Sem 2-Data Science & Big Data Analysis

Module-1- Introduction to Data Mining

 # Data Mining:-Data mining is the process of searching and analyzing a large number 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 Sometimes referred to as "**knowledge discovery in databases**" (KDD).
-It is also used in credit risk management, fraud detection, and spam filtering.
-The knowledge discovery process includes **Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.**

->Advantages of Data Mining
-Compared with other statistical data applications, data mining is a cost-efficient.
-Data Mining helps in the decision-making process of an organization.
-It automatically discovers the hidden patterns.

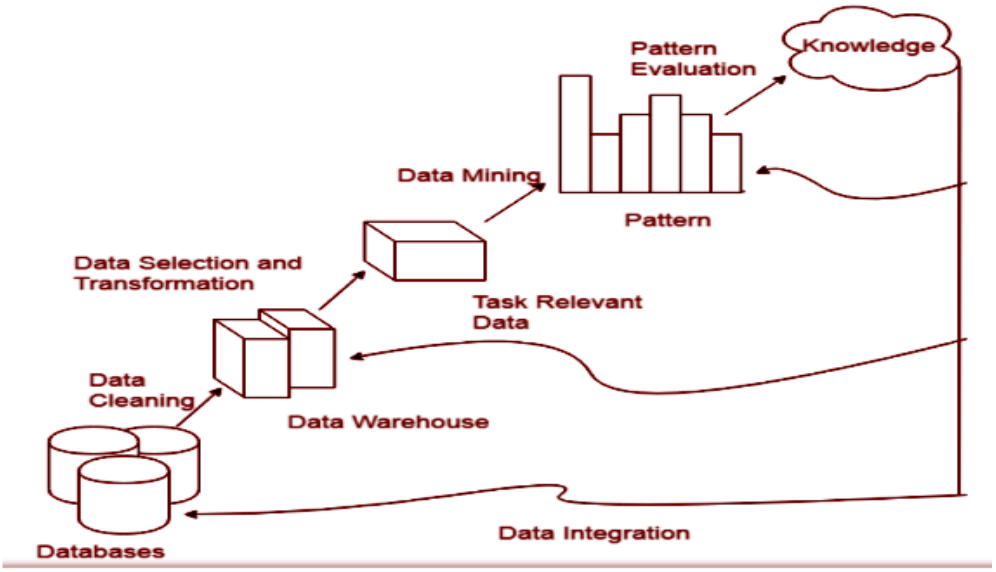
->Disadvantages of Data Mining
-There is a probability that the organizations may sell useful data of customers to other organizations for money.
-Many data mining software is difficult to operate.
-Different data mining software operate in different ways due to the different algorithms used in their design. Therefore, the selection of the right data for data mining is a very challenging task.

→**Steps of Data mining/ KDD process:**-The process begins with determining the KDD objectives.
-and ends with the implementation of the discovered knowledge.
-Steps Involved in KDD Process are;

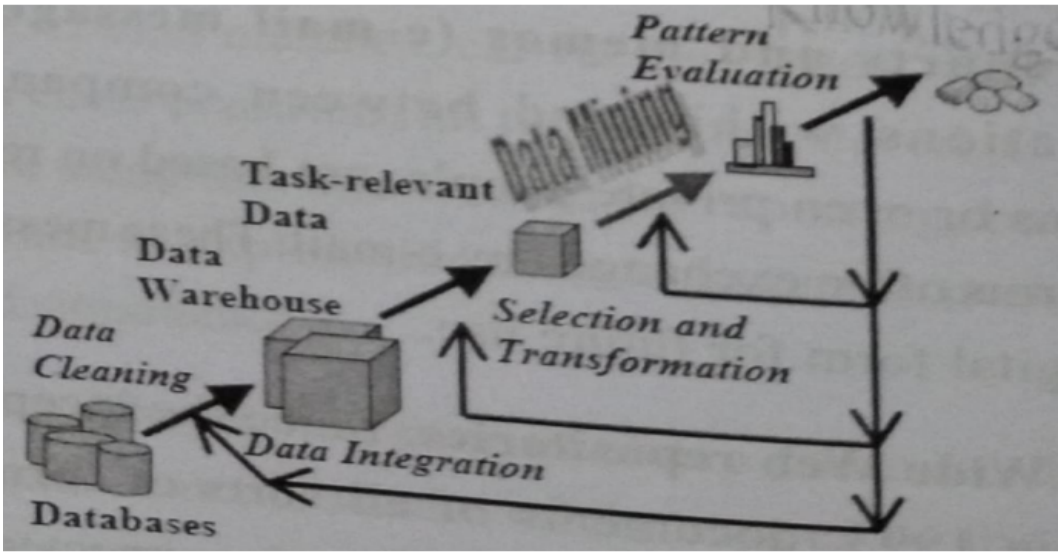
- **Data Cleaning**:- This is also known as data cleansing.
  -This is a phase in which noisy, inconsistent and irrelevant data are removed from the collection.
  ”noise is a random or variance error.”
- **Data integration**:- In this step ,multiple data source may be combined and put into a single source. (DataWarehouse)
- **Data selection**:-in this step , data relevant to the analysis is decided and retrieved from the data collection.
- **Data transformation**:-This is a phase in which the selected data is transformed into

appropriate forms for the mining procedure
Eg; performing summary or aggregation operation.

- **Data mining**:-It is the crucial step in which intelligent techniques are applied to extract patterns ,which are potentially useful.
- **Pattern evaluation**:- It Identifies strictly increasing patterns representing knowledge based on given measures.
- **Knowledge presentation**:-This is the final phase in which the discovered knowledge is visually represented to the user.
  -where visualization and knowledge representation techniques are used to present the mined knowledge to the user.



or

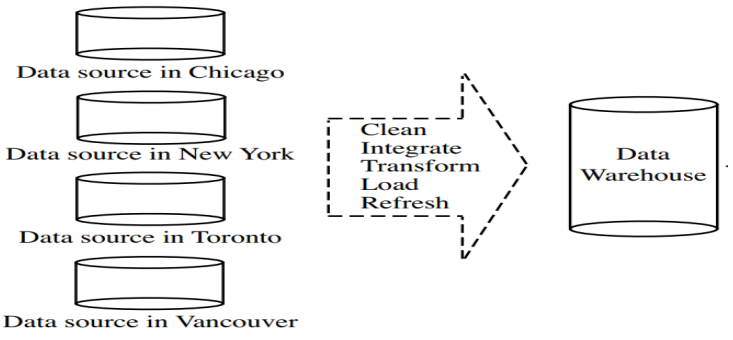


***Types of data that can be mined/types of Databases used in Data Mining:**-The most basic forms of data for mining applications are given below.
-Or Different kind of data can be mine. Some of the examples are mentioned below.

1. **Database Data** :-A database system, also called a database management system (DBMS).
   -Every DBMS stores data that are related to each other in a way or the other.
   -It also has a set of software programs that are used to manage data and provide easy access to it.
   - These software programs serve a lot of purposes, including defining structure for database, making sure that the stored information remains secured and consistent.

   -A relational database has tables that have different names, attributes, and can store rows or records of large data sets.
   -Every record stored in a table has a unique key. Entity-relationship model is created to provide a representation of a relational database that features entities and the relationships that exist between them.

2. **Data Warehouses**:- A data warehouse is a single data storage location that collects data from multiple sources and then stores it in the form of a **unified plan**.
   -When data is stored in a data warehouse, it undergoes cleaning, integration, loading, and refreshing.
   -Data stored in a data warehouse is organized in several parts.
   - If you want information on data that was stored 6 or 12 months back, you will get it in the form of a summary.
   -For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item.



3. **Transactional Data** :-Transactional database stores record that are captured as transactions.
   -These transactions include flight booking, customer purchase, click on a website, and others.

-Every transaction record has a unique ID.
-It also lists all those items that made it a transaction.

Or

-A transaction typically includes a unique transaction identity number (trans ID)
-and a list of the items making up the transaction (such as items purchased in a store).
-The transactional database may have additional tables associated with it.
-which contain other information regarding the sale, such as the date of the transaction,
the customer ID number

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

4. **Other Kinds of Data**:-We have a lot of other types of data as well that are known for their structure, semantic meanings, and versatility.
   -They are used in a lot of applications. Here are a few of those data types: data streams, engineering design data, sequence data, graph data, spatial data, multimedia data, and more.

***Data Mining Functionalities/Tasks/Types of Patterns Can Be Mined:-** It used to specify what kind of patterns to be found in data mining tasks.
-data mining tasks can be classified into two categories: **descriptive** and **predictive**.

- Descriptive mining tasks characterize the general properties of the data in the Database.
- Predictive mining tasks perform on the current data in order to make predictions.

-There are several data mining functionalities that are;

1. **Class/Concept Descriptions:-**Data can be associated with classes or concepts.
   - useful to describe individual classes and concepts in summarized, concise, and yet precise terms.
   -Such descriptions of a class or a concept are called class/concept descriptions.
   -There are two concepts here, one that helps with grouping and the other that helps in differentiating.
   - **1.1.Data characterization**:-This refers to the summary of general characteristics of the class that is under the study.
     -The output of data characterization can be presented in multiple forms.
   - **1.2.Data discrimination**:-It compares common features of class which is under study.
     -The output of this process can be represented in many forms.
     Eg., bar charts, curves and pie charts.

2.Mining frequent patterns:-One of the functions of data mining is finding data patterns.

-Frequent patterns are nothing but things that are found to be most common in the data

-Various types of frequency can be found in the dataset.

2.1.Frequent item set:-This term refers to a group of items that are commonly found together, such as milk and sugar.

2.2.Frequent substructure:- It refers to the various types of data structures that can be combined with an item set or subsequences, such as trees and graphs.

2.3.Frequent Subsequence:- A regular pattern series, such as buying a phone followed by a cover.

3.Association and correlation

->Association:-It analyses the set of items that generally occur together in a transactional dataset.

-It is also known as Market Basket Analysis for its wide use in retail sales.

-for example, it can be used to determine the sales of items that are frequently purchased together.

->Correlation:-Correlation is a mathematical technique for determining whether and how strongly two attributes is related to one another.

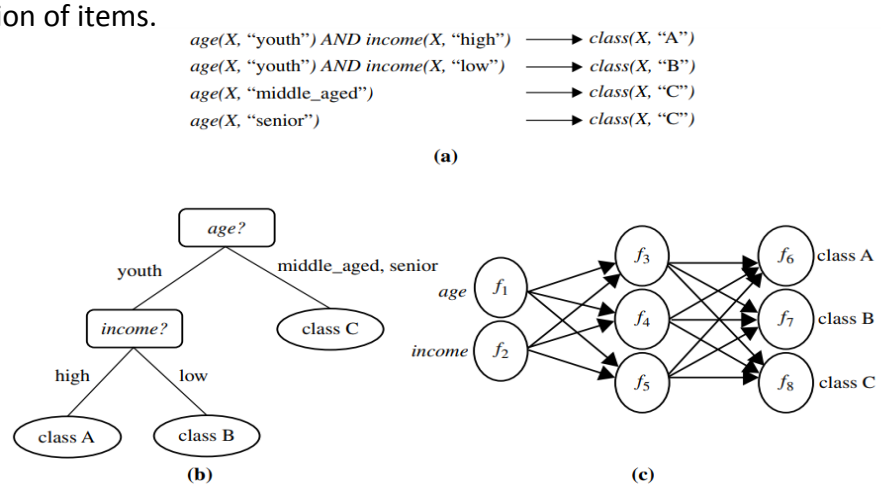
-For example, Highted people tend to have more weight.

4.Classification and Prediction

->Classification:-Classification is a data mining technique that categorizes items in a collection based on some predefined properties.

-It uses methods like if-then, decision trees or neural networks to predict a class or essentially classify a collection of items.

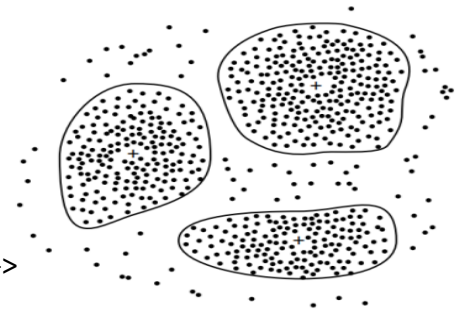
-A training set is used to train the system to predict the items from an unknown collection of items.



A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

->Prediction:-As implied in its name, this compelling data mining technique helps enterprises to match patterns based on current and historical data records for predictive analysis of the future. While some of the approaches involve Artificial Intelligence and Machine Learning aspects, some can be conducted via simple algorithms.

5. Cluster Analysis:-clustering is a popular data mining functionality in image processing, pattern recognition and bioinformatics.
-It is similar to classification but the classes are not predefined.
-Cluster fig —---------->

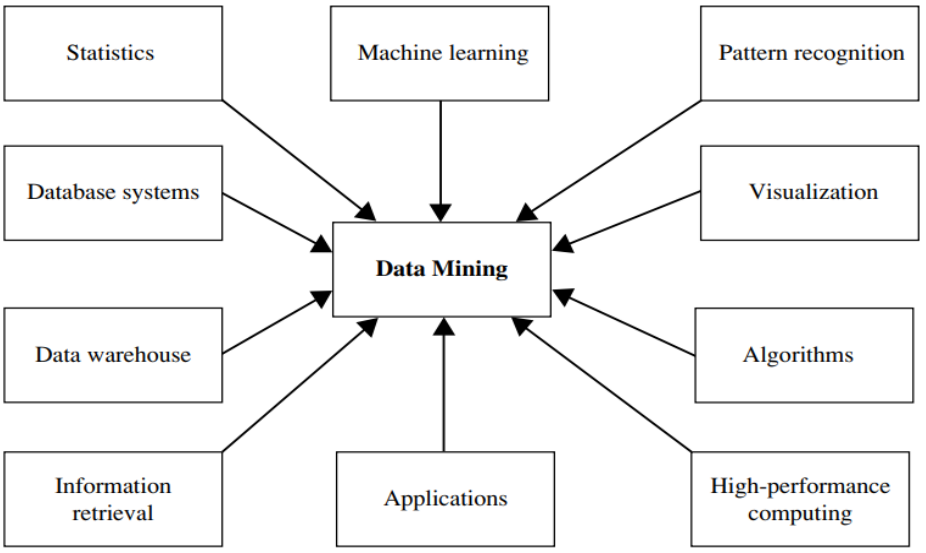


-Data attributes represent the classes.
-Clustering algorithms group data based on similarities and dissimilarities.

6. Outlier Analysis:-Outlier analysis is important to understand the quality of data.
-If there are too many outliers, you cannot trust the data or patterns.

Q)data Discrimination with Example?

***data mining Technologies:-**Data mining has incorporated many techniques from other domain fields like machine learning, statistics, information retrieval, data warehouse, pattern recognition, algorithms, and high-performance computing.



-The major technologies utilized in data mining are;

1. **Machine Learning**: -It canautomatically learn based on the given input data and make intelligent decisions.
   -There are similarities and interrelations between machine learning and data mining.
   -For classification and clustering approaches, machine learning is often applied to predict accuracy.
   -This are the problems in machine learning that are highly related to data mining.
   - Supervised learning
   - Unsupervised learning
   - Semi-supervised learning
   - Active learning
2. **Information Retrieva**l: -The technique searches for the information in the document, which may be in text, multimedia, or residing on the Web.
   -The most widely used information retrieval approach is the **probabilistic model.**
   -Information retrieval combined with data mining techniques is used for finding out any relevant topic in the document or web.
3. **Statistics**:-Data mining has a natural connection with statistics.
   -Statistics are useful for pattern mining.
   -When the statistical model is used on large data set, it increases the complexity cost.
   -When data mining is used to handle large real-time and streamed data, computation costs increase dramatically.
4. **Database System & Data warehouse**:- google

***Data mining Applications:**-Here is the list of areas where data mining is widely used

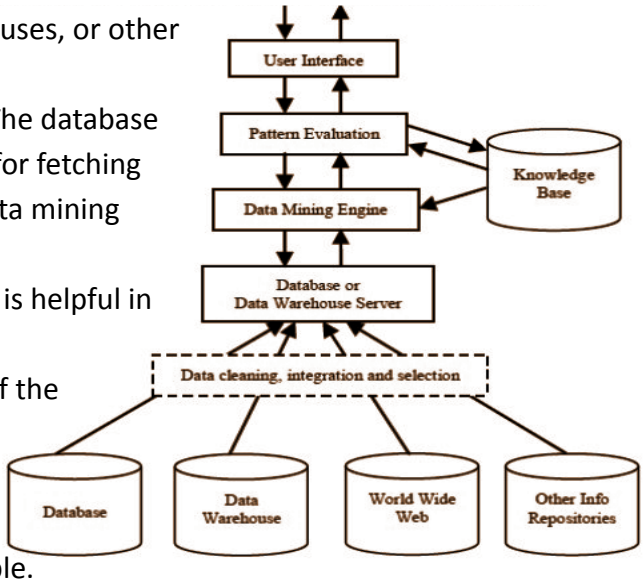
- **Healthcare**:-Data mining has a lot of promise for improving healthcare systems.
  -It identifies best practices for improving treatment and lowering costs using data and analytics.
  -machine learning, soft computing, data visualization, and statistics are among the data mining techniques used by researchers.
  Patients receive appropriate care at the correct place and at the right time thanks to the development of processes.
  -Healthcare insurers can employ data mining to detect fraud and misuse.
- **Banking and Finance:-**The banking industry is now dealing with and managing massive volumes of data and transaction information as a result of digitalization.
  -With its capacity to detect patterns, casualties, market risks, and other connections that are critical for managers to be aware of, data mining applications in banking can easily be the suitable answer.
- **Market Basket Analysis:**-Market Basket Analysis is a method for analyzing the purchases made by a consumer in a supermarket.

-This notion identifies a customer's habit of regular purchases.

- **Criminal Investigation**:-Data mining activities are also used in Criminology, which is a study of crime characteristics.
  -First, text-based crime reports need to be converted into word processing files.
  -Then, the identification and crime-machining process would take place by discovering patterns in massive stores of data.

***Data mining architecture:**-Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.
-Based on this view, the architecture of a typical data mining system may have the following major components.

● **Database, data warehouse, World Wide Web, or other information repository:-**This is one or a set of databases, data warehouses, or other kinds of information repositories.



● **Database or data warehouse server:-**The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

● **Knowledge base:**-The knowledge base is helpful in the entire process of data mining.
-It might be helpful to guide the search of the result patterns.
-The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable.

● **Data mining engine:**-The data mining engine is a major component of any data mining system.
-It contains several modules for operating data mining tasks they including association, characterization, classification, clustering, prediction, time-series analysis, etc.
-In other words, we can say data mining is the root of our data mining architecture.

● **Pattern evaluation module:**-They are responsible for finding interesting patterns in the data
-and sometimes they also interact with the database servers for producing the result of the user requests.

● **User interface:-**user interface module communicates between the data mining system and the user.
-This module helps the user to easily and efficiently use the system without knowing the complexity of the process.

***Data mining issues:**-The major issues in data mining regarding mining methodology, user interaction, performance, and diverse data types.

1.Mining methodology and user interaction issues

1.1.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 range of knowledge to discover a task.

1.2.Interactive mining of knowledge at multiple levels of abstraction:- it is difficult to know exactly what can be discovered within a database.

-The data mining process needs to be interactive.

-Interactive mining allows users to focus the search for patterns.

-Specifically, knowledge should be mined by drilling down, rolling up method.

-The user can interact with the data mining system to view data and discovered patterns at multiple levels and from different angles.

1.3.Incorporation of background knowledge:-To guide discovery process and to express the discovered patterns, the background knowledge can be used.

-Background knowledge may be used to express the discovered patterns at multiple levels of abstraction also.

1.4.Data mining query languages and ad hoc data mining:-Data Mining Query language that allows the user to describe ad hoc mining tasks.

-It should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

1.5. 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.

1.6.Handling noisy or incomplete data:-The data cleaning methods are required to handle the noise and incomplete objects while mining.

-If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

1.7.Pattern evaluation-the interestingness problem:-A data mining system can uncover thousands of patterns.

-Many of the patterns discovered may be uninteresting to the given user.

-because they represent common knowledge or lack novelty.

-The interestingness measures is used to guide the discovery process.

2.Performance issues:- These include efficiency, scalability, and parallelization of data mining algorithms.

2.1.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.

-knowledge discovery, efficiency and scalability are key issues in the implementation of data mining systems.

2.2.Parallel, distributed, and incremental mining algorithms:-The huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms.

3.Issues relating to the diversity of database types

3.1.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.

3.2. Mining information from heterogeneous databases and global information systems:-The data is available at different data sources on LAN or WAN.

-These data source may be structured, semi structured or unstructured.

-Therefore mining the knowledge from them adds challenges to data mining.

***Data Pre-processing**:-Data preprocessing is an important step in the data mining process.

-It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis.

-The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

-The data processing is carried out automatically or manually.

-Nowadays, most data is processed automatically with the help of the computer, which is faster and gives accurate results.

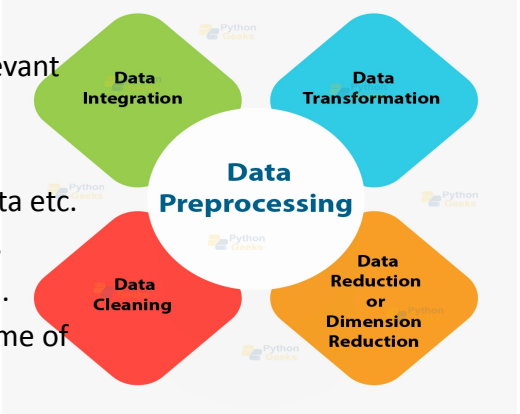
-Thus, data can be converted into different forms. It can be graphic as well as audio ones. It depends on the software used as well as data processing methods.

-Data processing is crucial for organizations to create better business strategies and increase their competitive edge.

-By converting the data into a readable format like graphs, charts, and documents, employees throughout the organization can understand and use the data.

-The most commonly used tools for data processing are **Storm, Hadoop, HPCC, Statwing, Qubole, and CouchDB**.

-The major steps involved in data preprocessing, namely, data cleaning, data integration, data reduction, and data transformation.



1. **Data Cleaning**:-The data can have many irrelevant and missing parts.
   -To handle this part, data cleaning is done.
   -It involves handling of missing data, noisy data etc.
   1.1 Missing Data:-This situation arises when some data is missing in the data.
   -It can be handled in various ways. Some of

them are:

- **Ignore the tuples**:-This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
- **Fill the Missing values**:-There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

1.2 Noisy Data: Noisy data is a meaningless data that can't be interpreted by machines.

-It can be generated due to faulty data collection, data entry errors etc.

-It can be handled in following ways :

- **Binning Method:** This method works on sorted data in order to smooth it.
  -The whole data is divided into segments of equal size and then various methods are performed to complete the task.
  - Each segmented is handled separately.
  -One can replace all data in a segment by its mean or boundary values can be used to complete the task.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

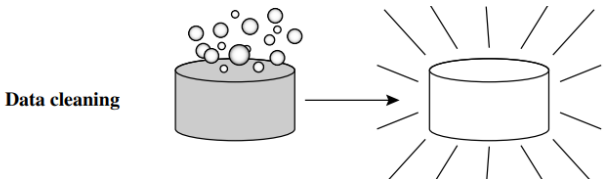
Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

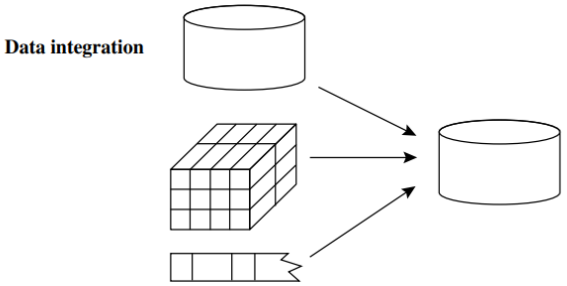
OR

-This method is to smooth or handle noisy data.
-First, the data is sorted then, and then the sorted values are separated and stored in the form of bins.
-There are three methods for smoothing data in the bin.
-**Smoothing by bin mean method**: In this method, the values in the bin are replaced by the mean value of the bin;
-Smoothing by bin median: In this method, the values in the bin are replaced by the median value;
-Smoothing by bin boundary: In this method, the using minimum and maximum values of the bin values are taken, and the closest boundary value replaces the values.

- **Regression**: Here data can be made smooth by fitting it to a regression function.
  -The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
- **Clustering**: This approach groups the similar data in a cluster.
  -The outliers may be undetected or it will fall outside the clusters.



2. **data integration**:-Data integration combines data from multiple sources to form a coherent data store.
   -These sources may include multiple databases, data cubes, or flat files.
   -There are multiple issues to consider during data integration. They are;
   2.1 **Schema integration and object matching**:-Schema integration is used to merge two or more existing database schemas into a single schema.
   -Schema integration and object matching can be complex.
   - For example, matching the entity identification (emp_id in one database and emp_no in another database), such issues can be prevented using metadata.
   2.2 **Redundancy**:-Redundancy is another issue.
   -An attribute may be redundant if it can be derived or obtaining from another Attribute.
   -Some redundancies can be detected by correlation analysis.
   2.3 **Detection and resolution of data value conflicts:**-This is the third important issue in data integration.
   -Attribute values from different sources may differ for the same entity.



3. **Data reduction**:--Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume.
   -Data reduction techniques are used to obtain a reduced representation of the dataset that
   is much smaller in volume by maintaining the integrity of the original data.
   -By reducing the data, the efficiency of the data mining process is improved.
   -Data reduction does not affect the result obtained from data mining.
   -That means the result obtained from data mining before and after data reduction is the
   same or almost the same.
   -Strategies or techniques or methods of data reduction in data mining, they are;
   3.1 Data cube aggregation:-This technique is used to aggregate data in a simpler form.

-Aggregation operations are applied to the data in the construction of a data cube.

3.2 Attribute subset selection:-where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.

3.3 Dimensionality reduction:-This mechanisms are used to reduce the data set size.

3.4 Numerosity reduction:-In this reduction technique the actual data is replaced with mathematical models or smaller representation of the data instead of actual data.

-This technique includes two types parametric and non-parametric numerosity reduction.

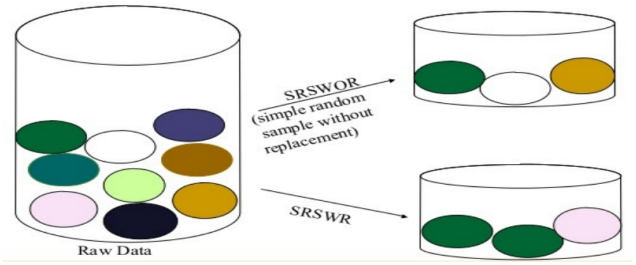
-Parametric numerosity reduction storing only data parameters instead of the original data.

-non-parametric method such as clustering, histogram, sampling.

->sampling method they are;

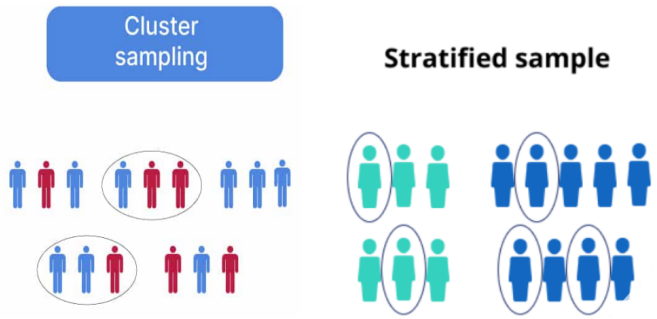
1.Simple random sample without replacement (SRSWOR)

2.Simple random sample with replacement (SRSWR)



3.Cluster sample

4.Stratified sample



4. **data transformation**:--Data transformation is a technique used to convert the raw data into a suitable format.

   -Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form.

   -Data transformation is a preprocessing technique that must be performed on the data before data mining, it make patterns easier to understand.

-Data transformation changes the format, structure, or values of the data and converts them into clean, usable data.

-The data transformation involves steps that are:

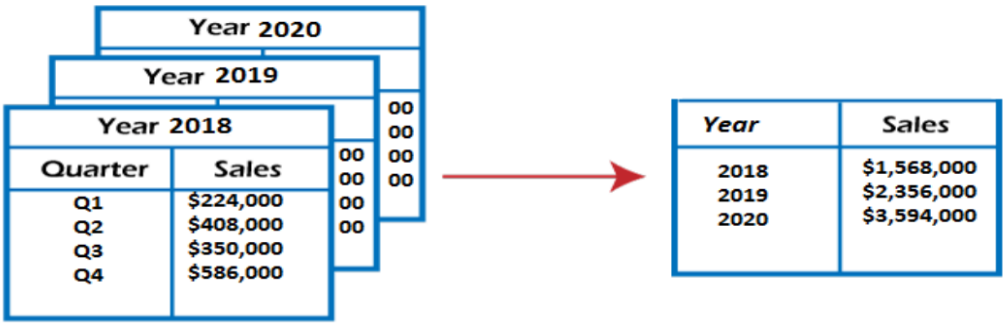
4.1 Smoothing:-smoothing is a process that is used to remove noise using some algorithms.

- It helps in predicting the patterns.

-the noise is removed from the data using the techniques such as

4.2 Aggregation:-aggregation is the method of storing and presenting data in a summary format.

-For example, we have a data set of sales reports of an enterprise that has quarterly sales of each year. We can aggregate the data to get the enterprise's annual sale report.



4.3 Generalization:-It converts low-level data attributes to high-level data attributes using concept hierarchy.

-This conversion from a lower level to a higher level is useful to get a clearer picture of the data.

-For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old).

-For example, attributes, like street, can be generalized to higher-level concepts, like city or country.

4.4 Normalization:-Normalizing the data refers to scaling the data values to a much smaller range such as [-1, 1] or [0.0, 1.0].

-There are different methods to normalize the data.

- **<u>Min-max normalization</u>**:-his method implements a linear transformation on the original data.

  -Let us consider that we have minA and maxA as the minimum and maximum value for attribute A.

  -Vi is the value for attribute A that has to be normalized.

  -The min-max normalization would map V_i to the V'_i in a new smaller range[new_minA, new_maxA].

  -The formula for min-max normalization is given below:

$$v'_i = \frac{v_i - min_A}{max_A - min_A}\left(new_{max_A} - new_{min_A}\right) + new\_min_A$$

- **z-score normalization (or zero-mean normalization):-**This method normalizes the value for attribute A using the mean and standard deviation.

  -The following formula is used for Z-score normalization:
  -Here $\bar{A}$ and σ_A are the mean and standard deviation for attribute A.
  -For example, we have a mean and standard deviation for attribute A as $54,000 and $16,000. And we have to normalize the value $73,600 using z-score normalization.

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}$$

$$\frac{73600-5400}{1600} = 1.225$$

- **Normalization by decimal scaling:**-This method normalizes the value of attribute A by moving the decimal point in the value.
  -This movement of a decimal point depends on the maximum absolute value of A.
  -The formula for the decimal scaling is given :

$$v'_i = \frac{v_i}{10^j}$$

*Difference between OLAP and OLTP

OLTP	OLAP
operational processing transaction	informational processing analysis
OLTP system is customer-oriented	OLAP system is market-oriented
OLTP system manages current data	OLAP system manages large amounts of historical data
It is an online transactional system.	It is used for data analysis
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 snowflake model and a subject oriented database design.
read/write	OLAP systems are mostly read-only

***Data Discretization in data mining**:-Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy.

-In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss.

-eg:we have an attribute of age with the following values.

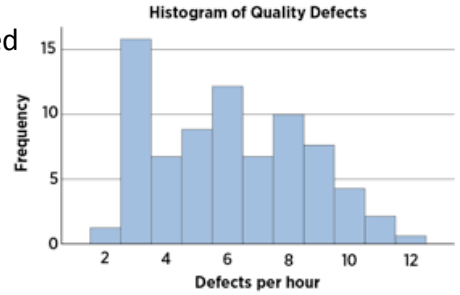
Age	1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77

-Table before Discretization and after Discretization

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

-This are the Famous techniques of data discretization:

- **Histogram analysis**:-The histogram is old method used to plot the attributes in a graph.
  -It is used to summarize discrete or continuous data that are measured on an interval scale.



- **Binning**:-it is a data smoothing technique and its helps to group a huge number of continuous values into a smaller number of bins.
  -For example, if we have data about a group of students, and we want to arrange their marks into a smaller number of marks intervals by making the bins of grades. One bin for grade A, one for grade B, one for C, one for D, and one for F Grade.

- **Cluster Analysis**:-Cluster analysis is commonly known as clustering.
  -Clustering is the task of grouping similar objects in one group, commonly called clusters.
  -All different objects are placed in different clusters.
- **Decision Tree**
- **Correlation Analyses**

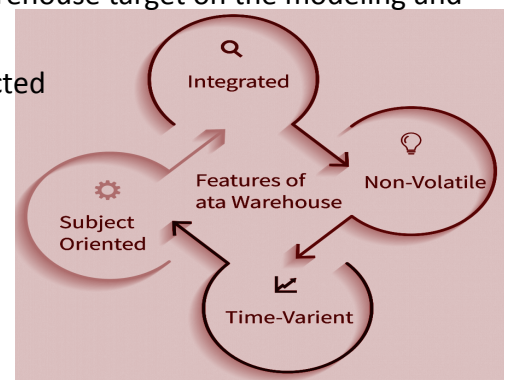
→**Data discretization and concept hierarchy generation**:- google important annu

***data warehouse**:--Data warehousing provides architectures and tools for business for organize, understand, and use their data to make strategic decisions.

-"Data Warehouse is a subject-oriented, integrated, and time-variant store of information in support of management's decisions."
-The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems

->Data Warehouse Properties / features;

- **Subject-oriented**:- A data warehouse is organized around major subjects, such as customer, supplier, product, and sales.A data warehouse target on the modeling and analysis of data for decision-makers.
- **Integrated**:-A data warehouse is usually constructed by integrating multiple heterogeneous sources.
  -Data cleaning and data integration techniques are applied.
- **Time-variant**:-Historical information is kept in a data warehouse.
  -For example, it can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse.
- **Non-volatile**:-Nonvolatile means that, once entered into the warehouse, data should not change.
  -This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.



***decision Support system in data mining:-**A decision support system (DSS) is a computerized program used to support determinations, judgments, and courses of action in an organization or a business.
-A decision support system helps in decision-making but does not necessarily give a decision itself.
-The decision makers compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions.
→**Benefits of DSS**

- Improves efficiency and speed of decision-making activities.
- Increases the control, competitiveness and capability of futuristic decision-making of the organization.

***Data objects:-**

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples , examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

***Differentiate data warehouse and database**

DATABASE	DATA WAREHOUSE
An organized collection of related data which stores data in a tabular format	A central location which stores consolidated data from multiple databases
Contains detailed data	Contains summarized data
Uses Online Transactional Processing (OLTP)	Uses Online Analytical Processing (OLAP)
Helps to perform fundamental operations of a business	Helps to analyze the business
Less fast and less accurate	Faster and accurate
Application oriented	Subject oriented
Tables and joins are complex because they are normalized	Tables and joins are simple because they are denormalized
Design is helped by entity relationship modelling	Design is helped by data modelling technique

Database	Data Warehouse
An organized collection of data.	A central repository of integrated data from one or more sources.
Usually tied to a single application such as a ticketing system	Usually store data from any number of applications
Primarily insert/write data	Primarily read/retrive data
Data is normalized to allow quick response times.	Data is denormalized for analytical and reporting efficiencies.
Current/Point-in-time data	Historical data
Online Transactional Processing	Online Analytical Processing
Provides a detailed relational view	Provides a summarized multidi-mensional view
For many concurrent transactions	Not for a large amount of concurrent transactions

Module -2

#Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods are;

1. **Market Basket Analysis:**-Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.
   -It involves analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.
   -This is a technique that gives the careful study of purchases done by a customer in a supermarket.
   -This concept identifies the pattern of frequent purchase items by customers.
   -**Market Basket Analysis is modelled on Association rule mining, i.e., the IF {}, THEN {} construct.**
   -For example, IF a customer buys bread, THEN he is likely to buy butter as well.

Association rules are usually represented as: {Bread} -> {Butter}

   -Algorithms that use association rules include **AIS, SETM and Apriori**.
   -The Apriori algorithm is commonly used for market basket analysis.
   -With the help of the Apriori Algorithm, we can further classify and simplify the item sets which are frequently bought by the consumer.
   -There are three components in APRIORI ALGORITHM:
   - ➢ SUPPORT:-It is been calculated with the number of transactions divided by the total number of transactions made,

Support = freq(A,B)/N

   - ➢ CONFIDENCE:-It is been calculated for whether the product sales are popular on individual sales or through combined sales.
     -That is calculated with combined transactions/individual transactions.

Confidence = freq(A,B)/freq(A)

   - ➢ LIFT:-Lift is calculated for knowing the ratio for the sales.

Lift= confidence percent/ support percent

   -When the Lift value is below 1 means the combination is not so frequently bought by consumers.
   →Some terminologies to familiarize yourself with Market Basket Analysis are:
   - Antecedent:Items or 'itemsets' found within the data are antecedents.
     -In simpler words, it's the IF component, written on the left-hand side.
     -In the above example, bread is the antecedent.

- Consequent:A consequent is an item or set of items found in combination with the antecedent.
  - It's the THEN component, written on the right-hand side.
  - In the above example, butter is the consequent.

→Benefits of Market Basket Analysis

- Enhanced Customer Understanding
- Better Pricing Strategies
- Sales Growth

→Applications of Market Basket Analysis

- Retail
- E-commerce
- Finance
- Telecommunications
- Manufacturing

Q) . Define itemset and frequent item sets
-itemset:-A collection of items or An itemset consists of two or more items.
-For example, all items bought by one customer during one visit to a department store.

2. **Frequent Itemsets:**--A set of items is referred to as an itemset.
   -An itemset that contains k items is a k-itemset.
   -The set {computer, antivirus software} is a 2-itemset.
   -The occurrence frequency of an itemset is the number of transactions that contain the itemset.
   -This is also known, simply, as the frequency, support count, or count of the itemset.
   -The set of frequent k-itemsets is commonly denoted by Lk.
   -The rule that satisfy both a minimum support threshold and a minimum confidence threshold are called strong.
   -The support and confidence value occur between 0% and 100%.

3. **Closed Itemsets :**-google imp (3 mark)
4. **Association Rules:**-If we think of the universe as the set of items available at the store,
   - each item has a Boolean variable representing the presence or absence of that item.
   -Each basket can be represented by a Boolean vector of values assigned to these variables.
   - The Boolean vectors can be analyzed items that are frequently associated or purchased together.
   -These patterns can be represented in the form of association rules.
   -For example, the customers who purchase computers also tend to buy antivirus software at the same time is represented in Association Rule is

Q)strong association rule?
An association rule having support and confidence greater than or equal to a user-specified minimum support threshold and respectively a minimum confidence threshold.

computer ⇒ antivirus software [support = 2%,confidence = 60%]

-A support of 2% for Association Rule means that 2% of al customers buy computer and antivirus software together.
-A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.
-Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min conf) are called **Strong Association Rules**.

***Apriori Algorithm:**-Apriori is an algorithm proposed by **R.Agrawal and R.Srikant** in 1994 for mining frequent itemsets for boolean association rules.
-This algorithm is used to find the frequent itemset.
-The key concepts in apriori algorithm

- Join:-join operation used to join the itemsets.
- Prune:- pruning step is used to execute the infrequent itemsets.

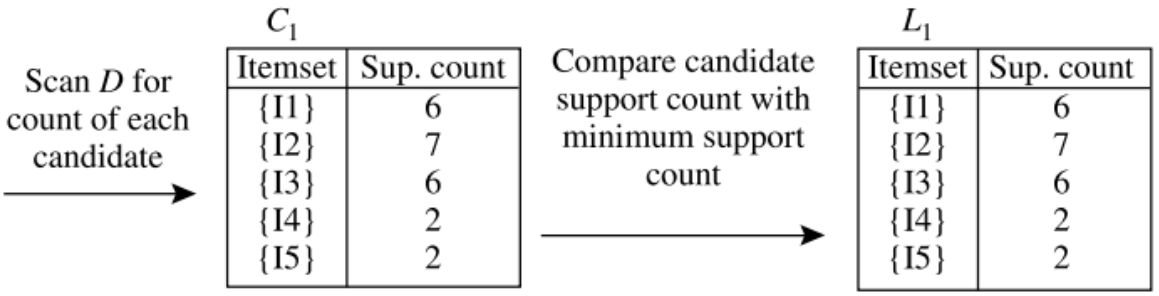
-following are the important steps in Apriori algorithm.

- Step1:-Tabulate set of frequent itemsets by scanning the database and enter the count of each itemset denoted as 'c1'.
- Step2:-collecting all itemset that satisfy minimum support denoted by L1.
- Step3:-Tabulate frequent two itemset with the help of L1 and denote it as c2.
- Step4:-collecting all itemset that satisfy minimum support denoted by L2.
- Step5:-repeat the process until no more frequent itemset can be found.

-Example consider database D contain 9 transactions. Each translation is represented as itemset. Minimum support required is 2 and minimum confidence required is 70 %.

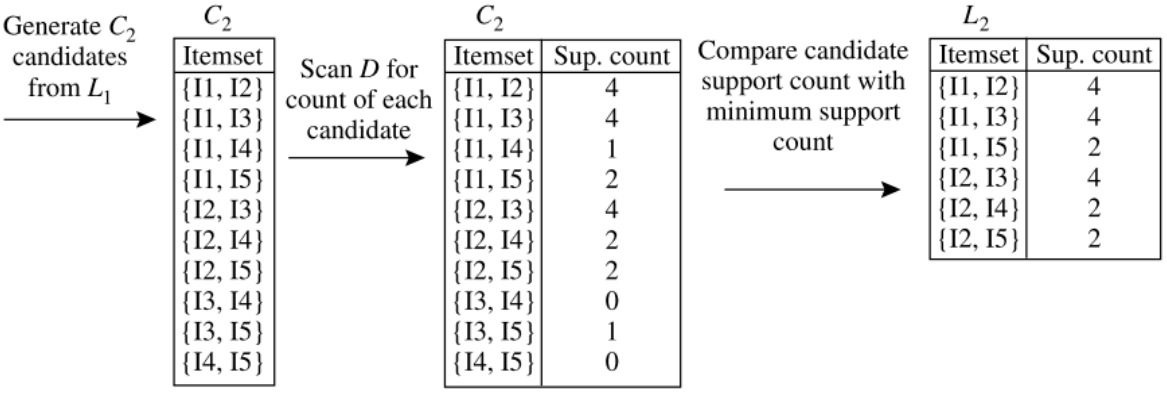
TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

-Scan database D for creation of c1 and compare candidate support count (sup.count) With minimum support count (which is 2).
-so there is no support count (sup.count) which is less than 2.
-so collecting all itemset than is greater than or equal to minimum support(which is 2) denoted by L1.

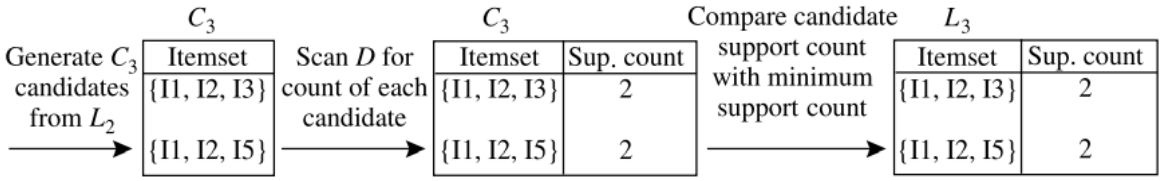


-Generate table c2 from L1 (with two item set).
-Scan the table c2 with database D for support count.
(eg:the combination of {I1,I2} is in database D is 4)
-compare support count of c2 with minimum support count (which is 2)
-There are {I1,I4},{I3,I4},{I3,I5} and {I4,I5} which support count (sup-count)is less than minimum count 2.
-so remove it and form L2.

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3



-Generate C3 from L2 (with 3 itemset)
-Scan the table C3 with database D for support count.
-compare support count of C3 with minimum support count (which is 2).
-there is no support count (sup.count) which is less than 2.
-collecting all itemset than is greater than or equal to minimum support(which is 2) denoted by L3.



Q) hash-based technique:- itcan be uesd to reduce the size of the candidate k-itermsets, Ck, for k>1.
-or it overcoming some of the weaknesses of the Apriori algorithm by reducing the number of candidate k-itemsets.

***Frequent Pattern(FP) Growth Algorithm:**-The FP-Growth Algorithm proposed by **Han in**.

-The two primary drawbacks of the Apriori Algorithm are:

- At each step, candidate sets have to be built.
- To build the candidate sets, the algorithm has to repeatedly scan the database.

-These two properties inevitably make the algorithm slower.

-To overcome these redundant steps, a new association-rule mining algorithm was developed named Frequent Pattern Growth Algorithm.

-A frequent pattern is generated without the need for candidate generation.

-FP growth algorithm represents the database in the form of a tree called a **frequent pattern tree or FP tree.**

-Eg:To start the FP growth algorithm we need only a translation table and a minimum support count.

Transaction ID	Items
1	{b, a}
2	{b, c, d}
3	{c, d, e, a}
4	{a, d, e}
5	{c, b, a}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{d, b, a}
10	{b, c, e}

Transaction Table

- Minimum Support Count=2

-The first step is to scan the database to find the occurrences of the itemsets in the database.

-This step is the same as the first step of Apriori.

- The count of 1-itemsets in the database is called support count or frequency of 1-itemset.

Transaction ID	Items
1	{b, a}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{c, b, a}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

Transaction Table

Items	Count
a	8
b	7
c	6
d	5
e	3

Count of Distinct Items in Transaction Table

-according to the count ,which have highest number must be first in the transaction items

Eg: consider first transaction items {b,a} here a value is 8 and b value is 7 .

-so rearrange them to {b,a} to {a,b}

-we need to rearrange them all according to there count number.

Transaction ID	Items	Rearranged items
1	{b, a}	{a, b}
2	{b, c, d}	{b, c, d}
3	{a, c, d, e}	{a, c, d, e}
4	{a, d, e}	{a, d, e}
5	{c, b, a}	{a, b, c}
6	{a, b, c, d}	{a, b, c, d}
7	{a}	{a}
8	{a, b, c}	{a, b, c}
9	{a, b, d}	{a, b, d}
10	{b, c, e}	{b, c, e}

Items	Count
a	8
b	7
c	6
d	5
e	3

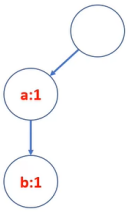
Count of Distinct Items in Transaction Table

-Then onwards we do not use transaction items column but instead we use transaction rearrange items column.

Transaction ID	Items	Rearranged items
1	{b, a}	{a, b}
2	{b, c, d}	{b, c, d}
3	{a, c, d, e}	{a, c, d, e}
4	{a, d, e}	{a, d, e}
5	{c, b, a}	{a, b, c}
6	{a, b, c, d}	{a, b, c, d}
7	{a}	{a}
8	{a, b, c}	{a, b, c}
9	{a, b, d}	{a, b, d}
10	{b, c, e}	{b, c, e}

-The next step is to construct the FP tree from translation table.
-First create the root for the tree. The root is represented by null.
-so we start the tree by translation 1.
-after completing translation is 1.
We have a:1 node and b:1 node
-the tree starts from null to a:1 because a is the first element in the Translation item.
-a:1 and b:1 indicate that a is one and b is one.
-after completion we go to translation Id 2.

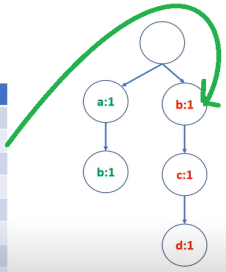
Transaction ID	Rearranged items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}



After Transaction id 1

-Here we have items {b,c,d}
-so we start the node which is close to the root As b ,because its start with b .

Transaction ID	Rearranged items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

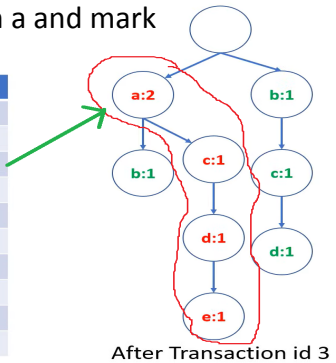


After Transaction id 2

-Next we have {a,c,d,e}.

-And we already have a node that is starting from a so we draw from a and mark a:1 to a:2 because we have 2 a value.

Transaction ID	Rearranged items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

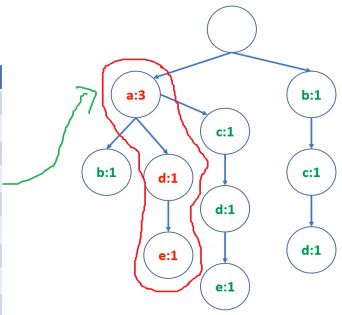


After Transaction id 3

-next items are {a,b,e}

-It also start from a so we mark a:2 to a:3

Transaction ID	Rearranged items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}



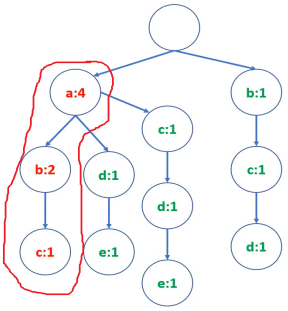
After Transaction id 4

-next item are {a,b,c}

-here we already have node a and b .

-so mark a:3 to a:4 and b:1 to b:2.

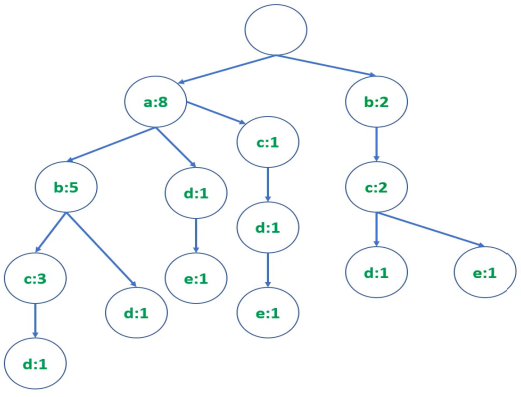
Transaction ID	Rearranged items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}



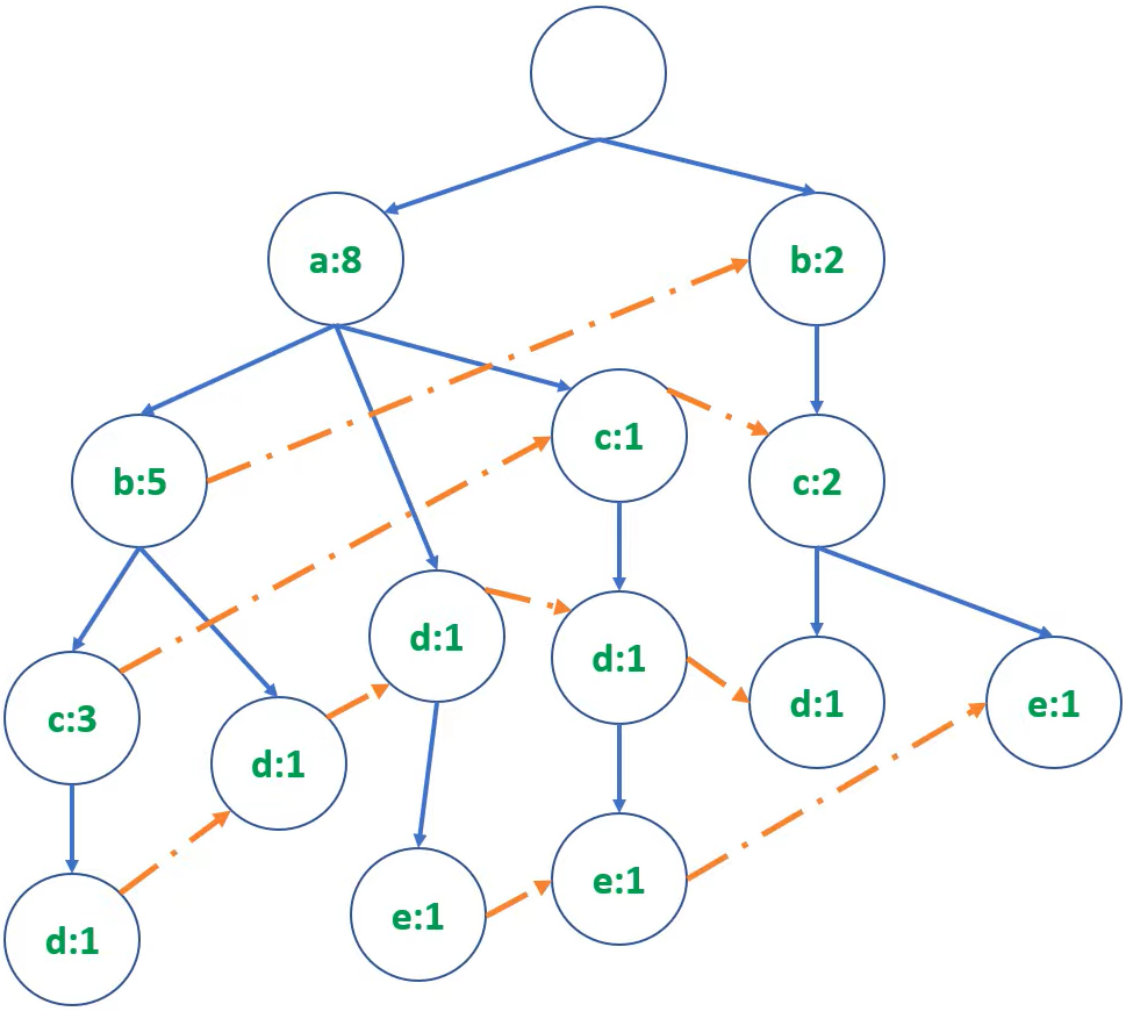
After Transaction id 5

-After the completion of all translations in the translation table we get FP tree.

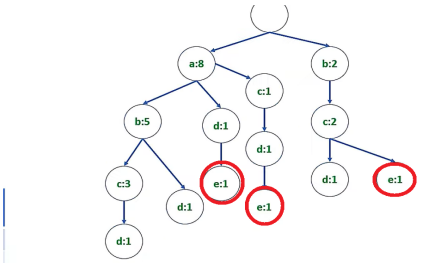
Transaction ID	Rearranged items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}



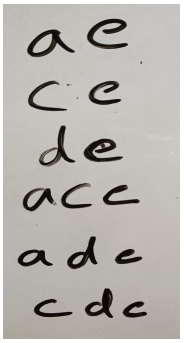
-we can draw arrows to between corresponding nodes.



-so once we obtain the FP tree and we find the frequent itemset which is ended with e.
-To find the itemset which are ending the e ,we need to consider the path in the FP tree which are ending with e.
-This are the paths that ending with e:
{acd:1},{ad:1},{bc:1}
-with the path we count the items in the path
For example we have 3 a in the path
-and 1 b,2 c,and 2 d.
-our support cout is 2 so we cancel less than 2 items ,here b:1 so we cancel it.
-in next column we try all the possibilities from The “count of each item in path” column that Ends with e.
-This are the combinations that can be formed From column 3.



Ending with	Paths	Count of each item in path	Candidate itemset with count of each w. r. t. transaction table	Frequent Itemset
e	acd:1, ad:1, bc:1	a:2, b:1, c:2, d:2	ae: 2, ce: 2, de: 2 , ace: 1, ade: 2 , cde: 1, e: 3	ae: 2, ce: 2, de: 2 , ade: 2 , e: 3



-and we cancel all the items which are less than support count.
And at last we form frequent itemset which is ending with e.
-next we form a frequent item set which is ending with d,c,b and a.

Ending with	Paths	Count of each item in path	Candidate itemset with count of each w. r. t. transaction table	Frequent Itemset
e	acd:1, ad:1, bc:1	a:2, **b:1,** c:2, d:2	ae: 2, ce: 2, de: 2 , ace: 1, ade: 2 , cde: 1, e: 3	ae: 2, ce: 2, de: 2, ade: 2 , e: 3
d	abc:1, ab:1, a:1, ac: 1, bc:1	a:2, b:3, c:3	ad:4, bd:3, cd:3, abd:2, acd:2, bcd:2, abcd:1	ad:4, bd:3, cd:3, abd:2, acd:2, bcd:2
c	ab:3, a:1, b:2	a:4, b:5	c:6, ac: 4, bc:3, abc:2	c:6, ac: 4, bc:3, abc:2
b	a:5	a:5	ab:5, b:7	ab:5, b:7
a	-	-	a:8	a:8

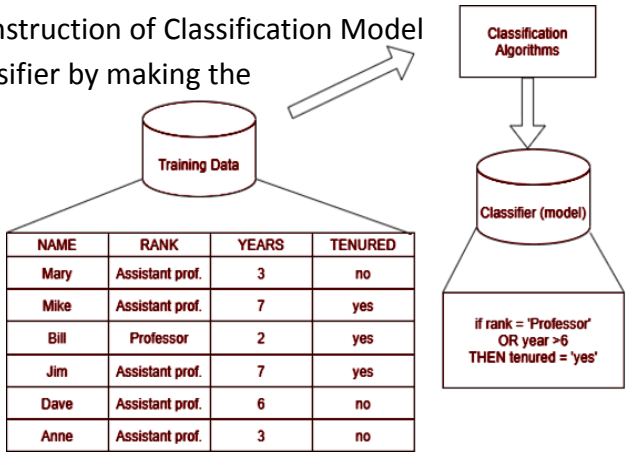
-from this table we can form the table

Ending with	Frequent Item sets
e	{ae, ce, de, ade , e}
d	{ad, bd, cd, abd, acd, bcd}
c	{c, ac, bc, abc}
b	{ab, b}
a	{a}

***Classification in Data mining**:-classification is a task of assigning objects to one of several predefined categories.
-Eg: see if you have student records then the student records can be assigned to one of the class labels or the categories like first class ,second class, third class or fail.
-The student records have a predefined categories like first class ,second class ,or a fail .
-Data classification systems have become a regular aspect of our everyday lives.
-The AI behind the spam folder in your inbox, for instance, uses data classification techniques to filter emails.
-The use of classification has become crucial to maintaining a clean and efficient data environment.

-There are 2 steps involved in Data Mining Classification they are;

- **Step 1: Learning Phase(Training Phase):** Construction of Classification Model Different Algorithms are used to build a classifier by making the model learn using the training set available.



NAME	RANK	YEARS	TENURED
Mary	Assistant prof.	3	no
Mike	Assistant prof.	7	yes
Bill	Professor	2	yes
Jim	Assistant prof.	7	yes
Dave	Assistant prof.	6	no
Anne	Assistant prof.	3	no

Or

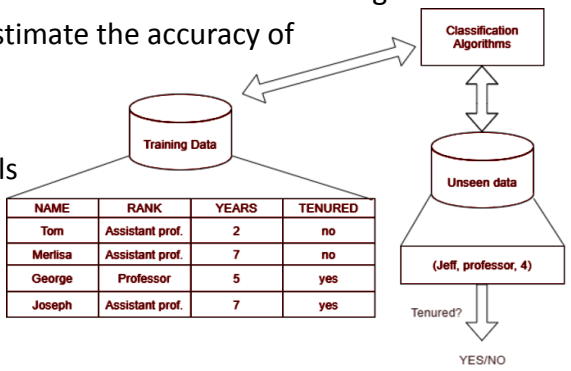
-This phase of Data Mining Classification mainly deals with the construction of the Classification model based on different algorithms available.

-This step requires a training set for the model to learn.

-The trained model gives accurate results based on the target dataset.

-When the test data is added to the model it provides accuracy to the Classification Model created.

- **Step 2: Classification Phase**:-Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.



NAME	RANK	YEARS	TENURED
Tom	Assistant prof.	2	no
Merlisa	Assistant prof.	7	no
George	Professor	5	yes
Joseph	Assistant prof.	7	yes

Or

-This phase of Data Mining Classification deals with testing the model that was created by predicting the class labels.

-This also helps in determining the accuracy of the model in real test cases.

→Importance of classification in data mining:-The data classification work process begins with a learning step, where training data is fed to an algorithm. Afterward, a classification rule or a set thereof is developed, which will be used by the algorithm to analyze the test data and produce new results.

-This are the Classifiers used they are;

1. **Decision Trees**:-This is the most robust Classification Technique for Data Mining.
   -It follows a flowchart similar to the structure of a tree.
   -The leaf nodes hold the classes and their labels.
   -The internodes have a Decision algorithm that routes it to the nearest leaf node. There can be multiple internal nodes to do this.
   -The horizontal and vertical phases can be prediction boundaries.
   -The only challenge is that it is complex, and requires expertise to create and ingest data into it.

2. **Bayesian Classifiers:**-The Naive Bayes Algorithm makes the assumption that every independent parameter will equally affect the outcome and has almost equal importance.
   -It calculates the probability of the event occurring, given that an event has already occurred.
   -Naive Bayes requires smaller training sets to learn.
   -It is faster in predicting when compared to other models.
   -It is plagued with the poor estimation issue where all the parameters have equal importance.
   - It doesn't provide results that are true in the real world.
   Or
   -This algorithm calculates the probability of a particular piece of data belonging to a category, and then classifies it accordingly.
   -It can be used to analyze vast amounts of data to find particular snippets that relate to a certain subject.

***Cluster:-**group of data into clusters so that the objects belong to the same group.
-Clustering helps to splits data into several subsets.
-Each of these subsets contains data similar to each other, and these subsets are called clusters.
-clustering analysis is widely used, such as data analysis, market research, pattern recognition, and image processing.
-Clustering analysis has been an solve problem in data mining due to its variety of applications.

→**Application of clustering:-** Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
-cluster can used In the field of biology, it can be used to derive plant and animal taxonomies, categorize with similar functionalities.
-It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
-Clustering also helps in classifying documents on the web for information discovery.
-Clustering is also used in outlier detection applications such as detection of credit card fraud.

→**Requirement of clusters / properties**

- **Scalability −** We need highly scalable clustering algorithms to deal with large Databases.
- **Ability to deal with different kinds of attributes −** Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.

- **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

*Main memory-based clustering algorithms typically operate on either of the following two data structures.

- **Data matrix:**-This represent n object , such as person , with p variable such as age, height, gender and so on.
  -the structure is in the form of a relational table or n-by-p matrix (n object * p variables)

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- **Dissimilarity matrix:-**This stores a collection of proximities that are available for all pairs of n objects.
  -It is often represented by an n-by-n table:--->

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

***Clustering Methods:-**The clustering methods can be classified into the following categories:

1. **Partitioning Method:-*Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data.***
   -Each partition will represent a cluster and k ≤ n.
   -It means that it will classify the data into k groups, which satisfy the following requirements
   - **Each group contains at least one object.**
   - **Each object must belong to exactly one group.**

   -There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications) etc.

2. **Hierarchical Method:**-A Hierarchical clustering method works via grouping data into a tree of clusters.
   -Hierarchical clustering stats by treating each data points as an individual cluster.
   -where each cluster is different from the other cluster, and the objects within each cluster
   are the same as one another.

-There are two types of hierarchical clustering

● **Agglomerative Hierarchical Clustering**:-Agglomerative clustering is one of the most common types of hierarchical clustering used to group similar objects in clusters.

-Agglomerative clustering is also known as AGNES (Agglomerative Nesting).

-In agglomerative clustering, each data point act as an individual cluster.

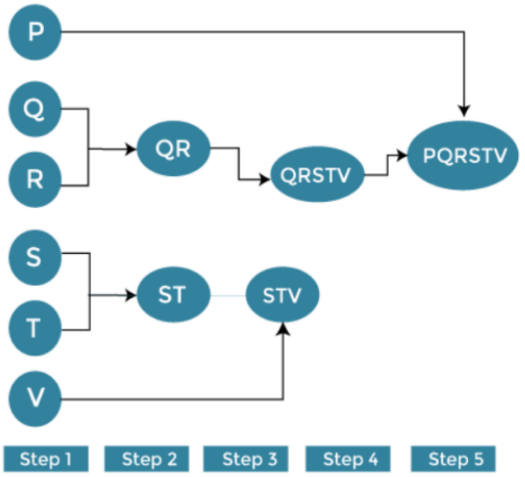
-at each step, data objects are grouped in a bottom-up method.

-At each iteration, the clusters are combined with different clusters until one cluster is formed.

->example

-Let's suppose we have six different data points P, Q, R, S, T, V.

-step1:-Consider each alphabet (P, Q, R, S, T, V) as an individual cluster and find the distance between the individual cluster from all other clusters.



-step2:-Now, merge the comparable clusters in a single cluster. Let’s say cluster Q and Cluster R are similar to each other so that we can merge them in the second step. Finally, we get the clusters [(P), (QR), (ST), (V)]

-step3:-Here, we recalculate the proximity as per the algorithm and combine the two closest clusters [(ST), (V)] together to form new clusters as [(P), (QR), (STV)]

-step4:-Repeat the same process. The clusters STV and PQ are comparable and combined together to form a new cluster. Now we have [(P), (QQRSTV)].

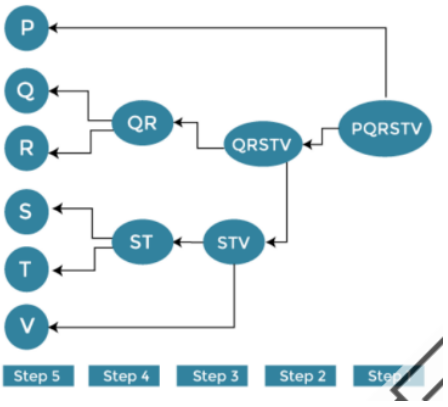
-step5:-Finally, the remaining two clusters are merged together to form a single cluster [(PQRSTV)]

● **Divisive Clustering:-**Divisive hierarchical clustering is exactly the opposite of Agglomerative Hierarchical clustering.

- In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster.

-The separated data points are treated as an individual cluster.

-Finally, we are left with N clusters.



->Advantages of Hierarchical clustering

-It is simple to implement.

-It is easy.

-It does not need us to pre-specify the number of clusters.

->Disadvantages of hierarchical clustering

-It breaks the large clusters.

-It is Difficult to handle different sized clusters and convex shapes.

-It is sensitive to noise and outliers.

3. **Density-based Method:**-In this method of clustering in Data Mining, density is the main focus.

   -In this clustering method, the cluster will keep on growing continuously.

   -At least one number of points should be there in the radius of the group for each point of data.

   -It identifies clusters of any shape in a data set, it means it can detect arbitrarily shaped clusters.

   --It is a scan method.

   -It is used to manage noise in data clusters.

   -Density-based clustering is used to identify clusters of arbitrary size

4. **Grid-Based Method:**-In this type of Grid-Based Clustering Method, a grid is formed using the object together.

   -A Grid Structure is formed by quantifying the object space into a finite number of cells.

   →Advantage of Grid-based clustering method;

   - Faster time of processing: The processing time of this method is much quicker than another way, and thus it can save time.
   - This method depends on the no. of cells in the space of quantized each dimension.

***Compare Classification and Prediction. (3 marks)**

- **Accuracy**: Accuracy of the classifier can be referred to as the ability of the classifier to predicts the class label correctly, and the accuracy of the predictor can be referred to as how well a given predictor can estimate the unknown value.
- **Speed:** The speed of the method depends on the computational cost of generating and using the classifier/predictor.
- **Robustness**: Robustness is the ability to make correct predictions or classifications, in the context of data mining robustness is the ability of the classifier or predictor to make correct predictions from incoming unknown data.

- **Scalability**: Scalability is referring to an increase or decrease in performance of the classifier or predictor based on the given data.
- **Interpretability**: Interpretability can be referred to as how readily we can understand the reasoning behind predictions or classification made by the predictor or classifier.

→ Differentiate training data set and test data set? (3 marks)

- **Training data** is the biggest (in -size) subset of the original dataset, which is used to train .
  -Firstly, the training data is fed to the algorithms, which lets them learn how to make predictions for the given task.
- **Test data**:Once we train the model with the training dataset, it's time to test the model with the test dataset.
  -This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset.
  -The test dataset is another subset of original data, which is independent of the training dataset.

→ Compare supervised learning and unsupervised learning process (3 marks)

SUPERVISED LEARNING	UNSUPERVISED LEARNING
Uses Known and Labeled Data as input	Uses Unknown Data as input
Less Computational Complexity	More Computational Complex
Uses off-line analysis	Uses Real Time Analysis of Data
Number of Classes are known	Number of Classes are not known
Accurate and Reliable Results	Moderate Accurate and Reliable Results
Desired output is given.	Desired output is not given.
In supervised learning it is not possible to learn larger and more complex models than with supervised learning	In unsupervised learning it is possible to learn larger and more complex models than with unsupervised learning
In supervised learning training data is used to infer model	In unsupervised learning training data is not used.
Supervised learning is also called classification.	Unsupervised learning is also called clustering.
We can test our model.	We can not test our model.
Optical Character Recognition	Find a face in an image.

Module-3-Introduction to Data Science

***Data science:**-Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.
-Data science and big data are used almost everywhere in both commercial and noncommercial fields.
-In Commercial companies almost every industry use data science and big data to understanding their customers, processes, staff, completion, and products.
-Many companies use data science to offer customers a better user experience, as well as to cross-sell, up-sell, and personalize their offerings.

***Facets of data science**:-In data science and big data you'll come across many different types of data, and each of them tends to require different tools and techniques.
-The main categories of data are these: (7 types)

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming

1. **Structured:**-In structured data means data will be arranged in row and column.
   -Structured data is the data that depends on a data model and locate in a fixed field within a record.
   -It's often easy to store structured data in tables within data bases or Excel files.
   -SQL, or Structured Query Language, is the preferred way to manage and locate data in databases.
   -It may also come across structured data that might give you a hard time storing it in a traditional relational database.
   -Hierarchical data such as a family tree is one such example.

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%

2. **Unstructured:**-Unstructured data is data that isn't easy to fit into a data model.
   -because the content is context-specific or varying.
   -One example of unstructured data is your regular email.
   -A human-written email, is also a perfect example of natural language data.
3. **Natural language:**-Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific **data science techniques and linguistics.**
   -for example the conversation making through letters ,email, text,essay etc.. all this are represented in your natural language.
4. **Machine-generated:**-Machine-generated data is that's automatically created by a computer, process, application or other machine without human intervention.
   -Machine-generated data is becoming a major data resource and will continue to do so.
   -Due to the huge amount and speed of machine data, highly scalable technologies are required for analysis.
5. **Graph-based:**-"Graph data" can be a confusing because any data can be shown in a graph.
   -The graph structures use nodes, edges, and properties to represent and store graphical data.
   -Graph-based data is a natural way to represent social networks, and its structure allows you to calculate specific metrics such as the shortest path between two people etc.
   -Examples of graph-based data can be found on many social media websites.

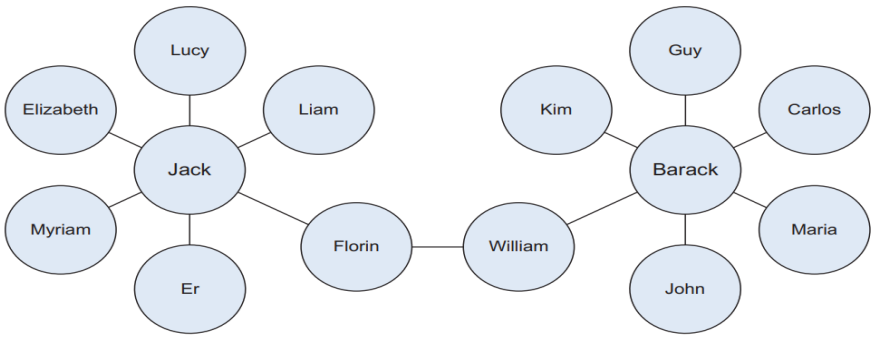
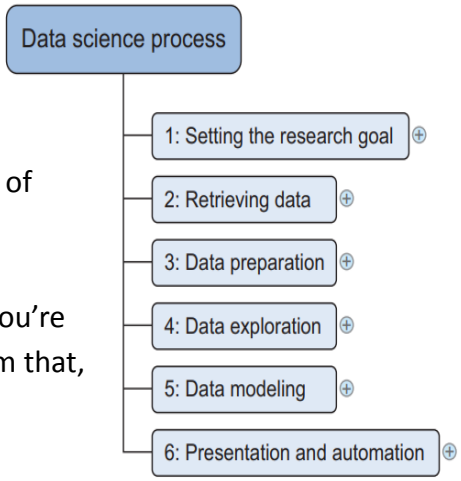


Figure 1.4 Friends in a social network are an example of graph-based data.

6. **Audio, video, and images:**-Audio, image, and video are data types that create specific challenges to a data scientist.
   -Tasks that easy for humans, such as recognizing objects in pictures, turn out to be challenging for computers.

7. **Streaming:**-streaming data can take almost any of the previous forms, it has an extra Property.

***Data Science process:**-The data science process typically consists of six steps they are;



1. **Setting the research goal:**- The first step of this process is setting a research goal.
   -The main purpose here is making sure all the stakeholders understand the what, how, and why of the project.
   -And prepare a project charter.
   -This charter contains information such as what you're going to research, how the company benefits from that, what data and resources you need, a timetable, and deliverables.
2. **Retrieving data:**-The second step is to collect data.
   -You've stated in the project charter which data you need and where you can find it.
   -In this step you ensure that you can use the data in your program, which means checking the existence of, quality, and access to the data.
   -Data can also be delivered by third-party companies
3. **Data preparation:**-Now that you have the raw data, it's time to prepare it.
   -This includes transforming the data from a raw form into data that's directly usable in your models.
   -To achieve this, you'll detect and correct different kinds of errors in the data, combine data from different data sources, and transform it.
   -If you have successfully completed this step, you can progress to next step.
4. **Data exploration:**-The fourth step is data exploration.
   -The goal of this step is to gain a deep understanding of the data.
   -The insights you gain from this phase will enable you to start modeling.
   - To achieve this you mainly use descriptive statistics, visual techniques, and simple modeling.
5. **Data modeling or model building:**-The fifth step is model building.
   -In this step, the actual model building process starts.
   -Building a model is an iterative process that involves selecting the variables for the model, executing the model, and model diagnostics.
6. **Presentation and automation:**-Finally, you present the results to your business.
   -This helps you decide if the project results are a success or a failure.

*Explain Web Analytics and Credit risk management (9 marks):-google

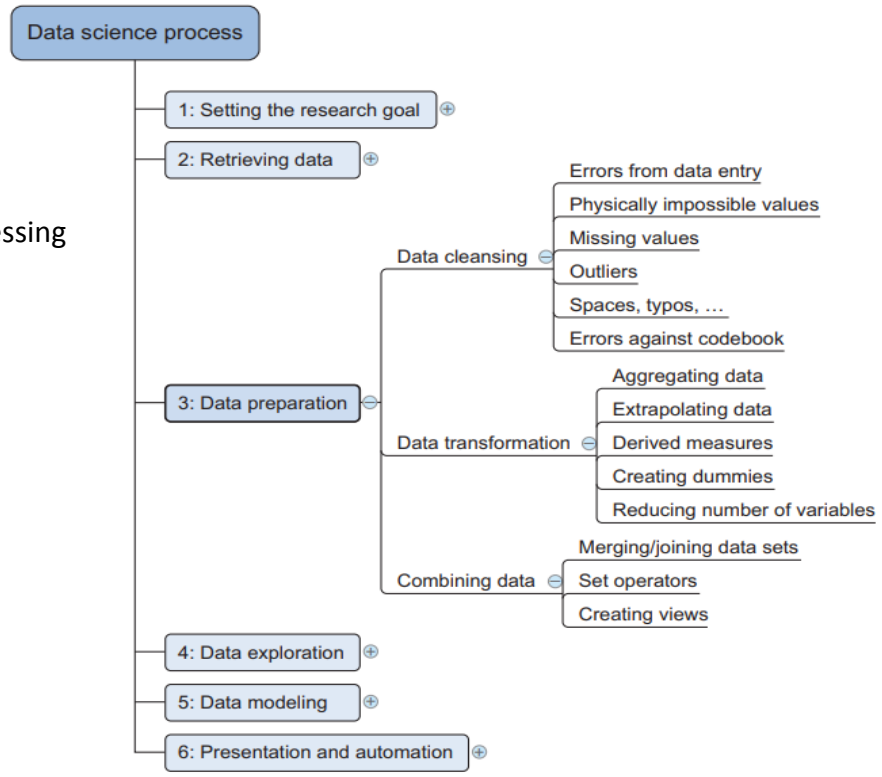
*** steps involved in data preparation in Data Science process** (9marks)**:-**

-Cleaning and transforming raw data before processing and analysis is known as **data preparation**.

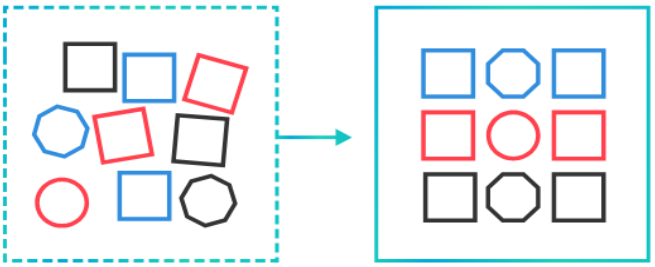
-Bakki front ill unde 👆

-The steps involved in preprocessing Are;

- **Data cleaning**
- **Data transformation**
- **Combining data**



1. **Data cleaning:**-Data cleansing is a subprocess of the data science process that focuses on removing errors in your data.
   -In Cleansing the data ensures that the data set can provide valid answers when the data is analyzed.
   -This step could be done manually for small data sets but requires automation for most realistically sized data sets.
   -There are software tools available for this processing.
   -There are many different problems possible with the data .
   -There could be **data entry errors ,missing values, out-of-range values, nulls,** and **whitespaces** that obfuscate values, as well as **outlier values** .
   - **Data entry errors**:-Data collection and data entry are error-prone processes.
     - Errors can arise from human sloppiness, and others are due to machine or hardware failure.
   - **White Spaces**:- White spaces tend to be hard to detect.
     Bakki thanne ezhuthukaa
2. **Data transformation:**-Data transformation is the process of converting, cleansing, and structuring data into a usable format.



-or the data transformation process converts raw data into a usable format by removing duplicates.

→**Reducing number of variables**:-Sometimes you have too many variables and need to reduce the number because they don't add new information to the model.

-Having too many variables in your model makes the model difficult to handle, and certain techniques don't perform well when you overload them with too many input variables.

→**Turning variables into dummies:**-Variables can be turned into dummy variables .

Customer	Year	Gender	Sales
1	2015	F	10
2	2015	M	8
1	2016	F	11
3	2016	M	12
4	2017	F	14
3	2017	M	13

M F

Customer	Year	Sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0
3	2016	12	1	0
3	2017	13	1	0
4	2017	14	0	1

Figure 2.13 Turning variables into dummies is a data transformation that breaks a variable that has multiple classes into multiple variables, each having only two possible values: 0 or 1.

-Dummy variables can only take two values: **true(1) or false(0)**.

-In this case you'll make separate columns for the classes stored in one variable and indicate it with 1 if the class is present and 0 otherwise.

-An example is turning one column named Weekdays into the columns Monday through Sunday.

-You use an indicator to show if the observation was on a Monday; you put 1 on Monday and 0 elsewhere.

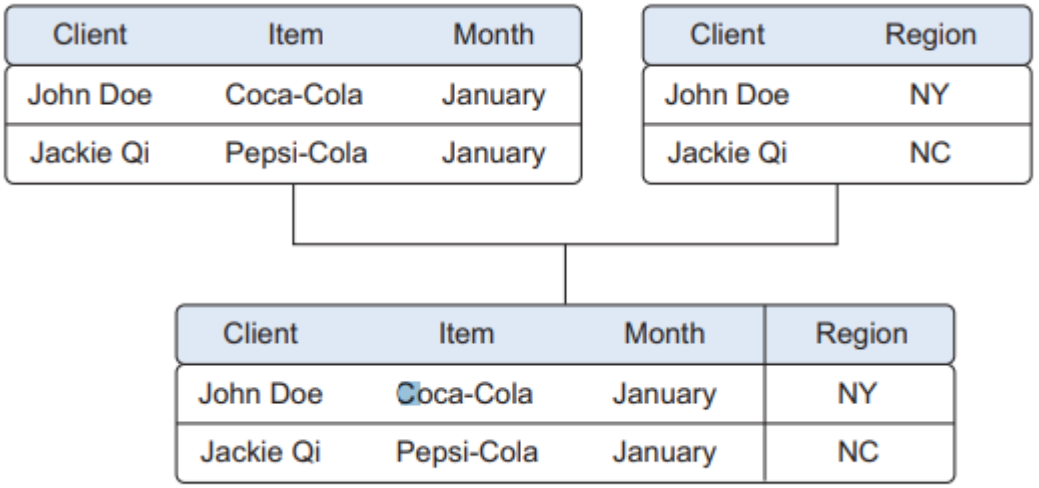
3. **Combining data:**-Your data comes from several different places, and in this substep we focus on integrating these different sources.

   -Data varies in size, type, and structure, ranging from databases .

   -You can perform two operations to combine information from different data sets.

   -The first operation is **joining**: It allows you to combine the information of one observation found in one table with the information that you find in another table.

-Eg:Let's say that the first table contains information about the purchases of a customer and the other table contains information about the region where your customer lives. Joining the tables allows you to combine the information .



-The second operation is **appending or stacking**: adding the observations of one table to those of another table.

-The below Figure shows an example of appending tables.

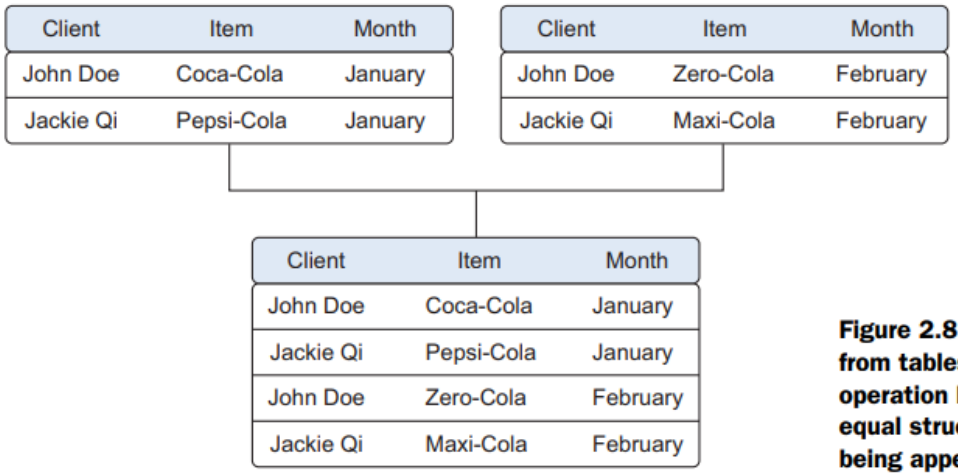


Figure 2.8 Appending data from tables is a common operation but requires an equal structure in the tables being appended.

-One table contains the observations from the month January and the second table contains observations from the month February.

The result of appending these tables is a larger one with the observations from January as well as February.

*** Big Data in healthcare** (9 marks) **:**-One of the most notable areas where data analysis is making big changes is healthcare.

-In fact, healthcare analytics has the potential to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases, and improve the quality of life in general.

-The average human lifespan is increasing across the world population, which poses new challenges to today's treatment delivery methods.
-Health professionals, just like business entrepreneurs, are capable of collecting massive amounts of data and looking for the best strategies to use these numbers.
-Big data is revolutionizing the healthcare industry and changing how we think about patient care.
-In this case, big data refers to the vast amounts of data generated by healthcare systems and patients, including electronic health records, claims data, and patient-generated data.
-With the ability to collect, manage, and analyze vast amounts of data, healthcare organizations can now identify patterns, trends, and insights that can inform decision-making and improve patient outcomes.
-big data also poses challenges and limitations that must be addressed.
-Such challenges and limitations include managing and analyzing vast amounts of data, and ethical considerations, such as patient privacy.

-Data collection and management are crucial aspects of using big data in healthcare decision-making.
-The types of data collected in healthcare include **electronic health records (EHRs), claims data, and patient-generated data.**
-EHRs contain a wide range of patient information, such as medical history, medications, and lab results, which can be used to identify patterns and trends in patient care.
-On the other hand, claims data includes information about insurance claims, such as the cost of treatments, and can be used to identify patterns in healthcare spending.
-Patient-generated data, including data from wearables, surveys, and patient-reported outcomes, can provide valuable insights into patients' experiences and preferences.

-However, collecting and managing all this data can be challenging.
-For instance, data may be stored in different systems, which makes it difficult to integrate and analyze.
-Data may also be incomplete or inaccurate, leading to inaccurate conclusions.
-Privacy and security concerns are also significant challenges, as healthcare organizations must protect patient data from unauthorized access.

-To overcome these challenges, healthcare organizations must have robust data management and security systems in place.
-This includes investing in data integration and warehousing tools to enable data to be easily integrated and analyzed.

→ **The three V's of Big Data (3 marks)**:-The 3 V's (volume, velocity and variety) are three defining properties of big data.
-Volume refers to the amount of data,

-velocity refers to the speed of data processing,
-and variety refers to the number of types of data.

→ **Explain Big Data and algorithmic trading** (3 marks):-Big data is a collection of data from many different sources and is often describe by v's : volume, variety and velocity.
-The data in the big data contains greater variety, arriving in increasing volumes and with more velocity.
-Algorithm trading is the use of computer programs for entering trading orders, in which computer programs decide on almost every aspect of the order, including the timing, price, and quantity of the order etc.

→ **Applications of Big Data** (3) :-The term Big Data is referred to as large amount of complex and unprocessed data.
-Now a days companies use Big Data to make business more informative and allow them to make business decisions.

1. **Financial and banking sector**:-Big data analytics help banks and customer behaviour on the basis of investment patterns, shopping trends, motivation to invest, and inputs that are obtained from personal or financial backgrounds.
2. **Healthcare**:-Big data has started making a massive difference in the healthcare sector, with the help of predictive analytics, medical professionals, and health care personnel. It can produce personalized healthcare and solo patients also.
3. **Government and Military**:-The government and military also used technology at high rates.
4. **E-commerce**:-E-commerce is also an application of Big data. It maintains relationships with customers that is essential for the e-commerce industry.
   -E-commerce websites have many marketing ideas to retail merchandise customers, manage transactions, and implement better strategies of innovative ideas to improve businesses with Big data.
5. **Social Media**:-Social Media is the largest data generator. The statistics have shown that around 500+ terabytes of fresh data generated from social media daily, particularly on Facebook.
   -The data mainly contains videos, photos, message exchanges, etc.
6. **Advertisement** (3 marks):-ig data allows your company to accumulate more data on your visitors so you can target consumers with tailored advertisements that they are more likely to view
   -Advertisers identify their target audience based on demographics, past customers, and other factors that might suggest the user would be interested in the ad.

Module - 4

***Big data technologies** :-This technology is primarily designed to analyze, process and extract information from a large data set and a huge set of extremely complex structures.
-This is very difficult for traditional data processing software to deal with.
-Big Data technology is primarily classified into the following two types:

1. **Operational Big Data Technologies**:-This type of big data technology mainly includes the basic day-to-day data that people used to process.
   -Typically, the operational-big data includes daily basis data such as online transactions, social media platforms etc.
   Eg:Online ticket booking system, e.g., buses, trains, flights, and movies, etc.
   -Online trading or shopping from e-commerce websites like Amazon, Flipkart, Walmart, etc.
2. **Analytical Big Data Technologies**:-Analytical Big Data is commonly referred to as an improved version of Big Data Technologies.
   -This type of big data technology is a bit complicated when compared with operational-big data.
   -Analytical big data is mainly used when performance criteria are in use, and important real-time business decisions are made based on reports created by analyzing operational-real data.
   Eg:-Stock marketing data
   -Weather forecasting data and the time series analysis
   (randum ezhuthanam technology ennu chothichal 👆👇)
   –Top Big Data Technologies:- We can categorize the leading big data technologies into the following four sections:

- **Data Storage**:-The leading Big Data Technologies that come under Data Storage are;
  - **Hadoop**:-When it comes to handling big data, Hadoop is one of the leading technologies that come into play.
    -This technology is based entirely on map-reduce architecture .
    -Hadoop is also best suited for storing and analyzing the data from various machines with a faster speed and low cost.
    -That is why Hadoop is known as one of the core components of big data technologies.
    -Hadoop is written in Java programming language.
  - **MongoDB**:-MongoDB is another important component of big data technologies in terms of storage.
    -No relational properties and RDBMS properties apply to MongoDb because it is a NoSQL database.
    -The structure of the data storage in MongoDB is also different from traditional RDBMS databases.
    -This enables MongoDB to hold massive amounts of data.

- **Data Mining**:-This are the leading Big Data Technologies that come under Data Mining:
  - ➢ **Presto:-** Presto is an open-source.
    -Presto is a Java-based query engine.
    -The size of data sources can vary from gigabytes to petabytes.
    - Companies like Repro, Netflix, Airbnb, Facebook and Checkr are using this big data technology and making good use of it.
- **Data Analytics**:-This are the leading Big Data Technologies that come under Data Analytics:
  - ➢ **Apache Kafka**:-Apache Kafka is a popular streaming platform.
    -This streaming platform is primarily known for its three core capabilities: publisher, subscriber and consumer.
    -It is referred to as a distributed streaming platform.
    -It is written in Java language.
    -Some top companies using the Apache Kafka platform include Twitter, Spotify, Netflix, Yahoo, LinkedIn etc.
  - ➢ **R-Language**:-R is defined as the programming language, mainly used in statistical computing and graphics.
    -It is a free software environment used by leading data miners, practitioners and statisticians.
    -Language is primarily beneficial in the development of statistical-based software and data analytics.
- **Data Visualization**:-This are the eading Big Data Technologies that come under Data Visualization:
  - ➢ **Tableau**:-Tableau is one of the fastest and most powerful data visualization tools used by leading business intelligence industries.
    -It helps in analyzing the data at a very faster speed.
    -Tableau is developed and maintained by a company named TableAU.
    -It is written using multiple languages, such as Python, C, C++, and Java.
  - ➢ **Plotly**:-As the name suggests, Plotly is best suited for plotting or creating graphs and relevant components at a faster speed in an efficient way.

*Explain:-

1. **Structuring Big data/Types Of Big Data**:-Big data structures can be divided into following three categories they are;
   - **Structured**:-Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data.
     -It is in a tabular form.
     -Structured Data is stored in the relational database management system.

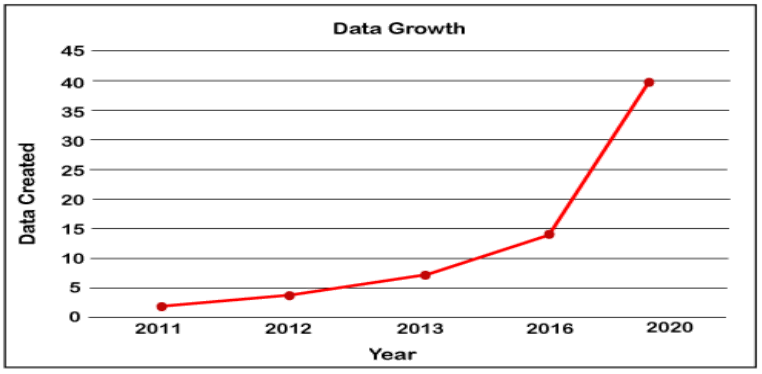
-Examples Of Structured Data

-An ‘Employee’ table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

- **Unstructured:**-Any data with unknown form or the structure is classified as unstructured data.
  In addition to the size being huge.
  -A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.
  -All the unstructured files, log files, audio files, and image files are included in the unstructured data
- **semi-structured:**-Semi-structured data can contain both the forms of data.
  -We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.
  -Example of semi-structured data is a data represented in an XML file.
  -JSON, XML, CSV, TSV, and email. OLTP (Online Transaction Processing) systems are built to work with semi-structured data.

2. **Elements of Big data/Characteristics Of Big Data**:-Big data can be described by the following characteristics/ four v’s
   - **Volume**:-The name Big Data itself is related to an enormous size.
     -Big Data is a vast 'volumes' of data generated from many sources daily, such as business processes, machines, social media platforms, networks, human interactions, and many more.
     -hether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.



- **Variety**:-Big Data can be **structured, unstructured,** and **semi-structured** that are being collected from different sources.
  -Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in **array for**ms, that are **PDFs, Emails, audios, SM posts, photos, videos**, etc
- **Velocity**:-Velocity plays an important role compared to others.
  -Velocity creates the speed by which the data is created in real-time.
  -Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc.
- **Variability**:-This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

3. **Big Data Analytics:**-front ill unde 👆

***Cloud computing and big data** (9 marks) :-Big Data and Cloud Computing as two mainstream technologies, are at the center of the IT field.
-Every day a huge amount of data is produced from different sources.
-This data is so big in size that traditional processing tools are unable to deal with them.
-Besides being big, this data moves fast and has a lot of variety.
-Big Data is a concept that deals with storing, processing and analyzing large amounts of data.
-Cloud computing on the other hand is about offering the infrastructure to enable such processes in a cost-effective and efficient manner.
-Rather than keeping files on a proprietary hard drive or local storage device, cloud-based storage makes it possible to save them to a remote database.
=As long as an electronic device has access to the web, it has access to the data and the software programs to run it.
-In cloud computing customers have to pay as per use.
-It is very flexible and can be resources can be scaled easily depending upon the requirement.
-Instead of buying any IT resources physically, all resources can be availed depending on the requirement from the cloud vendors.
Cloud computing has three service models;

- **Infrastructure as a Service (IaaS)**
- **Platform as a Service (PaaS)**
- **Software as a Service (SaaS)**

-**Features of Cloud computing**

1. **On-demand self-services**: The Cloud computing services does not require any human administrators, user themselves are able to provision, monitor and manage computing resources as needed.
2. **Security**: Cloud providers invest heavily in security measures to protect their users' data and ensure the privacy of sensitive information.
3. **Flexible pricing models**: Cloud providers offer a variety of pricing models, including pay-per-use, subscription-based, and spot pricing, allowing users to choose the option that best suits their need
4. **Huge Network Access**: A significant part of cloud services is that it's ubiquitous. Clients need any device with an internet connection to access cloud storage. In addition, the cloud service providers have ample access to the network that makes it easy for them to administer all the data uploaded on the Cloud through parameters like access time, latency, data output, and more.
5. **Economical**:Organization adopting Cloud Computing helps them to reduce the costs on IT expenses. In Cloud Computing, they have to just pay for the service which they have used. There are no extra charges which have to be paid.

–types of cloud computing: (simple ayittu ezhuthiyal mathii)

1. Public:- Public Cloud provides a shared platform that is accessible to the general public through an Internet connection.

-Public cloud operated on the pay-as-per-use model and administrated by the third party.

-In the Public cloud, the same storage is being used by multiple users at the same time.

-Public cloud is owned, managed, and operated by businesses, universities, government organizations, or a combination of them.

-Amazon Elastic Compute Cloud (EC2), Microsoft Azure, IBM's Blue Cloud, Sun Cloud, and Google Cloud are examples of the public cloud.

->Advantages of Public Cloud

-Public cloud has a lower cost than private, or hybrid cloud.

-Public cloud is location independent because its services are offered through the internet.

-In Public cloud, the cloud service provider is responsible for the manage and maintain data centers , so cloud user can save their time.

-easily buy public cloud on the internet.

-Public cloud offers scalable (easy to add and remove) and reliable (24*7 available) services to the users at an affordable cost.

->Disadvantages of Public Cloud

-Public Cloud is less secure because resources are shared publicly.

-In the public cloud, performance depends upon the speed of internet connectivity.

-Public cloud is less customizable than the private cloud.

2. Private:-Private cloud is also known as an internal cloud or corporate cloud.
-Private cloud provides computing services to a private internal network (within the organization) and selected users instead of the general public.
-Private cloud provides a high level of security.
-It also ensures that operational and sensitive data are not accessible to third-party providers.
-HP Data Centers, Microsoft, Elastra-private cloud, and Ubuntu are the example of a private cloud.

->Advantages of Private cloud

-Private clouds have more control over their resources and hardware than public clouds because it is only accessed by selected users.
-Security & privacy are one of the big advantages of cloud computing. Private cloud improved the security level as compared to the public cloud.
-Private cloud offers better performance with improved speed and space capacity.

->Disadvantages of Private Cloud

-The cost is higher than a public cloud because set up and maintain hardware resources are costly.
-Skilled people are required to manage and operate cloud services.
-Limited scalability

3.Hybrid/hetrogeneous:- Hybrid cloud is a combination of public and private clouds.

Hybrid cloud = public cloud + private cloud

-The main aim to combine these cloud (Public and Private) is to create a unified, automated, and well-managed computing environment.
-In the Hybrid cloud, non-critical activities are performed by the public cloud and critical activities are performed by the private cloud.
-Mainly, a hybrid cloud is used in finance, healthcare, and Universities.
-The best hybrid cloud provider companies are Amazon, Microsoft, Google, Cisco, and NetApp.

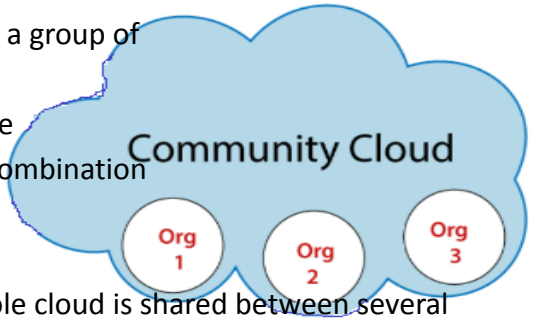
->Advantages of Hybrid Cloud

-It provides flexible resources because of the public cloud and secure resources because of the private cloud.
-Hybrid cloud costs less than the private cloud.
-It offers the features of both the public as well as the private cloud.
-Hybrid cloud is secure because critical activities are performed by the private cloud.

->Disadvantages of Hybrid Cloud

-In the Hybrid Cloud, networking becomes complex because of the private and the public Cloud.

4.Community Cloud :-Community cloud is a cloud infrastructure that allows systems and services to be accessible by a group of several organizations to share the information.



-It is owned, managed, and operated by one or more organizations in the community, a third party, or a combination of them.

->Advantages of Community Cloud

-Community cloud is cost effective because the whole cloud is shared between several organizations or a community.

- It allows the users to modify the documents as per their needs and requirement.

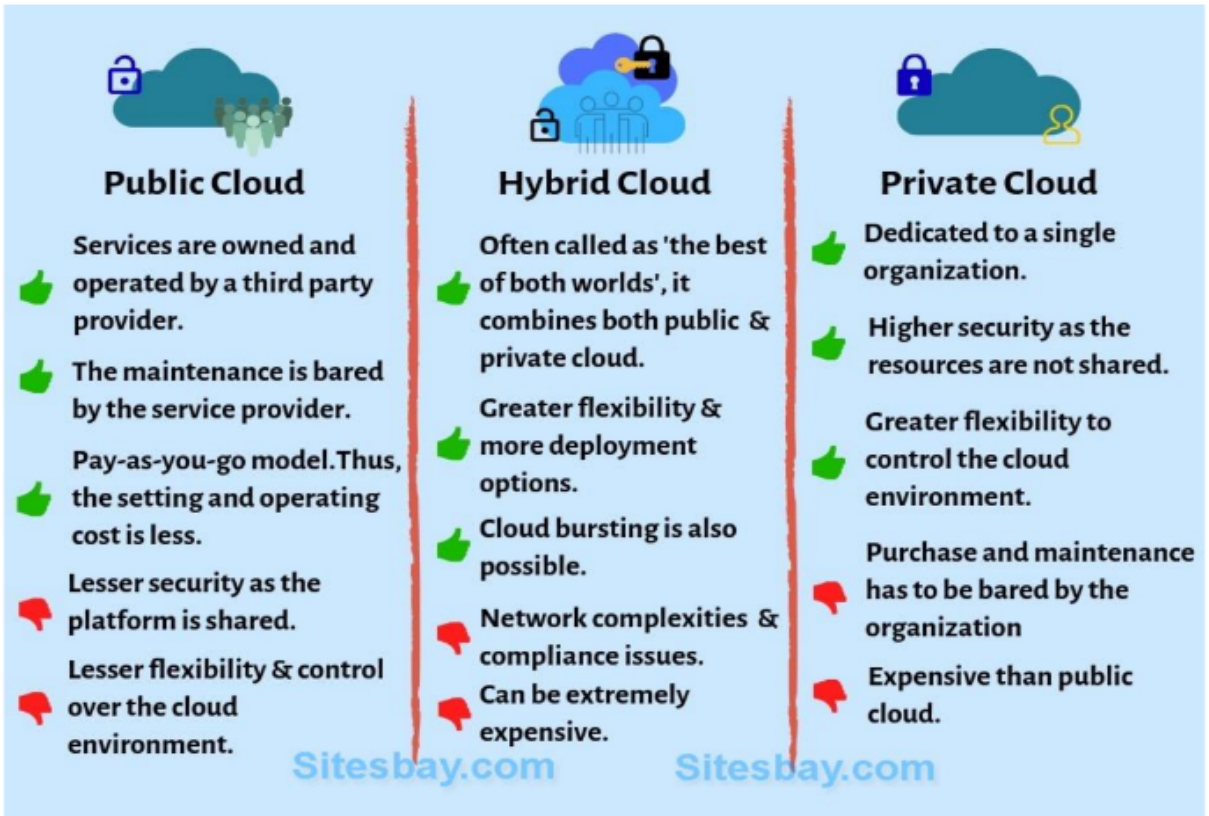
-Community cloud is more secure than the public cloud but less secure than the private cloud.

-Community cloud allows us to share cloud resources.

->Disadvantages of Community Cloud

-Community cloud is not a good choice for every organization.

-Community Cloud is costly than the public cloud.



***Distributed and Parallel computing** :-Computing:-The terms parallel computing and distributed computing are often used interchangeably, even though they mean slightly different things.

-The term parallel implies a tightly coupled system, whereas distributed refers to a wider class of system, including those that are tightly coupled.

->Parallel Computing:-In parallel computing multiple processors performs multiple tasks assigned to them simultaneously.

-Memory in parallel systems can either be shared or distributed.

-Parallel computing provides concurrency and saves time and money.

-There are multiple advantages to parallel computing. As there are multiple processors working simultaneously, it increases the CPU utilization and improves the performance.

- Moreover, failure in one processor does not affect the functionality of other processors.

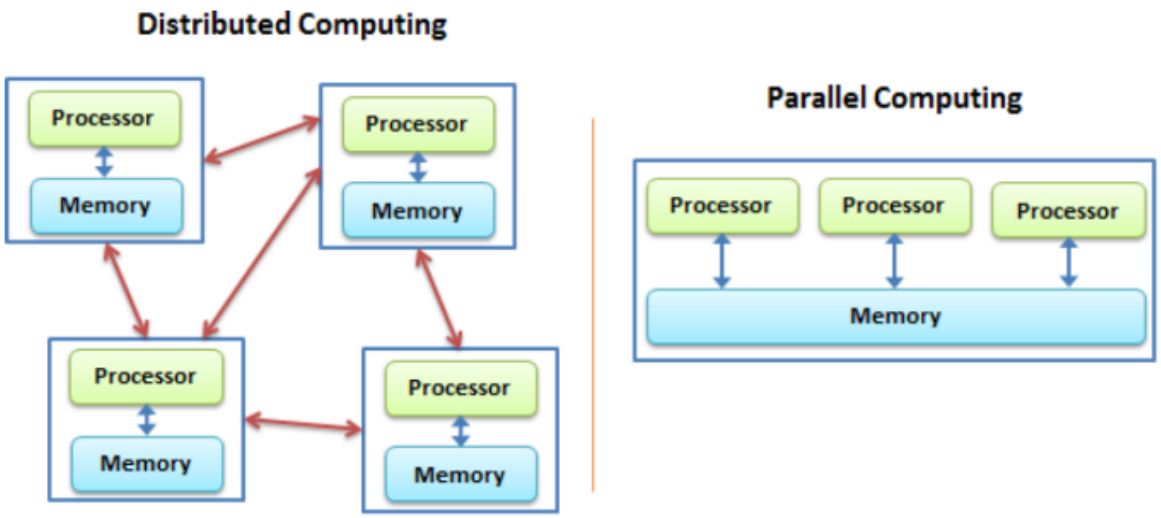
-Therefore, parallel computing provides reliability.

->Distributed Computing:-In distributed computing we have multiple autonomous computers which seems to the user as single system.

-In distributed systems there is no shared memory and computers communicate with each other through message passing.

-There are multiple advantages of using distributed computing. It allows scalability and makes it easier to share resources easily.

-It also helps to perform computation tasks efficiently.



*In-Memory Computing for Big data:- google

***Hadoop**:-Hadoop is an open source framework from Apache and is used to store, process and analyze data which are very huge in volume.

-Hadoop is written in Java and is not OLAP (online analytical processing).

It is used for offline processing.

-It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more.

-Moreover it can be scaled up just by adding nodes in the cluster.

→Features of hadoop:

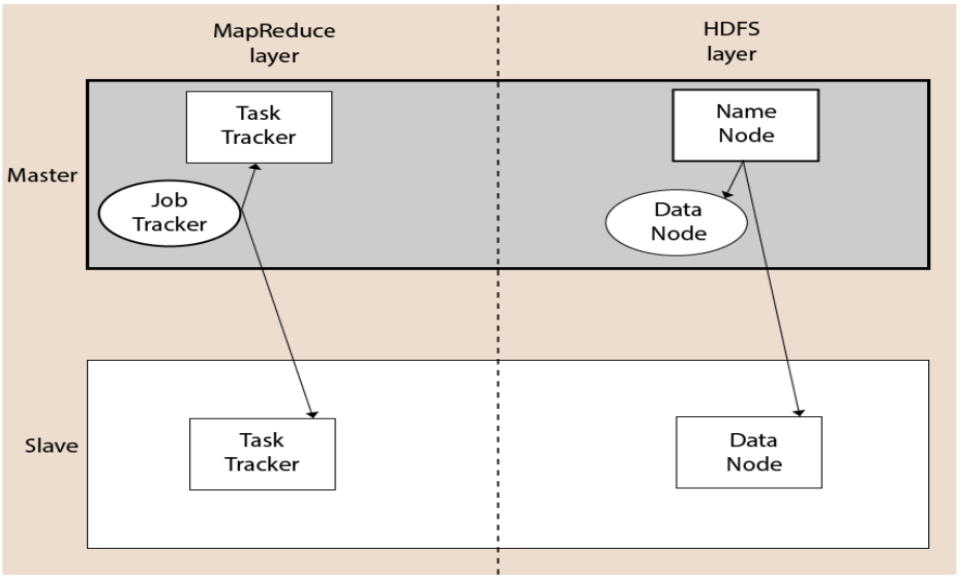
- it is fault tolerance.
- it is highly available.
- it's programming is easy.
- it have huge flexible storage.
- it is low cost.

→Hadoop Architecture:-

-The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System)

-A Hadoop cluster consists of a single master and multiple slave nodes.

-The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.



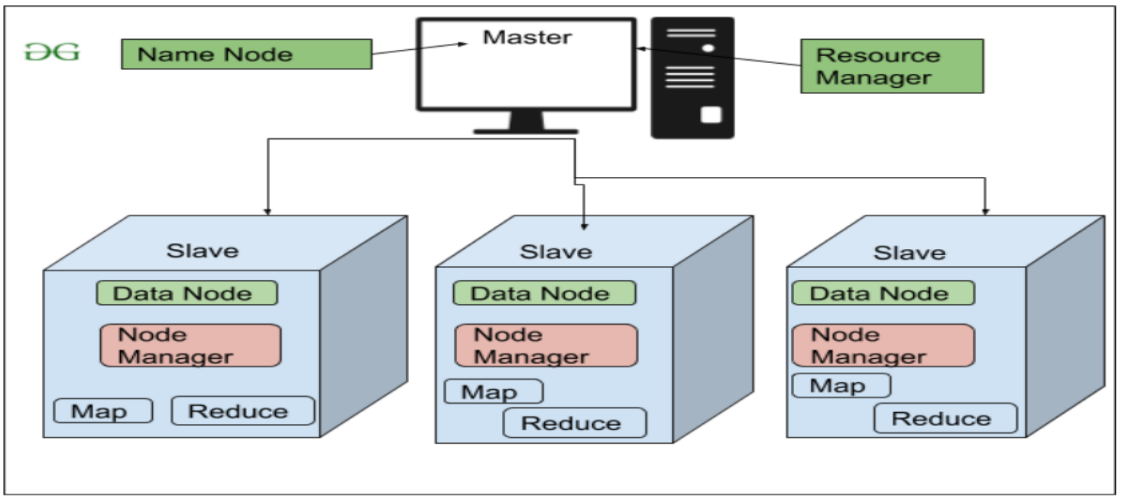
-There are three components of Hadoop:

1. **Hadoop HDFS** - Hadoop Distributed File System (HDFS) is the storage unit.
2. **Hadoop MapReduce** - Hadoop MapReduce is the processing unit.
3. **Hadoop YARN** - Yet Another Resource Negotiator (YARN) is a resource management unit.

***Hadoop Distributed File System**:-The Hadoop Distributed File System (HDFS) is a distributed file system for Hadoop.
-It contains a master/slave architecture.
-This architecture consist of a single NameNode performs the role of master, and multiple DataNodes performs the role of a slave.
-he Java language is used to develop HDFS.
-So any machine that supports Java language can easily run the NameNode and DataNode software.
-Data storage nodes in HDFS are;

- NameNode(Master)
- DataNode(Slave)

1. **NameNode(Master)**:-It is a single master server exist in the HDFS cluster.
   -As it is a single node, it may become the reason of single point failure.
   -It manages the file system namespace by executing an operation like the opening, renaming and closing the files.
   -It simplifies the architecture of the system.
2. **DataNode(Slave)**:-The HDFS cluster contains multiple DataNodes.
   -Each DataNode contains multiple data blocks.
   -These data blocks are used to store data.
   -It is the responsibility of DataNode to read and write requests from the file system's clients.
   -It performs block creation, deletion, and replication upon instruction from the NameNode.



***Hadoop MapReduce:**-It is a software framework and programming model for processing huge amount of data.

-It works into 2 phases mainly in map and reduce.

-Map task deal with splitting and mapping of data.

-reduce task shuffle and reduce the data .

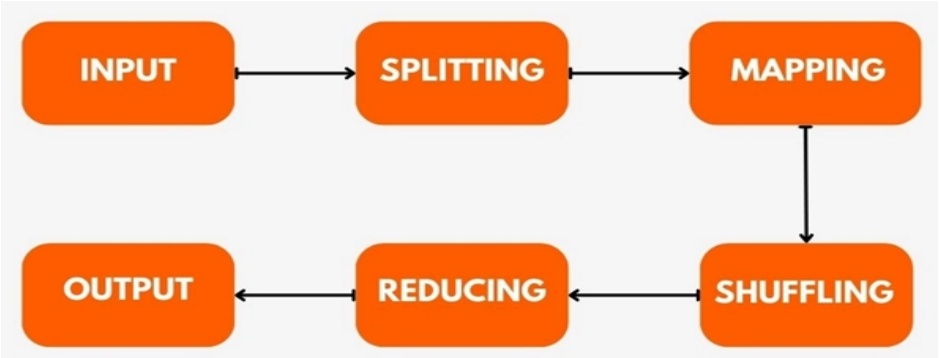
-Hadoop is capable of running map reduce program.

-The input of each phase is key value pair.

-In addition to it every programmers need to specify two function : **Map function** and **Reduce function.**

>**MapReduce Works**

-The entire process is divided into four steps of execution which are **splitting, mapping, shuffling and reducing.**



- **Input splitting**:-Map reduce in Big dta job is divide into fixed size pieces called
  -Input split is a chunk of input that is consumed buy a single map.
- **Mapping**:-It is the very first phase in the execution of map-reduce program.
  -In this phase each split is passed to a mapping function to produce output values.
- **Shuffling**:-It contain the output of mapping phase.
  -It duty is to gather the appropriate outcomes from the mapping step.
- **reducing:**-It have output of shuffling phase .
  -It combines values from shuffling p[hase and return a single output values.

>MapReduce Architecture:-The entire job is divided into tasks.

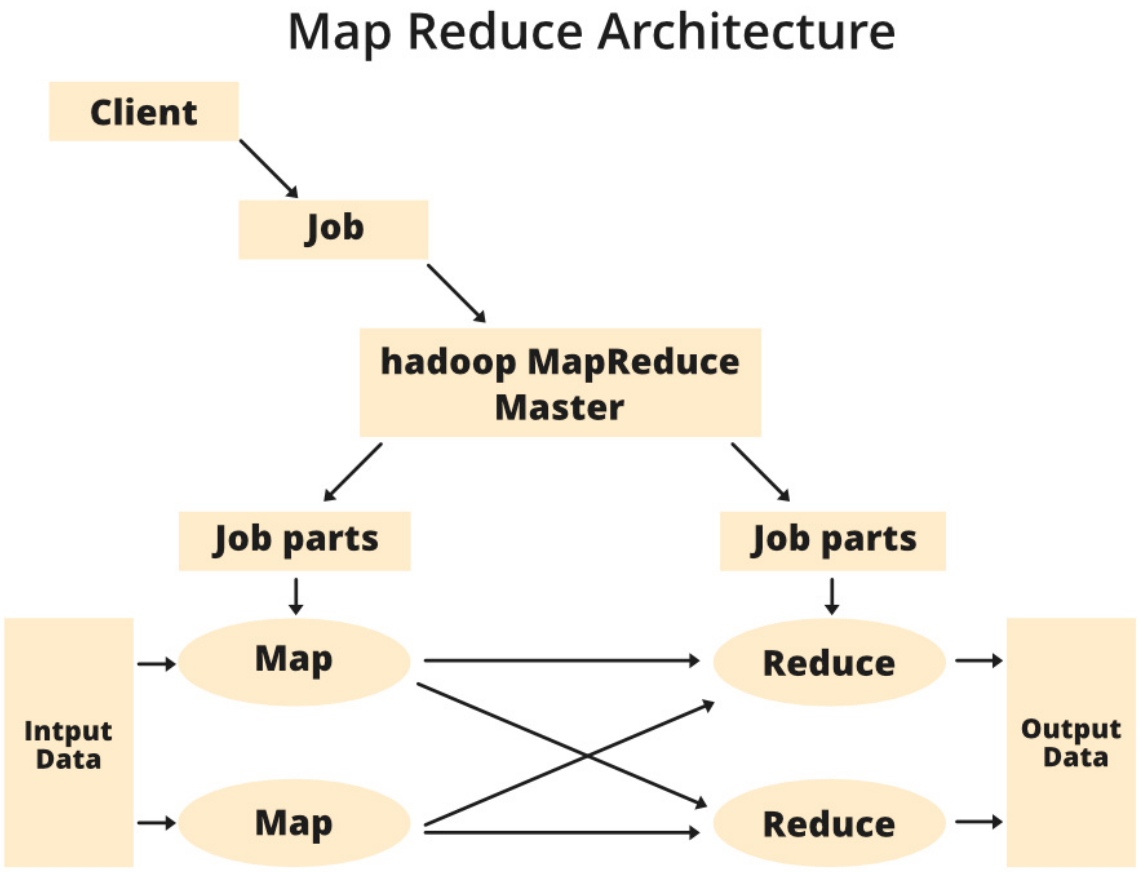
-There are two types of tasks namely, **Mapping tasks and Reducing tasks**.

-Mapping tasks splits the input data and performs mapping while reducing tasks performs shuffling and aggregates the shuffling values and returns a single output value, thereby reducing the data.

-The execution of these two tasks is controlled by two entities:

1. **Job Tracker** - It acts like a master and plays the role of scheduling jobs and tracking the jobs assigned to Task Tracker.
2. **Multiple Task Tracker** - It acts like slaves.
   -It tracks the jobs and reports the status of the jobs to the master (job tracker).

-In every execution there is only one job tracker and multiple tracker.



- **Client:** The MapReduce client is the one who brings the Job to the MapReduce for processing.
  -There can be multiple clients available that continuously send jobs for processing to the Hadoop MapReduce Manager.
- **Job**: The MapReduce Job is the actual work that the client wanted to do which is comprised of so many smaller tasks that the client wants to process or execute.
- **Hadoop MapReduce Master**: It divides the particular job into subsequent job-parts.
- **Job-Parts**: The task or sub-jobs that are obtained after dividing the main job.
  -The result of all the job-parts combined to produce the final output.
- **Input Data**: The data set that is fed to the MapReduce for processing.
- **Output Data**: The final result is obtained after the processing.

-In MapReduce, we have a client.
-The client will submit the job of a particular size to the Hadoop MapReduce Master.
-Now, the MapReduce master will divide this job into further equivalent job-parts.
-These job-parts are then made available for the Map and Reduce Task.
-The Map will generate intermediate key-value pair as its output.
-The output of Map i.e. these key-value pairs are then fed to the Reducer and the final output is stored on the HDFS.

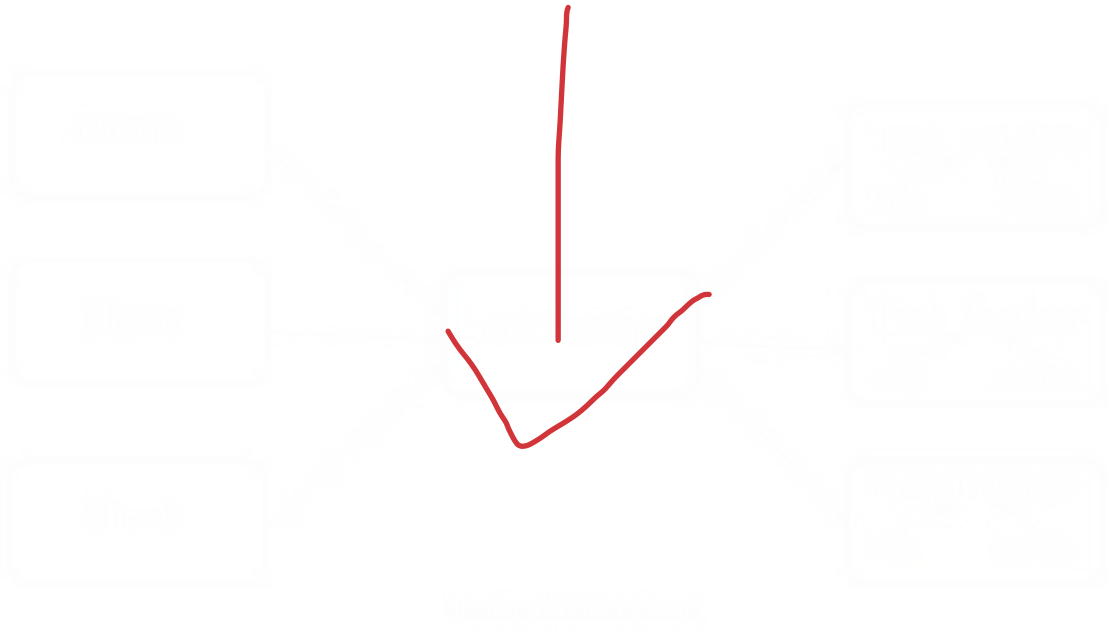
***Hadoop YARN**:-YARN stands for yet another resource negotiator.

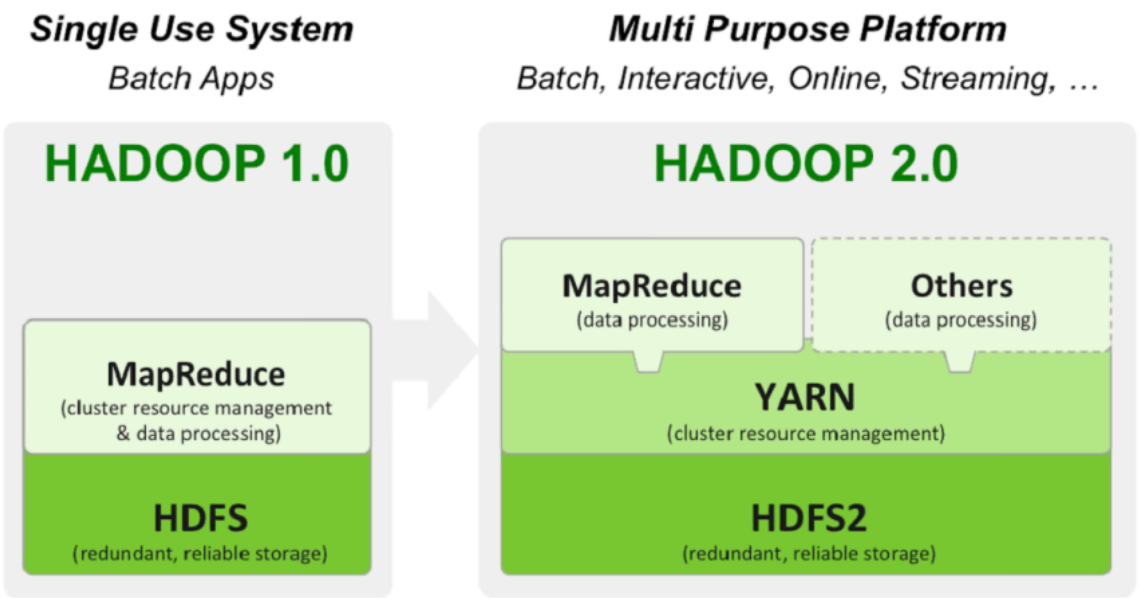
-It was introduced in hadoop 2.0 to remove the bottle neck on job tracker which is present in hadoop 1.0.

-It was launched for a **redesigned resource manager** but now it is **used in large scale distributed os used for big data.**

-YARN separates resource management layer rom the processing layer.

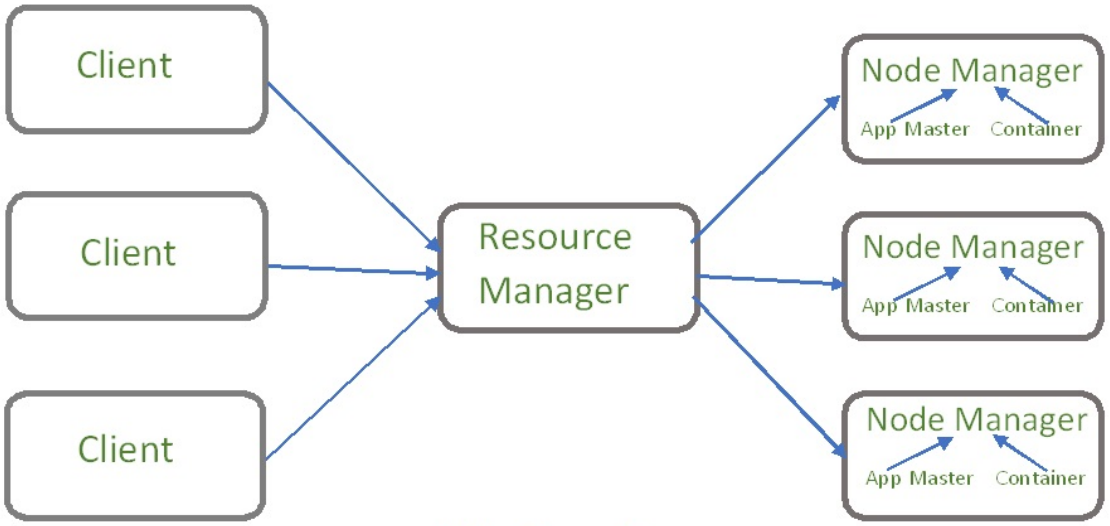
-In hadoop 1.0 is used by job tracker to split the resource manager and application manager.





→**Hadoop YARN architecture:**-YARN architecture basically separates resource management layer from the processing layer.
-In Hadoop 1.0 version, the responsibility of Job tracker is split between the resource manager and application manager.
-The main components of YARN architecture include:

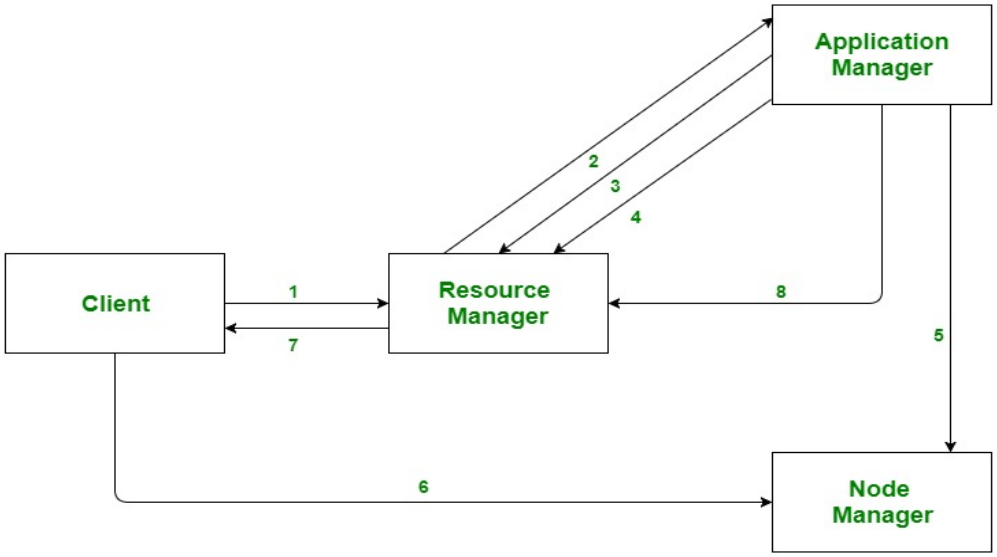
- **Client**: It submits map-reduce jobs.



Hadoop Yarn architecture

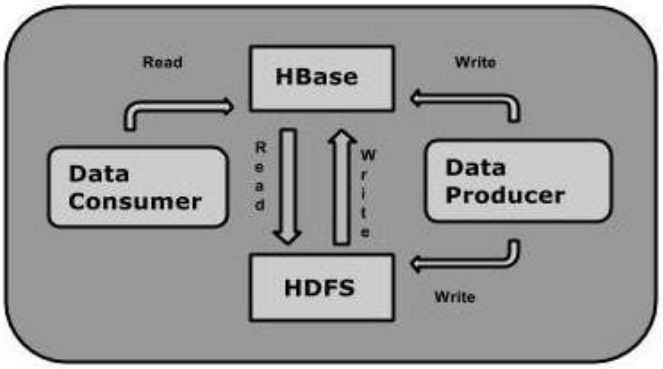
- **Resource Manager**: It is the master daemon of YARN and is responsible for resource assignment and management among all the applications.
  -Whenever it receives a processing request, it forwards it to the corresponding node manager and allocates the resources and fulfill the request .
- **Node Manager**:-It's primary job is to keep -up with the resource manager.

→**Application workflow in Hadoop YARN**:



1. Client submits an application
2. The Resource Manager allocates a container to start the Application Manager
3. The Application Manager registers itself with the Resource Manager
4. The Application Manager negotiates containers from the Resource Manager
5. The Application Manager notifies the Node Manager to launch containers
6. Application code is executed in the container
7. Client contacts Resource Manager/Application Manager to monitor application's status
8. Once the processing is complete, the Application Manager un-registers with the Resource Manager

***HBase:-**HBase is a distributed column-oriented database built on top of the Hadoop file system.
-It is an open-source project and is horizontally scalable.
-HBase is an essential part of our Hadoop ecosystem.
-HBase runs on top of HDFS (Hadoop Distributed File System).
-It can store massive amounts of data from terabytes to petabytes.
-HBase is a data model that is similar to Google's big table
-One can store the data in HDFS either directly or through HBase.
-Data consumer reads/accesses the data in HDFS randomly using HBase.
-HBase sits on top of the Hadoop File System and provides read and write access.



→**Storage Mechanism in HBase**:-HBase is a column-oriented database and the tables in it are stored by row.
-The table defines only column families, which are the key value pairs.
-A table have multiple column families and each column family can have any number of columns.
-in an HBase:

- Table is a collection of rows.
- Row is a collection of column families.
- Column family is a collection of columns.
- Column is a collection of key value pairs.

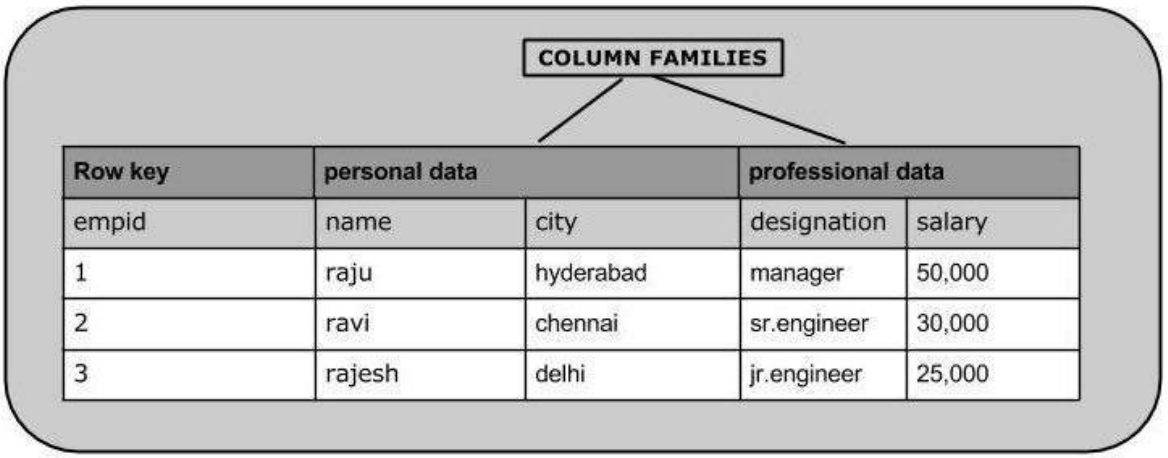
-Given below is an example schema of table in HBase.

Rowid	Column Family			Column Family			Column Family			Column Family		
	col1	col2	col3	col1	col2	col3	col1	col2	col3	col1	col2	col3
1												
2												
3												

→**Column Oriented and Row Oriented:**-Column-oriented databases that store data in the columns rather than as rows.

Row-Oriented Database	Column-Oriented Database
It is suitable for Online Transaction Process (OLTP).	It is suitable for Online Analytical Processing (OLAP).
Such databases are designed for small number of rows and columns.	Column-oriented databases are designed for huge tables.

-The following image shows column families in a column-oriented database:



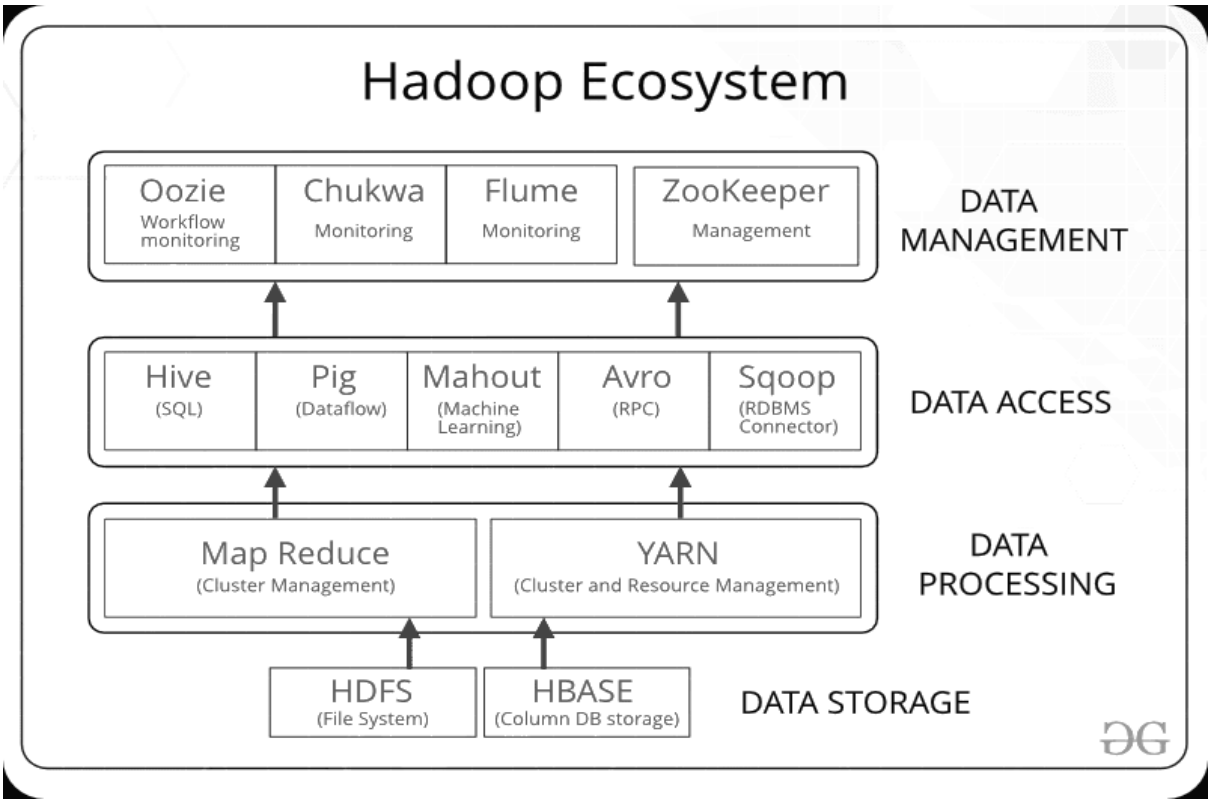
→**Features of HBase**

- HBase is linearly scalable.
- It has automatic failure support.
- It provides consistent read and writes.
- It integrates with Hadoop, both as a source and a destination.
- It has easy java API for client.
- It provides data replication across clusters.

→**Applications of HBase**

- It is used whenever there is a need to write heavy applications.
- HBase is used whenever we need to provide fast random access to available data.
- Companies such as Facebook, Twitter, Yahoo, and Adobe use HBase internally.

***Hadoop Ecosystem (**3 marks**)**:-Hadoop Ecosystem is a platform or a suite which provides various services to solve the big data problems.
-here are four major elements of Hadoop i.e. **HDFS, MapReduce, YARN, and Hadoop Common**.

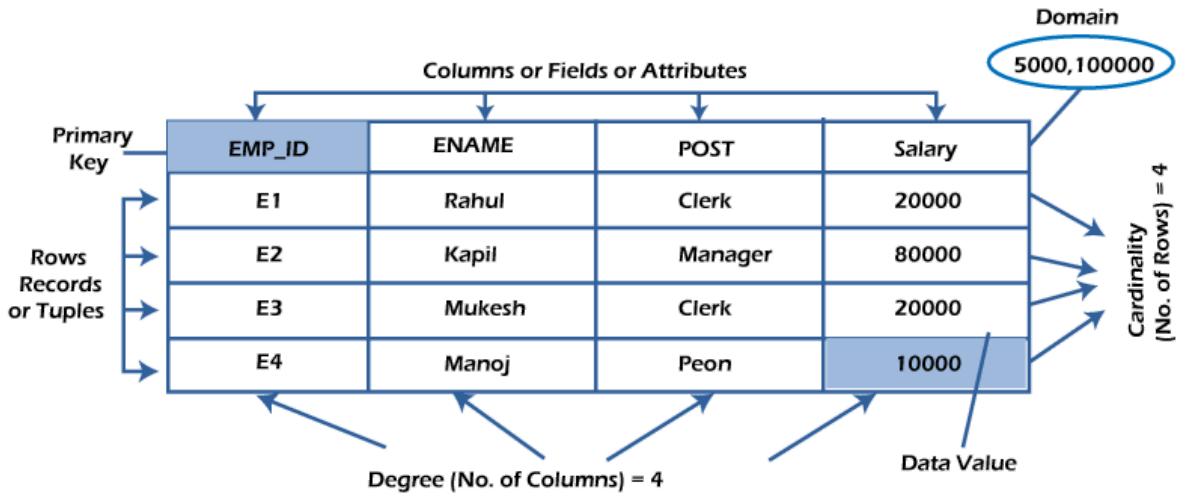


-Following are the components that collectively form a Hadoop ecosystem:

- **HDFS:** Hadoop Distributed File System
- **YARN:** Yet Another Resource Negotiator
- **MapReduce:** Programming based Data Processing
- **Spark:** In-Memory data processing
- **PIG, HIVE:** Query based processing of data services
- **HBase:** NoSQL Database
- **Mahout, Spark MLLib**: Machine Learning algorithm libraries
- **Solar, Lucene**: Searching and Indexing
- **Zookeeper**: Managing cluster
- **Oozie**: Job Scheduling

Module -5

#RDBMS:-RDBMS stands for Relational Database Management System.
-All modern database management systems like SQL, MS SQL Server, IBM DB2, ORACLE, My-SQL, and Microsoft Access are based on RDBMS.
-It is called Relational Database Management System (RDBMS) because it is based on the relational model **introduced by E.F. Codd**.
-In RDBMS Data is represented in terms of tuples (rows) in RDBMS.
-A relational database is the most commonly used database.
-It contains several tables, and each table has its primary key.
-Due to a collection of an organized set of tables, data can be accessed easily in RDBMS.
-Following are the various terminologies of RDBMS:



- **Table/Relation**:-Everything in a relational database is stored in the form of relations.
  -The RDBMS database uses tables to store data.
  -A table is a collection of related data entries and contains rows and columns to store data.
  -Each table represents some real-world objects such as person, place, or event about which information is collected.
- **Row or Record**:-A row of a table is also called a record or tuple.
  -It contains the specific information of each entry in the table.
  -It is a horizontal entity in the table.
  -For example, The above table contains 4 records.
- **Column/Attribute**:-A column is a vertical entity in the table which contains all information associated with a specific field in a table.
  -For example, "ENAME" is a column in the above table which contains all information about an employee's name.
- **Degree:**-The total number of attributes that comprise a relation is known as the degree of the table.
  -It's degree is 4.

- **Cardinality**:-The total number of tuples at any one time in a relation is known as the table's cardinality.
  -Here the cardinality is 4
- **NULL Values**:-The NULL value of the table specifies that the field has been left blank during record creation.
  -It is different from the value filled with zero or a field that contains space.

→**Advantage of RDBMS**

- Flexibility
- Ease of use
- Collaboration.
- Build - in security.
- Better data integrity.
- Multi user access.

→**DisAdvantage of RDBMS**

- Expensive
- Difficult to recover lost data
- Complex software.

→**Features of RDBMS:**

- Offers information to be saved in the tables.
- Numerous users can access it together which is managed by a single user.
- Virtual tables are available for storing the insightful data.
- In order to exclusively find out the rows, the primary key is used.
- The data are always saved in rows and columns.
- To retrieve the information the indexes are used.
- Columns are being shared between tables using the keys.

***Polyglot Persistence**:-

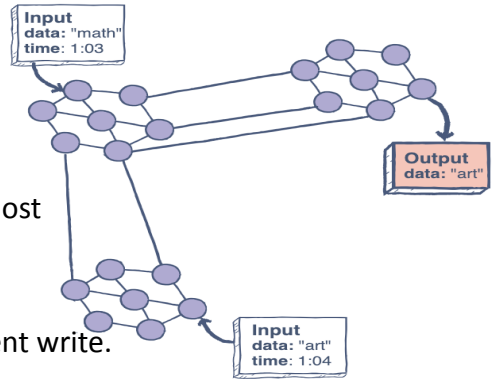
-Youtube

***CAP theorem** :-The three letters in CAP refer to three desirable properties of distributed systems with replicated data:

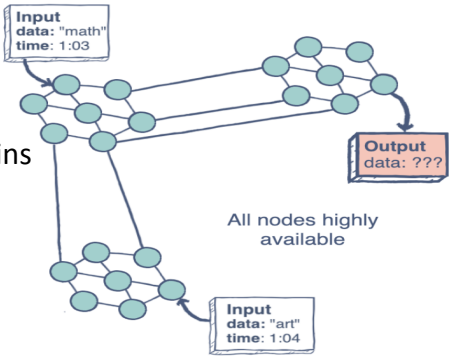
- **consistency** (among replicated copies),
- **availability** (of the system for read and write operations)
- **partition tolerance** (in the face of the nodes in the system being partitioned by a network fault).

-The CAP theorem states that it is not possible to guarantee all three of the desirable properties – consistency, availability, and partition tolerance at the same time in a distributed system with data replication.
-The theorem states that networked shared-data systems can only strongly support two of the following three properties:

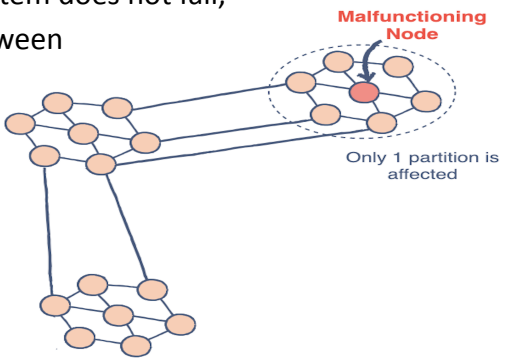
1. **Consistency**:-A system is said to be consistent if all nodes see the same data at thesame time.
   -Simply, if we perform a read operation on a consistent system, it should return the value of the most recent write operation.
   -This means that, the read should cause all nodes to return the same data, i.e., the value of the most recent write.



2. **Availability**:-Availability in a distributed system ensures that the system remains operational 100% of the time.
   -Note: this does not guarantee that the response contains the most recent write.



3. **Partition Tolerance**:-This condition states that the system does not fail, regardless of if messages are dropped or delayed between nodes in a system.
   -Partition tolerance has become more of a necessity than an option in distributed systems.
   -It is made possible by sufficiently replicating records across combinations of nodes and networks.



*** Non-relational database**:-It is also known as No sql.
-It is designed to handle and store large volumes of unstructured and semi-structured data.
-Unlike traditional relational databases that use tables with pre-defined schemas to store data.
-NoSQL databases are generally classified into four main categories:

- **Document databases**: These databases store data as semi-structured documents, such as JSON or XML, and can be queried using document-oriented query languages.
- **Key-value stores**: These databases store data as key-value pairs, and are optimized for simple and fast read/write operations.

- **Column-family stores**: These databases store data as column families, which are sets of columns that are treated as a single entity. They are optimized for fast and efficient querying of large amounts of data.
- **Graph databases**: These databases store data as nodes and edges, and are designed to handle complex relationships between data.

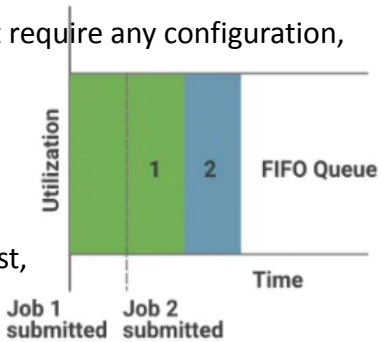
-Features of NoSQL :-

1. **Dynamic schema**: NoSQL databases do not have a fixed schema and can accommodate changing data structures without the need for migrations or schema alterations.
2. **Horizontal scalability**: NoSQL databases are designed to scale out by adding more nodes to a database cluster, making them well-suited for handling large amounts of data and high levels of traffic.
3. **Document-based**: Some NoSQL databases, such as MongoDB, use a document-based data model, where data is stored in semi-structured format, such as JSON or BSON.
4. **Key-value-based:** Other NoSQL databases, such as Redis, use a key-value data model, where data is stored as a collection of key-value pairs.
5. **Column-based**: Some NoSQL databases, such as Cassandra, use a column-based data model, where data is organized into columns instead of rows.
6. **Distributed and high availability**: NoSQL databases are often designed to be highly available and to automatically handle node failures and data replication across multiple nodes in a database cluster.
7. **Flexibility**: NoSQL databases allow developers to store and retrieve data in a flexible and dynamic manner, with support for multiple data types and changing data structures.
8. **Performance**: NoSQL databases are optimized for high performance and can handle a high volume of reads and writes, making them suitable for big data and real-time applications.

***YARN schedulers:**-A YARN scheduler determines how resources are allocated on a Hadoop cluster.

-There are three scheduling options available in YARN.

1. **FIFO Scheduler**:-In FIFO (FIRST IN FIRST OUT) scheduler, applications are placed in a queue and runs them in the order of submission.
   -This scheduling option is simple understand and doesn't require any configuration, but it is not suitable for shared clusters.
   -Eg:In the given figure , Job 1 is submitted first and it is a long running job with less priority, after sometime Job 2 which is a high priority job is submitted and takes less time than Job 1. But, since the Job 1 is submitted first, all the resources will be allocated to it and Job 2 has to



wait until Job 1 is done.

>Limitations:

- Jobs are executed based on FIFO principle and ignores the priority value.
- It is not suitable for shared clusters because large applications will use all the resources and other applications has to wait for it turn.

-On shared clusters it is better to use the Capacity Scheduler or the Fair Scheduler.

2. **Capacity Scheduler**:-The Capacity scheduler allows sharing of a Hadoop cluster within the organization.

   -hereby each team is allocated a certain capacity of the overall cluster.

   -Queues may be further divided in hierarchical fashion.

   -In Capacity scheduler, a separate dedicated queue allows the small jobs to start as soon as it is submitted.

   -Eg:In the above example there are two queues A and B, When Job 1 is submitted in Queue A the queue is empty and will take all the resources in the queue. When Job 2 is submitted in Queue B then it starts its execution instead of waiting for Job 1 to finish. Thus capacity scheduler ensures that the priority and small jobs are not starved for longer time when compared to that of FIFO scheduler.

   >Advantage:

   - Best for working with Multiple clients or priority jobs in a Hadoop cluster

   >Disadvantage:

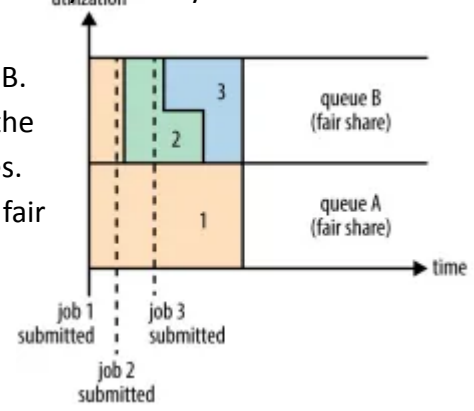
   - More complex

3. **Fair Scheduler**:-The Fair Scheduler is very much similar to that of the capacity scheduler.

   -The priority of the job is kept in consideration.

   -With the help of Fair Scheduler, the YARN applications can share the resources in the large Hadoop Cluster and these resources are maintained dynamically so there is no need for prior capacity.

   -Eg:In the above example there are two queues A and B. Job 1 is submitted to Queue A and it is observed that the cluster is empty so Job 1 utilize all the cluster resources. After sometime Job is submitted in Queue B, then the fair share preemption occurs and both the jobs 1 and 2 allocated equal resources in their respective queues. In meanwhile Job 3 is submitted in Queue B and since



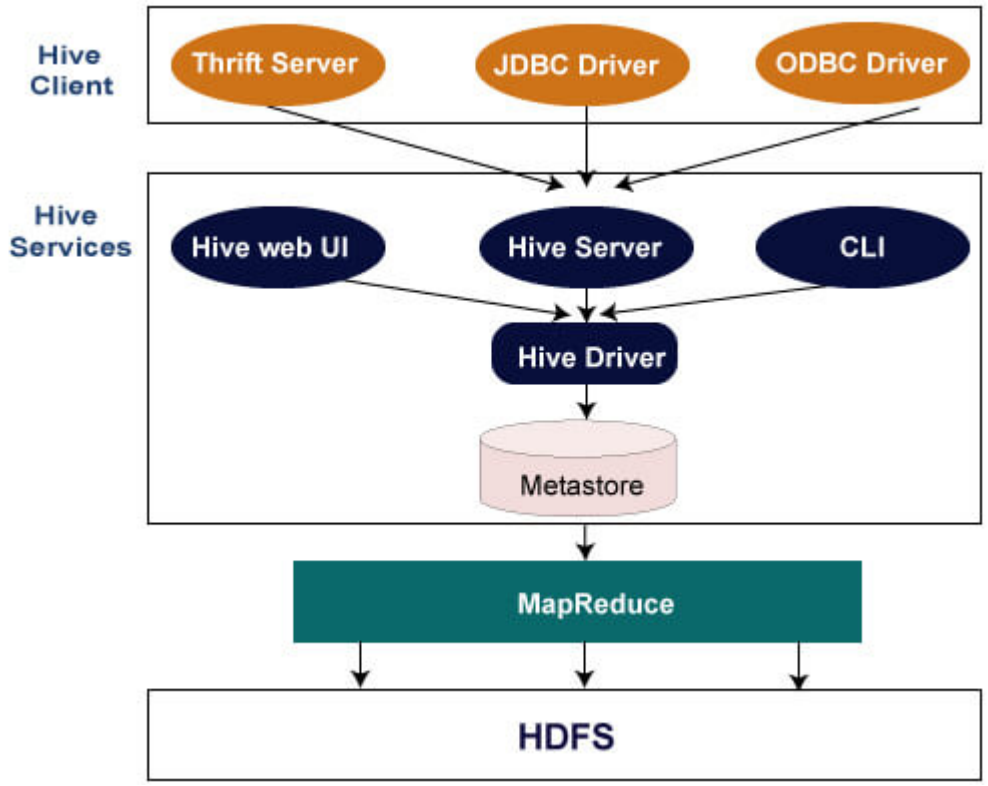
one job is already running the scheduler will assigned fair share to both thee Jobs in Queue B with equal resources. This way fair scheduler ensures that all the jobs are provided with required resources.

***Hive:**-Hive is a data warehouse and an ETL tool which provides an SQL-like interface between the user and the Hadoop distributed file system (HDFS).
-It is built on top of Hadoop.
-It facilitates reading, writing and handling wide datasets that stored in distributed storage and queried by Structure Query Language (SQL) syntax.
-It is not built for Online Transactional Processing (OLTP) workloads.
-Hive supports Data Definition Language (DDL), Data Manipulation Language (DML), and User Defined Functions (UDF).

→**Features of Hive:-**These are the following features of Hive:

- Hive is fast and scalable.
- It provides SQL-like queries (i.e., HQL) that are implicitly transformed to MapReduce or Spark jobs.
- It is capable of analyzing large datasets stored in HDFS.
- It allows different storage types such as plain text, RCFile, and HBase.
- It uses indexing to accelerate queries.
- It can operate on compressed data stored in the Hadoop ecosystem.
- It supports user-defined functions (UDFs) where user can provide its functionality.

→**Architecture of Hive:-**The following component diagram depicts the architecture of Hive:



>Hive Client:-Hive allows writing applications in various languages, including Java, Python, and C++. It supports different types of clients such as:-

- Thrift Server - It is a cross-language service provider platform that serves the request from all those programming languages that supports Thrift.
- JDBC Driver - It is used to establish a connection between hive and Java applications. The JDBC Driver is present in the class org.apache.hadoop.hive.jdbc.HiveDriver.
- ODBC Driver - It allows the applications that support the ODBC protocol to connect to Hive.

>Hive Services:-The following are the services provided by Hive:-

- Hive CLI - The Hive CLI (Command Line Interface) is a shell where we can execute Hive queries and commands.
- Hive Web User Interface - The Hive Web UI is just an alternative of Hive CLI. It provides a web-based GUI for executing Hive queries and commands.
- Hive MetaStore - It is a central repository that stores all the structure information of various tables and partitions in the warehouse. It also includes metadata of column and its type information, the serializers and deserializers which is used to read and write data and the corresponding HDFS files where the data is stored.
- Hive Server - It is referred to as Apache Thrift Server. It accepts the request from different clients and provides it to Hive Driver.
- Hive Driver - It receives queries from different sources like web UI, CLI, Thrift, and JDBC/ODBC driver. It transfers the queries to the compiler.
- Hive Compiler - The purpose of the compiler is to parse the query and perform semantic analysis on the different query blocks and expressions. It converts HiveQL statements into MapReduce jobs.
- Hive Execution Engine - Optimizer generates the logical plan in the form of DAG of map-reduce tasks and HDFS tasks. In the end, the execution engine executes the incoming tasks in the order of their dependencies.

→**built-in functions of Hive:**-These are functions that are already available in Hive.
-First, we have to check the application requirement, and then we can use these built-in functions in our applications.
-We can call these functions directly in our application.
-This are the some function

1. **Date Functions**:-It is used for performing Date Manipulations and Converting Date types from one type to another type.

Name	Return type	Description
year(string date)	int	It will return the year part of a date or a timestamp string. Example: year("2020-05-11 00:00:00") = 2000

month(string date)	int	It will return the month part of a date or a timestamp string. Example: month("2020-05-11 00:00:00")=05
Current_date	date	It will give the current date.
hour	int	
minute	int	
second	int	

2. **Mathematical Functions**:-These functions are used for Mathematical Operations.
   -This are the inbuilt mathematical functions in Hive.

Name	Return type	Description
floor(double a)	BIGINT	It returns the maximum BIGINT value that is equal or less than the double.
round(double a)	BIGINT	It returns the rounded BIGINT value of the double.

3. **String Functions**:-String manipulations and string operations these functions can be called.

Name	Return type	Description
reverse(string X)	string	It will give the reversed string of X
length(string A)	int	It will return the length of the string passed.

→aggregate functions of Hive:-

COUNT()	Returns the count of all rows in a table including rows containing NULL values When you specify a column as an input, it ignores NULL values in the column for the count. Also ignores duplicates by using DISTINCT. **Return:** BIGINT
SUM()	Returns the sum of all values in a column. When used with a group it returns the sum for each group. Also ignores duplicates by using DISTINCT. **Return:** DOUBLE
AVG()	Returns the average of all values in a column. When used with a group it returns an average for each group. **Return:** DOUBLE
MIN()	Returns the minimum value of the column from all rows. When used with a group it returns a minimum for each group. **Return:** DOUBLE
MAX()	Returns the maximum value of the column from all rows. When used with a group it returns a maximum for each group. **Return:** DOUBLE

Collect_set(col)	Returns a collection of elements in a group as a set by eliminating duplicate elements. **Return:** Array
Collect_list(col)	Returns a collection of elements in a group as a list including duplicate elements. **Return:** Array

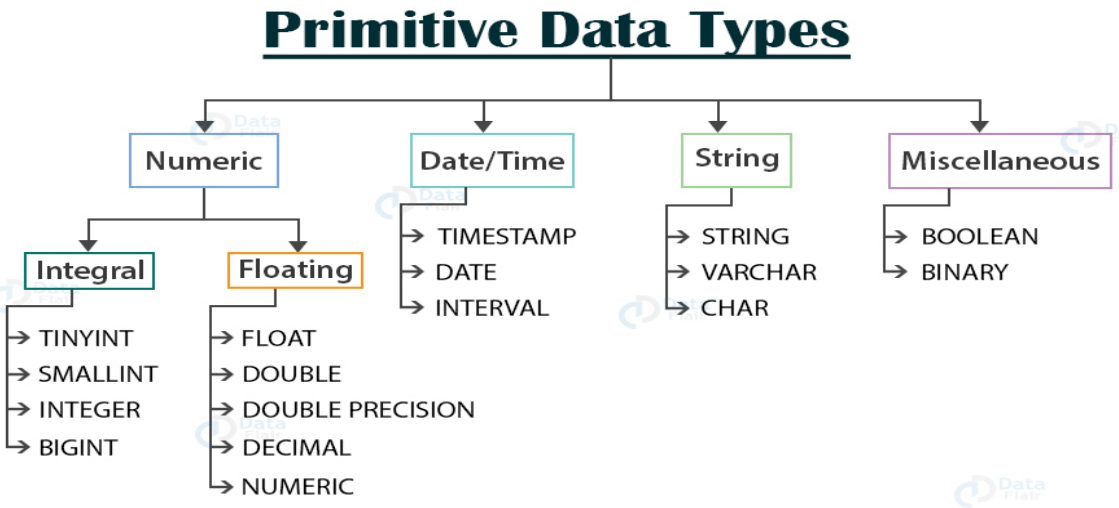
→**Limitations of Hive**

- Hive is not capable of handling real-time data.
- It is not designed for online transaction processing.
- Hive queries contain high latency.

→ **Data types in Hive**:-There are two categories of Hive Data types that are **primitive data type** and **complex data type**

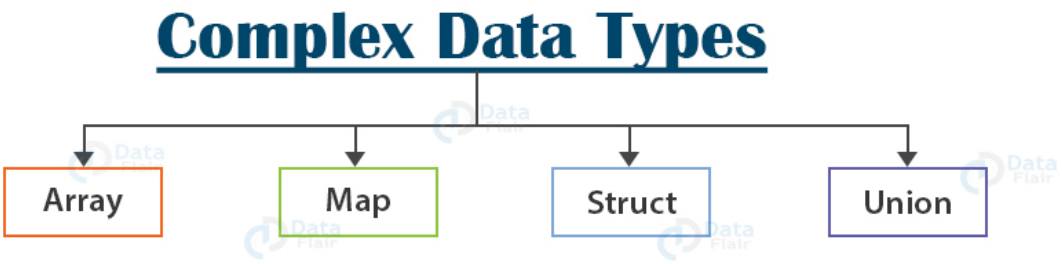
-Data Types in Hive specifies the column/field type in the Hive table.
-It specifies the type of values that can be inserted into the specified column.

1. Primitive Type



- Integral Types:-Integer type data can be specified using integral data types, INT. When the data range exceeds the range of INT, you need to use BIGINT and if the data range is smaller than the INT, you use SMALLINT. TINYINT is smaller than SMALLINT.
- String Types:-String type data types can be specified using single quotes (' ') or double quotes (" "). It contains two data types: VARCHAR and CHAR.
- Date/Time Types:-The Date value is used to specify a particular year, month and day, in the form YYYY--MM--DD. However, it didn't provide the time of the day. The range of Date type lies between 0000--01--01 to 9999--12--31.

2. Complex Type



Struct	It is similar to C struct or an object where fields are accessed using the "dot" notation.	struct('James','Roy')
Map	It contains the key-value tuples where the fields are accessed using array notation.	map('first','James','last','Roy')
Array	It is a collection of similar type of values that indexable using zero-based integers.	array('James','Roy')

*Variables, properties and queries in Hive:- google

3 mark questions

→**issues with relational model and non-relational model**

Relational Database	NoSQL
It is used to handle data coming in low velocity.	It is used to handle data coming in high velocity.
It gives only read scalability.	It gives both read and write scalability.
It manages structured data.	It manages all type of data.
Data arrives from one or few locations.	Data arrives from many locations.
It supports complex transactions.	It supports simple transactions.
It has single point of failure.	No single point of failure.
It handles data in less volume.	It handles data in high volume.
Transactions written in one location.	Transactions written in many locations.
support ACID properties compliance	doesn't support ACID properties
Its difficult to make changes in database once it is defined	Enables easy and frequent changes to database
schema is mandatory to store the data	schema design is not required
Deployed in vertical fashion.	Deployed in Horizontal fashion.

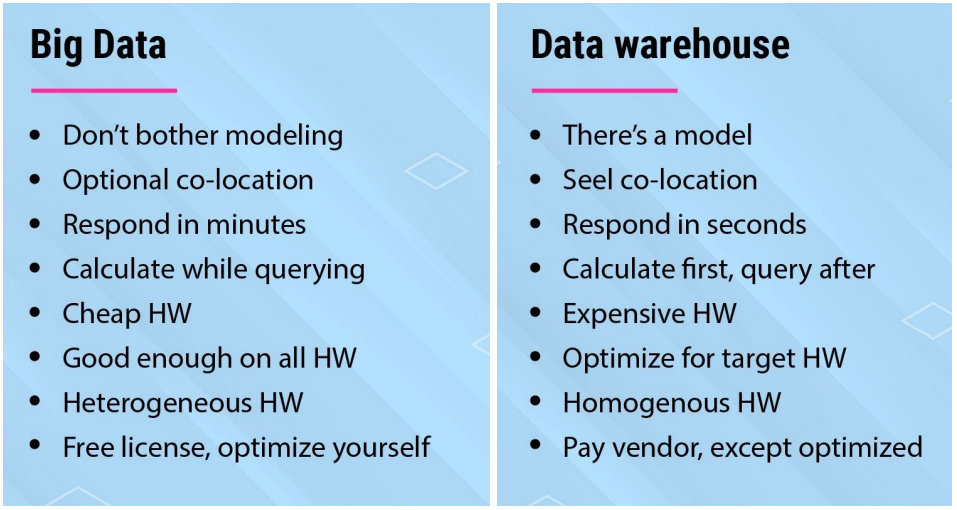
→**relationship between big data and a data warehouse**

Big Data

- Big data is a technology to store and manage large amount of data.
- It takes structured, non-structured or semi-structured data as an input.
- Big data doesn't follow any SQL queries to fetch data from database.
- When new data is added to big data, the changes are stored in files which are typically represented by tables.

Data Warehouse

- Data warehouse is an architecture used to organize the data.
- It only takes structured data as an input.
- In data warehouse we use SQL queries to fetch data from relational databases.
- In a data warehouse, new data does not impact the data warehouse directly, making it difficult to gain real-time insights from new data.



→**What are the functions of mapper and reducer**

-The mapper processes the data and creates several small chunks of data. Reduce stage – This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

→ **containers and node managers**

-The container refers to a collection of resources such as memory, CPU, disk and network IO. The number of containers on a node is the product of configuration parameter and the total amount of node resources. Node manager is the slave daemon of Yarn.

Join for more MCA short note : https://t.me/mgu_mca_shortnote

@MGU_MCA_SHORTNOTE