Basic Definitions

- **Database:**
  - A collection of related data.
- **Data:**
  - Known facts that can be recorded and have an implicit meaning.
- **Mini-world:**
  - Some part of the real world about which data is stored in a database. For example, student grades and transcripts at a university.
- **Database Management System (DBMS):**
  - A software package/system to facilitate the creation and maintenance of a computerized database.
- **Database system:**
  - The DBMS software together with the data itself. Sometimes, the applications are also included.

Impact of Databases and Database Technology

- Businesses: Banking, Insurance, Retail, Transportation, Healthcare, Manufacturing
- Service industries: Financial, Real-estate, Legal, Electronic Commerce, Small businesses
- Education : Resources for content and Delivery
- More recently: Social Networks, Environmental and Scientific Applications, Medicine and Genetics
- Personalized applications: based on smart mobile devices

Drawback of Flat File

- Data Redundancy
- Data Inconsistency
- Data Isolation
- Dependency in Application Programs
- Atomicity issues
- Data Security

Advantages of DBMS

- No redundant data
- Data Consistency and Integrity
- Data Security
- Privacy
- Easy access
- Easy Recovery
- Flexible

Keys in DBMS

- Keys play an important role in the relational database.
- It is used to uniquely identify any record or row of data from the table.
- It is also used to establish and identify relationships between tables.

Types of Keys

- **Super Key –** A super key is a group of single or multiple keys which identifies rows in a table.
- **Primary Key –** is a column or group of columns in a table that uniquely identify every row in that table.
- **Candidate Key –** is a set of attributes that uniquely identify tuples in a table. Candidate Key is a super key with no repeated attributes.

Types of Keys

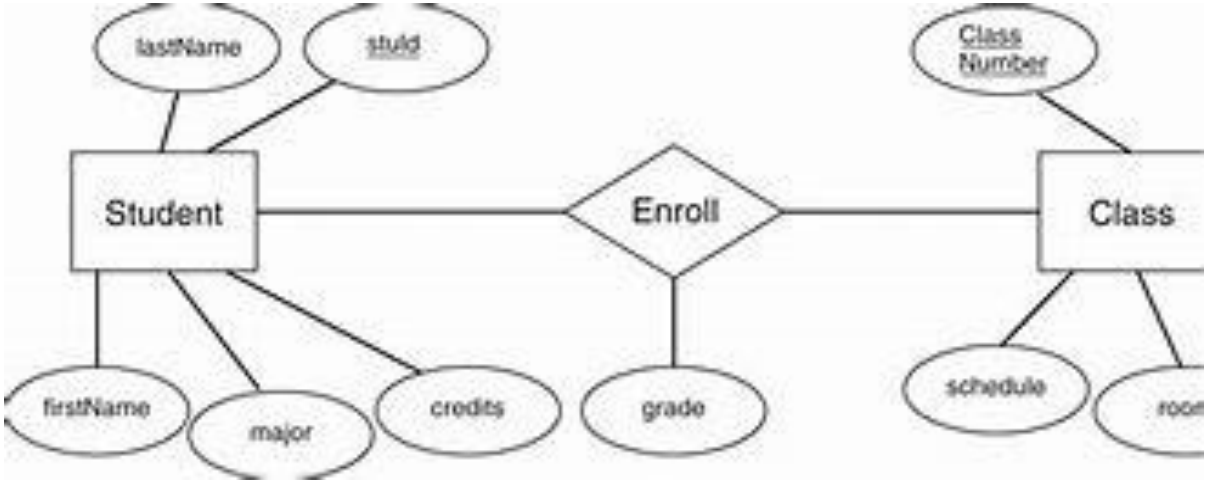
- **Alternate Key –** is a column or group of columns in a table that uniquely identify every row in that table.
- **Foreign Key –** is a column that creates a relationship between two tables. The purpose of Foreign keys is to maintain data integrity and allow navigation between two different instances of an entity.
- **Compound Key –** A compound key is a composite for which each attribute that makes it unique from its other attributes.

Data Models

- The ER Model
- The Object Oriented Model
- The Relational Model
- The Network Model
- The Hierarchical Model
- Physical Data Model

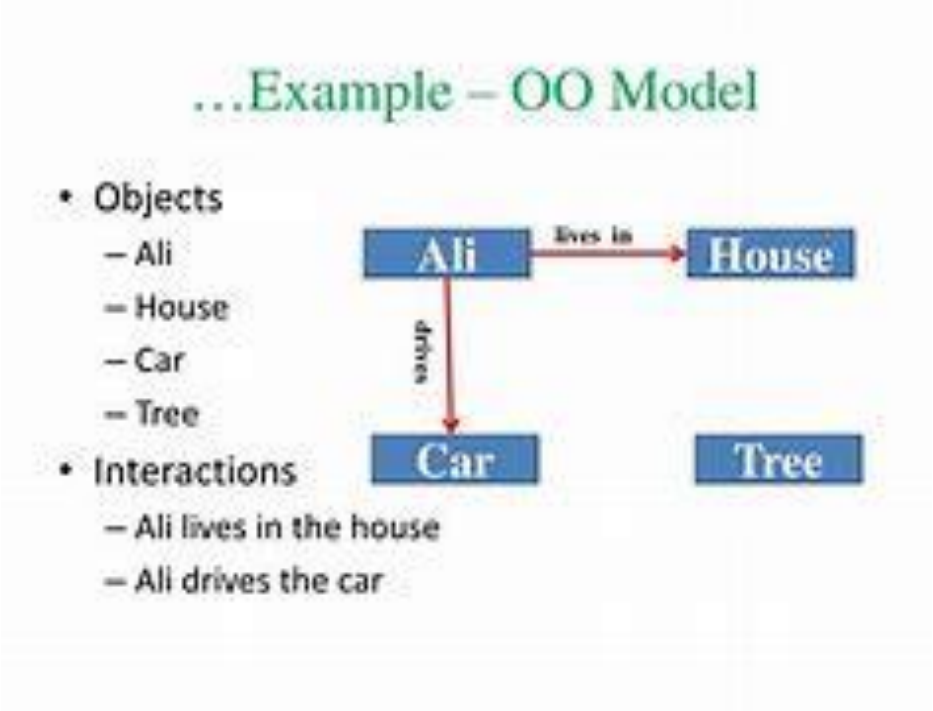
ER Model

- ER Model is a model which is based on real world objects simply termed as entities, attributes and their relationships. In ER model, Entity, attributes, and relationship play important role.



Object Oriented Model

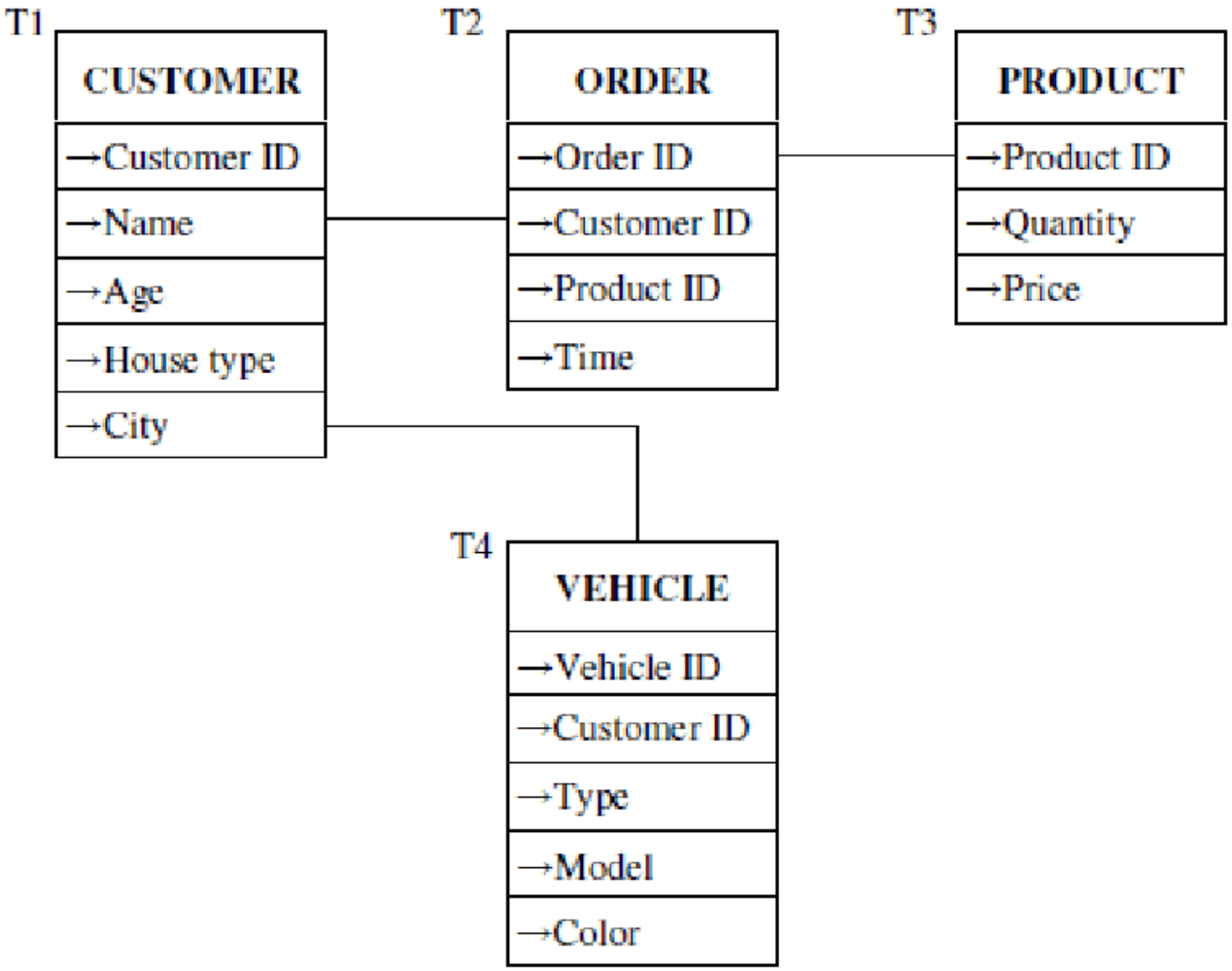
- OO Model is the data model in which data is stored in form of objects, which are instances of classes. It can manage complex data such as photo video etc.



Relational Model

- It is an approach to logically represent and manage the data stored in a database. In this model, the data is organized into a collection of two-dimensional inter-related **tables**, also known as **relations**.
- Each relation is a collection of columns and rows, where the column represents the attributes of an entity and the rows (or tuples) represents the records.

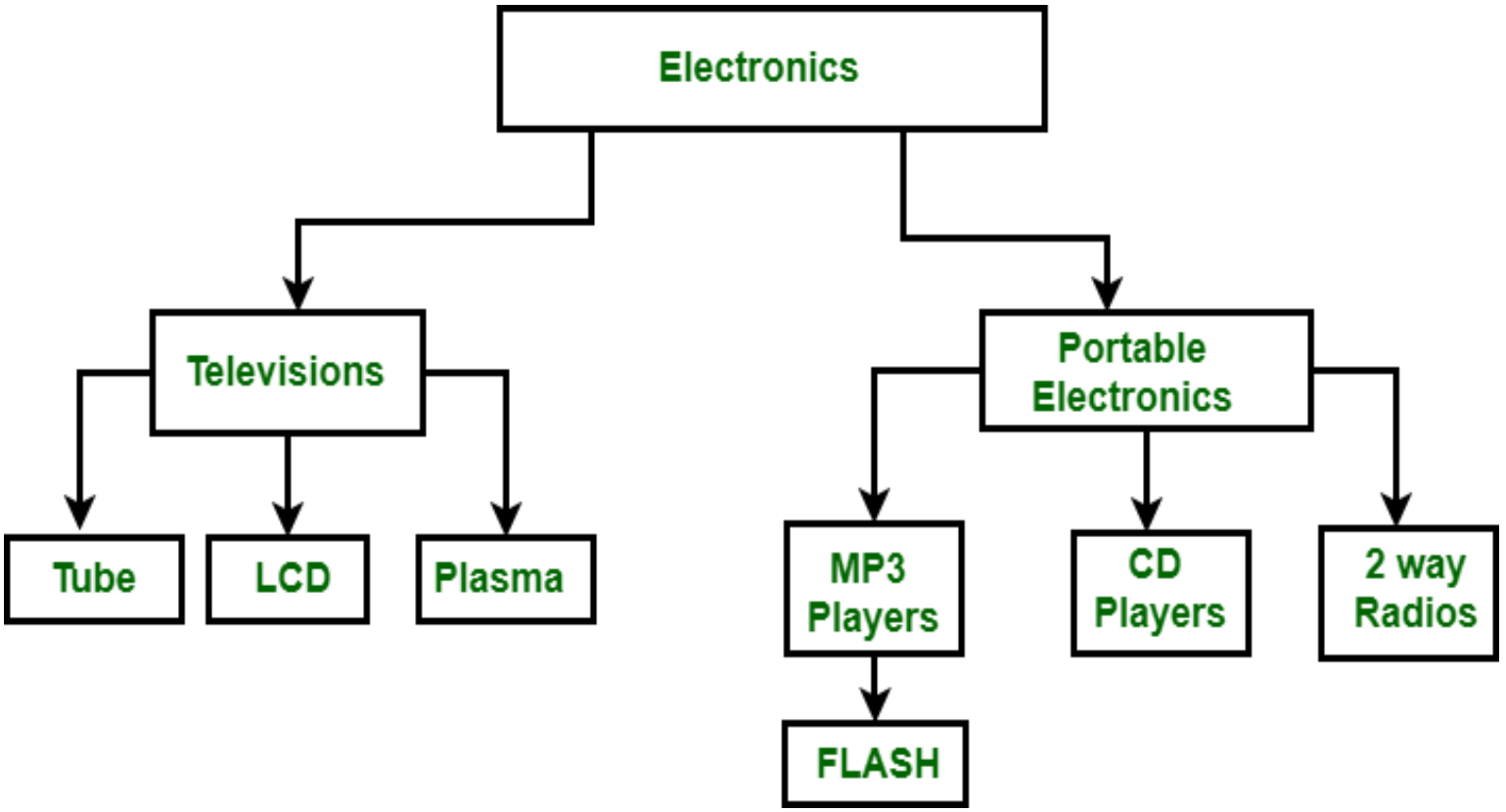
Relational Model



Hierarchical Model

- The data elements are linked in the form of an inverted tree structure with the root as the top and the branch formed below.
- There is a parent-child relationship among the data elements of a hierarchical database.
- A parent data element is one or more subordinate data elements.

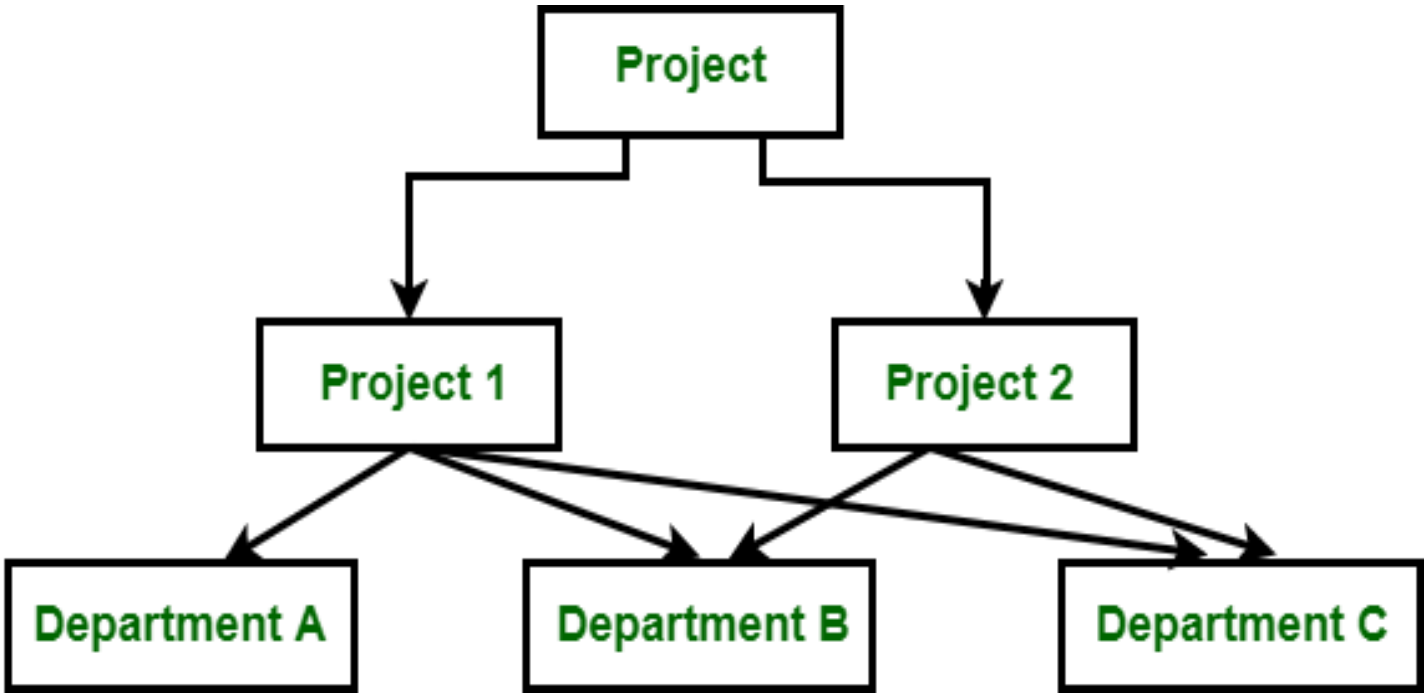
Hierarchical Model



Network Model

- A network database structure is an extension of the hierarchical database structure.
- In network database model a child data elements can have more than one parent or no parent at all.
- In this model the DBMS permits the extraction of the needed information by beginning from any data elements in the database structure instead of starting from the root data element.

Network Model



Objective of database approach

- Data Integration
  - The data in database may be located in different computers physically but it is connected through data communication links.
- Data Integrity
  - Integrity rules are designed to keep the data consistent and correct. These rules act like a check or the incoming data.
- Data Independence
  - The user can change data storage structures and operations without changing the application programs. The user can also modify programs without changing the application programs.

Objective of database approach

- Reduce Redundancy of Information
  - A duplication process may remove redundant blocks to reduce storage consumption within the volume or to minimize the volume of data that must be backed up.
- Security and User Privileges
  - Protection of data from unwanted users and giving users rights to what level they are to use the application.
- Ease of Application Development
  - It is a more comfortable workspace for the relevance of its factual purpose.

Data Repository

- A data repository is a structure consisting of one or more databases, containing data for the purpose of analysis.
- Data repositories are used in business to provide a centralized source of information.
- It might be used by business units to run reports or be used by analytics teams to study performance.
- Data repositories are also popular in academia, where they provide a reliable information to scientists and researchers.
- A data repository may also be referred to as a data library or a data archive.

Data Warehouse

- A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources.
- A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.
- A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users.
- It is not used for daily operations and transaction processing but used for making decisions.

Knowledge Discovery in Database

- KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets.
- New techniques and tools for extraction of knowledge from large data sets are the subject of the field called KDD.
- KDD is an umbrella term describing a variety of activities for making sense of data. It describes the overall process of finding useful patterns of data.

KDD Process

- **Selection**: Select a relevant subset of the data for analysis.
- **Pre-processing:** Clean and transform the data to make it ready for analysis. This may include tasks such as data normalization, missing value handling, and data integration.
- **Transformation:** Transform the data into a format suitable for data mining, such as a matrix or a graph.
- **Data Mining:** Apply data mining techniques and algorithms to the data to extract useful information and insights. This may include tasks such as clustering, classification, association rule mining, and anomaly detection.

KDD Process

- **Interpretation**: Interpret the results and extract knowledge from the data. This may include tasks such as visualizing the results, evaluating the quality of the discovered patterns and identifying relationships and associations among the data.
- **Evaluation**: Evaluate the results to ensure that the extracted knowledge is useful, accurate, and meaningful.
- **Deployment**: Use the discovered knowledge to solve the business problem and make decisions.

Data Mining

- Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information.
- Companies use data mining software to learn more about their customers. It can help them to develop more effective marketing strategies, increase sales, and decrease costs.
- Data mining is the process of analyzing a large batch of information to discern trends and patterns.
- Data mining can be used by corporations for everything from learning about what customers are interested in or want to buy to fraud detection and spam filtering.

Why data mining

- Establish relevance and relationships amongst data use this information to generate profitable insights.
- Business can make informed decisions quickly.
- Helps to find out unusual shopping patterns in grocery stores.
- Optimize website business by providing customize offers to each visitor.
- Creating and maintaining new customer groups for marketing purposes.
- Predict customer defection like which customers are more likely to switch to another supplier in the nearest.
- Differentiate between profitable and unprofitable customers.

OLTP

- OLTP (online transaction processing) is a class of software programs capable of supporting transaction-oriented applications.
- In computing, a transaction is a sequence of discrete information exchanges that are treated as a unit.
- Many everyday acts involve OLTP, including online banking, online shopping and even in-store shopping when the point of sale (POS) terminal is tied to inventory management software.
- OLTP (online transactional processing) enables the rapid, accurate data processing behind ATMs and online banking, cash registers and ecommerce, and scores of other services we interact with each day.

How OLTP Works

- OLTP involves taking transactional data, processing it and updating a back-end database to reflect the new input. While the applications may be complex, these updates are usually simple and involve only a few database records.
- A RDBMS is often used to manage OLTP. Relational databases are a good option for OLTP because it requires a database that can handle a large number of queries and updates while supporting fast response times.
- OLTP is used for executing online database transactions that frontline workers such as cashiers and bank tellers generate. Customer self-service applications like online banking, travel and e-commerce also generate database transactions and are tied into OLTP systems.

OLAP

- OLAP stands for Online Analytical Processing. OLAP systems have the capability to analyze database information of multiple systems at the current time. The primary goal of OLAP Service is data analysis and not data processing.
- Online Analytical Processing (OLAP) consists of a type of software tool that is used for data analysis for business decisions.
- OLAP provides an environment to get insights from the database retrieved from multiple database systems at one time.

OLAP Services

- OLAP services help in keeping consistency and calculation.
- We can store planning, analysis, and budgeting for business analytics within one platform.
- OLAP services help in handling large volumes of data, which helps in enterprise-level business applications.
- OLAP services help in applying security restrictions for data protection.
- OLAP services provide a multidimensional view of data, which helps in applying operations on data in various ways.

OLAP	OLTP
It is well-known as an online database query management system.	It is well-known as an online database modifying system.
Consists of historical data from various Databases.	Consists of only operational current data.
It makes use of data warehouse.	It makes use of DBMS.
It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc.	It is application-oriented. Used for business tasks.
In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized (3NF).
The data is used in planning, problem-solving, and decision-making.	The data is used to perform day-to-day fundamental operations.
Improves the efficiency of business analytics	Enhances the user's productivity.

Data Mart

- A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area.
- Data marts make specific data available to a defined group of users, which allows those users to quickly access critical insights without wasting time searching through an entire data warehouse.
- For example, many companies may have a data mart that aligns with a specific department in the business, such as finance, sales, or marketing.
- Organization can build a data mart at a much lower cost, time, and effort than that involved in building a data warehouse.

Meta Data

- Metadata is data about the data or documentation about the information which is required by the users. In data warehousing, metadata is one of the essential aspects.
- Metadata includes the following:
  - The location and descriptions of warehouse systems and components.
  - Names, definitions, structures, and content of data-warehouse and end-users views.
  - Identification of authoritative data sources.
  - Integration and transformation rules used to populate data.
  - Integration and transformation rules used to deliver information to end-user analytical tools.

ROLAP	MOLAP
ROLAP stands for Relational Online Analytical Processing.	MOLAP stands for Multidimensional Online Analytical Processing.
It usually used when data warehouse contains relational data.	It used when data warehouse contains relational as well as non-relational data.
It contains Analytical server.	It contains the MDDB server.
It creates a multidimensional view of data dynamically.	It contains prefabricated data cubes.
It is very easy to implement	It is difficult to implement.
It has a high response time	It has less response time due to prefabricated cubes.
It requires less amount of memory.	It requires a large amount of memory.

Advantages of Disadvantages of MOLAP

- Advantages
  - Superior Performance
  - Complex Computing
- Disadvantages
  - Only Limited Data can be Processed
  - Additional Investment Required

Advantages of Disadvantages of ROLAP

- Advantages
  - It can process a large amount of data
  - Relational database function can be used
- Disadvantages
  - Low Performance
  - QuiltSQL Restriction

Meta Data

- Metadata is data about the data or documentation about the information which is required by the users. In data warehousing, metadata is one of the essential aspects.
- Metadata includes the following:
  - The location and descriptions of warehouse systems and components.
  - Names, definitions, structures, and content of data-warehouse and end-users views.
  - Identification of authoritative data sources.
  - Integration and transformation rules used to populate data.
  - Integration and transformation rules used to deliver information to end-user analytical tools.

Drill Down and Roll Up Analysis

- Drill-down refers to the process of viewing data at a level of increased detail.
- while roll-up refers to the process of viewing data with decreasing detail.
- The control parameter is based on a measure of cluster sizes.
- For example, a drill down report that shows sales revenue by state can allow the user to select a state, click on it and see sales revenue by county or city within that state.
- For example, a rollup analysis on a factory element might use temperature attribute values for all pumps in the factory to calculate their average temperature.

Star Schema

- A star schema is a database organizational structure optimized for use in a data warehouse or business intelligence that uses a single large fact table to store transactional or measured data, and one or more smaller dimensional tables that store attributes about the data.
- It is called a star schema because the fact table sits at the center of the logical diagram, and the small dimensional tables branch off to form the points of the star.

Snowflake Schema

- The snowflake schema is a variant of the star schema. Here, the centralized fact table is connected to multiple dimensions.
- In the snowflake schema, dimensions are present in a normalized form in multiple related tables.
- The snowflake structure materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent tables.
- The snowflake effect affects only the dimension tables and does not affect the fact tables.

How data mining works

- Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends.
- It is used in credit risk management, fraud detection, and spam filtering.
- It is also used as market research tool that helps reveal the sentiment or opinions of a given group of people.

How data mining Process

- Data is collected and loaded into data warehouses on-site or on a cloud service.
- Business analysts, management teams, and information technology professionals access the data and determine how they want to organize it.
- Custom application software sorts and organizes the data.
- The end user presents the data in an easy-to-share format, such as a graph or table.

How data mining Phases

- Understand the Business
- Understand the data
- Prepare the data
- Build the model
- Evaluate the results
- Implement change and monitor

Data mining Techniques

- **Association rules**, also referred to as market basket analysis, search for relationships between variables.
- This relationship in itself creates additional value within the data set as it strives to link pieces of data.
- For example, association rules would search a company's sales history to see which products are most commonly purchased together; with this information, stores can plan, promote, and forecast.

Data mining Techniques

- **Classification** uses predefined classes to assign to objects.
- These classes describe the characteristics of items or represent what the data points have in common with each.
- This data mining technique allows the underlying data to be more neatly categorized and summarized across similar features or product lines.

Data mining Techniques

- **Clustering** is similar to classification.
- However, clustering identifies similarities between objects, then groups those items based on what makes them different from other items.
- While classification may result in groups such as "shampoo," "conditioner," "soap," and "toothpaste," clustering may identify groups such as "hair care" and "dental health."

Data mining Techniques

- **K-Nearest neighbor (KNN)** is an algorithm that classifies data based on its proximity to other data.
- The basis for KNN is rooted in the assumption that data points that are close to each other are more similar to each other than other bits of data.
- This non-parametric, supervised technique is used to predict the features of a group based on individual data points.

Data mining Techniques

- **Neural networks** process data through the use of nodes.
- These nodes are comprised of inputs, weights, and an output. Data is mapped through supervised learning, similar to the ways in which the human brain is interconnected.
- This model can be programmed to give threshold values to determine a model's accuracy.

Data mining Techniques

- **Predictive analysis** strives to leverage historical information to build graphical or mathematical models to forecast future outcomes.
- Overlapping with regression analysis, this technique aims at supporting an unknown figure in the future based on current data on hand.

Data Mining Algorithms

- Data Mining Algorithms are a particular category of algorithms useful for analyzing data and developing data models to identify meaningful patterns. These are part of machine learning algorithms.
- These are the examples, where the data analysis task is Classification Algorithms in Data Mining.
- A bank loan officer wants to analyze the data in order to know which customer is risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

ID3 Algorithm

- This Data Mining Algorithms starts with the original set as the root hub. On every cycle, it emphasizes through every unused attribute of the set and figures. That the entropy of attribute. At that point chooses the attribute. That has the smallest entropy value.
- The set is S then split by the selected attribute to produce subsets of the information.
- This Data Mining algorithms proceed to recurse on each item in a subset. Also, considering only items never selected before. Recursion on a subset may bring to a halt in one of these cases

ID3 Algorithm

- Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labeled with the class of the examples
- If there are no more attributes to select but the examples still do not belong to the same class. Then the node is turned into a leaf and labeled with the most common class of the examples in that subset.
- If there are no examples in the subset, then this happens. Whenever parent set found to be matching a specific value of the selected attribute.
- For example, if there was no example matching with marks >=100. Then a leaf is created and is labeled with the most common class of the examples in the parent set.

C4.5 Algorithm

- C4.5 is one of the most important Data Mining algorithms, used to produce a decision tree which is an expansion of prior ID3 calculation.
- That is by managing both continuous and discrete properties, missing values.
- The decision trees created by C4.5. that use for grouping and often referred to as a statistical classifier.
- The algorithm analyzes the training set and builds a classifier. That must have the capacity to accurately arrange both training and test cases.
- A test example is an input object and the algorithm must predict an output value. Consider the sample training data set S=S1, S2,…Sn which is already classified.

Naïve Bayes algorithm

- The Naive Bayes Classifier technique is based on the Bayesian theorem. It is particularly used when the dimensionality of the inputs is high.
- The Bayesian Classifier is capable of calculating the possible output. That is based on the input. It is also possible to add new raw data at runtime and have a better probabilistic classifier.
- This classifier considers the presence of a particular feature of a class. That is unrelated to the presence of any other feature when the class variable is given.

Naïve Bayes algorithm

- For example, a fruit may consider to be an apple if it is red, round.
- Even if these features depend on each other features of a class. A naive Bayes classifier considers all these properties to contribute to the probability. That it shows this fruit is an apple. Algorithm works as follows,
- Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier considers the effect of the value of a predictor (x) on a given class (c). That is independent of the values of other predictors.
- P(c|x) is the posterior probability of class (target) given predictor (attribute) of class.
- P(c) is called the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor of given class.
- P(x) is the prior probability of predictor of class.

The apiriori Algorithm

- The Apriori algorithm is widely used to find the frequent item sets from a transaction data set and derive association rules.
- To find frequent item sets is not difficult because of its combinatorial explosion. Once we get the frequent item sets, it is clear to generate association rules for larger or equal specified minimum confidence.
- Apriori is an algorithm which helps in finding routine data sets by making use of candidate generation. It assumes that the item set or the items present are sorted in lexicographic order.
- After the introduction of Apriori data mining research has been specifically boosted. It is simple and easy to implement.

The apiriori Algorithm

The basic approach of this algorithm is as below:

- **Join**: The whole database is used for the hoe frequent 1 item sets.
- **Prune**: This item set must satisfy the support and confidence to move to the next round for the 2 item sets.
- **Repeat**: Until the pre-defined size is not reached till, then this is repeated for each item set level.