DATA MINING

Data Mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. A primary reason for using data mining is to assist in the analysis of collections of observations of behaviour. Such data are vulnerable to co-linearity because of unknown inter-relations. An unavoidable fact of data mining is that the (sub-) set(s) of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviours that exist

across other parts of the domain. To address this sort of issue, the analysis may be augmented using experiment-based and other approaches, such as Choice Modelling for human-generated data. In these situations, inherent correlations can be either controlled for, or removed altogether, during the construction of the experimental design.

Data mining commonly involves four classes of tasks:

(i) **Clustering:** Is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

(ii) **Classification:** Is the task of generalizing known structure to apply to new data. **For example**, an e-mail program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbour, naïve Bayesian classification, neural networks and support vector machines.

(iii) **Regression:** Attempts to find a function which models the data with the least error.

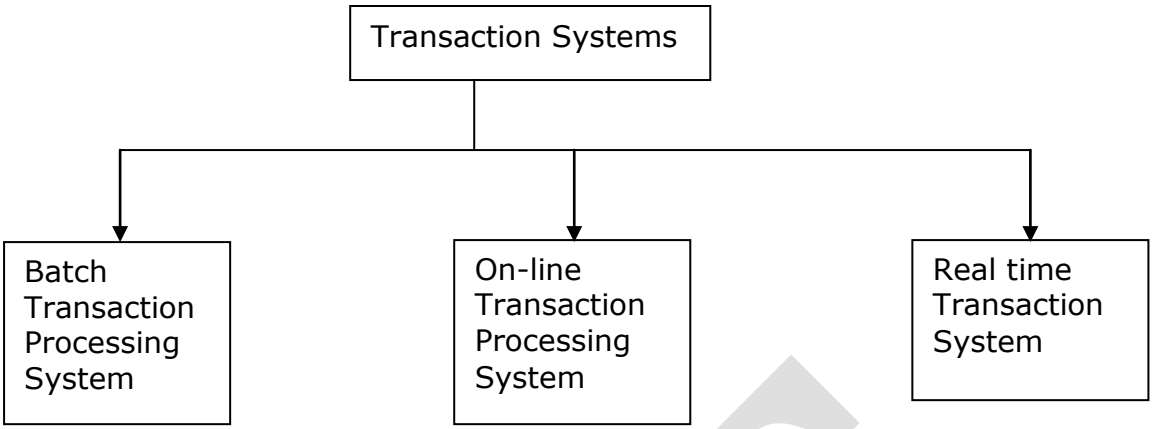
(iv) **Association rule learning:** Searches for relationships between variables. **For example,** a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

On-line Transaction Processing (OLTP): Every organization requires some on-line application system or server to manage their daily activities. These systems help in recording the transactions. The organization goes through with their employees, customer and vendors. It is impossible to imagine an enterprise without an online transaction system.

TRANSACTION

A transaction is nothing but an interaction between different users, or different systems or user and a system.

Transaction Systems: The transaction systems which mimic real life system like 'Salary processing, library, banking airline, defence missile systems.



1. **Batch Transaction Processing System:** Example: Salary Slip generation.

2. **On-line Transaction Processing System:** Example: Air-line reservation, Railway Reservation, Banking ATM machine.

3. **Real Time Transaction Processing System:** Example: Air traffic control system; Missile defense system.

Transaction Properties: These properties are called ACID.

Atomicity: Transaction should either **completely succeed or completely fail.** For any reasons, if the system crashes before the completion of the transaction, the database state should not change. The data, which was involved with the transaction, should be restored to the previous consistent state in the database. The transaction is indivisible or undividable which means it cannot be divided further into subtasks.

Consistency: Transactions must preserve database consistency or stability. A transaction transforms the database from one consistent state to another consistent state.

Isolation: A transaction's operations like INSERT, SELECT, UPDATE and DELETE should not interfere with other transactions, or in other words it should not interfere with transactions of other users of the database. The database system should reveal the individual changes made by a transaction only after a transaction completed successfully.

	OLTP	**OLAP**
Definition	On Line Transaction Processing	On Line Analytical Processing
Data	Dynamic (day to day transaction / operational data)	Static (historical data)
Data Atomicity	Data is stored at microscopic level	Data is aggregated or summarized and stored at the higher level
Normalization	Normalized Databases to facilitate insertion, deletion and updation	De-normalized Databases to facilitate queries and analysis
History	Old data is purged or archived	Historical data stored to enable trend analysis and future predictions
Queries	Simple queries and updates Queries use small amounts of data (one record or a few records) **Example:** update account	Complex queries Queries use large amounts of data **Example:** Total annual sales for north

	balance enroll for a course	region. Totally monthly sales for north region.
Updates	Updates are frequent	Updates are infrequent
Response Time	Fast response time is important Data must be up-to-date, consistent at all times	Transactions are slow Queries consume a lot of bandwidth
Joins in queries	Joins are more and complex as tables are normalized	Joins are few and simple as tables are de-normalized
	An OLTP system aims at one specific process **Example:** Ordering from an online store	An OLAP integrates data from different processes **Example:** Combines sales, inventory and purchasing data
Data Models	Complex data models, many tables	Simple data models, fewer tables
Focus	OLTP focuses on performance	OLAP focuses on flexibility and broader scope

Durability: Once a transaction completes (commits), the changes made to database are permanent and available to all the transactions that follow it.

Concurrency: Concurrency means allowing different transactions to execute simultaneously.

Dead Lock: Deadlock is a situation where one transaction is waiting for another transaction to release the resource it needs, and vice versa. Each transaction will be waiting forever for the other to release the resource.

If a deadlock occurs, one of the participating transactions must be rolled back to allow the other to proceed. There are various methods to choose which transaction to roll back when a deadlock is detected. Usually rollback action is decided on:

- How long the transaction have been running
- Data already updated by the transaction
- Data that remains to be updated by the transaction.

There are schemes available for preventing deadlock. Most of the RDBMS products allow deadlocks to occur and resolve them, when they are detected.

ON LINE ANALYTICAL PROCESSING (OLAP): An organization's success also depends on its ability to analyze data and to make intelligent decisions that would potentially affect its future. Systems that facilitate such analysis are called On Line Analytical Processing (OLAP) systems.

Data Warehouse: A data warehouse is a repository which stores integrated information for efficient querying and analysis. Data warehouse has data collected from multiple, disparate sources of an organization. It is the basis for decision support and data analysis systems.

Purpose of data Warehouse:

- Analysis requires millions of records of data which are historical in nature.
- Data is collected from heterogeneous sources (e.g. RDBMS, flat files etc.)

- Need to make quick and effective strategic decisions

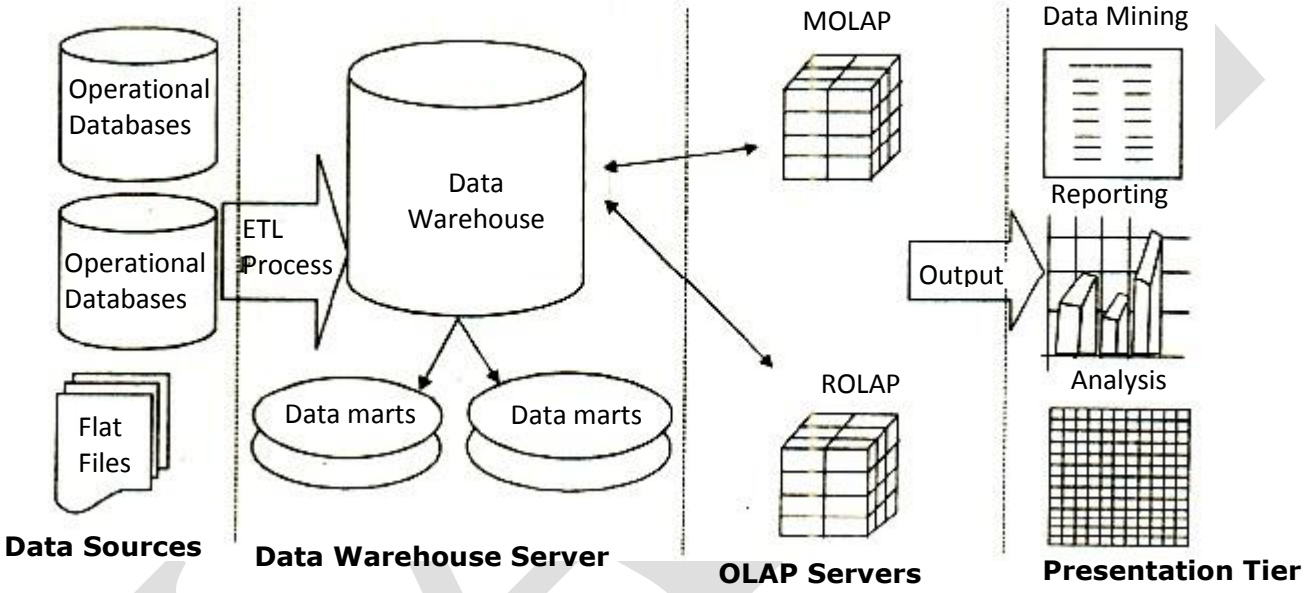
Characteristics of Data ware House:

(i) **Subject-oriented:** Means that all data pertinent to a subject / business area are collected and stored as a single unit.

(ii) **Integrated:** Means that data from multiple disparate sources are transformed and stored in a globally accepted fashion

(iii) **Static / Non-volatile:** Means data once entered into the warehouse does not change. It is periodically added if required.

(iv) **Time Variant:** Data warehouse maintains historical data which are used to analyze the business or market trends and facilitate future predictions.



Data Warehouse Architecture

Data Collection for Data Warehouse Applications:

(i) Extraction, transformation and loading (ETL): This is the most important step in Data Warehousing.

ETL: The process such as Extract, Transform and load can be described as the process of selecting, migrating, transforming, cleaning and cerverting mapped data from the operational environment to data warehouse environment.

Storing of data in Data Warehouse:

(i) **Dimensional Modeling:** The dimensional modeling is also known as star schema because in dimensional modeling there is a large central fact table with many dimension tables surrounding it.

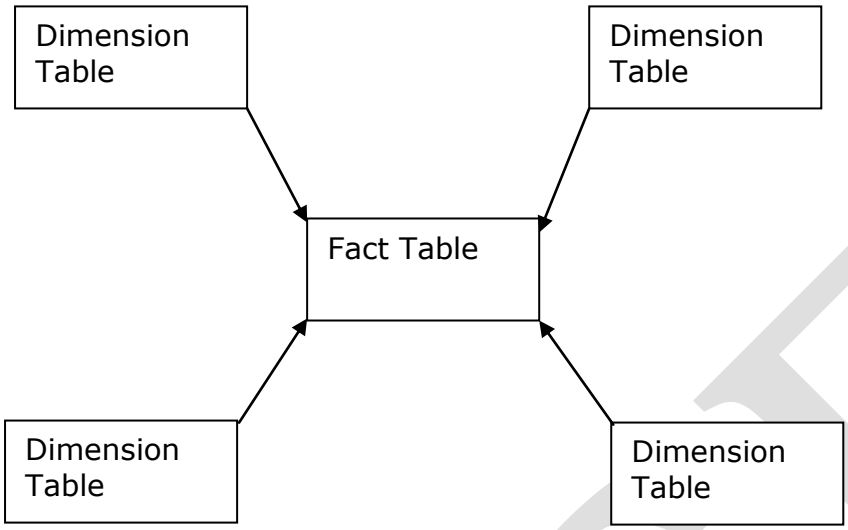
Available schemas for dimensional modeling:

(i) Star Schema (ii) Snowflake Schema

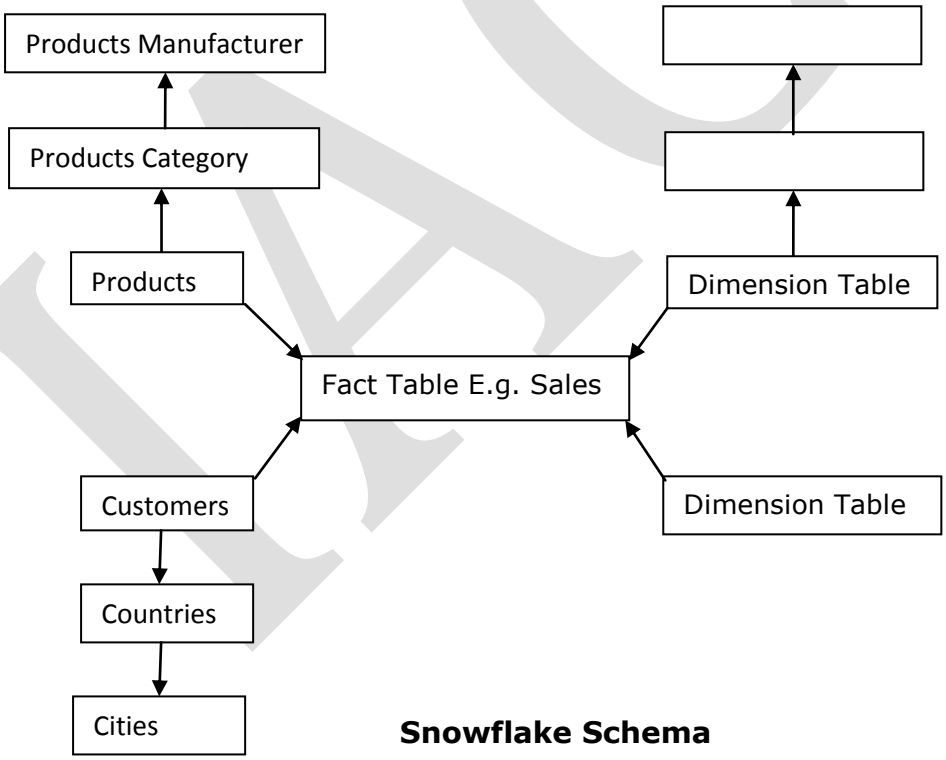
(iii) **Dimension Table:** The attributes in these tables describe the fact records in the fact table. It contains attributes which summarize the useful information required by the analyst.

Star Schema: It is the simplest data warehouse schema. It resembles a star. The center of star consists of at least one or more fact tables and the points radiating from the center the dimension tables.

Star Schema



Snowflake Schema: It is a complex data warehouse schema. The snowflake schema consists of a single, central fact table, which is surrounded by dimensional hierarchies which are normalized. Each level of the dimension is represented in a table.



Snowflake Schema

Disadvantages of Snowflake Schema:

(i) It increase the number of dimension tables (ii) It requires more foreign key joins

Difference between Data Warehouse and Data Mart:

DATA WAREHOUSE	DATA MART
A data warehouse is a repository which stores integrated information from multiple disparate sources for efficient querying and analysis	A data mart is a subset of a data Warehouse which focuses on a single area of data and it is organized for quick analysis.
It mainly focuses on the organization of data and offers little focus about the presentation of data.	It focuses mainly on the presentation of data to the customers rather than the way in which the data is organized in the data warehouse
There is usually a central data warehouse system	There can be several data marts that operate on the central data Warehouse
Data Warehouse is used on an enterprise level	Data Mart is used on a business division / department level
Data Warehouse contains data from heterogeneous sources for analysis	Data Mart only contains the required subject specific data for local analysis

******* IACE *******