mining the data again from scratch.

3. **Diverse Data Types Issues**

a) **Handling of relational and complex types of data** − the database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.

b) **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN.

❖ Data Warehouse Applications

As discussed before, a data warehouse helps business executives to organize, analyze, and use their data for decision making. A data warehouse serves as a sole part of a plan-execute-assess "closed-loop" feedback system for the enterprise management. Data warehouses are widely used in the following fields −

- Financial services
- Banking services
- Consumer goods
- Retail sectors
- Controlled manufacturing

❖ Data Mining Applications

Data mining is highly useful in the following domains −

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection