UNIT - 1

DATA WAREHOUSE

Data :- In computing, Data is information that has been translated into a form that is efficient for movement or processing. It's a collection of Information. Data is types Structured data & unstructured data.

In real time, 20% Structured data is stored in Database's (in the form of rows & columns)

Un-structured data means information not in the form of rows and tables. Ex: Newspaper,

Data Warehouse :- The term 'datawarehouse' was first coined by "Bill Inmon in 1990". According to Inmon, a data warehouse is a Subject oriented (state/fields), Integrated (Join/work together/combining), Time-varient (moment of observation) and Non-volatile (recollecting/retaining data) collection of data.

In support of management's decision making process.

*Datawarehouse is a addition to the DataBase, By using DB we can't generate reports and analyze the data. Users only can store and retrive data from database.

→ It's a collection of Datamarts representing historical data from different operation in the company.

→ It can also define in many ways.

* A decision Support Database that is maintained separately from organization operational Database. Supporting information processing by providing a solid platform of consolidated historical data for analysis.

Information :- Collection of data that has been processed, organized, structured or presented in a given context to make it useful, meaningfull, and understandable.

Data :- Collection of raw Data, uninterpri and isolated facts, science and symbol before they are given any context meaning.

Raw Data :- Milk

Isolated :- Milk, Bread, Jam .....
Milk, cookies, Rice.

DataBase :-- An organised collection of structured data stored an access digitally from a system it may be a relational database

Data wareHouse :- The word data warehouse coind by Bill Inmon in 1990 according to the bill Inmon data warehouse called as subject-oriented, integrated, time variant and Non-volatile

A centralised computing repository design specifically for reporting, Data analysis and Business intelligency

Instead of managing day-by-day operational

Transactions., it aggrigate large volume of current data and historical data from multiple sources into single unified source of truth to help the organization making decision making and Informati

Differences between DW and DBMS

Database Management System	Data Warehouse.
1. A common Database is based on operational or transactional processing. Each operation is an indivisble transaction.	1. A Data warehouse is based on analytical processing.
2. Generally, a Database stores current and up-to-date data which is used for daily operations.	2. A DW maintains historical data overtime. Historical data is the data kept over years and can be used for trend analysis, make future predictions and decision support.
3. A database is generally application specific. Ex: A database stores related data, such as the student details in a school.	3. A DW is integrated generally at the organization level, by combining data from different db. Ex: A DW integrates the data from one or more databases, so that analysis can be done to get results, such as the best performing school in a city.
4. Constructing a Database is not so expensive.	4. Constructing a Data Warehouse can be expensive.

Applications of Data Warehousing:- DW can be applied anywhere where we have a huge amount of data and we want to see statistical results that helps in decision making.

Ex: 1) Social Media Websites
2) Banking
3) Governments
4) E-Commerce
5) Telecommunications
6) Transportation Services
7) Marketing and Distribution.
8) Healthcare and Retail.

Difference between DataWarehouse(OLAP) and Operational Database (OLTP)

Data Warehouse	Operational Database
1. It involves historical processing of information	1. It involves day-to-day processing.
2. OLAP systems are used by knowledge workers such as executives, managers and Analyst	2. OLTP systems are used by clerks, DBA's or database professionals.
3. It is used to analyze the business	3. It is used to run the business
4. It focuses on information out	4. It focuses on Data In.
5. It is based on star schema, Snowflake schema, and fact Constellation schema.	5. It is based on Entity Relationship Model.
6. It focuses on information Out	6. It is application oriented
7. It contains historical data	7. It contains current data.
8. It provides summarized and consolidated data.	8. It provides details and flat relational view of data.
9. It provides summarized and multidimensional view of data.	9. It provides primitives and highly detailed data.
10. The no: of users is in 100's	10. The no: of users is in 1000's
11. The no: of records accessed is in millions	11. The no: of records accessed is in thousands
12. The database size is from 100 GB to 100 TB	12. The database size is from 100 MB to 100 GB.
13. These are highly flexible Online Analytical processing	13. It provides high performance Online Transaction processing.

Key	Datawarehouse	Opirtional Database.
Basic	A datawarehouse is a repository for structured, filtered data that has already been processed for a specific purpose.	Operational Database are those databases where data changes frequently.
Data Structure	Data warehouse has de-normalized schema.	It has normalized schema.
Performance	It is fast for analysis Queries.	It is slow for analytics queries.
Types of Data	It focuses on historical data.	It focuses on current transactional data.
Usecase	It is used for OLAP	It is used OLTP.

Functions of Datawarehouse :-

It works as a collection of data and here is organized by various communication/communities that endures the features to recover the data functions.

It has stocked facts about the tables which has high transaction levels which are observed so as to define the datawarehousing techniques and major functions which are involved in-this are:

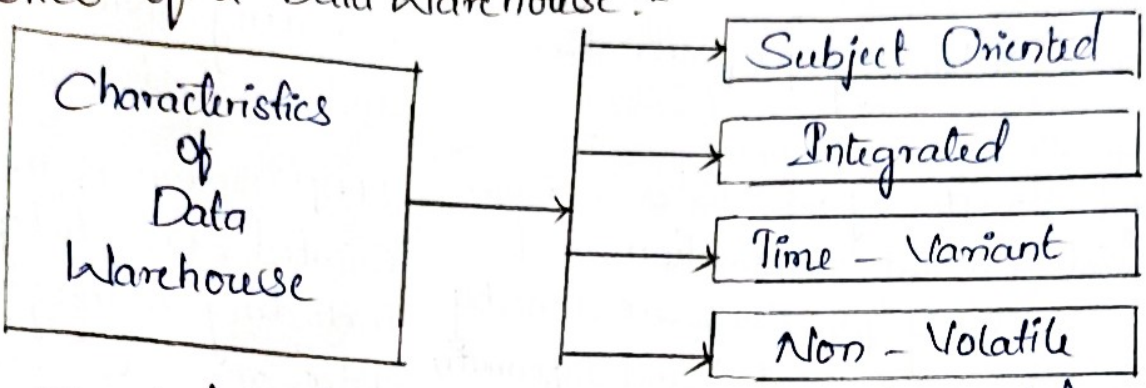
1. Data Consolidation
2. Data cleaning.
3. Data Integration

Operational Data Base	Data Warehouse.
1. Operational systems are designed to support high volume transaction processing. (OLTP)	1. Data warehousing systems are typically designed to support high-volume analytical processing (OLAP)
2. Operational systems are usually concerned with current data.	2. Datawarehousing systems are usually concerned with historical Data.
3. Data with in operational Systems are mainly update regularly according to need.	3. Non-volatile, New data may be added regularly. Once added rarely change.
4. It is designed for real-time business dealings and processes.	4. It is designed for analysis of business measures by subject area, categories, and attributes.
5. It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time pertable.	5. It is optimized for extent loads and high complex, unpredictable queries that access many rows pertable.
6. It supports thousands of con-current clients.	6. It supports a few con-current client relative to OLTP.
7. Operational systems are widely process-oriented.	7. Data warehousing systems are widely subject oriented
8. Data In	8. Data Out.
9. Less no: of Data accessed	9. Large no: of data accessed.
10. Relational Databases are created for online Transactional Processing (OLTP)	10. Data warehouse designed for on-line Analytical processing (OLAP).

Difference between OLTP and OLAP.		
Feature	OLTP	OLAP.
Characteristic	It is a system which is used to manage operational Data.	It is a system which is used to manage informational Data.
Users	Clerks, Clients and information technology professionals.	Knowledge workers including managers, executives and analysts.
System Orientation	OLTP system is a customer oriented, transaction and Query processing are done by clerks, clients, and information technology professionals.	OLAP system is market oriented, knowledge workers, including managers, do data analysts executive and analysts.
Data Contents	OLTP System manages current data that too detailed and are used for decision making.	OLAP system manages a large amount of historical data. Provides facilities for summarization and aggregation and stores and manages data at different levels of granularity.
Data Base	100MB - GB	100 GB - TB.
DataBase Design	OLTP system usually uses an entity relationship (ER) data model and application - oriented database - design.	OLAP system typically user either a star or snowflake. model and subject oriented database design.
Volume of Data	Not very large	Because of their large volume, OLAP data are stored on multiple storage media.
Access mode	Read / write	Mostly write
No: of records accessed	Tens	Millions.

Operational Database :- An operational database management system is software that is designed to allow users to easily define, modify, retrieve and manage data in real-time. While conventional database rely on batch processing, Operational database systems are oriented toward real-time, transactional operations.

Characteristics of a Data Warehouse :-



Subject-Oriented :- A datwarehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, Suppliers, Sales, revenue, etc.

A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.

Integrated :- A data-warehouse is constructed by integrating data from heterogeneous source such as relational database, flat files, etc. This integration enhances the effective analysis of data.

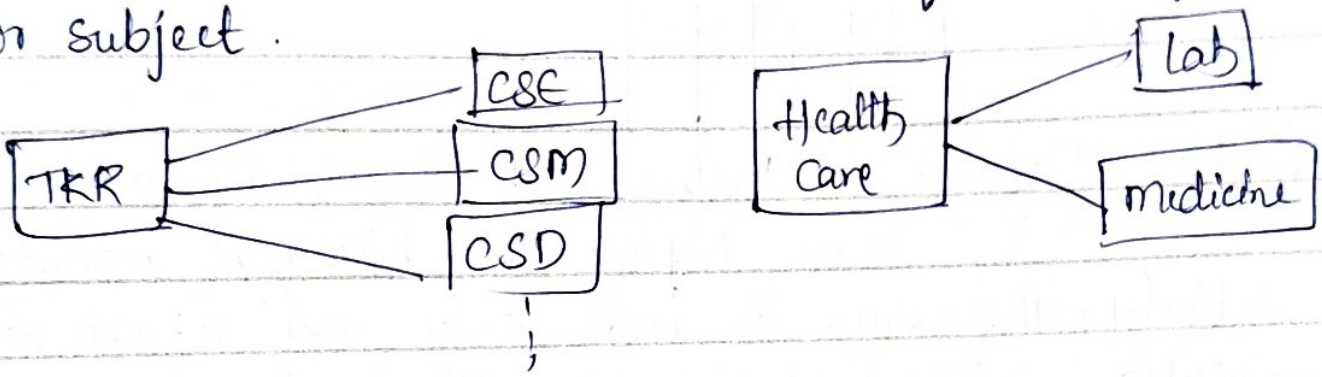
Time Variant :- The data collected in a data-warehouse is identified with a particular time period. The data in a data-warehouse provides information from the historical point of view.

(AUTONOMOUS)

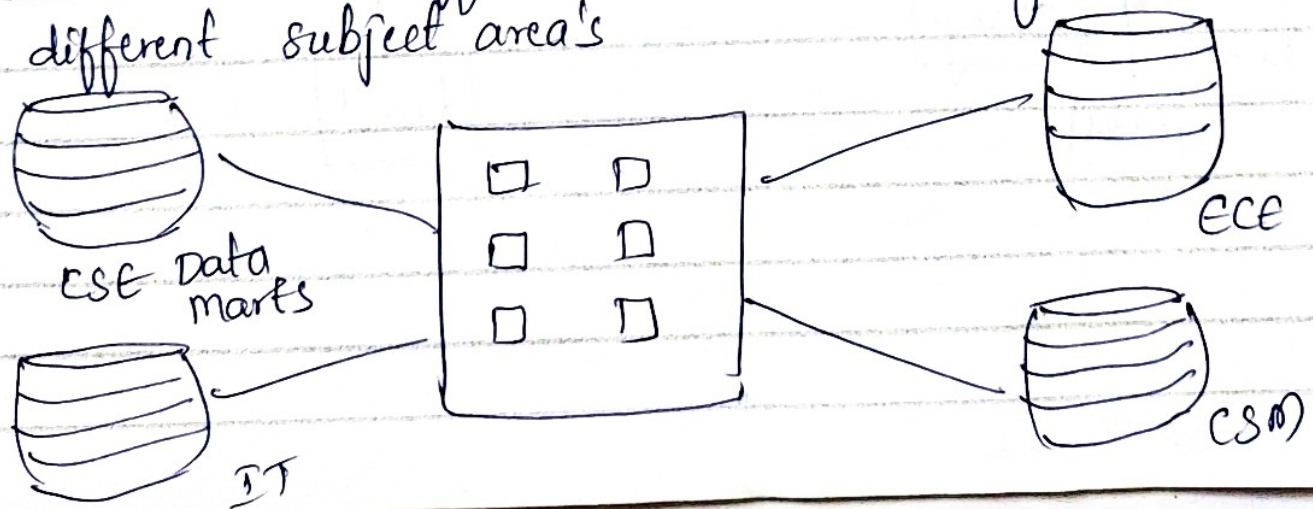
Characteristics of Data Warehousing

Subject Oriented :- It focus on specific, themes like Sales, health care, marketing or distribution rather than day-by-day operation the design its collect the information/data organised data related to a particular topic to support analysis and making decision.

It removes the unnecessary data and easier to get users clear and relavant data for the topics or subject.

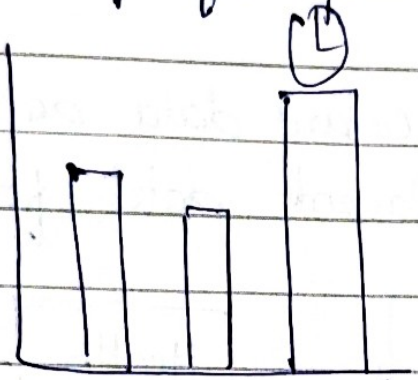


Integrated :- Combining data from different sources like, main frames and relational database into a consistent and reliable formate it ensures all the relavent data is unified for more accurate and efficient decision making across different subject area's

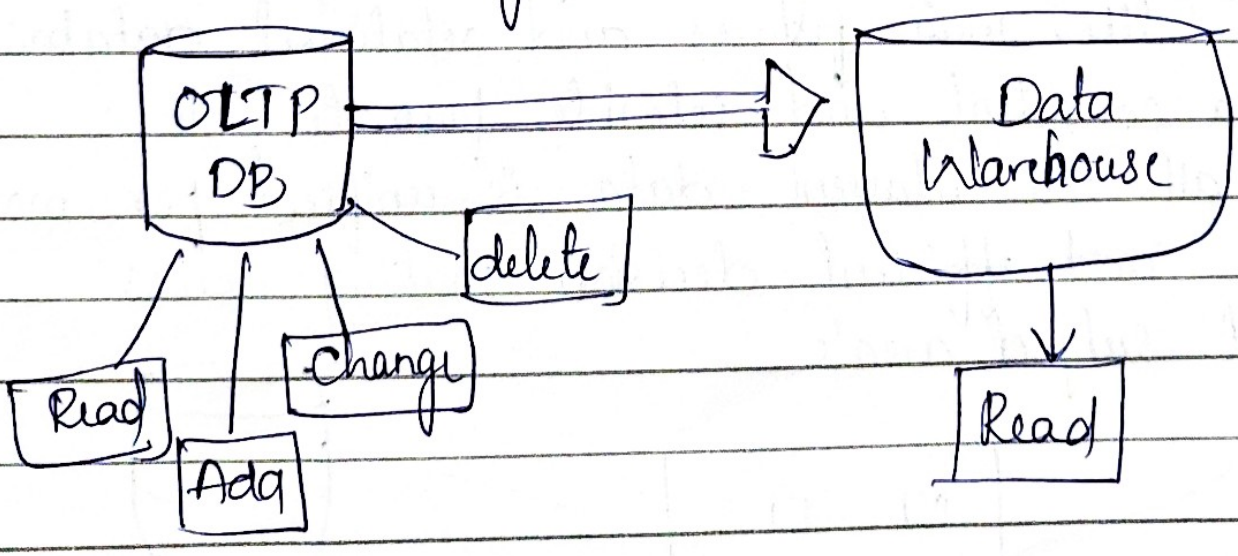


3) Time Varient :- The data which stored in data warehouse over a different time periods such as weekly monthly, Yearly.

Once the data is entered in datawarehouse it is not changed or updated, preserving the state of data at a specific point of Time.



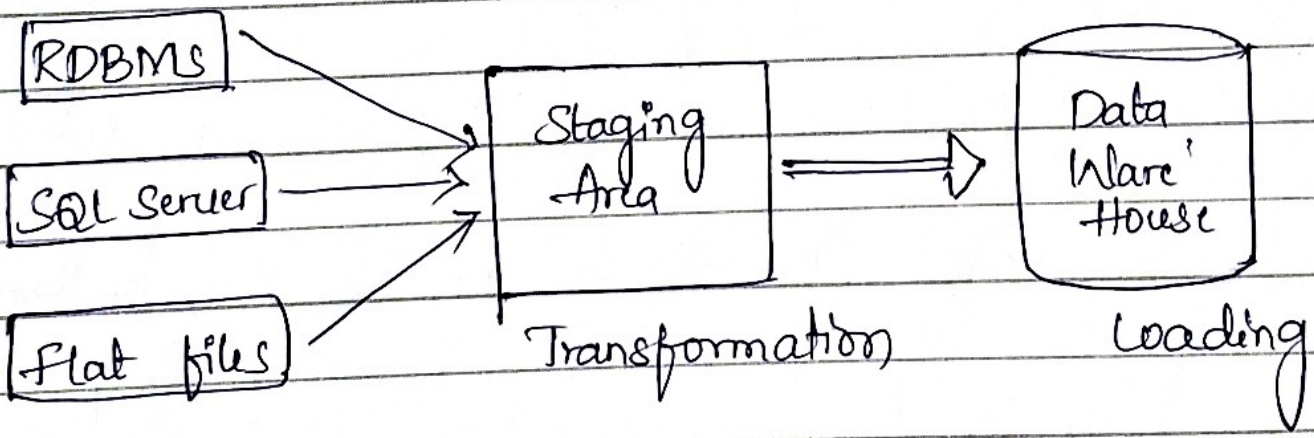
4) Non-Volatile :- Instead of new data added over time keeping the historical records actively. the data is read only and refresh at a specific interval by making it ideals for Analysing trends and long term performance.



Application of DW :-
Organization
Banking sectors
Colleges
Government sectors

Use of Data Warehouse :-
It acquire the knowledge which is already stored in database and perform analysis for better implementation

ETL Process in Data Warehouse :-
Extract Transform Load



Extraction :- The collection of data from different sources RDBMS, SQL, flat files, xml spread sheets, web pages etc. while loading in DW it has to be transform in staging area because the file formats are different from diff sources that can be corrupt / interrupt Data.

formats and can be corrupted also. Hence loading it directly into the datawarehouse may damage it and rollback will be much more difficult.

Therefore, this is one of the most important steps of ETL process.

2. Transformation :- Set of rules or functions are applied on the extracted data to convert it into a single standard format. It may involve processes / Tasks.

a) Filtering :- Loading only certain attributes into the DW

b) Cleaning :- Filling up the NULL values with some default values

Mapping U.S.A, United States, and America into USA, etc.

c) Joining :- Joining multiple attribute into one.

d) Splitting :- Splitting a single attribute into multiple attributes

e) Sorting :- Sorting tuples on the basis of some attributes (generally key-attribute)

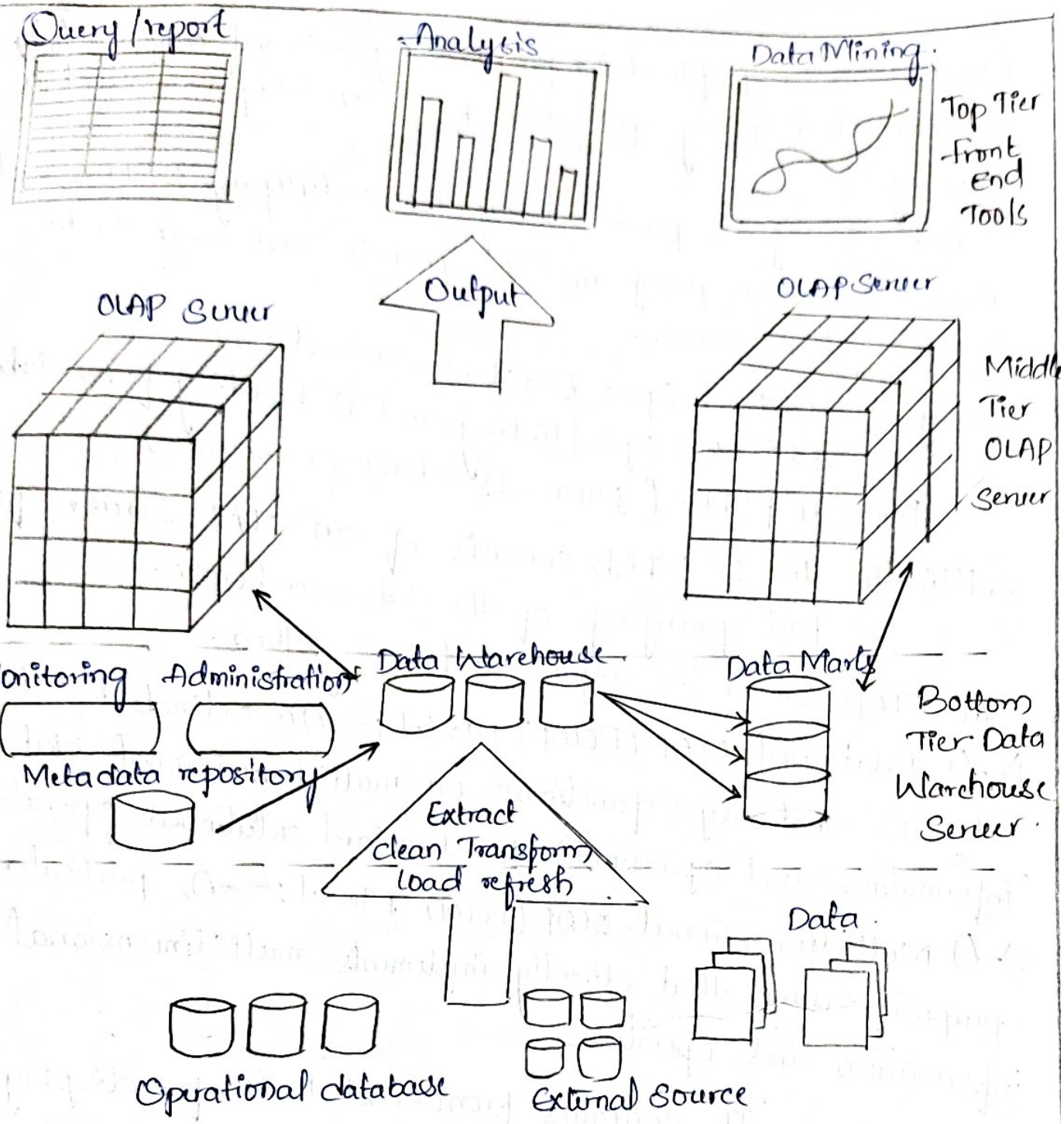
3. Loading :- The transformed data is finally loaded into the data warehouse. Sometimes the data is updated by loading into the DW very frequently and Sometimes it is done after longer but regular intervals.

The rate and period of loading solely depends on the requirements and varies from system to system.

∴ ETL process can also use the pipelining concept i.e., as soon as some data is extracted, it can transformed and during the period some new data can be extracted.

Advantages and Disadvantages of ETL process

Advantages in ETL	Disadvantages in ETL
1. Improved Data Quality :- ETL process ensures that the data in the data warehouse is accurate, complete, and up-to-date.	1. High-cost :- ETL process can be expensive to implement and maintain especially for organizations with limited resources.
2. Better Data Integration :- ETL process helps to integrate data from multiple sources and systems, making it more accessible and usable.	2. Complexity :- ETL process can be complex and difficult to implement, especially for organizations that lack the necessary expertise or resources.
3. Increased data security :- ETL process can help to improve data security by controlling access to the data warehouse and ensuring that only authorized users can access the data.	3. Limited flexibility :- ETL process can be limited in terms of flexibility, as it may not be able to handle unstructured data or real-time data streams
4. Improved Scalability :- ETL process can help to improve scalability by providing a way to manage and analyze large amounts of data.	4. Limited Scalability :- ETL process can be limited in terms of scalability, as it may not be able to handle very large amounts of data.
5. Increased Automation: ETL tools and technologies can automate and simplify the ETL process, reducing the time and effort required to load and update data in the ware-house.	5. Data privacy Concerns :- ETL process can raise concerns about data privacy, as large amounts of data are collected, stored, and analyzed.



Three-Tier Data Warehouse Architecture

The metadata Repository: Stores information that defines DW objects. It includes the parameters and information for the middle and Top-Tier applications.

1. A description of the DW structure, including the warehouse schema, dimensions, hierarchies, data mart locations, and contents, etc.



Data Marts - Contains a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to specific selected subjects.

ex: a marketing data mart may confine its subjects to customer, item & sales. The data contained in DM tends to be summarized

Three - Tier Architecture of Data Warehouse.

→ DW helps for making decisions in business using large datasets.

→ This widely used for it's clear structure, dividing data processing into three layers for efficient access and management.

1) Bottom Tier (Data Sources and Data Storage)
2) Middle Tier (OLAP Engine)
3) Top Tier (Front-End Tools)

Bottom Tier :- The foundation of DW, responsible for collecting, processing and storing data from multiple sources. It plays a critical role in preparing data for analysis.

Common challenges in Bottom Tier.

Integrating data from diverse sources presents several challenges.

1) Data Quality :- Inconsistent data can lead to errors and unreliable analytics.
2) Data Compatibility :- Different data formats and structures can complicate integration.
3) Scalability :- Handling increasing volumes of data efficiently

Solutions are: Implement Robust ETL Tools :-

Utilize powerful ETL tools like Informatica, Microsoft SSIS / Confluent to streamline the data integration process.

Standardize Data formats :- At the point of entry minimizes compatibility issues.

Continuous Data Quality Management :- Regularly check and clean data to maintain high quality.

Scalability Planning :- Design data storage solutions that can expand as data volume grows, ensuring that the architecture can handle future increase in data without performance degradation.

Middle Tier :- hosts the OLAP Server, which processes complex analytical queries. It acts as a bridge b/w the data storage layer (BT) and the user interface (TT), ensuring data is quickly retrieved, aggregated and ready for reporting and analysis. It is designed for high speed analytical processing. OLAP Server models come in 3.

Note :- OLAP is a powerful technology for complex calculations, trend analysis and data modeling.

1) ROLAP (Relational OLAP) :- This model uses a relational DB to store and manage warehouse data. It is ideal for handling large data volumes as it operates directly on relational databases

2) MOLAP (Multidimensional OLAP) :- This model stores data in a multidimensional cube. The storage and retrieval processes are highly efficient, making MOLAP suitable for complex analytical queries that require aggregation.

3) HOLAP (Hybrid OLAP) :- It is combination of relational and multidimensional online analytical processing paradigms. HOLAP is the ideal option for a seamless functional flow across the database system when the repository houses both the relational database management system and the multidimensional DBMS.

Common Challenges in Middle Tier :-

Data Latency :- Delays in data availability can impact decision-making.

Query Performance :- Managing large volumes of data can slow down query performance.

Data Integration :- Combining data from different sources with varying formats can be challenging.

Solution for Real-Time & Incremental Loading :-

Update data frequently using real-time and incremental loading to reduce latency and support faster decision-making.

Query Optimization :- Improve performance with indexing, partitioning and optimized SQL for faster data retrieval.

Standardization & Integration Tools :- Standardize data formats and use tools like Talend or Informatica for seamless integration and improved data quality.

Top Tier :- The front-end client layer, which is essential for interacting with the data stored and processed in the lower tiers. This layer includes a variety of business intelligence (BI) tools and techniques designed to facilitate easy access and manipulation of data for reporting, analysis, and decision-making.

BI tools are critical components of the Top Tier, providing robust platforms through which users can query, report, and analyze data. Popular BI Tools include :

IBM Cognos :- Offers comprehensive reporting capabilities.

Microsoft BI platform :- Integrates well with existing Microsoft products, providing a familiar interface for users.

SAP BW :- Specializes in managing large DB and integrating with other SAP products.

(AUTONOMOUS)

Crystal Reports :- known for its powerful reporting features

SAS Business Intelligences :- Provides advanced analytics

Pentaho :- A versatile tool for data integration and visualization.

Common Challenges in Top Tier :-

Usability Issues :- Complex tools can hinder user adoption and effectiveness.

Integration Difficulties :- Ensuring seamless integration with other tiers can be challenging.

Solutions are: User Training and Support :- Offering comprehensive training sessions to help users fully leverage the capabilities of BI tools

Choosing Integrative Tools :- Select tools that easily integrate with existing systems in the data Warcho architecture, ensuring consistency and reliability in data handling.

Logical Multi-Dimensional Data Model:-

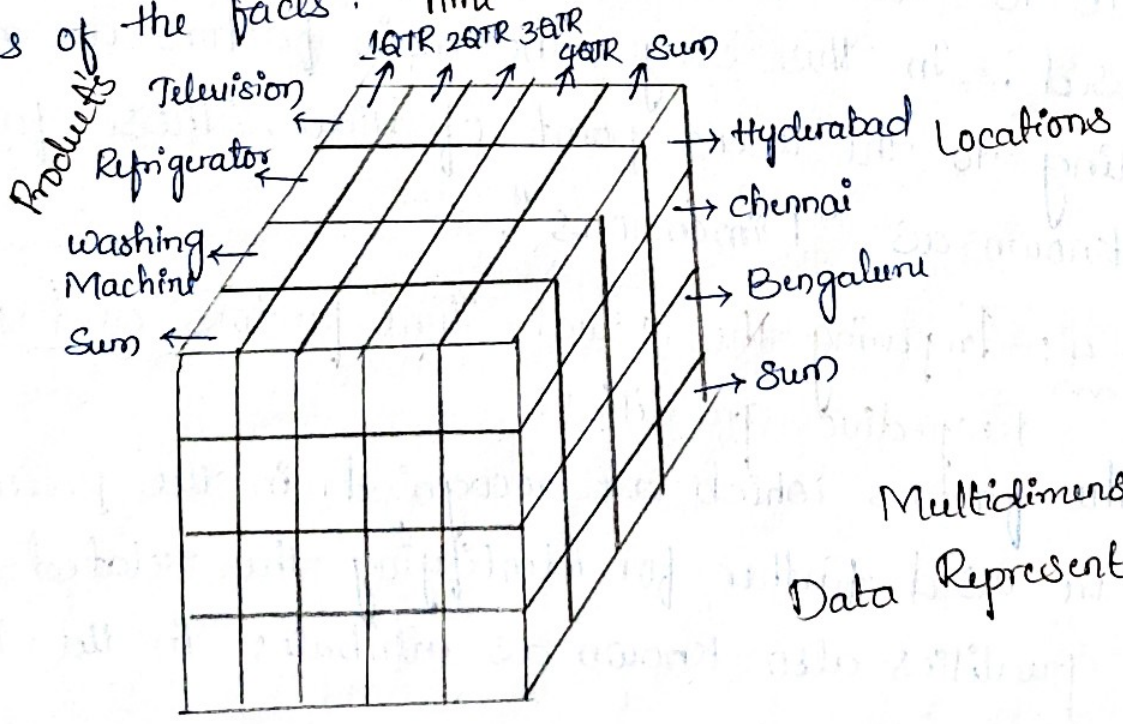
The multi-dimensional Data Model is a method which is used for ordering data in the database along with good arrangement and assembling of the contents in the database.

It allows customers to interrogate analytical questions associated with market / business trends, unlike relational database which allow customers to access data in the form of Queries.

They allow users to rapidly receive answers to the requests which they made by creating and examining that data comparatively fast.

OLAP (Online Analytical processing) and datawarehousing uses multi-dimensional database. It is used to show multiple dimensions of data to users.

It represents data in the form of data cubes. Data cubes allows to model and view the data from many dimensions and perspectives. It is defined by dimensions and facts and its represented by a fact-table. Facts are numerical measures and fact-tables contains measures of the related dimensional tables or names of the facts.



Multidimensional Data Representation.

Working on a Multi-dimensional Data Model:-

The following stages should be followed by every project for building a multi-dimensional Data Model.

Stage-1:- Assembling Data from the client:

A multi-dimensional data Model collects correct Data from the client. Mostly, Software professionals provides Simplicity to the client about about the range of data which can be gained with the selected technology and collect the complete data in detail.

Stage-2:- Grouping different Segments of the System:-

The multi-dimensional Data Model recognize and classifies all the data to the respective section they belong to and also build it problem-free to apply step-by-step.

Stage-3:- Noticing the different proportions:-

It is the basis on which the design of the system is based. In this stage, the main factors are recognized according to the user's point of view. These factors are also known as "Dimensions".

Stage-4:- Preparing the actual-Time factors and their Respective qualities.

The factors which are recognized in the previous step are used further for identifying the related qualities these qualities also known as attributes in the Database.

Stage-5:- Finding the actuality of factors which are listed previously and their qualities.

A multi-dimensional Data Model separates and differentiates the actuality from the factors which are collected by it. These actually play a significant role in the arrangement of a multi-dimensional Data Model.

Stage-6:- Building the Schema to place the data with respect to the information collected from the steps above:-

On the basis of the data which was collected previously a schema is built.

Let us take the example of the data of a factor which sells products per quarter in Banglore. The Data is represented in the table.

Location = "Banglore"				
Time (Quarter)	Type of Item			
	Jam	Bread	Sugar	Milk
Q1	350	389	35	50
Q2	260	528	50	90
Q3	483	256	20	60
Q4	436	396	15	40

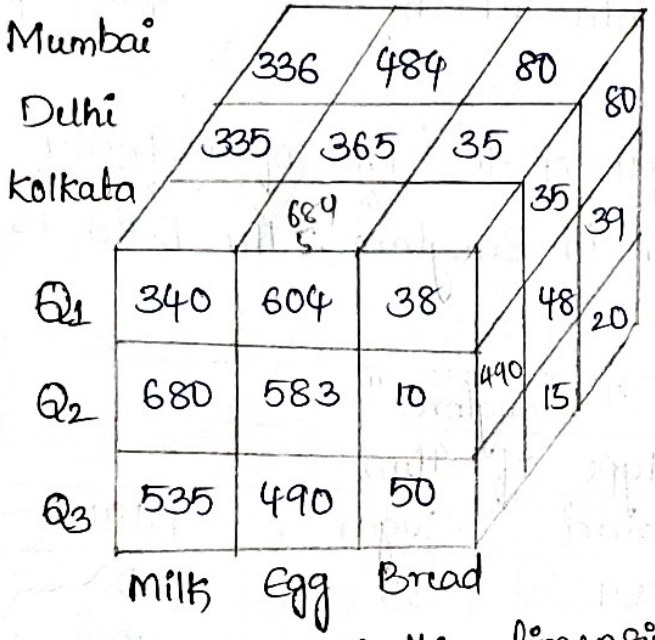
In the above given presentation. The factory's sales for banglore are, for the time dimension which items, which is sorted according to the kind of item which is sold.

If we desire to view the data of the sales in a three dimensional table then it is represented in a diagram, here the data of the sales is represented as a two-

dimensional table. Let us consider the data according to item, time and location.

	Location = "Kolkata"			Location = "Delhi"			Location = "Mumbai"		
	Item			Item			Item		
Time	Milk	Egg	Bread	Milk	Egg	Bread	Milk	Egg	Bread
Q_1	340	604	38	335	365	35	336	484	80
Q_2	680	583	10	684	490	48	595	594	39
Q_3	535	490	50	389	385	15	366	385	20

This Data Represented in the form of 3-dimensions conceptually



Advantages of Multi-dimensional Data Model :-

1) A multi-dimensional data-Model is easy to handle.
2) It is easy to maintain
3) It's performance is better than that of normal data base
4) The representation of data is better than traditional databases. that is because the multi-dimensional database are multi-viewed and carry different types of factors
5) It is workable on complex systems and applications,

Contrary to the simple one-dimensional database systems.

6) The compatibility in this type of database is an upliftment for projects having lower bandwidth for maintance staff.

Disadvantages of Multi-Dimensional Data Model :-

1) The multi-dimensional Data Model is slightly complicated in nature and it requires professional to recognize and examine the data in the database.

2) It is complicated in nature due to which the databases are generally dynamic in design

3) The path to achieving the end product is complicated most of the time

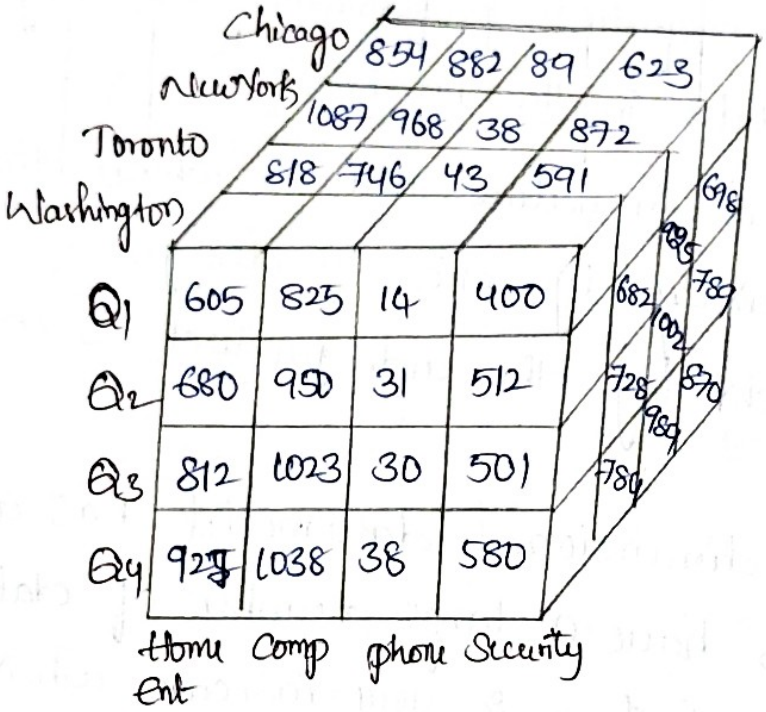
4) As the multi-dimensional data model has complicated systems, databases have a large number of databases due to which the system is very insecure when there is a security breach.

Example :- 2D view of Sales Data for all electronics according to time & Item.

Time	Location = Washington			
	Items			
	Home entertainment	Computer	Phone	Security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1032	38	580

3D view of the sales for All electronics According Time, Item, Location.

Location = Chicago					Location = New York				Location = Toronto				Location = Washington			
	Item				Item				Item				Item			
Time	Home Ent	Comp	phone	Sec	Home Ent	Comp	phone	Sec	Home Ent	Comp	phone	Sec	Home Ent	Comp	phone	Sec
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	19	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580



Data Modeling :-

Data Warehouse modeling is the process of designing the schemas of the detailed and summarized information of the data warehouse.

Data-Warehouse modeling is an essential stage of building a data-warehouse for two Reasons.

First, through the schema, data Warehouse clients can visualize the relationships among the warehouse data, to use them with greater ease.

Multi-Dimensional Data Model (MDM)

Organizes data into multiple dimensions such as time, product, location to support fast analytical queries in data warehouse and OLAP systems.

Key features of the Multidimensional Data Model.

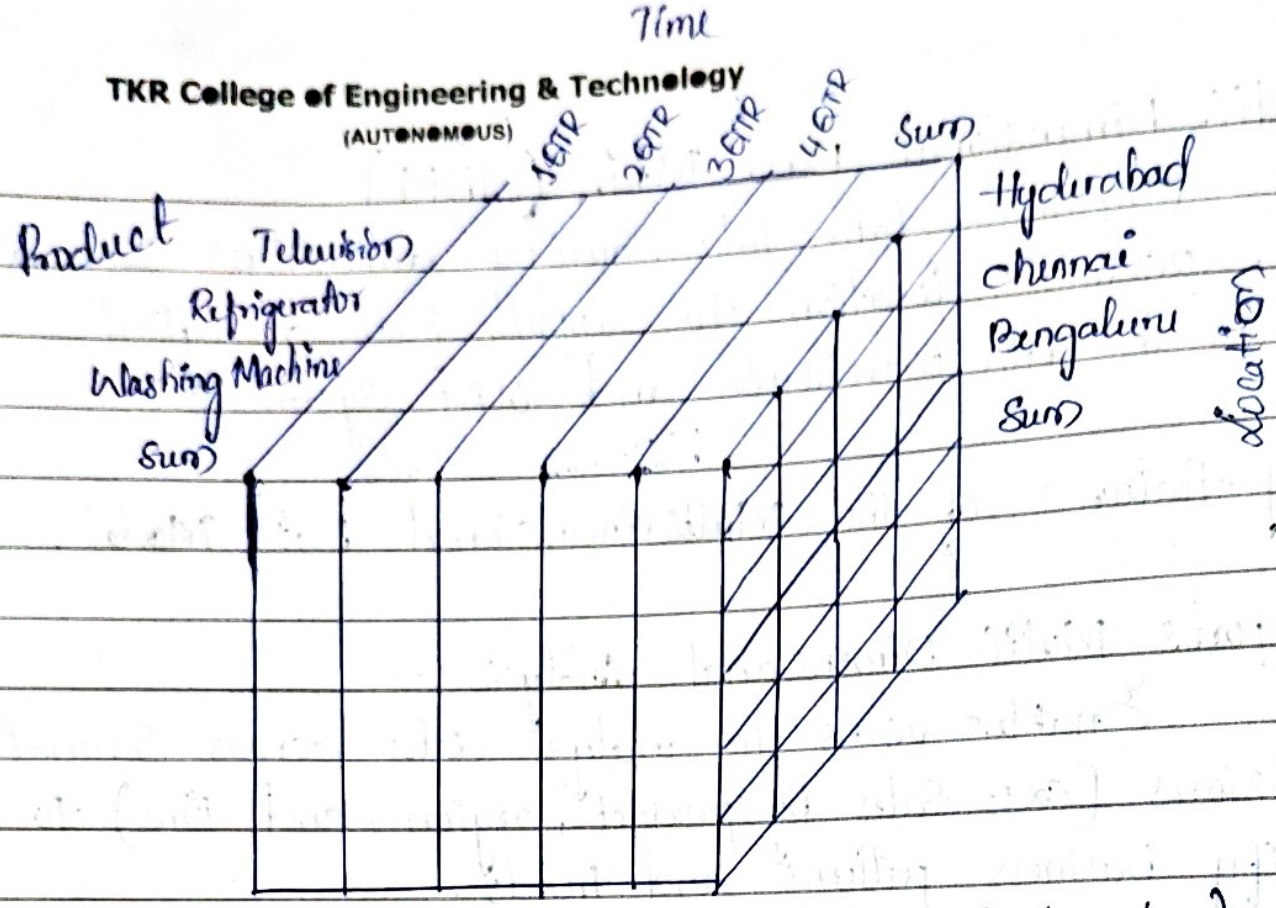
1) Supports Multi-Dimensional Analysis :-
Enables users to analyze data across several dimensions (ex:- Sales by product, region, and time) to identify business patterns and trends.

2) Used in OLAP and Data Warehousing :-
Forms the foundation of OLAP systems, where data is stored in multidimensional structures for rapid aggregation and reporting.

3) Data cubes :- Represents data as cubes that allow viewing information from different perspectives and performing complex analytical operations efficiently.

4) Facts :- Numerical measures (ex:- sales, revenue). Stored in a fact table.

5) Dimensions :- Descriptive attribute (ex: product, region, customer). Stored in dimension tables. The fact table links to multiple dimensions to support multi-axis analysis.



Multi-dimensional Data Representation (Data cubes)

Multi-dimensional Data Representation

Fact Table :- Contains measures and foreign keys to dimensions

Dimension Tables :- Contains attributes that describe context for facts

Cube Representation :- Organizes facts across multiple dimensions for faster aggregation and slicing.

Working on a Multi-dimensional Data Model.

1) Identifying Business Requirements :- Determine the key questions the business wants to analyze. (Ex: Sales, trends, regional performance).

2) Defining Dimensions :- Identify descriptive categories such as time, product, location, or customer that will provide context for analysis.

3) Defining facts :- Select measurable values such as Sales, cost, or units sold that will be stored in the fact table.

4) Identifying Hierarchies :- Establish natural roll-up paths eg: (Year → Quarter → Month → Day)

5) Designing the Schema :- Create a Star/Snowflake Schema linking fact and dimension tables.

6) Building Data Cubes :- Organize facts and dimensions into cube structures that support slicing, dicing, drill-down, roll-up and fast aggregation.

Example :- The data of a factory which sells products per Quarter in Bangalore.

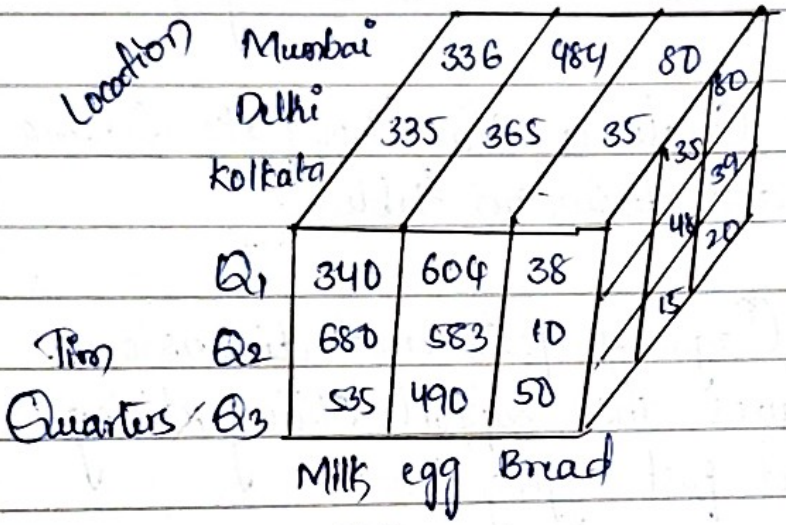
Location = "Bangalore"

Time (Quarter)	Type of Item: Jam	Bread	Sugar	Milk
Q_1	350	389	35	50
Q_2	260	528	50	90
Q_3	483	256	20	60
Q_4	436	396	15	40

	Location = "Kolkata"			Location = "Delhi"			Location = "Mumbai"		
Time	Item			Item			Item		
	Milk	Egg	Bread	Milk	Egg	Bread	Milk	Egg	Bread
Q1	340	604	38	335	365	35	336	484	80
Q2	680	583	10	684	490	48	595	594	39
Q3	535	490	50	389	385	15	366	385	20

2D representation of Data

The same data can also be represented in the form of three dimensions as go



3D Representation of the same 2D data

Core Components of a MDM

1) Measures :- Numerical values used for analysis
Sales, units sold, Revenue stored in fact tables

2) Dimensions :- Descriptive attributes providing context for measures (Ex:- product type, location, time)
Stored in Dimension Tables

3) Cubes :- Multidimensional Structures linking facts and

(AUTONOMOUS)

dimensions, enabling fast slicing, dicing, aggregation drill-down and roll-up.

4) Aggregation :- Summarizing data at higher levels. eg:- daily → monthly → yearly totals) for faster reporting and trend analysis.

5) Drill Down :- Move from summary to details (Year → Quarter → Month → Day)

6) Roll-Up :- Move from details to summary (Day → Mont

7) Hierarchies: Define levels within each dimension (Ex: Regio State → City) allowing structured Navigation of data.

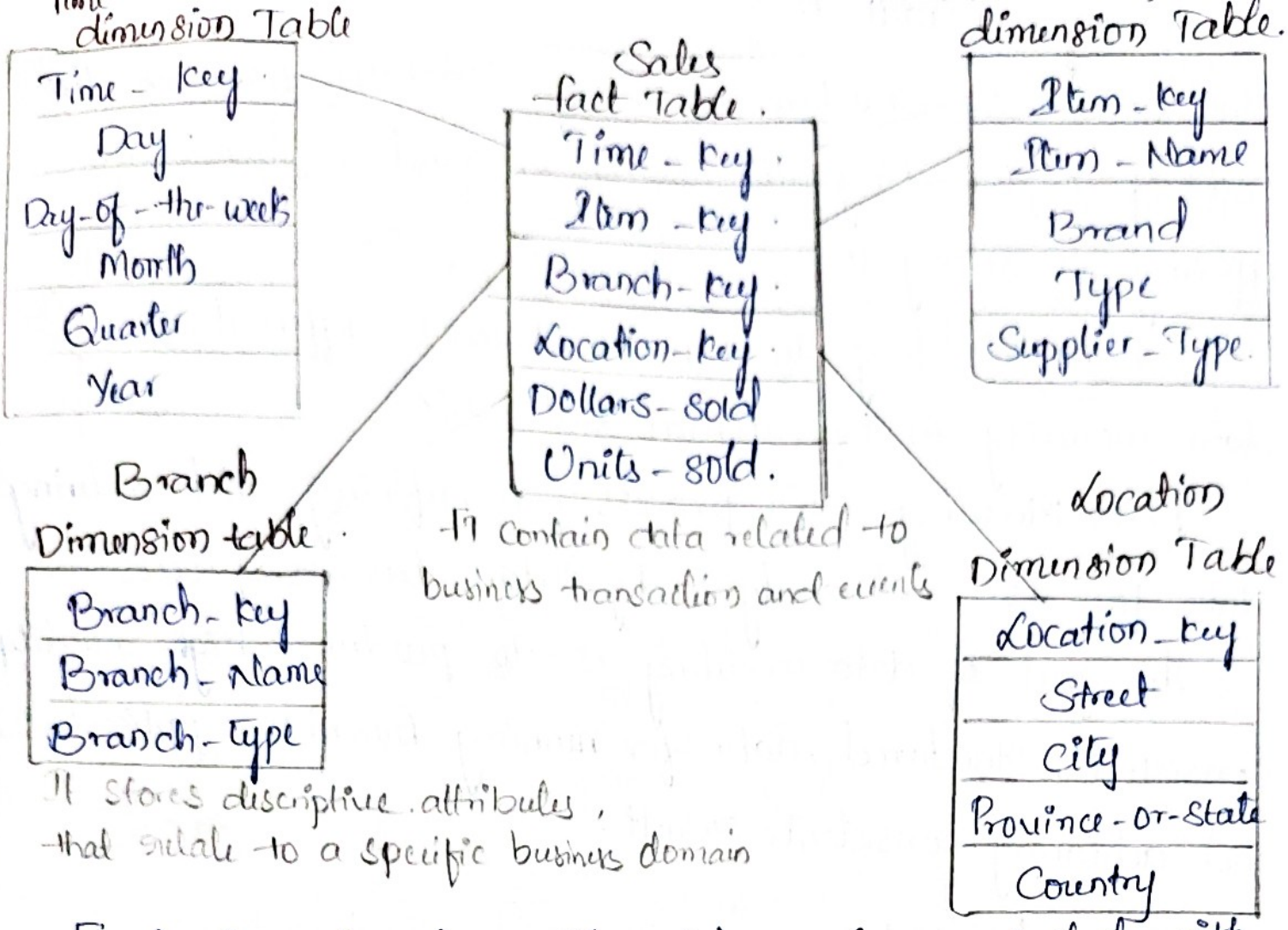
8) OLAP Integration :- OLAP System use multi-dimensional models to run complex analytical queries efficiently by leveraging cubes, hierarchies, and aggre

Schema Design :-

A Schema is a collection of database objects including tables, views, indexes, and synonyms. Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates.

A database, a data warehouse also requires to maintain a schema.

A data base uses relational model while a data warehouse uses star, Snowflake and fact constellation Schema.



* Each dimension in a Star Schema is represented with only One-dimension Table.
* This dimension table contains the set of attributes
* The diagrame shows the sales data of a company with respect to the four dimensions, namely Time, Item, Branch and Location
* There is a fact table at the center. It contains the keys to each of four dimensions.
* The fact table also contains two attributes, Namely dollars sold and Units sold,
* Each dimension has only one dimension table and each table holds a set of attributes.

⇒ A Star schema is a type of data modeling techniques used in data warehousing to represent data in a structured and intuitive way.

⇒ Data is organized into a central fact table that contains the measures of interest, surrounded by dimension tables that describe the attributes of the measures.

⇒ In a sales data warehouse, the fact table might contain sales revenue, units sold and profit margins.

⇒ Each record in the fact table represents a specific event or transaction, such as a sale or order.

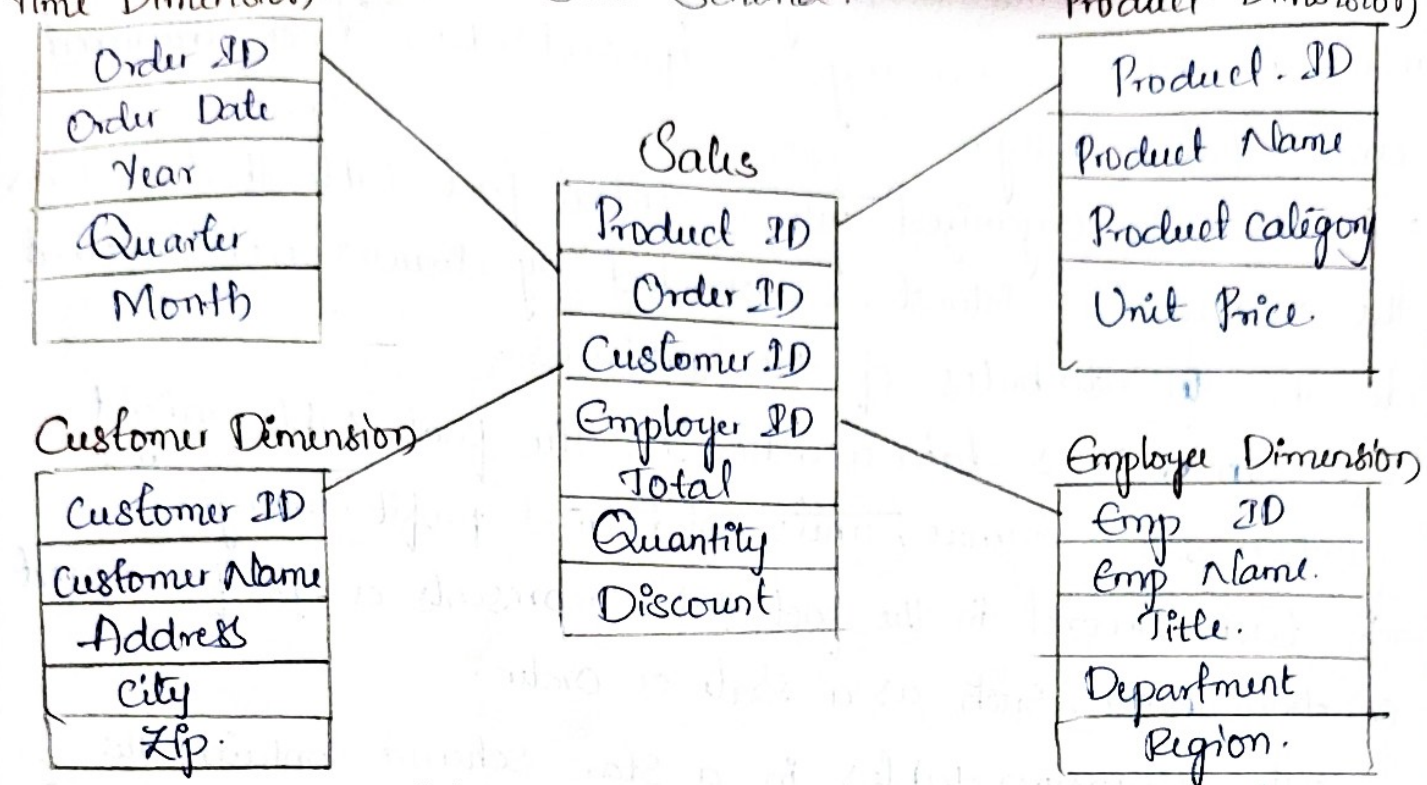
⇒ The dimension tables in a star schema contain the descriptive attributes of the measures in the fact tables.

⇒ These attributes are used to slice and dice the data in the fact table, allowing users to analyze the data from different perspectives. DT includes product, customer, time & locatn

⇒ In a star schema, each dimension table is joined to the fact table through a foreign key relationship.

⇒ This allows users to query the data in FT using attributes from the dimension tables.

* Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any no: of dimensional tables. The star schema is a necessary cause of the snowflake schema.



Model of Star Schema :-

Business process data, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact table data. Sales price, Sale quantity, distant, speed, weight, and weight measurements are few ex: of fact data in Starschema.

A star schema having multiple dimensions is termed as Centipede Schema. It is easy to handle a star schemas which have dimension of few attributes.

Advantages of Star Schema	Dis-advantages
1. Simpler Queries - Join logic star schema is quite in Comparsion to other join logic which are needed to fetch data from a transactional Schema that is highly normalized.	1. Data Integrity is not enforced well since in a highly de-normalized Schema state.
2. Simplified Business Reporting Logic - In comparison to transactional schema highly normalized but SS makes Simpler common business reporting and period-over-period.	2. Not flexible in terms if analytical needs as a normalized data model.
3. Feeding cubes :- It's widely used by all OLAP sys. to design cubes efficiently. In OLAP systems deliver a ROLAP mode of operations.	3. SS don't reinforce busines entities - atleast not frequ

Features of Star Schema :-

1. Central Fact - Table :- It contains the numerical data being analyzed. It also contains foreign keys to link to dimension tables.

2. Dimension Tables :- It contains descriptive attributes about the data being analyzed. These attributes provide context to the numerical data in the fact table. Each DT is linked to the FT through a foreign key.

3. Denormalized Structure :- It means that redundancy is allowed in the schema design to improve query performance. This is because it is easier and faster to join a small no! of tables than a large no! of tables.

4. Simple Queries :- It is designed to make queries simple and fast. Queries can be written in a straightforward manner by joining the fact table with the appropriate DT.

5. Aggregated Data :- The numerical data in the FT is usually aggregated at different levels of granularity, such as daily, weekly, or monthly. This allows for analysis at different levels of detail.

6. Fast Performance :- Designed for fast query performance, because the schema is denormalized and data is pre-aggregated making queries faster and more effecient.

7. Easy to Understand :- It is also interpret, even for non-technical users, because the schema is designed to provide context to the numerical data through the use of DT.

Snowflake Schema in Data Warehouse Model :-

The Snowflake Schema is a variant of the Star Schema. Here, the centralized fact table is connected to multiple dimensions. In the Snowflake schema, dimensions are present in a normalized form in multiple related tables. The Snowflake structure materialized when the dimensions of a star schema are detailed and highly structured, having several levels of relationship, and the child tables have multiple parent tables.

A Snowflake schema is a type of data modeling technique used in datawarehousing to represent data in a structured way that is optimized for querying large amounts of data efficiently.

In a Snowflake schema, the dimension tables are normalized into multiple related tables, creating a hierarchical or "Snowflake" structure.

What is Snowflaking ? (multi-dimensional data model that's used to store data in a complex format)

It is result of further expansion and normalization of the dimension table, is said to be Snowflaked if the low-cardinality attribute of the dimensions has been divided into separate normalized tables. These tables are then joined to the original dimension table with referential constraints (foreign key constraint)

Characteristics of Snowflake Schema :-

1) The Snowflake Schema uses small disk space.
2) It is easy to implement the dimension that is added to the Schema.

3) There are multiple tables, so performance is reduced.
4) The dimension table consists of two or more sets of attributes that define information at different grains
5) The sets of attributes of the same dimension table are populated by different source systems.

Advantages :-

1) It provides structured data which reduces the problem of data integrity
2) It uses small disk space because data are highly structured

Disadvantages :-

1) Snowflaking reduces space consumed by dimension tables but compared with the entire data warehouse the saving is usually insignificant.
2) Avoid snowflaking or normalization of a dimension table, unless required and appropriate
3) Do not snowflake hierarchies of dimension table into separate tables. Hierarchies should belong to the dimension table only and should never be snowflake.
4) Multiple hierarchies that can belong to the same dimension have been designed at the lowest possible details.

Features of the Snowflake Schema :—

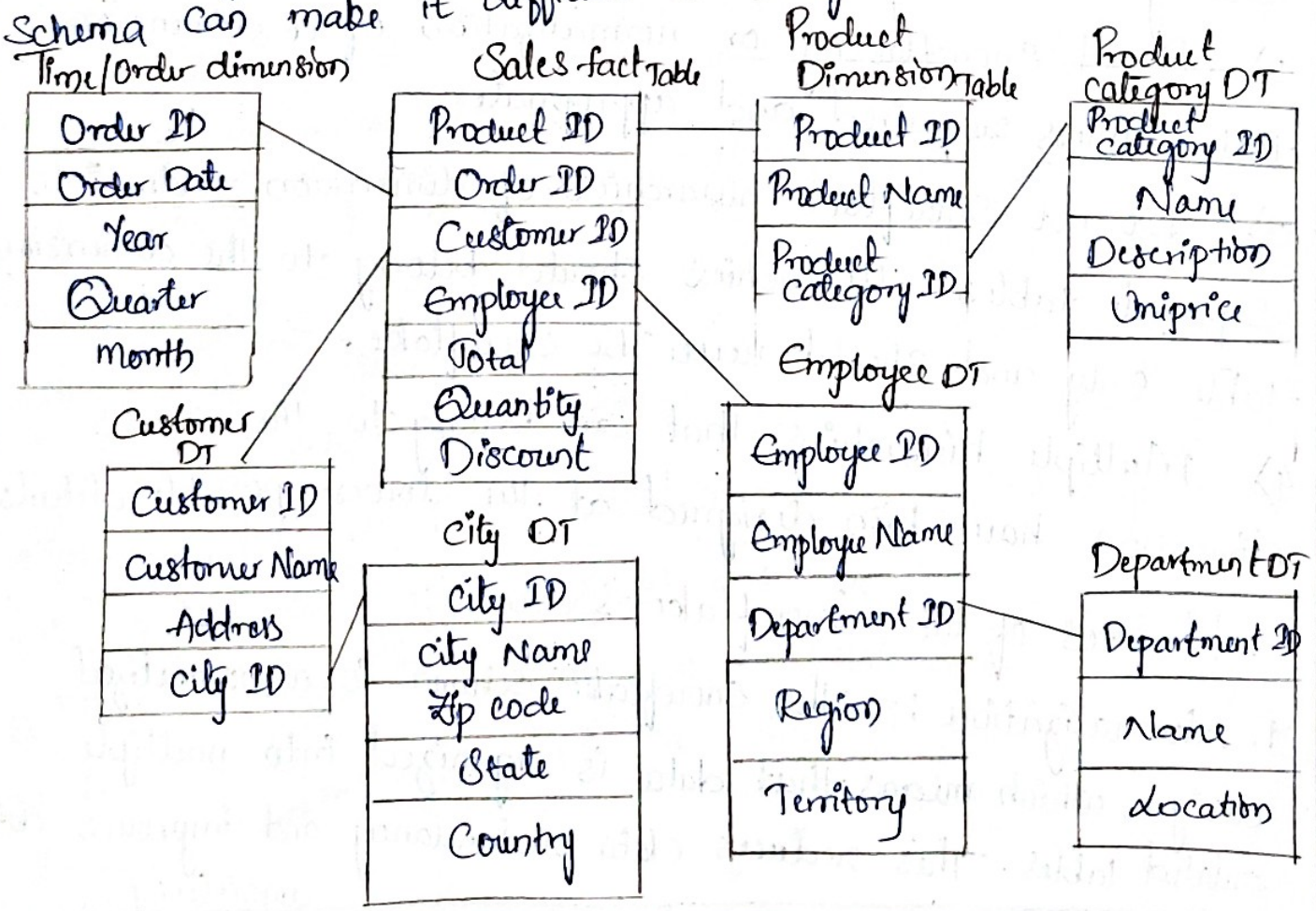
1. Normalization :— The snowflake schema is a normalized design, which means that data is organized into multiple related tables. This reduces data redundancy and improves data consistency.

2) Hierarchical Structure :- That is organized around a central fact table. The fact table contains the measures or metrices of interest, and the dimension tables contain the attributes that provide context to the measures.

3) Multiple Levels :- The Snowflake schema can have multiple levels of dimension tables, each related to the central fact table. This allows for more granular analysis of data and enables users to drill down into specific subsets of data.

4) Joins :- The Snowflake schema typically requires more complex SQL queries that involve multiple tables joins. This can impact performance, especially when dealing with large data sets.

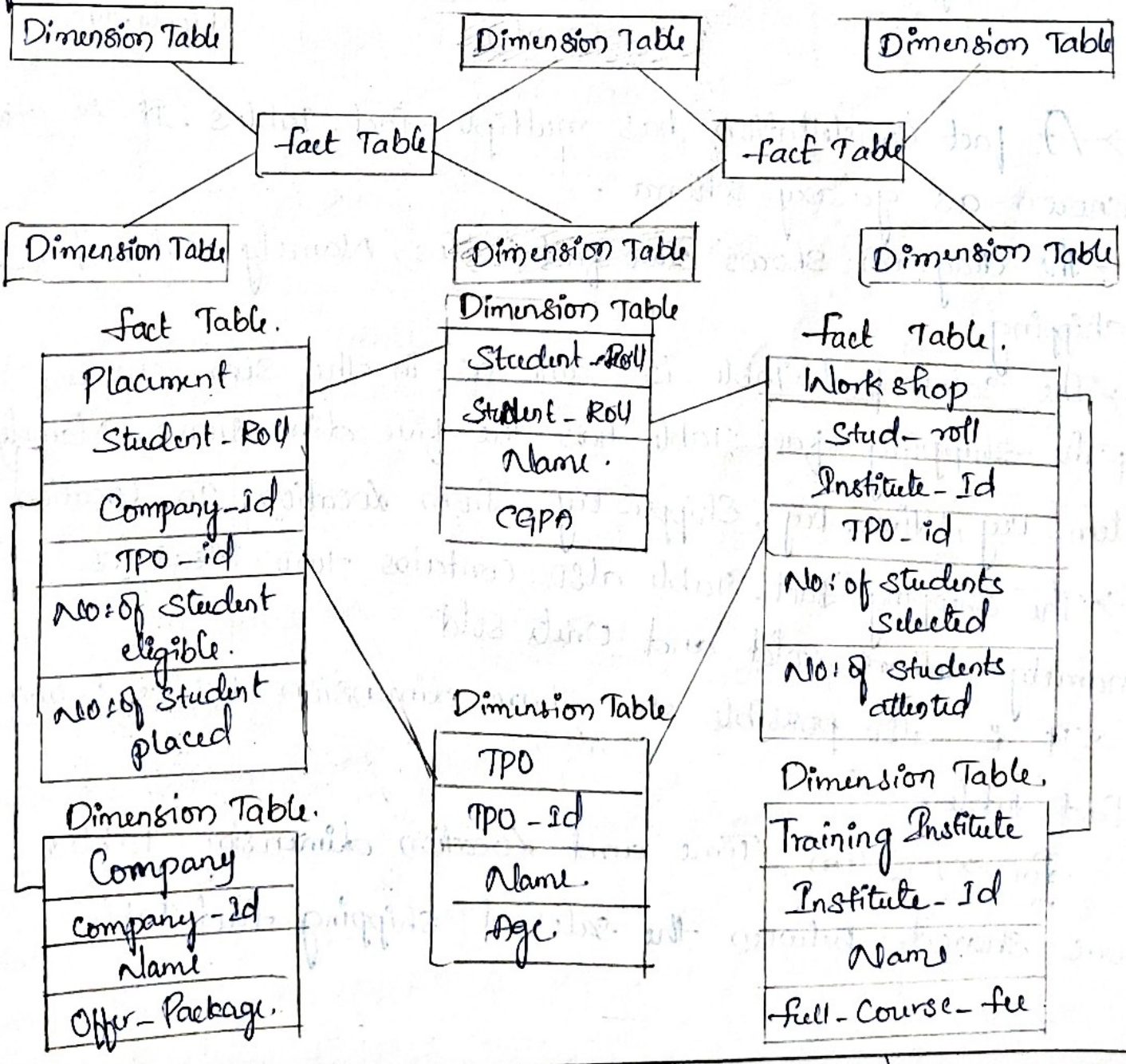
5) Scalability :- The Snowflake Schema is scalable and can handle large volumes of data. However, the complexity of the Schema can make it difficult to manage and maintain.

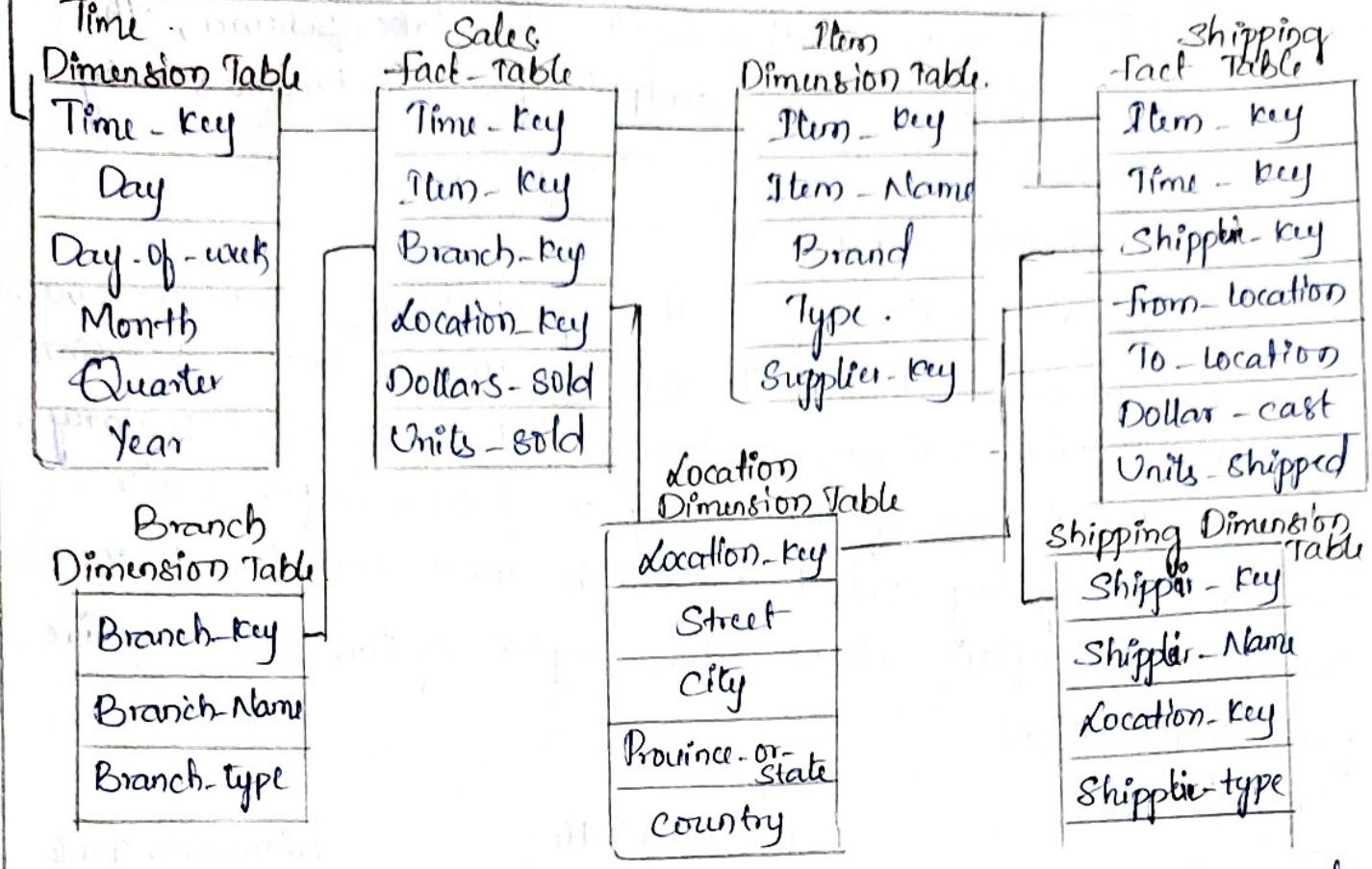


NOTE :- Due to normalization in the Snowflake Schema, the redundancy is reduced and therfore, it becomes easy to maintain and the save storage space.

- Fact Constellation Schema :-

It is a collection of multiple fact tables having some common dimension tables. It can be viewed as a collection of several star schemas and hence, also known as Galaxy Schema. It is one of the widely used schema for Data warehouse designing and it is much more complex than Star and Snowflake Schema. For complex systems, we require fact constellations.





→ A fact Constellation has multiple fact_Tables. It is also known as galaxy schema.

→ The diagram shows two fact_Tables, Namely Sales & Shipping.

→ The Sales fact_Table is same as in the Star Schema.

→ The shipping fact Table has the five dimensions, Namely Item_key, Time_key, Shipper_key, from_location, To_location.

→ The shipping fact_Table also contains two Measures namely dollars sold and Units sold

→ It is also possible to share dimension tables between fact Table.

for ex: Item, Time and Location dimension Tables are shared between the sales & shipping fact Table.

Q: What is a fact ?

In DW, a fact represents a business event or transaction that can be measured and analyzed.

Ex:- Product sale, Bank Transaction, a Student Examination result, Airline Booking

Ex:- Sales fact Table

Date_key	Product_key	customer_key	store key	Quantity	Sales_Amount	Discount
101	501	1001	201	5	5,000	500

Date, Product, Customer, Store → Foreign key.

Quantity, Sales_Amount → Measures

Q: What is a Measure ?

A measure is a numerical value stored in a fact table that represents a Business performance indicator

Measures used :- Calculation, Decision-Making, Aggregation, Analysis, Reporting.

Ex:-

Business Area	Measure
Sales	Sales Amount, Quantity, sold, Profit
Banking	Transaction_Amount, Balance
Education	Marks, Attendance
Hospital	Treatment cost, No:of Patients
E-commerce	Order_value, Shipping Cost

Fact Measures :- Numerical data values that describe the performance or activity of a business process

Ex:- Sales Amount = Quantity x Unit Rice

If: Quantity = 10, Unitprice = 500

Sales-Amount = 10 x 500 = 5,000

Types of Fact Measures :- 1) Fully Additive Measures
2) Semi-Additive "
3) Non-Additive "
4) Derived "
5) Stored "
6) Factless "

1) Fully Additive Measures :-

Can be added across all dimension

Ex:-

Date	Product	Region	Sales
Jan1	Laptop	North	10,000
Jan2	Laptop	South	15,000
Jan3	Laptop	East	20,000

Total sales :-

10000 + 15,000 + 20,000 = 45,000

Sales can be added by :- Time, product, Region, customer

∴ Sales Amount is an additive measures

Sales Revenue, Quantity sold, Total cost, Profit, Non-fo

Advantages :- 1) Easy to aggregate
2) Useful for reports
3) Supports SUM operations

(AUTONOMOUS)

2) Semi-Additive Measures:- Can be added across some dimensions but not across all dimensions. The most common is Account Balance.

Ex:-

Date	Account	Balance
Jan 1	A1	10,000
Jan 2	A1	12,000
Jan 3	A1	15,000

It is incorrect to calculate: 10,000 + 12000 + 15000 = 37000 because these are balance on different dates

Ending balance, Beginning balance, Average balance

Ex:-

Date	Branch	Account Balance
Jan 1	Hyd	1,00,000
Jan 2	Vijayawada	2,00,000

Bal can be added across Branches

1,00,000 + 2,00,000 = 300000

But, bal should not normally be added across time

Ex:- Bank Account Bal, Inventory Level, Stock Quantity, Employee Headcount, Available Capacity.

3) Non-Additive Measures:- Cannot be added across any dimension.

Ex:-

Product	profit Margin
Laptop	20%
Mobile	30%

We cannot calculate: 20% + 30% = 50%

Instead, the percentage should be calculated using the underlying values :-

$$\text{profit Margin} = \frac{\text{Total Profit}}{\text{Total Sales}} \times 100$$

Ex:- Percentage, Ratio, Avg, Unit price. Profit Margin, Conversion Rate.

Ex:- $\text{profit Margin} = \frac{20,000}{10,000} \times 100 = 20\%$.

* Fact less fact :- A factless fact table contains no numerical measures

It records the occurence of an event or relationship.

Ex:-

Student_key	Date_key	class_key
101	26/07/25	10
102	28/7/25	10

The row itself represents an event ; The stu attended class
∴ There may be no numerical measure such as marks or fees

Uses :- Attendance tracking
Student enrollment
Event participation
Product Promotion tracking
Customer visits

Dimension Table Characteristics :-

Dimension Table key :- Primary key of the dimension Table uniquely Identifies each row in the Table.

Table is wide :- A dimension Table has many columns attributes

It is not un-common for some dimension tables to have more than 50 attributes. Therefore we say that the dimension tables is wide. If you lay it out as a table with columns and rows, the table is spread out horizontally.

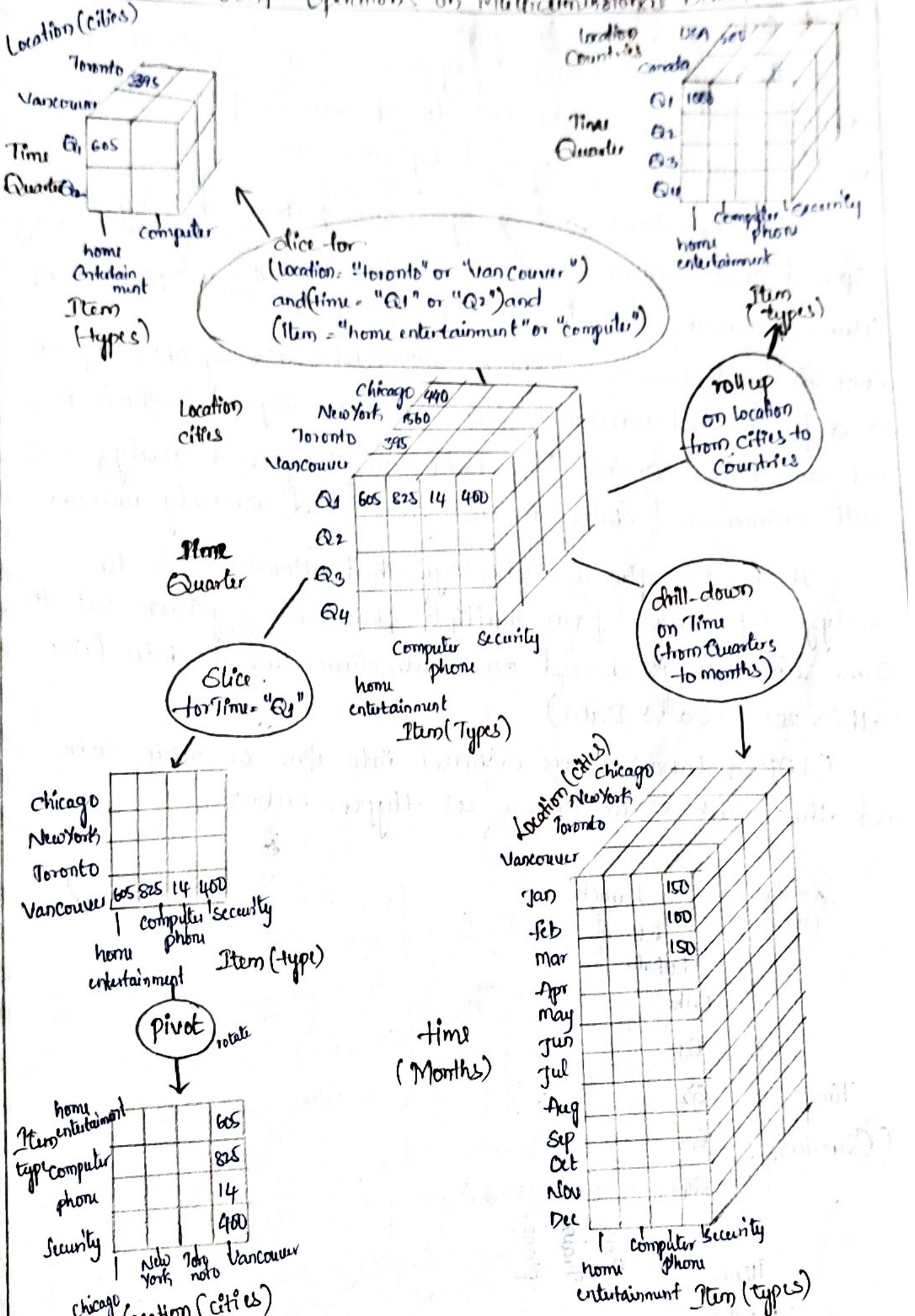
Dimension Table contain textual information that represents the attributes of the business.

It contain relatively static data.

Are joined to a fact table through a foreign key reference.

Fact Table	Dimension Table.
1) Fact table contains the measuring on the attributes of a dimension table.	1) It contains the attribute on that truth table calculate the metric.
2) There is less attributes than dimension table.	2) While in DT there is more attributes than fact-Table.
3) There are more records than dimension table.	3) There is less records than fact table.
4) Fact table forms a verfical Table	4) forms a horizontal Table.
5) In numerical format and text format	5) The attribute format of DT is in text format.
6) It comes after DT	6) It comes before fact Table.
7) fact Table is less than dimension Table in a Schema	7) The no: of dimension is more than fact table in a schema.
8) It is used for Analysis purpose and decision making	8) The main task of dimension table is to store the information about a business and its process.

OLAP Operations on Multidimensional Data



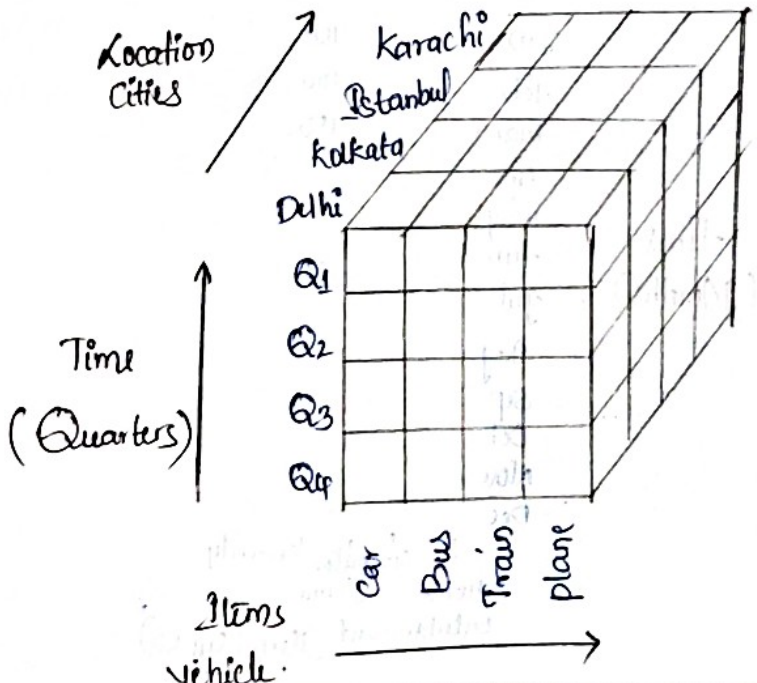
OLAP Cube :- Online Analytical Processing Server.

It is a data structure optimized for very Quick data analysis. The OLAP cube consists of numeric facts called measures which are categorized by dimensions.

Data Operations and analysis are performed using single Spread sheet, where data values are arranged in a row and Column format. This is ideal for 2-dimensional data. OLAP Contains multidimensional data, with data usually obtain from a different and unrelated source. Using a Spread sheet is not an optimal option. The cube can store and analyze multi-dimensional data in a logical and orderly manner.

It is a software technology that allows users to analyze information from multiple database systems at the same time. It is based on multi-dimensional data (Ex: Delhi → 2018 → Sales Data)

OLAP databases are divided into One or more cubes and these cubes are known as Hyper-cubes.

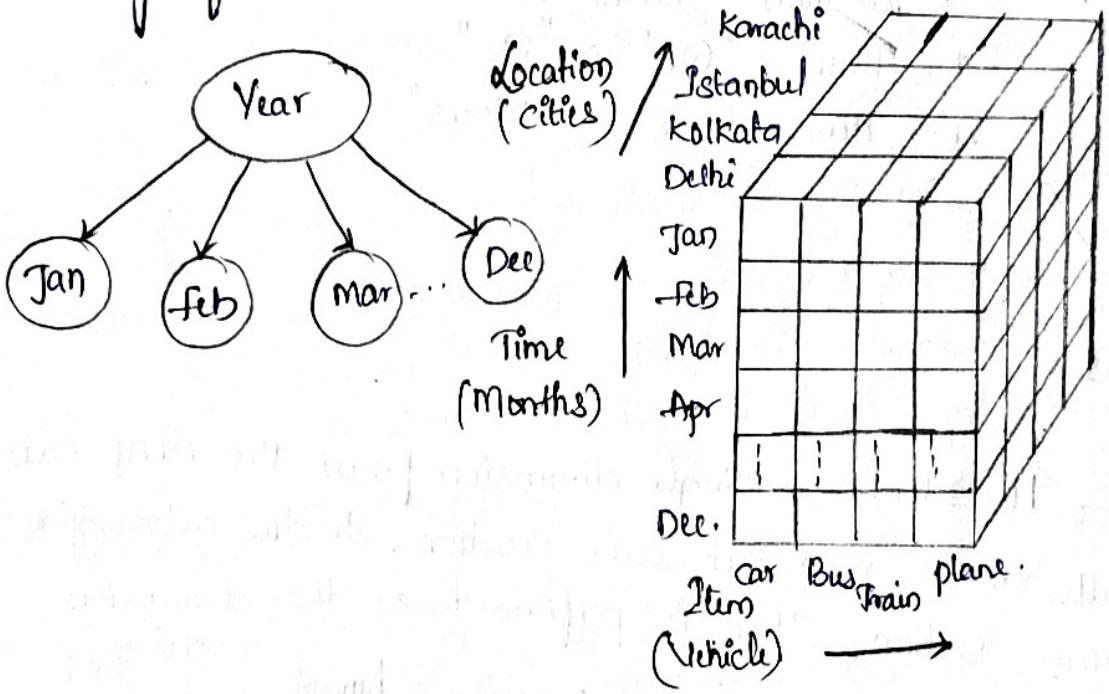


OLAP Operations:- There are five basic analytical operations that can be performed on an OLAP cube:

1. Drill Down:- In drill-down operation, the less detailed data is converted into highly detailed data is converted into highly detailed data. It can be done by:

* Moving down in the concept hierarchy
* Adding a new dimension.

In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of Time dimension (Quarter → Month)

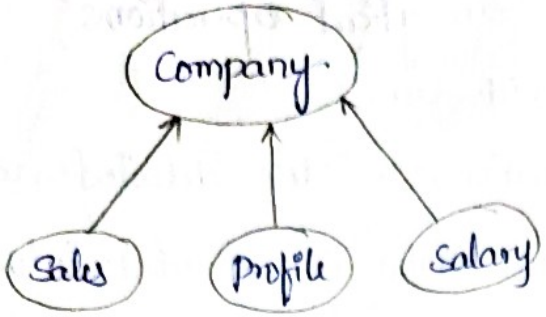


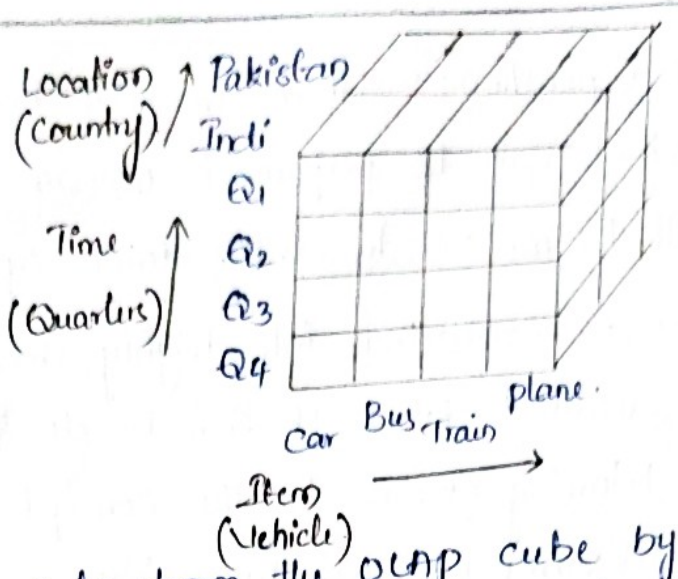
2. Roll-Up:-

It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:

* Climbing up in the concept Hierarchy.
* Reducing the dimensions

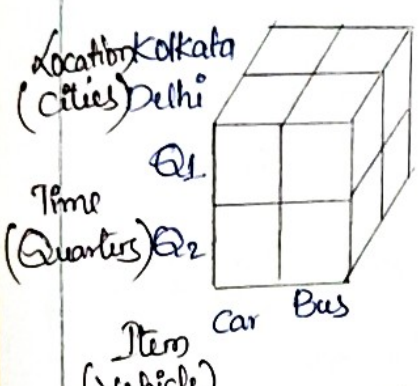
In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of location dimension (city → country)



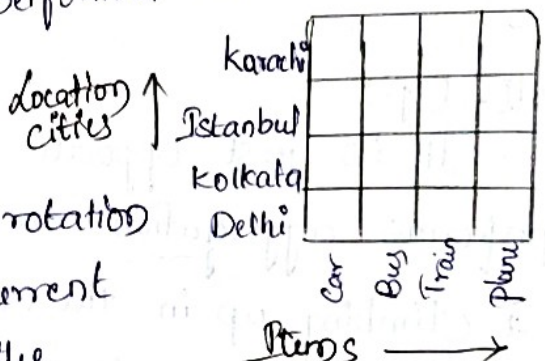


3. Dice :- It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:

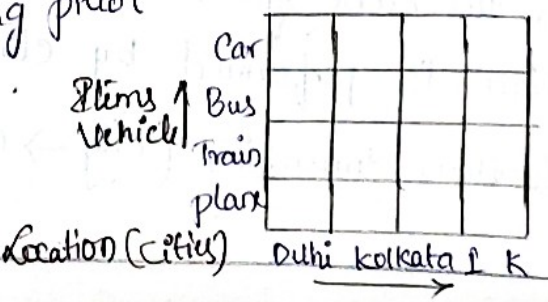
* Location = "Delhi" or "Kolkata"
* Time = "Q1" or "Q2"
* Item = "Car" or "Bus"



4. Slice :- It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = "Q1".



5. Pivot :- It is also known as rotation operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the Slice operation, performing pivot operation gives a new view of it.



OLAP Servers :- Online Analytical Processing (OLAP) refers to a set of software tools used for data analysis in order to make business decisions.

⇒ OLAP provides a platform for gaining insights from database retrieved from multiple database systems at the same time.

⇒ It is based on a multidimensional data model, which enables users to extract and view data from various perspectives.

⇒ A multidimensional database is used to store OLAP data. Many Business Intelligence (BI) applications rely on OLAP technology.

Types of OLAP Servers :- There 3-Major types of OLAP Servers are : 1) ROLAP [Relational Online Analytical Processing]
2) MOLAP [Multi-dimensional Online Analytical Processing]
3) HOLAP [Hybrid Online Analytical Processing].

1) Relational OLAP (ROLAP) :-

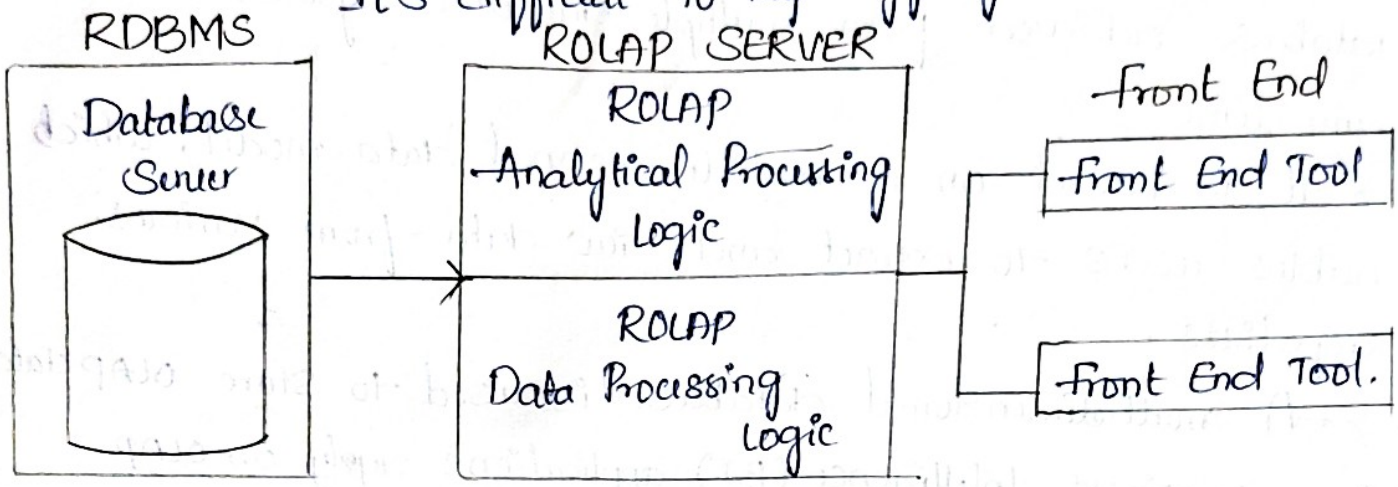
* Relational On-line Analytical Processing ROLAP is primarily used for data stored in a relational database, where both the base data and dimension tables are stored as relational tables.
* ROLAP Servers are used to bridge the gap between the relational back-end Server and the client's front-end tools.
* ROLAP Servers store and manage warehouse data using RDBMS, and OLAP middleware fills in the gaps.

Benefits :

It is compatible with data warehouses and OLTP systems.

* The data size limitation of ROLAP technology is determined by the underlying RDBMS. As a result, ROLAP does not limit the amount of data that can be stored.

Limitations :- SQL functionality is constrained.
It's difficult to keep aggregate table upto date.



Multi-Dimensional OLAP (MOLAP) :-

* Through array-based multi-dimensional storage engines, Multi-dimensional On-line Analytical Processing (MOLAP) supports multi-dimensional views of data.
* Storage utilization in multi-dimensional data stores may be low if the data set is sparse
* MOLAP stores data on discs in the form of a specialized multi-dimensional array structure. It is used for OLAP, which is based on the array's random access capability.
* Dimension instances determine array elements, and the data or measured value associated with each cell is typically stored in the corresponding array element.
* The multi-dimensional array is typically stored in MOLAP in a linear allocation based on nested traversal of the axes in some predetermined order.

However, unlike ROLAP, which stores only records with non-zero facts, all array elements are defined in MOLAP, and as a result, the arrays tend to be sparse, with empty elements occupying a large portion of them.

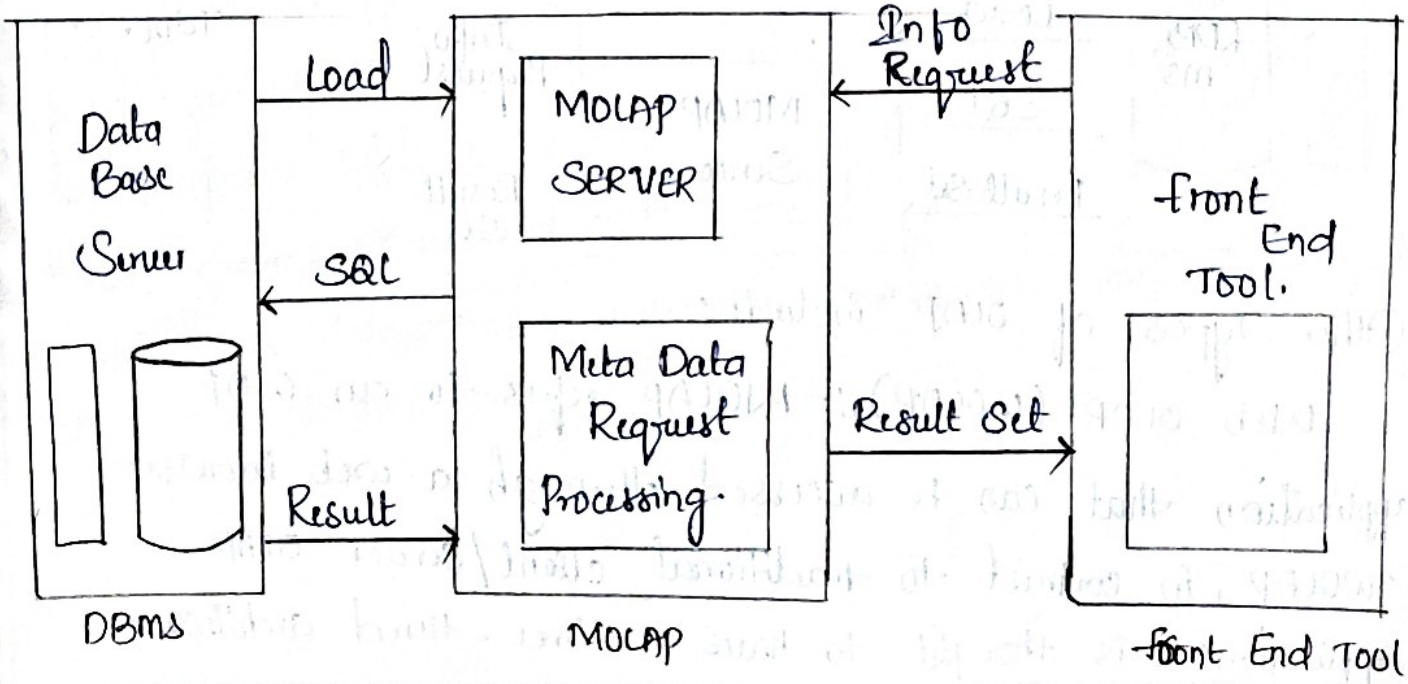
MOLAP systems typically include provisions such as advanced indexing and hashing to locate data while performing queries for handling sparse arrays, because both storage and retrieval costs are important when evaluating online performance.

MOLAP cubes are ideal for slicing and dicing data and can perform complex calculations. When the cube is created, all calculations are pre-generated.

Benefits :- * Suitable for slicing and dicing operations.
* Outperforms ROLAP when data is dense.
* Capable of performing complex calculations.

Limitations :- It is difficult to change the dimensions without re-aggregating.

Since all calculations are performed when the cube is built, a large amount of data cannot be stored in the cube itself.



Hybrid OLAP (HOLAP) :-

ROLAP and MOLAP are combined in Hybrid On-line Analytical Processing (HOLAP). HOLAP offers greater scalability than ROLAP and faster computation than MOLAP.

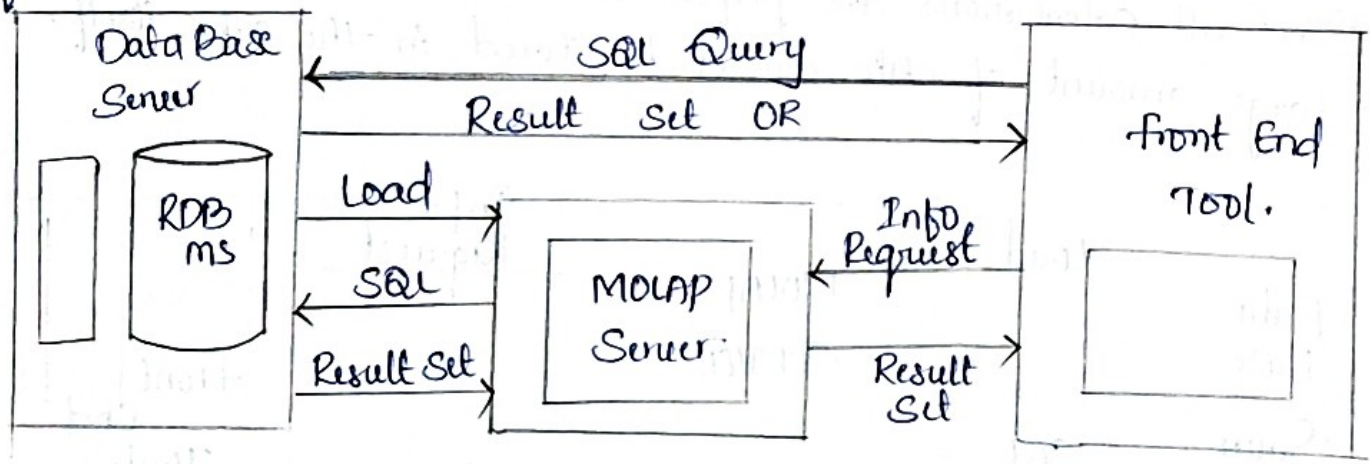
HOLAP is a hybrid of ROLAP and MOLAP. HOLAP servers are capable of storing large amounts of detailed data.

On the hand, HOLAP benefits from ROLAP's greater scalability. HOLAP, on the other hand, makes use of cube technology for faster performance and summary-type information, Because detailed data is stored in a relational database, cubes are smaller than MOLAP.

Benefits :* HOLAP combines the benefits of MOLAP and ROLAP.
* Provide quick access at all aggregation levels.

Limitations :- Because it supports both MOLAP and ROLAP servers HOLAP architecture is extremely complex.

There is a greater likelihood of overlap, particularly in their functionalities.



Other Types of OLAP include :-

Web OLAP (WOLAP) :- WOLAP refers to an OLAP application that can be accessed through a web browser.

WOLAP, in contrast to traditional client/server OLAP applications, is thought to have a three-tired architecture

Consisting of three components : a client, middleware, and a database server.

2) Desktop OLAP (DOLAP) :- In that case, the user can download the data from the source and work with it on their desktop or laptop. In comparison to other OLAP application functionality is limited. It is less expensive.

3) Mobile OLAP (MOLAP) :- Wireless functionality or mobile devices are examples of MOLAP. The users is working and accessing data via mobile devices.

4) Spatial OLAP (SOLAP) :- SOLAP egress combines the capabilities of geographic information System (GIS) and OLAP into a single user interface. SOLAP is created because the data can be alphanumeric, image, or vector. This allows for the quick and easy exploration of data stored in a spatial database.

Data-Warehouse Components and Architecture :-

We build a data-warehouse with Software and hardware components. to suit the requirements of our organizations we arrange these building we may want to boost up another part with extra tools and services.

Source Data Component :-

Source data coming into data warehouse may be grouped into four broad categories

UNIT - II

INTRODUCTION TO DATA - MINING

DATA - MINING :-

Data mining is a process of extracting useful information from large sets of data. It involves using various techniques from statistics, machine learning, and database systems to identify patterns, relationships, and trends in the data. This information can then be used to make data-driven decisions, solve business problems, and uncover hidden insights.

Applications of data mining include customer profiling and segmentation, market basket analysis, anomaly detection, and technologies are widely used in various industries, including finance, healthcare, retail, and telecommunications.

In general terms, "Mining" is the process of extraction of some valuable material from the earth ex:- coal mining, diamond mining, etc.

In the context of computer science, "Data Mining" can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

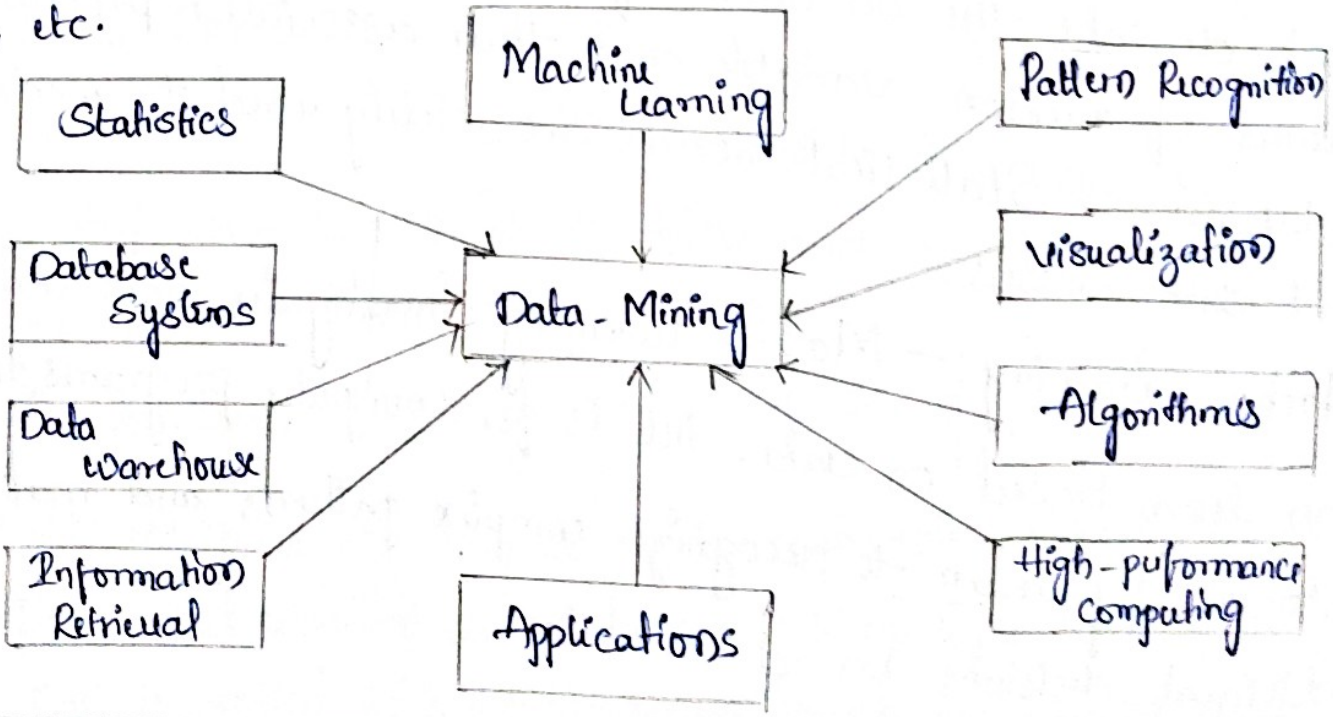
It is basically the process carried out for the extraction of useful information from a bulk of data or data-warehouses. One can see that the term itself is a little confusing. In the case of coal or diamond mining, the result of the extraction process is coal or diamond. In mining, the result of the

ex) x

But in the case of Data Mining, the result of the extraction process is not data!! Instead, data mining results are the patterns and knowledge that sense, we can think of Data Mining as a step in the process of knowledge Discovery or knowledge Extraction.

Gregory Piatetsky-Shapiro coined the term "knowledge Discovery in Databases" in 1989. However the term 'data mining' became more popular in the business and press communities. Currently, Data Mining and knowledge Discovery are used interchangeably.

Now-a-days, data mining is used in almost all places where a large amount of data is stored and processed. for ex:- Banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards personal loans, or insurance as well, Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.



Basically, Data Mining has been integrated with many techniques from other domains such as statistics, machine learning, pattern recognition, database and data-warehouse systems, information retrieval, visualization, etc. to gather more information about the data and to helps predict hidden patterns future trends, and behaviors and allows businesses to make decisions.

Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing summarizing it into meaningful information.

Data Mining can be applied to any type of data ex: Data-warehous Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-Series Databases, world wide web. Data Mining as a whole process

Statistics :- Statistics studies the collection, analysis, interpretations or explanation and presentation of data. Data Mining has an inherent connection with statistics.

A statistical model is a set of mathematical functions that describe the behaviour of the objects in a target class interms of random variable and their associated probability distributions. Statistical models are widely used to model data and data classes.

Machine Learning :- Machine learning investigates how computer can learn based on data. ML is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data.

a) Supervised learning :- Comes from the labeled examples in the training dataset.

b) Un-supervised learning :- Input examples are not class labelled. We may have clustering to discover classes with in the data.

c) Semi-Supervised learning :- using both labeled & un-labeled

d) Active learning :- The algorithm is designed in such a way that the desired output should be decided by the algorithm itself

Algorithm :- Can be written for any application in datamining

Application :- Can be built and used as a Technology.

Pattern Recognition :- By observing various patterns we can perform data mining.

Visualization :- To depending analysis of data by plotting into graph we do visualization

Database Technology :- Using database SQL we extract data from database.

DataMining on what kind of Data ?

1) Relational Database :-

A database system (DBMS) consists of a collection of inter-related data known as Database and a set of software programs to manage and access the data

A relational Database is a collection of tables, each of which is assigned a unique name each table consists of set of attributes and usually stores a large set of tuple (records or Rows) when DM is applied to RDBMS we can search for trends or data patterns

(AUTONOMOUS)

What is data Mining

Data Mining is the process of discovering useful patterns, trends and relationships from large sets of data using statistical, mathematical, and machine learning techniques. Its goals to turn raw data into meaningful information that supports better Decision-making.

Def: DM is the process of extracting hidden and valuable knowledge from large Databases.

Common Tasks in DM:-

1) Classification :- Assigning data to predefined categories
ex: Spam / Not Spam.

2) Clustering :- Grouping similar data without predefined labels

3) Association Rule Mining :- finding relationships b/w items (ex: Customer who buy bread also buy butter)

4) Prediction :- forecasting future outcomes based on historical data.

5) Anomaly Detection :- Identifying unusual patterns or outliers, such as fraudulent transactions.

Ex: A supermarket analyzes customer purchase records and discovers that customers who buy milk often also buy bread.
→ Place these items closer together.
→ Offer bundle discounts.
→ Improve marketing strategies

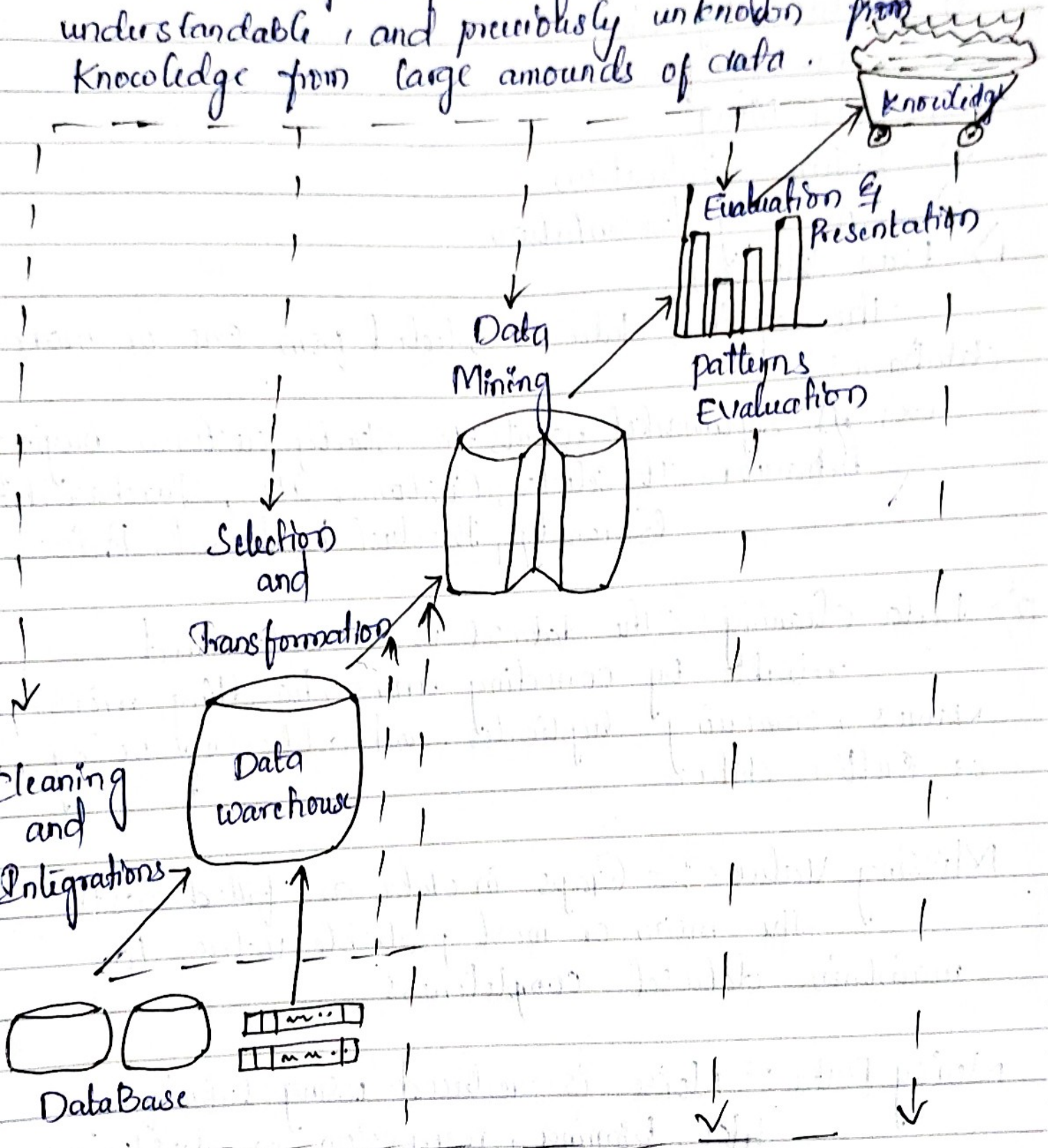
Applications of DM :-
1) Banking :- Fraud detection and credit risk analysis
2) Healthcare :- Disease prediction and patient diagnosis
3) Retail :- Customer segmentation and recommendation systems.
4) Education :- Student performance analysis.
5) Marketing :- Targeted advertising & customer behavior analysis

NOTE :- DM is the process of analyzing large Datasets to discover hidden patterns, relationships, and insights that help organization make informed decisions.
after 1000 records
Hidden Pattern :- Customer who buy bread often also buy butter. Data mining discovers this relationship. This is a hidden pattern

KDD Process in Data Mining:

↳ Knowledge Discovery in DataBase

It is the complete process of discovering useful, valid, understandable, and previously unknown from Knowledge from large amounds of data.



Steps of the KDD Process:

1) Data Selection
2) Data cleaning
3) Data Integration
4) Data Transformation
5) Data Ming
6) Pattern Evaluation
7) Knowledg Presentation

1) Data Selection:

The required data is selected from one or more databases.

Ex: A supermarket wants to study customer buying behavior. It selects: Customer ID, Purchase date, Quantity, Product Purchased, Price.

2) Data Cleaning:- The dataset is accurate and reliable by correcting errors, handling missing values, removing duplicates, and addressing noisy or outlier data.

Missing Values:- Gaps in data are filled with the mean or most probable value to maintain dataset completeness

Noisy Data:- Noise is reduced using techniques like binning, regression, or clustering to smooth or group the data.

Removing Duplicates:- Duplicate records are removed to maintain consistency and avoid errors in analysis.

Data cleaning in KDD to enhance the quality of the data and improve the effectiveness of data mining

Ex:-

Customer ID	Product	Quantity
C101	milk	2
C102	Bread	—
C103	milk	2

Data Integration :- Data from different sources is combined into one common dataset

Ex: Sales Database + Customer Database + product database + Integrated data.

DATA Transformation and Reduction :- It includes converting data into a format that is more suitable for analysis.

Normalization :- Scaling data to a common range for consistency across variables.

Discretization :- Converting continous data into discrete categories for simpler analysis

Data Aggregation :- Summarizing multiple data points (ex: averages or totals) to simplify analysis

Concept Hierarchy Generation :- Organizing data into hierarchies for a clearer, higher level view.

Data Reduction helps simplify the dataset while preserving key information.

Dimensionality Reduction :- (ex: PCA)
Reducing the no: of variables while keeping essential data

Numerosity Reduction :- Reducing data points using methods like sampling to maintain critical patterns.

Data Compression :- Compacting data for easier storage and processing

Together, these techniques ensure that the data is ready for deeper analysis and mining

DATA Mining :- The process of discovering valuable previously unknown patterns from large datasets through automatic or semi-automatic means dg It involves exploring vast/large amounts of data to extract useful information that can drive decision-making.

key characteristics of DM patterns include.

Validity :- Patterns that hold true even with new data!

Novelty :- Insights that are non-obvious and surprising

Usefulness :- Information that can be acted upon for practical Outcomes.

Understandability :- Patterns that are interpretable and meaningful to humans.

PATTERN Evaluation :- The discovered patterns are evaluated to determine whether they are useful, Intersting, valid, Meaningful

Ex :- Suppose the system discovers
→ 80% of customers who buy bread also buy butter
This may be considered a strong and useful pattern
However, a pattern such as :
→ Only 1 customer bought apples and Oranges together may not be useful because it is based on very little data.

Knowledge Presentation :-
The final knowledge is presented to users in an understandable form.
This may include :- Reports, Charts, Graphs, Tables, Dashboards, Rules.

ex:- Bread ⟶ Butter
strong buying relationship.

Major Challenges in Data Mining
The major issues can be in mining methodology, user interaction, performance/scalability and data types.

1) Loco-Quality Data :- Many of the values within the database can be noisy, incomplete, incorrect, inaccurate, or unreliable. It can also be a poor representation of the population.

2) Complex And Huge Data :- Processing large, complex and unstructured data into a structured format can take alot of time and computational cost.

3) Distributed Data :- Data is not always available in one place. It can be very challenging to bring all the data from heterogeneous sources to a centralized repository so tools and algorithms that enable the mining of distributed data in high demand.

4) Domain knowledge :- A detector that tells us where to dig. Without background knowledge it is almost impossible to get interesting information from an enormous amount of data.

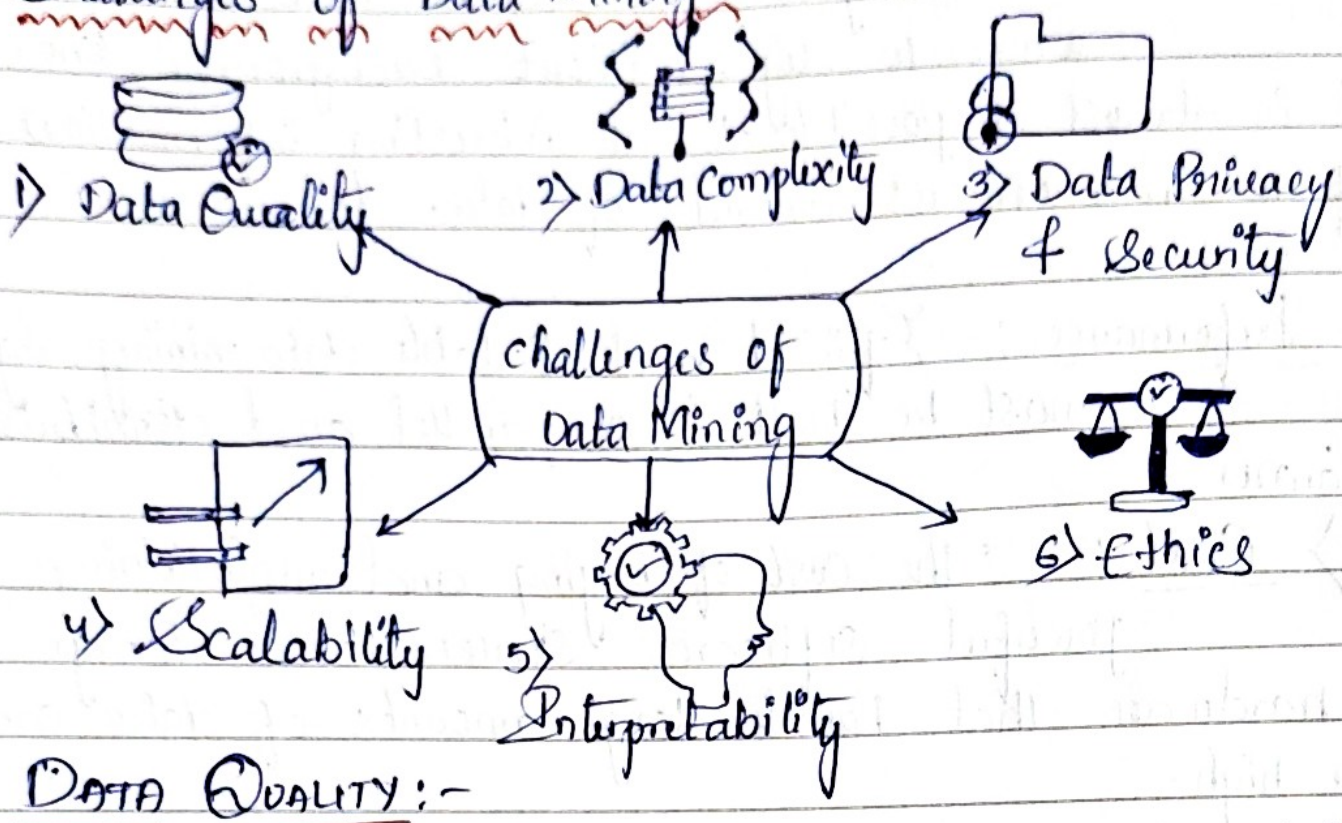
5) Performance :- Efficient and scalable data mining algo must be built in a parallel and distributed manner.

6) Cost :- The cost of buying and maintaining powerful software, servers, and storage hardware that handle large amounts of data can be high.

7) Security and Privacy Concerns :- Individuals, organizations, and governments are concerned about data security and privacy. Ex: If you want to use student data for an education project, you can have a lot of difficulties in getting this kind of data because of privacy issues.

(AUTONOMOUS)

Challenges of Data Mining



1) **DATA QUALITY:-**

High-quality data is the foundation of successful DM - but real-world data is rarely perfect

Issues Include:- Missing values
Duplicate Records
Incorrect or Out dated Information
Inconsistent formats (ex: India, IND, IN

Because DM alg rely heavily on input data, poor-quality data leads to misleading or completely wrong results.

Q: Why does Data Quality Suffer?

→ Human errors during data entry
→ System integration problems
→ Faulty Sensors
→ Data stored in incompatible formats.

Q:- How it is handled ?

1) Data cleaning :- fixing or removing errors
2) Data preprocessing :- Transforming data into a usable form

2) DATA COMPLEXITY :-

Today's data is not massive - It's also messy, multi-structured and constantly changing.

We deal with : Social media posts
IoT sensor readings
Images, videos, logs
Textual and Transactional data

These are often stored in different formats and from different sources, making integration extremely difficult.

Q:- How Experts deal with this ?

They use advanced DM technique :- Clustering
Classification
Neural Networks
Association rule Mining

These helps uncover hidden patterns even in high-dimensional data.

3) DATA PRIVACY & SECURITY :- As organizations collect more data, protecting that data becomes a bigger challenge

Many datasets contain :- Personal details
Health Information
Financial Records
Behavioral Data

Leakage of such data can damage user trust and violate laws like GDPR, CCPA, HIPAA etc.

Security Risks in DM : ⇒ Cyber-attacks
→ Insider - threads
→ Unauthorized sharing
→ Weak - access controls

How Privacy is Protected
Data anonymization :- Removing identifiable info

Data Encryption :- Securing data using Cryptographic techniques

Access controls :- Allowing only authorized users to view sensitive data.

> SCALABILITY :- Modern datasets can reach terabytes or petabytes - far beyond what a single computer can handle.

Data mining systems must :
→ Process huge datasets
→ Scale as data grows
→ Work in real-time for streaming data
eg : Stock prices, IOT

How scalability problems are solved : Using distributed computing frameworks like.

→ Apache Hadoop , Apache Spark
These systems split data across multiple machines so tasks can run in parallel.

5) Interpretability :- A big challenge in DM is that many models behave like blackboxes

Ex: * Deep learning models with millions of parameters
* Complex Decision Tree
* Ensemble models
They may produce accurate predictions, But explaining why they reached a conclusion is often difficult.

Why interpretability matters :-
→ Businesses need understandable insights
→ Health and finance require transparency
→ Regulators demand explanations

How experts Improve Interpretability
→ Visualization
→ Feature importance charts
→ Model explanation tools like LIME / SHAP

6) ETHICS :- With great analytical power comes great responsibility

DM can unintentionally :

→ Reinforce social biases
→ Enable discrimination
→ Violate user privacy
→ Be used to monitor or manipulate users

Unethical or careless data usage can lead to serious consequences.

Ethical challenges include :→ fairness
→ Transparency
→ Informed consent

→ Bias in algorithms → Responsible use of insights

(AUTONOMOUS)

TASKS AND FUNCTIONALITIES OF DATA MINING

DM functions are used to define the trends or correlations contained in DM activities.

TASK IN DATA MINING

I. Descriptive DM	II. Predictive DM
1) Cluster Analysis	1) Decision trees
2) Association rule mining	2) Neural networks
3) Visualization	3) Regression Analysis

I. Descriptive Data Mining :-

Finding patterns and relationships in the data that can provide insights into the underlying structure of data. Descriptive DM is often used to summarize or explore the data, and it can be used to answer questions such as: 1) What are the most common patterns or relationships in the data ? 2) Are there any clusters or groups of data points that share common characteristics ?

3) What are the Outliers in the data, and what do they represent ?

Cluster Analysis :- This technique is used to identify groups of data points that share similar characteristics.

Clustering can be used for segmentation, anomaly detection and summarization.

2) Association Rule Mining :- Used to identify relationships b/w variables in the data, It can be used to discover co-occurring events or to identify patterns in transaction data.

3) Visualization :- This technique is used to represent the data in a visual format that can help users to identify patterns or trends that may not be apparent in the raw data.

II. Predictive DM :- Is concerned with developing models that can predict future behavior or outcomes based on historical data, PDM is used for classification or regression tasks. and it can be used to answer questions such as:

1) What is the likehood that a customer will churn?
2) What is the expected revenue for a new product launch?
3) What is the probability of a loan defaulting?

1) Decision Tree :- This technique is used to create a model that can predict the value of a target variable based on the values of several input variable. DT are used for classification Tasks

2) Neural Networks :- Used to create a model that can learn to recognize patterns in the data. NN used for image recognition, Speech recognition, and Natural language processing (NLP)

3) Regression Analysis :— Used to create a model that can predict the value of a target variable based on the values of several input variable. RA is used for prediction tasks.

Both Descriptive and Predictive Data Mining techniques are important :- for gaining and making better decision. Descriptive DM can used to explore the data and identify patterns, while predictive DM can used to make predictions based on those patterns.
Together, help organizations to understand their data and make informed decisions based on that understanding.

DATA MINING FUNCTIONALITY :-

1) Class /concept Descriptions :- It can be helpful to define individual groups and concepts
   ↳ Data characterization :- The output of DC can be presented in forms of pie charts, bar charts, curves, multi-dimensional Data cubes
   ↳ Data Discrimination :— It comparison of the general feature of the target class data object against the general features of objects from one or multiple contrasting classes.

2) Mining frequent Patterns, Associations and Correlations :- Frequent patterns are nothing but things that are found to be most common in the data.

↳ frequent itemset :- This applies to a no: of items that can be seen together regularly
Ex:- milk and sugar.

↳ frequent Subsequence :- This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.

↳ frequent Substructure :- It refers to the different kinds of data structure such as trees and graphs that may be combined with the itemset or subsequence.

1 Association Analysis :- The process involves uncovering the relationship b/w data and deciding the rules of the association.
It is a way of discovering the relationship b/w various items.

DATA MINING TASK PRIMITIVES

Refa to the basic building blocks or components that are used to construct a data mining procss.

These primitives are used to represent the most common and fundamental tasks that are performed during the DM process.

The use of DM task primitive can provide a modulars, and reusable approach, which can improve the performance, efficiency, and understandability of the DM process.



Task relevant Data
1) Database Name
2) Database Tables
3) Relevant attributes
4) Data grouping criteria



Type of knowledge to be mined
1) classification
2) clustering
3) Prediction
4) Discrimination
5) Correlation Analysis



Background Knowledge.
1) concept Hierarchy
2) User beliefs about relationships in data.



Measures of Patterns
1) Simplicity
2) Novelty
3) Certainty
4) Utility



Visualization of patterns
1) Visualization of discovered patterns
2) Cubes
3) charts
4) Tables
5) Graphs

1) The Set of Task relevant data to be Mined :—

→ It refers to specific data that is relevant and necessary for a particular task or analysis being conducted using DM techniques.

→ This data may include specific attributes, variables or characteristics that are relevant to the task at hand, such as customer demographics, sales data, or website usage statistics.

→ The data selected for mining is typically a subset of the overall data available, as not all data may be necessary or relevant for the tasks.

Ex:- Extracting the database name, DB Tables and relevant required attributes from the dataset from the provided input DB.

2) Kind of Knowledge to be mined :—

→ It refers to the types of information or insights that are being sought through the use of DM Tech

→ This describes the data mining tasks that must be carried out.

→ It includes various tasks such as classification, clustering, discrimination, characterization, association and evolution analysis.

Ex:- It determine the tasks to be performed on the relevant data in order to mine useful information such as classification, clustering, prediction, discrimination, outlier detection, and correlation analysis.

3) Background knowledge - to be used in the discovery process :-

→ It refers to any prior information or understanding that is used to guide the DM process.

→ This include domain-specific knowledge, such as industry-specific terminology, trends, or best practices, as well as knowledge about the data itself.

→ The use of background knowledge can help to improve the accuracy and relevance of the insights obtained from the dm process

Ex:- The use of background knowledge such as concept hierarchies, and user beliefs about relationships in data order to evaluate and perform more efficiently

4) Interestingness Measures and Thresholds for pattern Evaluation :-

→ It refers to the methods and criteria used to evaluate the quality and relevance of the patterns or insights discovered through DM,

→ Interestingness measures are used to quantify the degree to which a pattern is considered to be interesting or relevant based on certain criteria, such as its frequency, confidence, lift.

→ These measures are used to identify patterns that are meaningful or relevant to the tasks.

→ Thresholds for pattern evaluation, on the other hand, are used to set a minimum level of interestingness

that a pattern must meet in order to be considered for further analysis or action.

Ex:- Evaluating the interestingness measures such as utility, certainty, and novelty for the data and setting an appropriate threshold value for the pattern evaluation.

5) Representation for visualizing the discovered pattern:-
It refers to the methods used to represent the patterns or insights discovered through data mining in a way that is easy to understand and interpret. Visualization techniques such as charts, graphs and maps are commonly used to represent the data and can help to highlight important trends, patterns or relationships within the data.
Visualizing the discovered pattern helps to make the insights obtained from the data mining process more accessible and understandable to a wider audience, including non-technical stakeholders.

Ex:- Presentation and visualization of discovered pattern data using various visualization techniques such as boxplot, charts, graphs, tables, etc.

Advantages of Data Mining Task Primitives

1) Modularity:— Data mining task primitives provide a modular approach to DM, which allows for flexibility and the ability to

easily modify or replace specific steps in the process.

2) Reusability :- DM task primitives can be used across different DM projects, which can save time and effort.

3) Standardization :- Data Mining task primitives provide a standardized approach to DM, which can improve the consistency and quality of the DM process.

4) Understandability :- DM task primitives are easy to understand and communicate, which can improve collaboration and communication among team members

5) Improved Performance :- DM task primitives can improve the performance of the DM process by reducing the amount of Data that needs to be processed, and by optimizing the data for specific DM algorithms.

6) Flexibility :- DM task primitives can be combined and repeated in various ways to achieve the goals of the ~~data~~ DM process, making it more adaptable to the specific needs of the project.

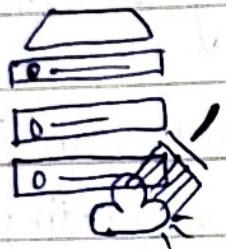
7) Efficient use of resources :- Data mining task primitives can help to make more efficient use of resources, as they allow the right tools, avoiding unnecessary steps and reducing the time and computational power needed.

DATA PREPROCESSING IN DATA MINING

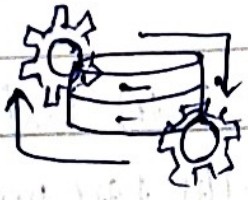
Real-world data is often incomplete, noisy and inconsistent, which can lead to incorrect results if used directly. Data preprocessing in DM is the process of cleaning and preparing raw data so it can be used effectively for analysis and model building.

→ Real data contains missing and incorrect values.
→ Data may come from multiple sources
→ Large data sets often have irrelevant information
→ Clean data gives better mining results

Data Preprocessing.



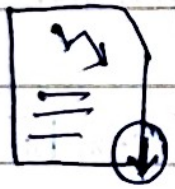
Data cleaning



Data Integration



Data Transformation



Data Reduction.

DATA Cleaning :— It is the process of identifying and correcting errors or inconsistencies in dataset.

- Handling missing values
- Removing duplicate records
- Correcting wrong or inconsistent data
- Handling Outliers

(AUTONOMOUS)

Data cleaning :-

1) Missing data
   - Ignore the tuple
   - fill missing value
2) Noisy data
   - Binning Method
   - Regression
   - clustering

Inconsistent data :- The data which does not match with existing data

Noisy data :- which does not belong to the given range

Missing data :- Data is missing

Binning Method :- To make the analysis easier and continous values will be created as a single value

3 types methods :- Smoothing data by Mean
Smoothing data by Median
Smoothing data by Boundaries (Min & Max

CGPA Data :- 3.5, 3.2, 3.0, 3.4, 3.7, 3.4, 3.4, 4.0, 3.5

Sorted :- 3.0, 3.2, 3.4, 3.4, 3.4, 3.5, 3.5, 3.7, 4.0

Creating bins and labeling

Bin low :- 3.2, 3.2, 3.2

Bin median :- 3.4, 3.4, 3.4

Bin high :- 3.7, 3.7, 3.7

Bin low :- 3.0, 3.2, 3.4
Bin med :- 3.4, 3.4, 3.5
Bin high :- 3.5, 3.7, 4.0

Boundaries (min & max)

Bin low - 3.0, 3.0, 3.4
Bin med - 3.4, 3.4, 3.5
Bin High - 3.5, 3.5, 4.0

If there are unequal values (odd numbers) then we use another method.

Ex:- 3.0, 3.2, 3.4, 3.4, 3.5, 3.7, 4.0

$$\frac{\text{Max} - \text{Min}}{\text{No : of Bins}} = \frac{4.0 - 3.0}{3} = \frac{1}{3} = 0.3$$

Bin-1 low :- 3.0
$+0.3$
$3.3 \Rightarrow 3.0, 3.2.$

Bin-2 Medium :- 3.4
$+0.3$
$3.7 \Rightarrow 3.4, 3.4, 3.5$

Bin-3 High :- 3.7
$+0.3$
$4.0 \Rightarrow 3.7, 4.0$

Techniques used:

Mean Imputation :— Replaces missing values with the average of the attribute

Median Imputation :— Replaces missing values with the middle value; useful when outliers exist.

Mode Imputation :— Replaces missing values with the most frequent value.

Deletion Method :— Removes records that contains missing values

Interquartile Range (IQR) :— Detects outliers using the range b/w Q1 and Q3

Z-Score Method :— Identifies outliers based on standard deviation from the mean.

Binning :— Smooths noisy data by grouping values into bins

Regression Smoothing :— Uses regression to predict and smooth noisy values

Duplicate Detection :— Identifies and removes repeated records.

Ex:— Replacing missing age value with the average age

Removing repeated rows in a datasets.

II. DATA INTEGRATION :—

It involves merging data from various sources into a single, unified dataset. It can be challenging due to differences in data formats, structures, meanings.

→ Used when data comes from database, files, APIs
→ Removes redundancy b/w datasets
→ Resolves conflicts in data values

Techniques used :-

Schema Matching :- Aligns attributes from different data sources

Entity Resolution :- Identifies records that refers to the same real-world entity.

Correlation Analysis :- Finds and removes redundant attributes

Data Conflict Resolution :- Resolves inconsistencies in units or data values

Duplicate Elimination :- Removes overlapping records after integration

Ex:- Merging customer data from sales and marketing database

III : DATA TRANSFORMATIONS :-

Data transformation converts data into a suitable form so that data mining alg can work effectively

Bring data into a common format
Improve mining efficiency
Make data suitable for modeling

Techniques used :-

Min-Max Normalization :- Scales data into a fixed range, usually 0 to 1.

Z-Score Normalization :- Transforms data using mean and standard deviation.

Decimal Scaling :- Normalizes data by moving the decimal point

Log Transformation :- Reduces data skewness using logarithmic scaling.

One-Hot Encoding :- Converts categories into binary columns

Label Encoding :- Assigns numeric labels to categorical values

Aggregation :- Combines detailed data into summarized form.

Ex :- Converting salary values into a fixed range (0-1)
Changing text labels like Male/female into numeric values

IV. DATA REDUCTION :-

It reduces the datasets' size while maintaining key information. This can be done through feature selection which chooses the most relevant features and features extraction which transforms the data into a lower-dimensional space while preserving important details

→ Improves processing speed
→ Saves storage space
→ Makes analysis easier

Techniques Used:-

Principal Component Analysis (PCA):-
Reduces dimensions by projecting data onto principal components

Linear Discriminant Analysis (LDA):-
Reduces dimensions while maximizing class separation

filter Methods:- Select features based on statistical measures

Wrapper Methods:- Select features using model performance

Embedded Methods:- Perform feature selection during model training

Simple Random Sampling:- Selects data points randomly from the dataset

Stratified Sampling:- Samples data proportionally from each class.

Benefits of Data Preprocessing:-
Improves data quality
Increases accuracy of mining results
Reduces errors in models
Makes data easier to understand

Advantages:-

1) Improved Data Quality:- Ensures data is clean, consistent, and reliable for analysis

(AUTONOMOUS)

2) Better Model Performance :— Reduces noise and irrelevant data, leading to more accurate predictions and insights.

3) Efficient Data-Analysis :— Streamlines data for faster and easier processing

4) Enhanced Decision-Making :— Provides clear and well-organized data for better business decisions.

Data preprocessing is used to transform the raw data in a useful and efficient format.

(Data Re-processing)

- Data cleaning
  - → Missing Data
    1. Ignore the Tuple
    2. fill the Missing values (manually, by mean or by most probable value)
  - → Noisy Data.
    1. Binning Method
    2. Regression
    3. Clustering
- Data Transformation
  - → Normalization
  - → Attribute Selection
  - → Discretization
  - → Concept Hierarchy Generation
- Data Reduction
  - → Data Cube Aggregation
  - → Attribute subset Selection
  - → Numerosity Reduction
  - → Dimensionality Reduction

Steps Involved in Data preprocessing:

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.

a) Missing Data :- This situation arises when some data is missing in the data. It can be handled in various ways.

Data cleaning :-

Processing of removal of incorrect, incomplete, inaccurate data, and replaces missing data.

OR

Data cleaning is to improve the quality of data by filling the missing values, reconfiguring the data to make data in consistent format.



a) Missing Data :- This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are :

1. Ignore the tuples :- This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

2. 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.

Methods for filling Missing values :

→ If any attribute value is mistakenly not recorded, then it will lead to inaccurate results.

→ place of missing values, we can replace with "NA" Not applicable.

* Mean value.
* Median value. (unknown or $-\infty$)

Mean value :- Estimating avg of attribute that contains missing values and replacing those values with avg value.

In student DB, missing values in marks attribute can be filled by calculating the avg marks of all student. i.e., all missing value tuples will contain identical mean value.

In median all tuples with missing value contains identical constant, but when data analysis is being performed these value are treated as special value by the data mining process, results. inaccurate results.

Sometimes replace with most-probable/most occuring value in DB

Missing values can be filled in 2 ways.

manual
↓
small DB

Automatic
more efficient for large data sets.

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 handle in following ways:

1. 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 tasks.

Noisy Data / Variance Error / Variance in measured values:

If we check our weight multiple times at some time, it will change like 60.90, 60.99, 60.89 → Same we but diff values.

Ex: If we try to check for weight of a person, it will give different weight at each time (on same day) i.e., called Noisy data.

→ To remove noisy data, Binning method is used.

Binning :- If we take age column.

20
21
22
30
31
32

we create bins

20	30
21	31
22	32

Label the bins.

Now, Replace bins with the label of the Bin

20
20
20
30
30
30

* Here we don't loose the data, because our survey results will be like > 50 age are likely to fall diabetic.

Same kind of * 30 are eligible for job.

Data are kept * So labels all used for queries.

is Bins

Handling Noisy Data.

↓

Error data, Random error (or) Variance in a measured value.

Methods to Handle Noisy Data.

1) Smoothing by bin Mean
2) Smoothing by bin Median
3) Smoothing by bin Boundary.

- 1st data is sorted along with the error value.
- 2nd Sorted data is stored in bins or buckets (with equivalent width & frequency).

1) Smoothing by Bin Mean :-

→ Calculate average of data values of individual buckets.
Substitute existing value with avg value.

Ex1:- BIN A : 2, 3, 4, 5 value in BIN

Avg of $\frac{2+3+4+5}{4} = 3.5$

Replace with avg value.

BIN A : 3.5, 3.5 3.5 3.5.

6, 5, 10, 7, 9, 5, 8, 4

Ex-2:- 2, 3, 5, 6, 8, 9, 9, 10

Sort data in buckets or bins with equal bandwidth.

BIN-A : 2, 3, 5, 6 avg 4

BIN-B : 8, 9, 9, 10 avg 9

$\frac{3+5}{2} = \frac{8}{2} = 4$

Replace buckets with avg value.

BIN-A : 4 4 4 4

BIN-B :- 9 9 9 9

Smoothing by Bin Median :-

→ Arrange data in ascending or des order and then take middle no: or medium no:

Ex :- 1, 2, 3, 4, 5

BIN-A : 3 3 3 3 3 Replaced with middle value.

Smoothing by Bin Boundary :-

* Highest & lowest values in individual bucket are considered values are treated as boundaries of a bucket.

* Each bin value is then replaced by the closed boundary value.

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

→ Sort the given data.

BIN-A : 4, 8, 9, 15
BIN-B : 21, 21, 24, 25
BIN-C : 26, 28, 29, 34

} Divides into N intervales, each containing same no: of samples.

Smoothing by BIN Means

Avg
BIN-A : 9, 9, 9, 9
BIN-B : 23, 23, 23, 23
BIN-C : 29, 29, 29, 29.

BIN median : Middle value.

BIN-A : 9, 9, 9, 9
BIN-B : 24 24 24 24
BIN-C : 29 29 29 29

BIN Boundary : Substituted with closest Min or Max value.

BIN-A : 4 4 4 15
BIN-B : 21 21 25 25
BIN-C : 26 26 26 34

Boundary value remains same. (min & max)

2) 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.)

Ex:→ For a building, based on sq.ft rooms etc, price is predicted (Numeric prediction)

→ Sales person will give no: of prediction of sales before he starts the day.

Clustering :- Similar data items are grouped at one place.

Dissimilar items - Outside the cluster

Theown out of the cluster dissimilar items are error items

3) Clustering :- This approach group the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

II. Data Transformation :-

This step is taken in order to transform the data in appropriate forms suitable for mining process.

1) Normalization :- It is done in order to scale the data values in a specified range (-1.0 to 1.0. or 0.0 to 1.0)

2) Attribute Selection :- In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3) Discretization :- This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4) Concept Hierarchy Generation :- Here attribute by are converted from lower level to higher level in hierarchy. for ex: The attribute City can be converted to Country.

Data Transformation :-

The data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

1) Smoothing - To remove noisy data (Binning, Regression, clustering Previous topics)
2) Attribute Construction
3) Aggregation
4) Discretization (Explained in DR)
5) Concept Hierarchy Generation
6) Normalization.

2) Attribute Construction :- New attribute are created from an existing set of attribute.

C-name C-ID address

→ These attribute are used to construct another DB set that contains information about the employee who joined in Yr 2019

→ This method of reconstruction make mining more efficient and helps in creating new data sets quickly

3) Aggregation :-

Summary and Aggregation function can be applied on the data for constructing data cube for data analysis.

Explained, in data redundency.

6) Normalization :- Normalization in ML is the process of translating data into the range [0,1] or any other range or simply transforming data into unit sphere (closed set of pt equal

→ Normalization and Standardization change the co-ordinate syste so that all variables have the same scale.

Benefites of Normalization :-

→ Whenever you are dealing with features (or) parameters that different from each other in terms of range of values (which is the case in most of data sets) then you have to normalize the data so, that the difference in there range of values doesn't affec your outcome.

Ex :- You want to asse how speed of running of an attribute depends on his weight & height.

Height ranges from 4 - 7 feets & diff scales.
weight ranges from 50kgs - 120 kgs.

→ If we feed these factors in the model, Height will less values
weight will large values
→ This may not give us desired result.
→ To overcome, rescale or standardise the 2 features so that their range is comparable and new values will now be fed to the model.

Ex:- Calculating CGPA for marks and SGPA (ranging all students marks on CGPA scale 1–10)

ID	Gendu	age	Salary
1	f	27	19000
2	m	51	64
3	m	52	100000
4	f	33	55000
5	m	45	45000

Min – Max Normalization :-

Max distance for Salary 100000 – 19000 = 81000

Max distance for age : 52 – 27 = 25.

-Here age MD and sales are 25, 81000 low to high range, which didn't give correct result while analysing.

So, make them to fall in same scale range (0 – 1)

New Min for age & Salary = 0

New max for age & Salary = 1

$$V_i' = \frac{V_i - \min x_i}{\max x_i - \min x_i} (\text{New_Max}_A - \text{New_Min}_A) + \text{New_Min}_A$$

$$V_1 = \frac{27 - 27}{52 - 27} \times (1 - 0) + 0$$

$$= 0.00.$$

$$V_1 = \frac{51 - 27}{52 - 27} \times (1 - 0) + 0$$

$$= 0.96.$$

After normalization age & Sal - they fall under same scale.

ID	Gender	age	Sal.
1	1	0.80	800
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32

Definition:- Normalization is a technique often applied as part of data preparation for Data Mining.

The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

Min-Max Normalization: is one of the most common ways to normalize data.

For every feature, the minimum value of that feature gets transformed into a 0 and the maximum value of, transformed into 1 and every other value gets transformed into a decimal value b/w 0 & 1

Ex: Use the min-max normalization method to normalize the data. 200 (Min), 300, 400, 600 and 1000 (max). by setting min = 0. & max = 1.

$$X_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \times (New_Max - New_Min) + New_Min$$

$$= \frac{x - 200}{1000 - 200} \times (1 - 0) + 0)$$

$$= \frac{200 - 200}{1\,800} \times 1$$

$$= 0.$$

Original data	200	300	400	600	1000
[0,1] normalized	0	0.125	0.25	0.5	1.

Ex-②:- Given Data min = 1000, max = 9000

$$V = \frac{x - min}{max - min}$$

$$V = \frac{1000 - 1000}{9000 - 1000} = 0$$

$$V = \frac{2000 - 1000}{9000 - 1000} = 0.125$$

$$V = \frac{3000 - 1000}{9000 - 1000} = 0.25, \quad V = \frac{5000 - 1000}{9000 - 1000} = 0.5, \quad V = \frac{9000 - 1000}{9000 - 1000} = 1$$

Data (v)	Normalized Data (v)
1000	0
2000	0.125
3000	0.25
5000	0.5
9000	1

Z-Score Normalization :- $V' = \frac{V - A}{\sigma A}$

→ Is to understand how far the data pt is from the mean. Z-score is useful to those kinds of data analysis where in there is a need to compare value mean (avg) such as result from tests or surveys.

Ex:- A person wt is 150 pounds Now, if there is a need to compare that value with the avg wt of the population listed in a vast table of data.

→ Z-score normalization is needed to study such values especially if someone's wt. is recorded in kgs.

→ As data comes from diff sources its very common to have diff attribute in any batch of data there normalization in data missing is like pre-processing & preparating the data for analysis.

Marks – 8, 10, 15, 20

$$\text{Mean value} = \frac{8 + 10 + 15 + 20}{4} = 13.25$$

$$\text{Standard deviation} = \sqrt{\frac{\Sigma (\text{Individual marks} - \text{mean})^2}{n}}$$

$$= \sqrt{\frac{(8 - 13.25)^2 + (10 - 13.25)^2 + (15 - 13.25)^2 + (20 - 13.25)^2}{4}}$$

$$= 5.31$$

Z-Score Normalization of 8, 10, 15, 20.

$$8 = \frac{8 - 13.25}{5.37} = -0.97$$

$$10 = \frac{10 - 13.25}{5.37} = 0.60$$

$$15 = \frac{15 - 13.25}{5.37} = 0.32$$

$$20 = \frac{20 - 13.25}{5.37} = 1.4$$

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

v = Original value.
$\bar{A}$ = Mean value.
σ_A = Standard deviation value.

Ex-2: $Z = \frac{x - \mu}{\sigma}$ μ = Mean, σ = Standard deviation

Data(v)	Z-score Data
200	-1.06
300	-0.707
400	-0.354
600	0.354
1000	1.77

$$\text{Mean} = \frac{200 + 300 + 400 + 600 + 1000}{5} = 500$$

$$\text{Standard deviation} = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

$$= \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}}$$

$= 282.8$

$$Z = \frac{x - \mu}{\sigma}$$

$$v = \frac{200 - 500}{282.8} = -1.06$$

$$v = \frac{300 - 500}{282.8} = -0.707$$

$$v = \frac{400 - 500}{282.8} = -0.354$$

$$v = \frac{600 - 500}{282.8} = 0.354$$

$$v = \frac{1000 - 500}{282.8} = 1.77$$

Z-Score Normalization or Scaling

Z-Score Normalization - Mean Absolute Deviation.

$$Z = \frac{x - \mu}{A}$$

μ = Mean
A = Mean Absolute Deviation

200
300
400
600
1000

$$\text{Mean} = \frac{(200 + 300 + 400 + 600 + 1000)}{5} = 500$$

$$\text{Mean Absolute Deviation} = \frac{|200-500| + |300-500| + |400-500| + |600-500| + |1000-500|}{5}$$

$= 240$.

$$v = \frac{200-500}{240} = -1.25$$

$$v = \frac{300-500}{240} = -0.833$$

$$v = \frac{400-500}{240} = -0.417$$

$$v = \frac{600-500}{240} = 0.417$$

$$v = \frac{1000-500}{240} = 2.08$$

Mean

Data(v)	Normalized Data(v)
200	-1.25
300	-0.833
400	-0.417
600	0.4117
1000	2.08

Z-Score Absolute deviation:-

Helps you understand where a specific observation falls within a distribution.

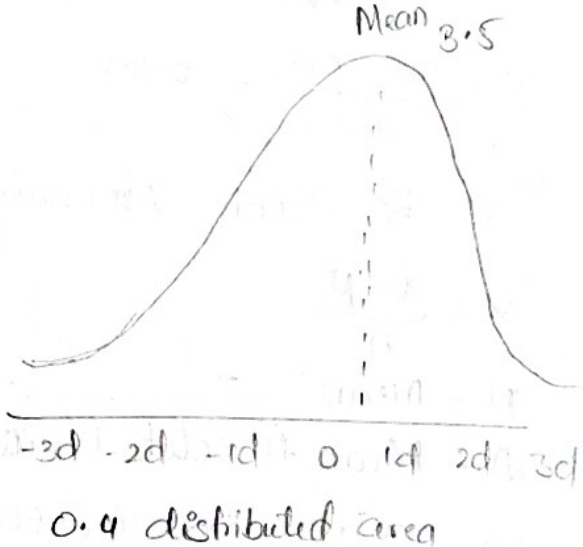
Z-score are often used in a medical setting to analyze how a certain newborn's compares to the mean wt of all babies

Ex: wt of new born babies are normally distributed with mean of 3.5 kgs & standard deviation 0.5

If a certain new born wt 3.6kg & their Z-score is 0.4 this means 0.4 standard deviation above mean.

Z-score longer from -3 to +3

New born 0.4 SD from mean represented in distributed area.



0.4 distributed area

$$Z = \frac{x-M}{\sigma}$$

Absolute Deviation = $|x - \text{mean}|$ → 10, 12, 14, 16, 18 unit deviation from Mean

$|10-14| = 4$

Mean Absolute Deviation = $\frac{4+2+0+2+4}{5} = 2.4$

Decimal scaling :-

Min - Max range : 12,000 - 98000

New range : 0 to 1.0 (0 to 1 $\quad Vi' = \frac{Vi}{10^d}$

$10^d = d =$ No of decimals in

4000 $\Rightarrow d = 4 \Rightarrow 10^4$

$Vi' = \frac{4000}{10^4}$

$= \frac{4000}{10000} = 0.4.$

8, 10, 15, 20 marks Consider max no of 20.

So, $d = 2$

$= 10^2$

$Vi'(8) = \frac{8}{100} = 0.08 \quad Vi(15) = \frac{15}{100} = 0.15$

$Vi'(10) = \frac{10}{100} = 0.10 \quad Vi(20) = \frac{20}{100} = 0.20$

Suppose the recorded values of A range from -845 to 945

Max value = 3 decimals

$= 10^3 = 1000.$

845 normalized to $= 0.845$

945 " " $= 0.945$

Normalization using Decimal scaling

Find value of j,

The smallest integer j such that $\max\left(\frac{Vi}{10^j}\right) < 1 = \frac{200}{10^3} < 1.$

$\frac{200}{10^3} = 0.2 \quad \frac{1000}{10^3} = 1.$

$\frac{300}{10^3} = 0.3$

$\frac{400}{10^3} = 0.4$

$\frac{800}{10^3} = 0.6$

Data (v)	decimal Scaled (v)
200	0.2
300	0.3
400	0.4
800	0.6
1000	1

What is attribute selection ?

Def:- Attribute selection is the process of choosing the important attributes that contribute to the prediction or analysis while discarding irrelevant, redundant or less useful attributes.

OR

It is the process of selecting only the most relevant attributes (features) from a dataset and removing unnecessary or irrelevant ones. This helps improve the accuracy and efficiency of DM alg.

Ex:- StudentID Name Age Gender Marks Attendance Result

= After selection

StudentID Age Gender Marks Attendance Result

Why is Attribute Selection Important ?-

→ Reduce the size of the dataset
→ Removes irrelevant and duplicate attribute
→ Improves the performance and accuracy of DM a
→ Reduce processing time and storage requirements
→ Makes the data easier to understand

Advantages :-

1) Improves model accuracy
2) Reduces computation time
3) Eliminates irrelevant and redundant data
4) Simplifies the dataset
5) Helps build efficient data mining models

What is discretization ?

Def :- Discretization is the process of transforming continuous data into discrete (categorical) value by grouping similar values into intervals.

(OR)

To converts continuous numerical values into a small no: of Intervals or Categories (Bins)

Why is Discretization Important ?

Simplifies large datasets
Makes Data easier to understand and analyze.
Improves the performance of some data mining ML alg

→ Reduces the effect of small measurement differences and noise.

Ex:- Student's Marks Dataset

Student	Marks
A	32
B	48
C	67
D	81
E	94

Discretization →

Student	Marks Category
A	Poor (0-40)
B	Average (41-60)
C	Good (61-80)
D	Excellent (81-100)
E	Excellent (91-100)

∴ Values 32, 48, 67 are converted into categories

Common Methods of Discretization

1) Equal-width Binning :- Divide the range of values into intervals of equal size.

2) Equal-frequency (equal-depth) Binning :- Each interval contains approximately the same no. of data values

3) Concept Hierarchy - Groups values based on predefined levels (ex:- city → state → country)

Advantages

1) Reduces data complexity
2) Makes patterns easier to identify
3) Improves the efficiency of some alg
4) Helps in better visualization and interpretation

Disadvantage :-

1) Some information may be lost when exact values are replaced by categories
2) Choosing inappropriate intervals can reduce accuracy

What is Concept Hierachy generation ?

Def:- CHG is the process of replacing low-level, detailed data with higher-level more general concepts based on a predefined hierarchy

OR

Organize data into different levels of abstraction, from detailed (low level) to general (high level)

It helps summarize and simplify data for easier analysis.

Ex:- Location Hierarchy

City	City → State → Country
Vijayawada	Vijayawada → AP → India
Visakhapatnam →	Visakhapatnam → AP → IND
Hyderabad	Hyd → TG → IND

Ex: Age Hierarchy

Age	Age Category
8	Child
20 →	Young
45	Adult
70	Senior citizen

(AUTONOMOUS)

Q:- Why is Concept Hierarchy Generation Important ?

Simplifies large datasets
Makes data easier to analyze and understand
Supports data summarization and data mining
Helps discover patterns at different levels of details

Advantages:-

Reduces data complexity
Improves data summarization
Makes reports easier to understand
Helps identify patterns at different levels of abstraction

Dis-advantages:-

Too much generalization may hide important details
Building an appropriate hierarchy can be time-consuming.

4. DATA REDUCTION

Used to reduce the size of a dataset while preserving the important information it contains.
The main goal is to make data storage, processing, and analysis faster and more efficient without significantly affecting the quality of the results.

Objectives

Reduce storage space.
Speed up data mining and ML alg
Eliminate redundant or unnecessary data
Maintain the accuracy of analysis

Common Techniques

1. Dimensionality Reduction
Reduces the no: of features (attributes) in DB
Ex: Using Principal Component Analysis (PCA) to reduce 100 features to 10

2. Numerosity Reduction
Replaces the original data with a smaller representation
Methods — Sampling
Clustering
Regression models
Histograms

3. Data Compression:
Compression data into a smaller formal
Lossless :- No information is lost
Lossy :- Some less important info is discarded

4. Data Cube Aggregation
Summarizes detailed data into higher-level aggregates
Ex:- Daily Sales data can be aggregated into monthly sales totals.
Ex:- A company has 10,00,000 customer records with 100 features :- After applying data reduction:
→ No: of records may be reduced to 1,00,000 sampling
→ No: of features may be reduced to 20 using PCA
→ Analysis becomes much faster while retaining most of the useful information

DATA CUBE AGGREGATION
Def:- Data is summarized values across one or more dimension. It reduce the amount of data while preserving useful information for analysis

Ex:-

Date	Product	City	Sales
1Jan	Rice	Hyd	5,000
2Jan	Rice	Hyd	6,000
3Jan	Rice	Hyd	4,000

Instead of storing/analyzing every daily record, the data can aggregated into a monthly summary.

Month	Product	City	Total Sales
jan	Pen	Hyd	15,000

3 daily records are replaced by one monthly record, reducing the amount of data while keeping the important info

Level of Aggregation

Data :- Data → Month → Year

City → State → Country

Product → Category

Advantages :-

1) Reduces data volume
2) Improves query performance
3) Simplifies reporting and analysis
4) Saves storage space
5) Supports faster Decision-making

Dis-advantages :-

1) Some detailed information is lost
2) Not suitable when detailed record-level analysis required

Attribute Subset Selection :-

Selects only the most relevant attributes (features) from a datasets and removes irrelevant or redundant attributes. This reduces the size of the data while maintaining or improving the accuracy of DM and ML tasks.

(AUTONOMOUS)

Student ID	Name	Age	Address	phoneNo	Marks	Grade

The goal is to predict a student performance, attributes like Address and phoneNo not be useful

Student ID	Age	marks	Grade

Only the student relevant attributes are retained

Methods.

Forward selection :-
Starts with no attributes
Adds the most useful attributes one by one

Backward Elimination :-
Starts with all attributes
Removes the least useful attributes one by one

Stepwise Selection
Combines forward selection and backward Elimina

Advantages :-
Reduces data size
Speeds up DM and ML
Removes irrelevant and redundant attributes
Improves model accuracy in many cases.
Makes the dataset easier to understand

Drs-Advantages

1) Important attributes may be removed if the selection is not done carefully
2) Finding the best subset can be computationally expensive for very large datasets.

Numerosity :-

Df:- Reduces the amount of data by replacing the original dataset with a smaller representation while preserving its essential characteristics; Instead of storing every data record, it stores a compact summary or model of the data.

Ex: A company has 10,00,000 customer records. Instead of using all records for analysis, it selects a representative sample of 10,000 records

Orginal Data: 10,00,000 records
Reduced Data: 10,000 representative records

The smaller dataset still represents the overall characteristics of the original data.

Methods

1) Sampling:
Selects a representative subset of the data

Ex: selecting 1,000 students from 50,000 students for a survey

2) Regression:-
Replaces data with a mathematical model or equation.
Ex:- Predicting house price using a regression equation instead of storing every price

3) Clustering
Groups similar data objects into clusters
Ex:- Grouping customers into High-value, Medium-value and low-value customer clusters

4) Histograms:-
Divides data into intervals (bins) and stores only the frequency of each interval
Ex:- Grouping ages into ranges such as 0-10, 11-20 and 21-30

Advantages:-
Reduces storage requirements
Improves processing speed
Makes data analysis more efficient
Retains the important characteristics of the data
Disadvantages:-
Some detailed information may be lost

The quality of results depends on the reduction method used

Used:-
Reduces the no:of data records
Saves storage space
Speeds up data processing and DM
Maintains the overall information with less data

Dimensionality Reduction

Def: Reduce the no:of attributes (features or dimensions) in a dataset while preserving the most important information. It simplifies the dataset, making it easier and faster to analyze.

Methods:-

1) Attribute subset selection
Selects only the most relevant attributes and removes unnecessary ones

(AUTONOMOUS)

e) Principal Component Analysis (PCA)

Transforms the original features into a smaller no: of new features (called principal components) that retain most of the information.

Used:-

Reduce the no:of features

Removes irrelevant, redundant, or highly correlated attribut

Reduces storage space

Improves the speed and efficiency of data mining & ML

Helps improve model performance and reduces overfitting

Stud-ID	Name	Age	Gender	Attendance	marks	Grade	Ad

Stud-ID	Age	Attendance	marks	Grade

Feature Subset Selection process and Approaches in Data Mining

Def:- Features Subset selection (FSS) also known as Attribute subset selection, is a data reduction techniques that selects the most relevant feature/attributes from a dataset and removes irrelevant, redundant, or noisy features. This reduces the dimensionality of the data while maintaining or improving the accuracy of DM model

Original Dataset
↓
Feature Evaluation
↓
Search Strategy
Forward / Backward / Stepwise
↓
Best feature Subset
↓
Data Mining Algorithm
↓
Faster & Accurate Results

Step-1: Original Dataset:-
The dataset contains all available features

Ex:-

Stud-ID	Name	Age	Gender	Attendance	Marks	Address	ph	Grade

Step-2:- Feature Evaluation

Each feature is evaluated to determine how important it is for mining task

Useful features:

Age	Attendance	Marks	Grade

Irrelevant features

Name	Address	phone Number

Step-3: Search Strategy:-

A search method is used to find the best combination of features selection

Common search methods:- forward/Backward Elimination stepwise Selection

Step-4:- Best feature subset

Only the important features are selected

Selected features:

Age	Attendance	Marks	Grade

Step-5: Data Mining model

The selected features are used to build a classification, clustering or prediction model.

Approaches of feature Subset Selection :-
3 main approaches : 1) Filter - Approach
2) Wrapper - Approach
3) Embedded - Approach

1) filter - Approach :-
→ Selects features before training the model.
→ Uses statistical techniques such as correlation, Information Gain, etc...
→ Independent of the learning algorithm.

Ex:- Student dataset, Marks and Attendance are selected because they strongly affect performance, while Address & phone Number are removed.

Advantage:- Fast, Simple, Suitable for large Datasets

2) Wrapper Approach :-
Evaluates different feature subsets by training and testing a ML model
Selects the subset that gives the highest accuracy

Ex:- A bank tests different combinations of Age, Income, Loan Amount, and Credit Score to find the best features for predicting loan approval

Advantage:- High accuracy Better feature selection

Disadvantage:- Time-consuming

(AUTONOMOUS)

3) Embedded Approach :-
Feature selection is performed during model training
The learning alg automatically selects important features

Ex:- 1) decision tree automatically chooses Income and Credit Score while building the model

Advantages :- Easier than the wrapper approach
Good accuracy, Efficient

Ex:- 1 bank wants to predict whether a customer will repay a loan

| Customer ID | Name | Age | Income | occupation | phoneNo: | Loan Amount | Credit Score |

After feature subset selection

| Age | Income | Loan Amount | Credit Score |

Advantages
1) Reduces Data size
2) Removes irrelevant and redundant features
3) Improves processing speed
4) Saves Storage Space
5) Can improve prediction accuracy

Disadvantages
1) Important features may be removed if not selected carefully
2) Finding the best subset may required additional computation

Discretization and Binarization :-

Binning / Discretization :-

Purpose:- Binning is primarily used to convert continous numerical values into discrete bins or categories.

This is often beneficial when there's no linear relationship b/w the numerical values and the target variable or when certain ranges of values hold specific, meaningful groupings.

How it works:- Binning divides the range of a numerical feature into a fixed no:of intervals or bins and assigns each data point to a bin based on its value.

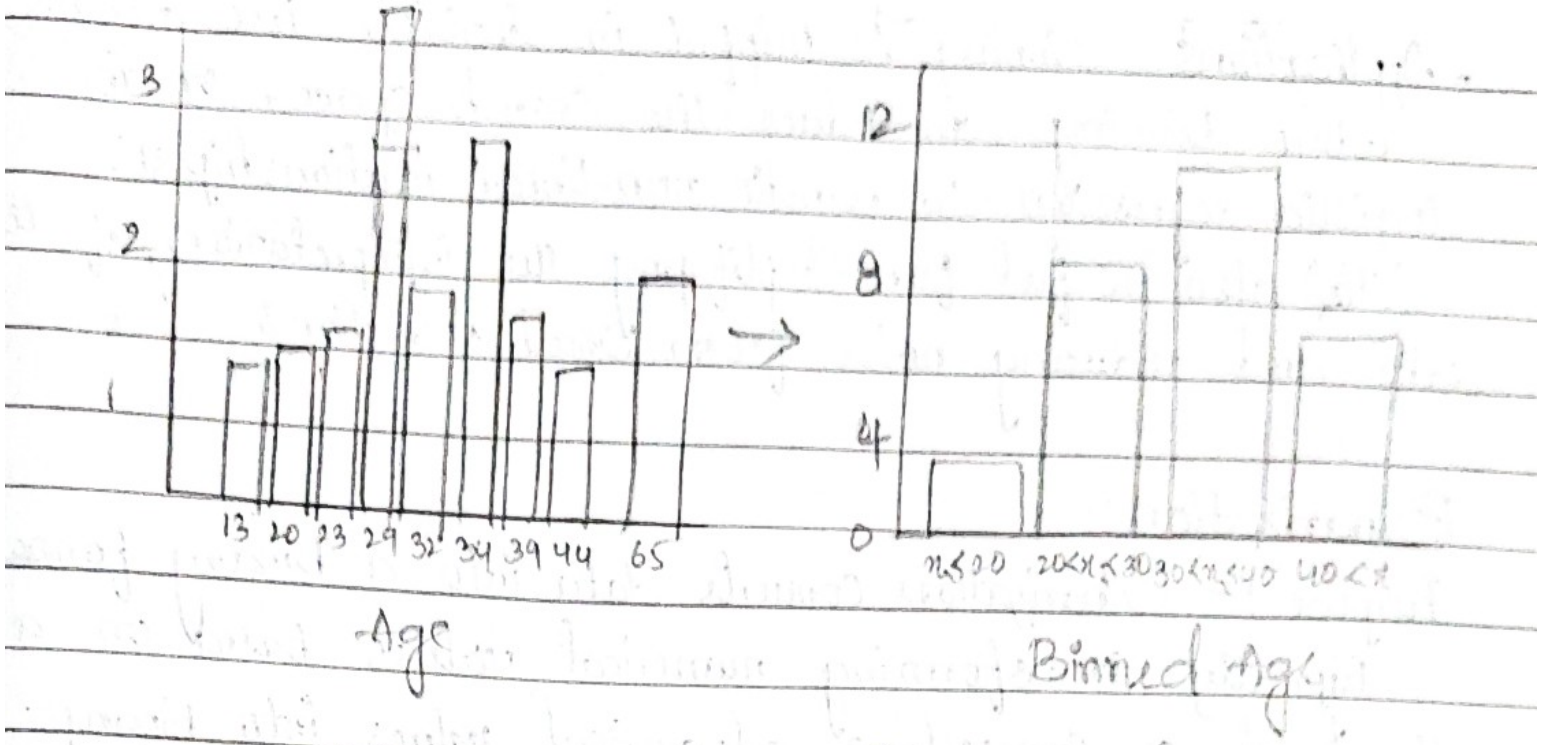
Types of Binning :-

Equal-width binning :- Divides the data range into intervals of the same width

Equal-frequency binning :- Divides the data so that each bin contains approximately the same no:of observations.

Custom Binning :- Allows manual definition of intervals based on domain knowledge or other considerations.

Ex:- For a datasets with age [20,23,25,30,45,60,80] we might create bins as [20-29], [30-39], [40-49], [50+..] Categorizing each age into the appropriate bin



actual ages [13, 20, 23, 29, 32, 34, 39, 44, 65, etc]

grouped into intervals (Bins):

$x \leq 20 \longrightarrow 13, 20$

$20 < x \leq 30 \longrightarrow 23, 29$

$30 < x \leq 40 \longrightarrow 34, 39$

$x > 40 \longrightarrow 44, 65$

The height of each bar now represents the total no: of ages in that interval

Suppose the age Data :-

13, 18, 20, 22, 24, 28, 31, 34, 36, 39, 42, 44, 50, 65

Bin	values	count
$x \leq 20$	13, 18, 20	3
$20 < x \leq 30$	22, 24, 28	3
$30 < x \leq 40$	31, 34, 36, 39	4
$x > 40$	42, 44, 50, 65	4

Instead of remembering 14 individual ages we summarize them into 4 groups.

Applications- Binning is helpful in decision-tree models, where binning can reduce the search space, or in logistic regression to handle non-linear relationships. It's also useful for simplifying the interpretation of the data and reducing noise from small variations.

Binarization:-

Purpose:- Binarization converts data into a binary format, typically transforming numerical values based on a threshold or converting categorical values into binary indicators

How it works:-

for numerical features, binarization applies a threshold
Ex: all values above the threshold becomes 1, Others 0

for categorical features, it transforms categories into a binary format using one-hot encoding

A, B, C into [1,0,0], [0,1,0], [0,0,1]

Numerical data like [2.5, 3.7, 1.8, 4.0] setting a threshold of 3.0 could produce a binary output [0,1,0,1]

Applications- Binarization is widely used in alg that benefit from binary inputs. Such as Naive Bayes / NeuralNW It's also helpful in handling categorical data for alg that cannot operate on non-numerical data directly.

Finding no: of binary attribute required:

Let's have n be the no: of binary attributes $n = \lceil \log_2 m \rceil$

Conversion into binary:

No: assigned is converted to its respective binary value.

Ex: if no: of binary attribute is 3 in no's then

2 - 010

if no: of binary attribute is 4 in no's then

2 - 0010

Ex: Lets us consider {awful, poor, ok, good, great}

1) Assigning numerical value:

2) Identifying no: of binary attribute:

$n = \lceil \log_2 m \rceil$

i.e., $n = \lceil \log_2 5 \rceil \approx 3$

3) Binary Conversion:

Attribute values	Integer value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
ok	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Binarization.

Measures of Similarity and Dis-similarity :-

More important - used in clustering, some classification, anomaly detection

Similarity :

Numerical measure of how alike two data objects are.

If it is higher when objects are more alike.

Often falls in the range [0,1].

Dis-Similarity :-

Numerical measure of how different are two data objects.

lower when objects are more alike.

Minimum dissimilarity is often 0

Upper limit varies.

Proximity refers to a similarity or dissimilarity.

Similarity / Dis-similarity for objects with Single Attribute:
P and q are the attribute values for two data objects.

Attribute Type	Dis-Similarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } P = q \\ 1 & \text{if } P \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } P = q \\ 0 & \text{if } P \neq q \end{cases}$
Ordinal	$d = \frac{\lvert P-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the no: of values)	$s = 1 - \frac{\lvert P-q \rvert}{n-1}$, $s = 1-d$
Interval OR Ratio	$d = \lvert P-q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - \min\text{-}d}{\max\text{-}d - \min\text{-}d}$

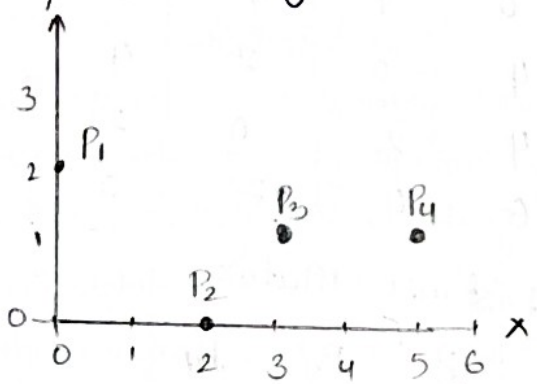
Dissimilarities b/w Data objects with multiple Numeric Attributes.

Euclidean Distance :-

$$dist = \sqrt{\sum_{k=1}^{n} (P_k - q_k)^2}$$

Where n is the no: of dimensions
P_k, q_k are respectively,
k^{th} attributes of data objects p and q

Standardization is necessary, if scales differ.



Point	x	y
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1

$P_1 - P_2$ / $P_2 - P_1$

$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

$= \sqrt{(0-2)^2 + (2-0)^2}$

$= \sqrt{4+4}$

$= \sqrt{8} = 2.828$

Distance Matrix.

Points	P_1	P_2	P_3	P_4
P_1	0	2.828	3.162	5.099
P_2	2.828	0	1.414	3.162
P_3	3.162	1.414	0	2
P_4	5.099	3.162	2	0

Minkowski Distance:-

Minkowski Distance is a generalization of Euclidean Distance. Given 2 object P and Q

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r\right)^{\frac{1}{r}}$$

where r is a parameter, n is the no: of dimensions and p_k and q_k are respectively k^{th} attributes of data objects P & Q.

Ex:-

If r=1. City block (Manhattan, taxicab, L_1 norm) distance.

A common example of this is the Hamming distance, which is just the no: of bits that are different b/w 2 binary vectors.

If r=2. Euclidean Distance.

If $r \to \infty$. "Supremum" (L_{max} norm, L_∞ norm) distance.

This is the maximum difference b/w any attribute of the vectors.

$$d(x,y) = \lim_{r\to\infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{\frac{1}{r}}$$

Do not confuse r with n i.e., all these distances are defined for all no: of dimensions.

Points	x	y
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1

L_1	P_1	P_2	P_3	P_4
P_1	0	4	4	6
P_2	4	0	2	4
P_3	4	2	0	2
P_4	6	4	2	0

Distance Matrix.

Manhattan Distance:-

$P_1 = (0,2) = (x_1, y_1)$

$P_2 = (2,0) = (x_2, y_2)$

$d = |x_1 - x_2| + |y_1 - y_2|$

$= |0-2| + |2-0|$

$= 2+2$

$= 4$

Similarly for the other points.

L_2	P_1	P_2	P_3	P_4
P_1	0	2.828	3.162	5.099
P_2	2.828	0	1.414	3.162
P_3	3.162	1.414	0	2
P_4	5.099	3.162	2	0

Difference Between KDD and Data Mining.

Parameter	KDD	Data Mining.
Definition	KDD refers to a process of identifying valid, novel, potentially useful, and ultimately understandable patterns and relationships in data.	Data Mining refers to a process of extracting useful and valuable information or patterns from large data sets.
Objective	To find useful knowledge from data	To extract useful informatio from data.
Techniques Used.	Data cleaning, data integration, data selection, data transformatio data mining, pattern evaluation, and knowledge representation and visualization.	Association rules, classificatio clustering, regression, decision trees, neural network and dimensionality reductio
Output	Structured information, such as rules and models, that can be used to make decision or predictions.	Patterns, associations, or insights that can be used to improve decision-making or understanding.
Focus	Focus is on the discovery of useful knowledge, rather than simply finding patterns in data.	Focus is on the discovery of patterns or relationships in data.

UNIT-III

Association Rules :- Problem Definition, frequent Itemset Generation, The APRIORI principle, Support & Confidence Measures, Association Rule generation, APRIORI Algorithm, The partition Algorithm FP- Growth, Compact Representation of frequent Item set - Maximal frequent Item set, closed frequent Item set.

Association Rule:-

→ It finds interesting associations & R/s among large sets of data items.

→ This rule shows how frequently a itemset occurs in a trans^n

→ Association rules are simple If/THEN stmts that help discover R/s b/w seemingly independent relation DB or other data repositories.

Association rule has 2 parts

* An antecedent (if) &
* A Consequent (then).

If — Something that's found in data

then — Found in combination.

Frequent item set — Set of items that frequently appear together in transactional data set.

* If a customer buys bread, he's 70% likely of buying Milk.

Bread - antecedent

Milk - Consequent

→ Retailer store's association rule to target customers better.

→ Association rules are created by thoroughly analyzing data & looking for frequent <u>if/then</u> patterns.

→ Depending on two parameters, imp R/s are observed.

Support - Indicates how frequent if/then R/s appears in the DB.

Confidence - No. of times these R/s have been found to be true.

Support - Measures of interestingness

Areas where ARM has helped.

* Market Based Analysis.
  - to adjust store layout & store catalog to place.

* Medical Diagnosis:-
  Rls blw new signs & corresponding disease.

(etc.)

USE CASE

Business Problem:- A retailer store manager wants to conduct Market Basket analysis to comeup with a better strategy of product placement & product bundling.

Benefit:- Based on the rules generated, the store manager can strategically place the products together to growth in sales & revenue of store.
offers Buy 1 & get 2 free & 1% off

on this can be generated based on designated the rules generated.

Association Rule — frequent item set.

Interesting R/S

If Mom buys {Milk, Diaper} — Items frequently occur together.

If a lady buys Milk. chance of buying Diapers.

If father goes shop {Diapher, Beer}

Diaper - Beer.

Interesting R/s that don't buy not frequent.

In Amazon, we can see "frequently Brought together".

Association Mining:- Searches for frequent items in data sets.

→ Frequent mining shows which items appear together in a transaction or relation.

Need of Association Mining:-

→ Frequent mining is generation of association rules from Transactional [dataset.

AR — Simple If/Then statements.

Eg:- If 2 items X & Y purchased frequently then its good to put them together in store or provide discount offer on one item on purchase of other item.

If — Item found within the data.

THEN.— Item found in combination with If (that item)

Association Rule

An implication expression of form

$$X \rightarrow Y$$

where X & Y are any 2 Itemsets

Milk → Bread

$$X \rightarrow Y$$

Milk, Diaper → Bread

X → Y

→ Association rules has no. of applications

Eg ~~transacti~~ & ~~ated~~ used to ~~body~~ discover relations in Transactional data or in Medical data sets.

Medicine: There are many variables to consider when making diagnosis, as many diseases share same symptoms.

Retail: - Market Base Analysis.

Entertainment: - Netflix & Spotify can use association rules to grow their "Content recomendations engines".

ML models analyze past user behaviour

data for frequent patterns, develop association rules & organize content in a way to put most interesting content for a given user first

Rule :- $x \Rightarrow y$

$$\text{Support} = \frac{frq(x, y)}{N}$$

$frq(x, y)$ – No. of times x & y brought together.

N – No. of Transaction.

$$\text{Confidence} = \frac{frq(x, y)}{frq(x)}$$

$frq(x)$ – Total no. of times item x has brought in the total trans^n.

$$\text{confidence}(x \to y) = \frac{\text{Support}(x \cup y)}{\text{Support}(x)}$$

→ Association rule is imp. in ML & employed in Market Basket analysis, web usage Mining, etc.

* Association rule can be divided into 2 types of algorithms
  1. Apriori
  2. FP Growth Algorithm

Apriori

→ This algo. use frequent datasets to generate association rule.

→ It is designed to work on the DB's that contains transactions.
Mainly used to MBA to understand the products that can be brought together.

JOIN (Apriori Principle)

→ Step generate (K+1) itemset from K-itemset by joining each item with itself.

Prune:- Scans the count of each item in the DB. if candidate item doesn't meet min support then it is removed.

Apriori Algorithm – Association Rule Mining

Eg:-

Transaction ID	Items Bought
T1	{Bread, Butter, Milk}
T2	{Bread, Butter}
T3	{Beerquile, Cookie, Diapee}
T4	{Milk, Diapher, Bread, Butter}
T5	{Biscuite, Diapher}

→ In a supermarket, T1 means bill made a customer & items he/she purchased.

for Easy understanding, assign no.s to items.

Bread – 1

Butter – 2

Milk – 3

Biscuite – 4

Cookie – 5

Diapher – 6

Total no. of items = 6

Total no. of transaction = 5.

* Apply Apriori Alg, to discover strong association rule

Assume, Min - support = 40%.

Min - confidence = 70%.

→ support & confidence can be declared by user.

* Generate association rule from the frequent itemsets & calculate the confidence of each rule & identify all strong association rules.

* Generate candidate itemsets (Ck) & Qualified frequent itemsets step by step until the larget frequent itemsets generated.

Itemset	Support count
1	3
2	3
3	2
4	2
5.	1 ✗ < 2
6	3

Support count = no. of items times particular item purchased by all customers / transactions.

Generate frequent item sets from above table.

Assume · Min_support = 40%

Min_support_count =

min_support x itemset_count

= 40% x 5

= 2

Min support = 2

Item 5 : support count ≤ 2.

$1 \leq 2$

Discard. 5 (Item)

Now, Remaining items are

1, 2, 3, 4, 6. (5 unique items)

Generate 2 Item-sets from above

{1,2} {1,3} {1,4}, {1,6}

{2,3} {2,4} {2,6}, {3,4}

{3,6} {4,6}

Item-set	Frequency / support count	Min support 2
{1,2}	3	
{1,3}	2	
{1,4}	0	< 2 X
{1,6}	1	< 2 X
{2,3}	2	
{2,4}	0	X

{2,6} 1 < 2 X

{3,4} 0 < 2 X

{3,6} 1 < 2 X

{4,6} [illegible]

Discard {2,6} {3,4} {3,6} {1,4}
{1,6} {2,4}.

Remaining - 1, 2, 3, 4, 6. (unique items)

3 - Item Set (Generate).

{1,2,3} {1,2,4} {1,2,6}
{2,3,4} {2,3,6} {1,3,4} {1,3,6}
{1,4,6} {3,4,6} {2,4,6}.

Item-set	Support-count.
{1, 2, 3}	2
{1, 2, 4}	0 X < 2 X
{1, 2, 6}	1 X
{1, 3, 4}	0 X
{1, 3, 6}	1 X
{1, 4, 6}	0 X
{2, 3, 4}	0 X
{2, 4, 6}	0 X
{3, 4, 6}	0 X

final : 1 item-set

{1, 2, 3} — 2 <

Generated frequent Item-sets.

write Associate rules, select those association rules based on confidence.

Bread, Butter, Milk.

confidence = 70%.

$x \rightarrow y = P(y/x)$

$= \boxed{\frac{P(x \cup y)}{P(x)}}$.

$P(x \cup y)$ = combination of x & y items occurred frequently

$P(x)$ = no. of times x appeared.

5 frequent itemsets

{Bread, Butter} {1, 3} {2, 3}

1, 2

{4, 6}

1 → 2 → Here How many no. of purchased times butter has brought

2 → 1 — others customer takes Bread.

2-1 → Vice versa

Bread → Butter = $3/3 = 100\%$. (strong)

Butter → Bread = $\frac{3}{3} = 100\%$ (")

Bread, Milk

Bread → Milk = 2/3 = 67%.

Milk → bread = 2/2 = 100%. (Steong).

{1, 3} — 2

count of bread = 3

Bread → Milk = combination count of Bread & Milk / count of only Bread.

{Diapers, Biscuits}

diapers → Biscuits = 2/3 = 67.%.

bees — diapers = 2/3 = 100%. (Steong)

{Bread, Butter & Milk}

Bread, butter → Milk = 2/3 = 67%. X

Bread, Milk → butter = 2/2 = 100%. ✓

Milk, butter → bread = 2/2 = 100 %. ✓

bread → butter, milk = 2/3 = 67 %.

butter → bread, milk = 2/3 = 67 %.

Milk → bread, butter = 2/2 = 100 %. ✓

So, finally Association b/w Bread, Milk, butter are steong !

Example 2

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

Min support = 50%

4 = No. of trans

Items — 1, 2, 3, 4, 5

Frequency for each Item

Itemset	Frequency	Min. Support
1	2	2/4 = 50%
2	3	3/4 = 75%
3	3	3/4 = 75%
4	1	1/4 = 25% X
5	2	2/4 = 50%

New Itemset.

(1, 2, 3, 5)

Pairs

(1,2) (1,3) (1,5) (2,3) (2,5) (3,5)

Itemset	Frequency	Min. Support
(1,2)	1	1/4 = 25% X
(1,3)	2	2/4 = 50%
(1,5)	1	1/4 = 25% X
(2,3)	2	2/4 = 50%
(2,5)	3	3/4 = 75%
(3,5)	2	2/4 = 50%

→ observe pair (1,2) occured together & count the occurance.

New Itemset = (1,3) (2,3) (2,5) (3,5)

(1,2,3,5)

form Triplets

(1,2,3) (1,2,5) (1,3,5) (2,3,5)

Itemset	Frequency	support
(1,2,3)	1	25% = 1/4 X
(1,2,5)	1	25% X
(1,3,5)	1	25% X
(2,3,5)	2	50%

New Itemset = (2, 3, 5)

Now calculate support & confidence:

$$\text{confidence} = \frac{\text{support}(A \cup B)}{\text{support of } (A)}$$

OR

$$\frac{P(A \cup B)}{P(A)}$$

Generate association Rules for (2, 3, 5)

Rule	Support	Confidence
$2 \wedge 3 \rightarrow 5$	2	$2/2 = 100$
$3 \wedge 5 \rightarrow 2$	2	$2/2 = 100$
$2 \wedge 5 \rightarrow 3$	2	$2/3 = 66$ ✗
$2 \rightarrow (3 \wedge 5)$	2	$2/3 = 66$ ✗
$3 \rightarrow (2 \wedge 5)$	2	$2/3 = 66$ ✗
$5 \rightarrow (2 \wedge 3)$	2	$2/3 = 66$ ✗

Confidence

$2 \wedge 3 \rightarrow 5 = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$

or

$\frac{2,3 \rightarrow 5}{A \rightarrow B}$

$= \frac{s((2 \wedge 3) \cup 5)}{s(2 \wedge 3)}$

$= \frac{s(2,3,5)}{s(2,3)} = 2/3$

Support of $2,3,5 = 2$

$2 \rightarrow \frac{(3 \wedge 5)}{A \rightarrow B} = \frac{s(2 \cup (3 \wedge 5))}{s(2)}$

$= \frac{s(2,3,5)}{s(2)} = 2/3$

$(2,3) \rightarrow 5$ are association
$(3 \wedge 5) \rightarrow 2$ rules.

2, 3, 5 & 3, 5, 2 are strong.

THE PARTITION ALGORITHM

FP GROWTH

→ FP Growth, employs Divide & Conquer Strategy.

→ Huge DB can be divided into smaller DB's & then frequent itemsets can be mined.

→ Initially, DB is compressed into Frequent Pattern Tree (FP-Tree).

FP - frequent pattern.

FP-Tree

→ the compressed form of i/p is called as FP tree.

→ Tree is constructed by sequentially reading each transⁿ from the dataset & then mapping each of it to a path of FP-Tree.

→ the paths will get overlapped when transⁿ have common items.

Item Set

→ collection of zero or many item in association analysis is termed as an "Item set".

If there are n items in itemsets — n-item set

null item in itemsets — null item set.

Finding frequent Itemsets using FP-Growth Algorithm.

Step 1 :- list items

Step 2 :- write frequency/support count of each item

Step 3 :- write priorities ← High/

Step 4 :- More frequency - More priority

Same frequency - FCFS.

Step 5 : Construct FP Tree.

TID	Items-bought
T_1	K, A, D, B
T_2	D, A, C, E, B
T_3	A, C, B, E
T_4	B, D, E
T_5	A, D, B

Support count = 3

List Items

A, B, C, D, E, K.

Support count:

A → 4

B → 5

C → 2 X

D → 4

E → 3

K → 1 X

discard.

final items in Itemset

A, B, D, E.

Give Priorities

B A D E

(B:5) (A:4) (D:4) (E:3).

A & D have same priorities so follow FCFS.

FP Tree

Root = NULL.

First Transaction $T_1 = \{A, D, B\}$

Acc. to priority = B, A, D.

→ New branch is created for 1st trans^n

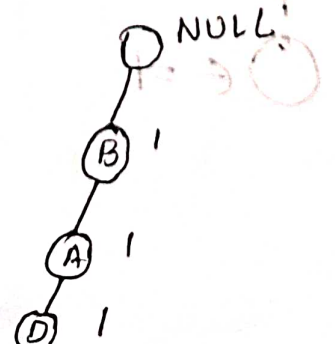
→ B is connected to the root node as a left child

→ A is linked to B as left child

→ D " " " A " "

→ After every transaction SC of respective transaction item is incremented by 1.

i.e., B=1, A=1, D=1.



FP tree for T1

for T2 = { D, A, C, B }

Acc. to priority = { B, A, D, E }

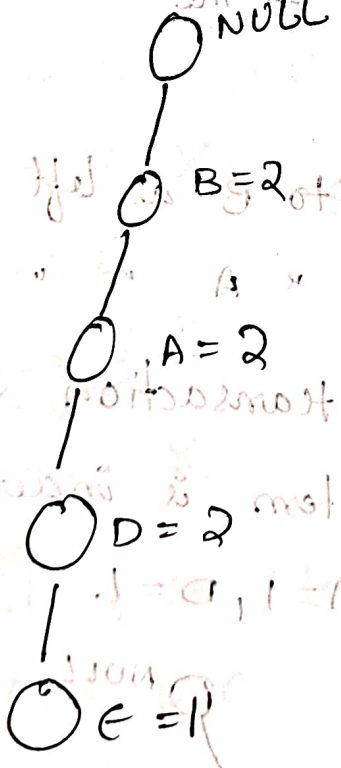
∴ B - A - D path already exist, there is no need to create new branch for these items.

→ Increment SC of respective items by 1.

B = 2, A = 2, D = 2.

→ E is linked as the left child of D.

E = 1. (SC)



Transaction T_2 = {A, B, C}

Acc. to priority = {B, A, C}

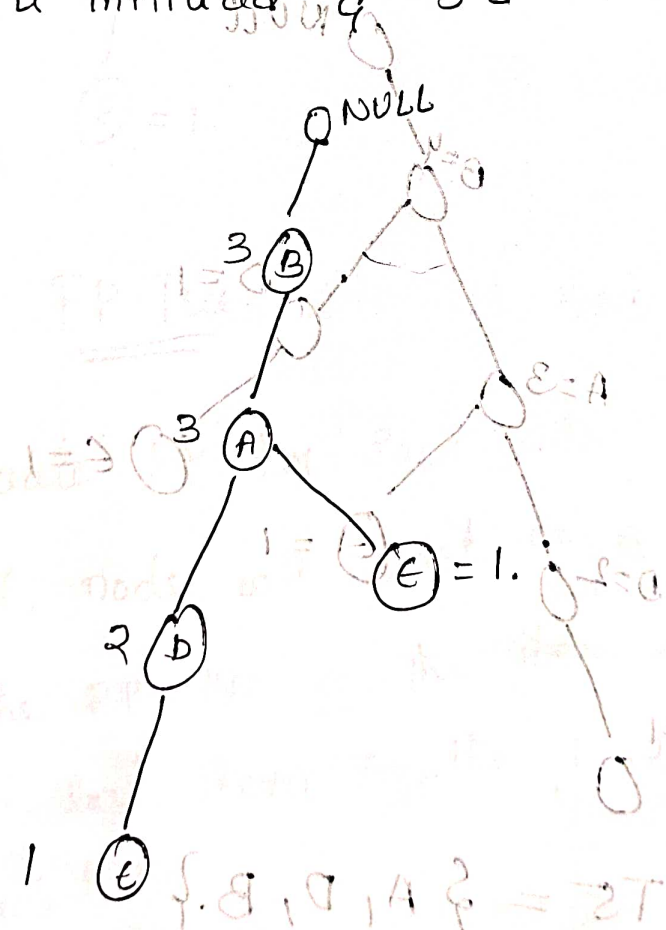
* path B – A already exist
* no need to create new branch for these items:

  SC incremented by 1

  i.e., B = 3, A = 3.

* C is linked as right child of 'A' &

  C is initilized & SC = 1.



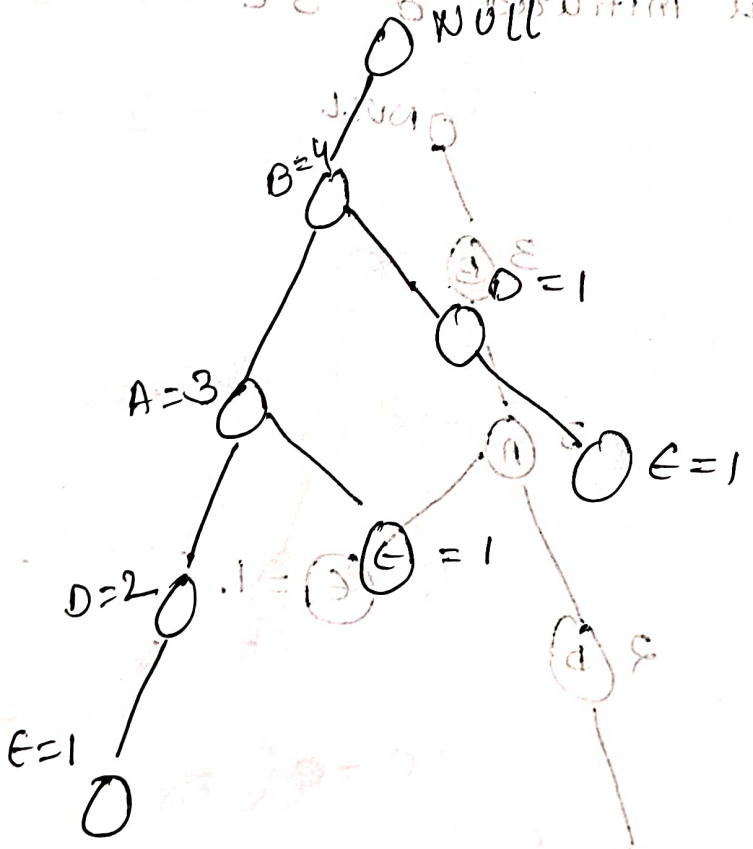
for T4 = { B, D, E }

priority = { B, D, E }

→ B, D, E results in creation of new branches from B to D, where D is linked as a right child of 'B'.

→ E linked as left child of D.

→ S.C. of B, D, E are 4, 1, 1.



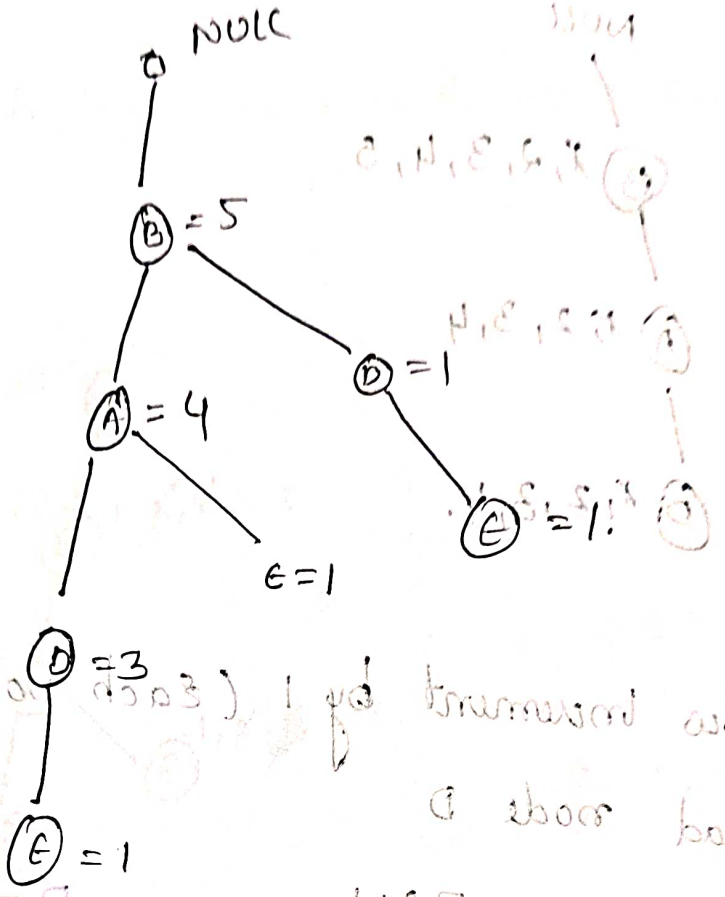
for T5 = { A, D, B. }

Priority = { B, A, D }

Path B – A – D already exists.

SC of B, A, D is inverted.

B = 5, A = 4, D = 3



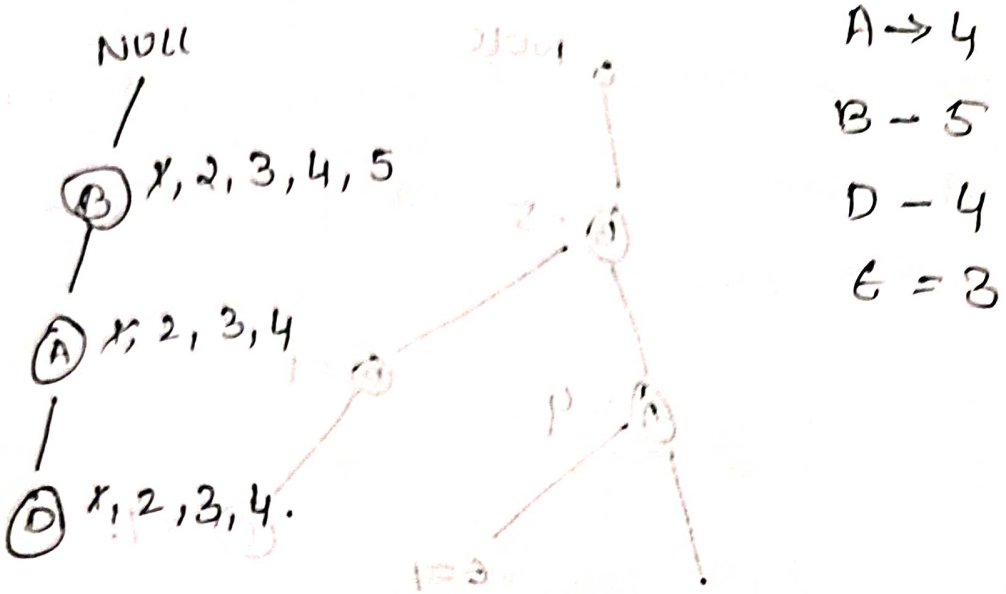
Mining FP Tree

→ Generate FP for each node.

→ First, node with least no. of occurrence in the FP-tree or the item selected i.e., last items in the order is selected.

→ Now select Transaction & write the path to each node in the tree.

T1:- BAD



A → 4
B - 5
D - 4
E = 3

* Now increment by 1 (each node). To read node D

D → BA:1

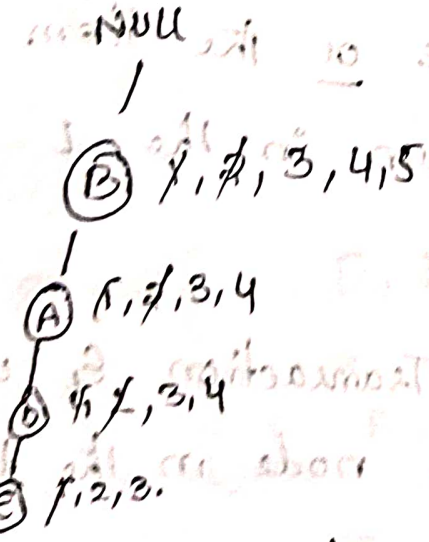
To reach node A

A → B:1

D → BA:1

A → B:1

T2:- BADE



E → BAD:1

D → BA:1

A → B:1

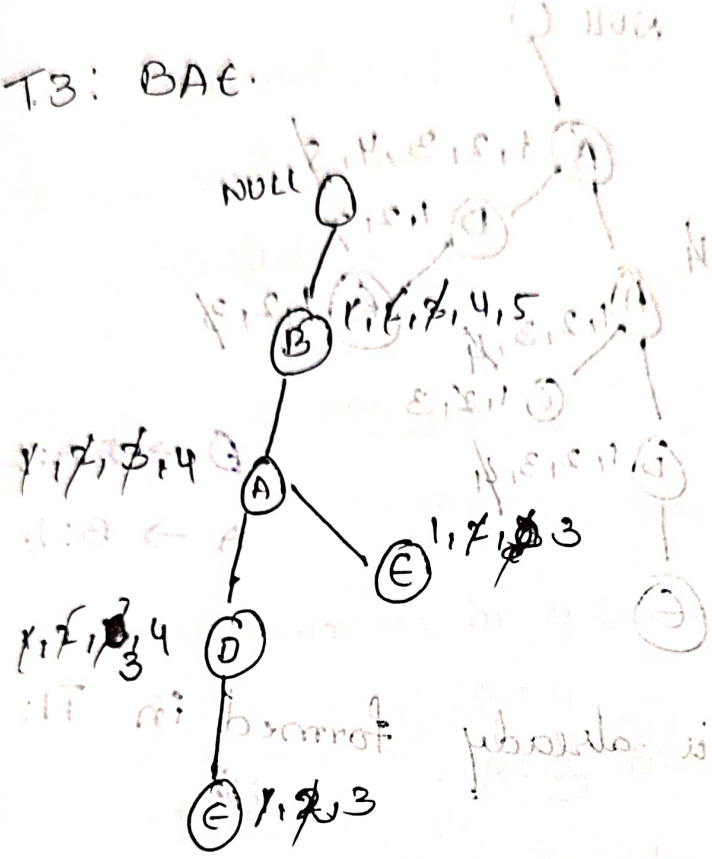
In T2, increment B, A, D by 2, because while visiting this

path B or B, A, D or occurring 2nd time in 2nd transn.

→ To reach reach 'ϵ' we need to visit BAD once

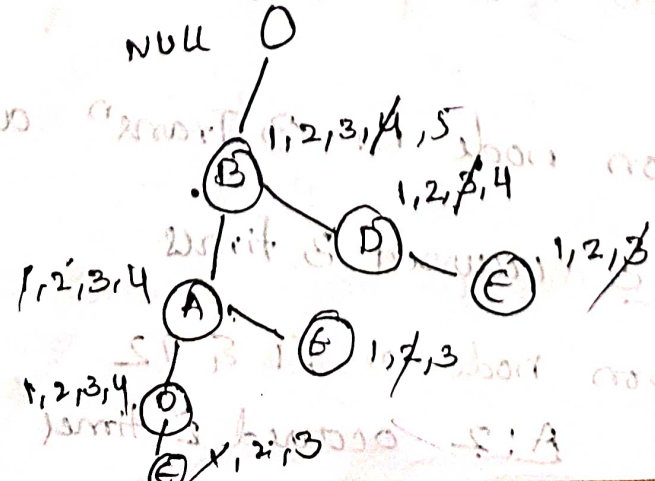
→ To reach 'D', we need to visit BA once.

T3: BAϵ.



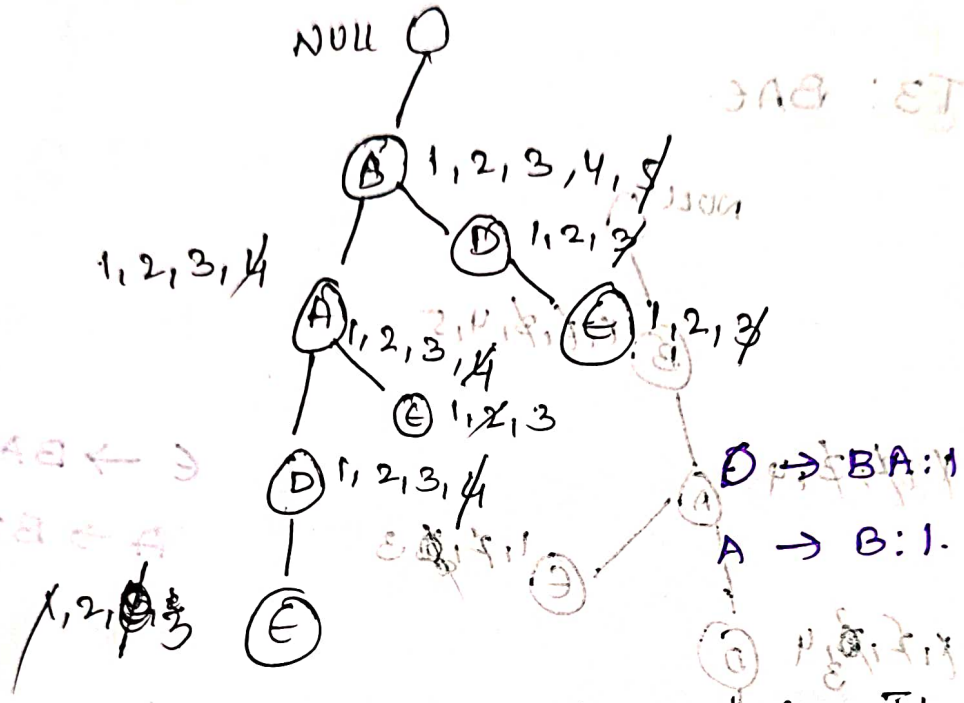
→ To reach 'ϵ' in T3, we need to visit path BA once & for A, B once.

T4:- B, D, ϵ



In T4, To visit E, need to visit BD & for D, visit B.

T5: BAD.



→ BAD path is already formed in T1.

→ Now, select last node.

Item — conditional pattern Base

E — E → BAD:1 (T1)

E → BA:1 (T2)

E → BD:1 (T4).

→ common node in 3 Trans^n are B.

B:3 occurred 3 times

A common node in T1 & T2

A:2 occured 2 times

'D' common node in T1 & T4.

D : 2 occurred 2 times

support count = 3

'AD' support count is (2,2)

SC < 3, so Discard A & D.

∴ conditional FP-tree consists of only single node i.e., B:3. only one FP is generated is {B, E : 3}.

Item	Conditional Pattern Base
D	BA:1, BA:1, B:1, BA:1

common in 4 trans is B.

B:4

A:3 ~~SC(A) < 3~~

Now form the combination with B&D & A&D

BD : 4, AD : 3.

Item	Conditional Pattern Base
A	B:1 (T_1), B:1 (T_2), B:1 (T_3), B:1 (T_4)

B:4

form combination, BA

BA : 4

frequent patterns are

{B, E : 3}

{B, D : 4}, {A, D : 3}

{B, A : 4}

Drawbacks of FP Growth Algo

→ The size of an FP-tree ↑ with the ↑ in the size of dataset due to which memory space may become unavailable.

→ If an item-sequence is present in a single record, then it's represented in the FP-tree.

Prob 2

RID	Items
100	F, A, C, D, G, I, M, P
200	A, B, C, F, L, M, O
300	B, F, H, J, O, W
400	B, C, K, S, P
500	A, F, C, E, L, P, M, N.

step 1:- SC = 3.

F : 4 (boundary)
A : 3
C : 4
B : 3
M : 3
P : 3
L : 2 X

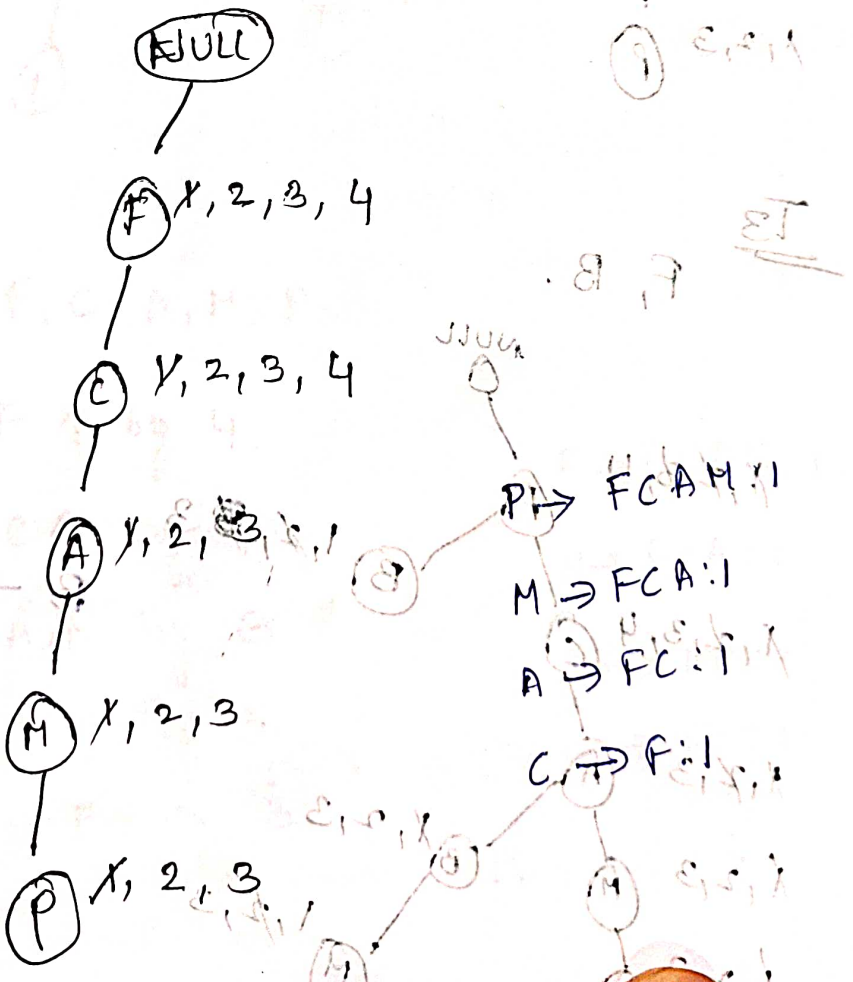
O : 2 X
D : 1 X
E : 1 X
G : 1 X
H : 1 X
I : 1 X
J : 1 X

W : 1 X
K : 1 X
S : 1 X
N : 1 X

∴ SC < 3

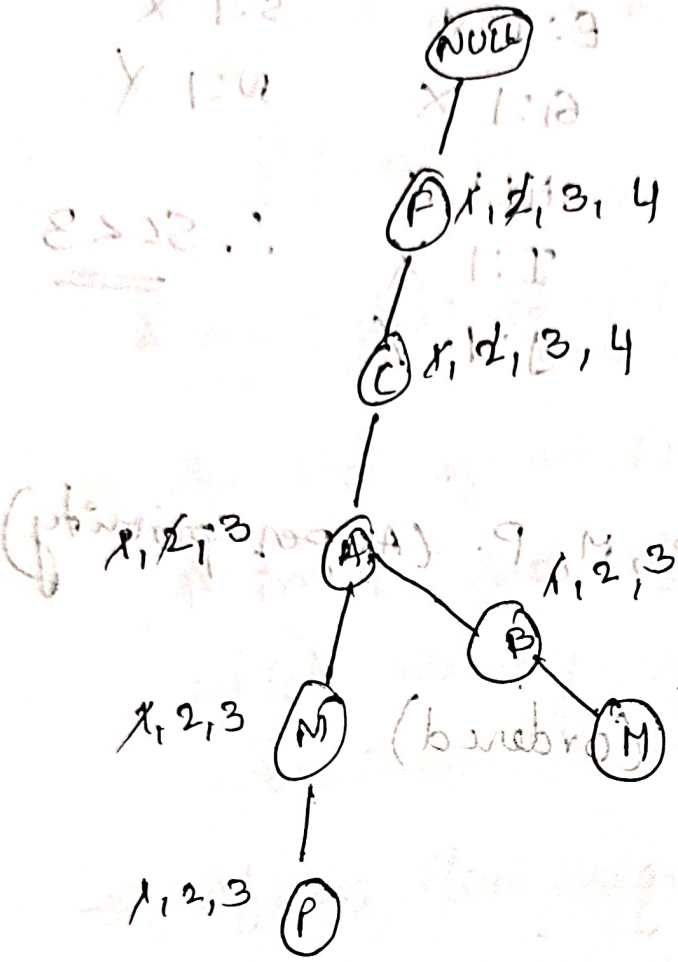
F, C, A, B, M, P. (As per priority)

T1 = FCAMP. (ordered)



T_2

F, C, A, B, M (ordered)



M :→ FCAB ; 1

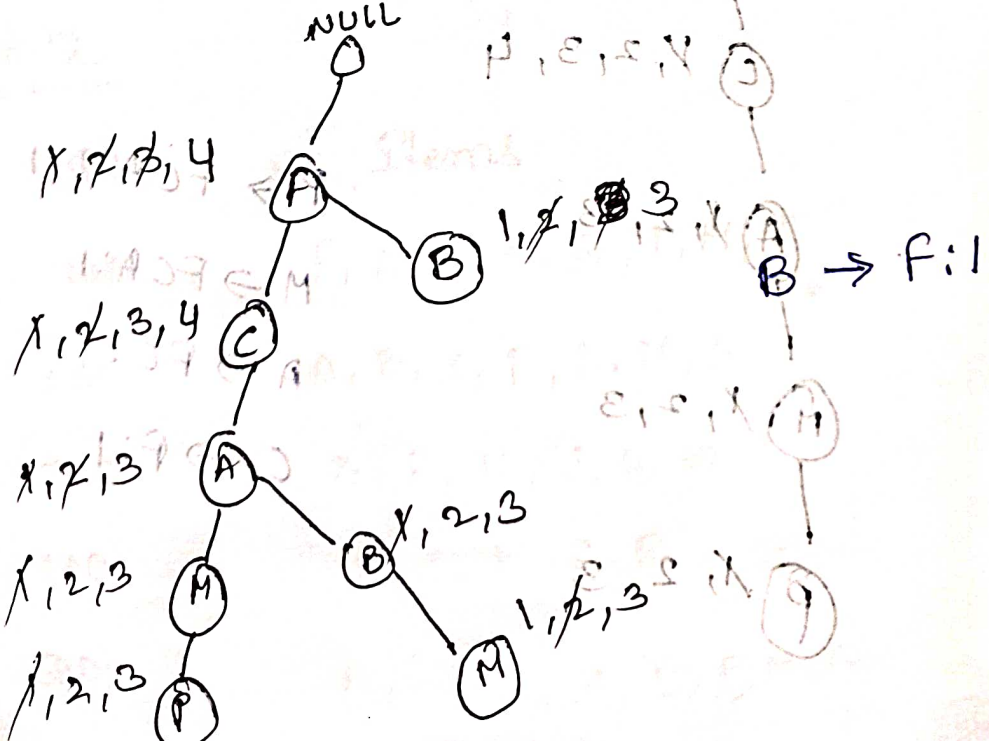
B :→ FCA ; 1

A :→ FC : 1

C ↦ F : 1.

T_3

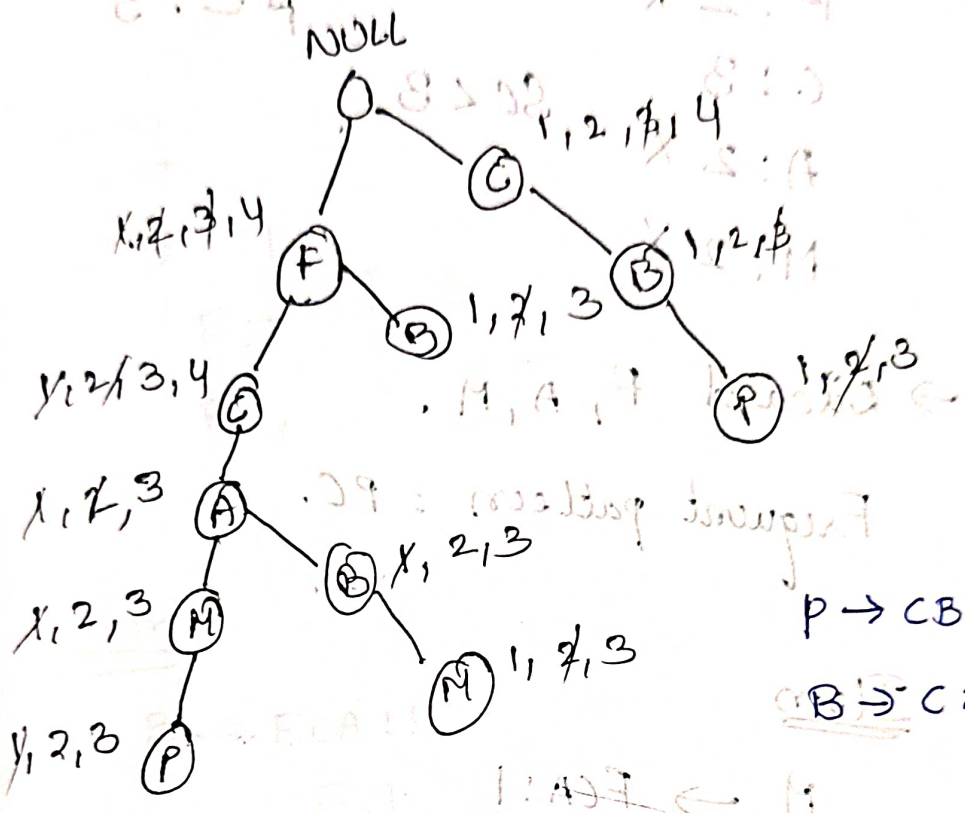
F, B.



B → F : 1

T4 C, B, P.

→ A new branch right side of the root node. because transaction does not start from F.



P → CB : 1

B → C : 1

T5

F, C, A, M, P.

F ↑ by 4.

C ↑ by 4

A ↑ by 3

M ↑ by 3

P ↑ by 3.

P → FCAM : 1

M → FCA : 1

A → FC : 1

C → F : 1

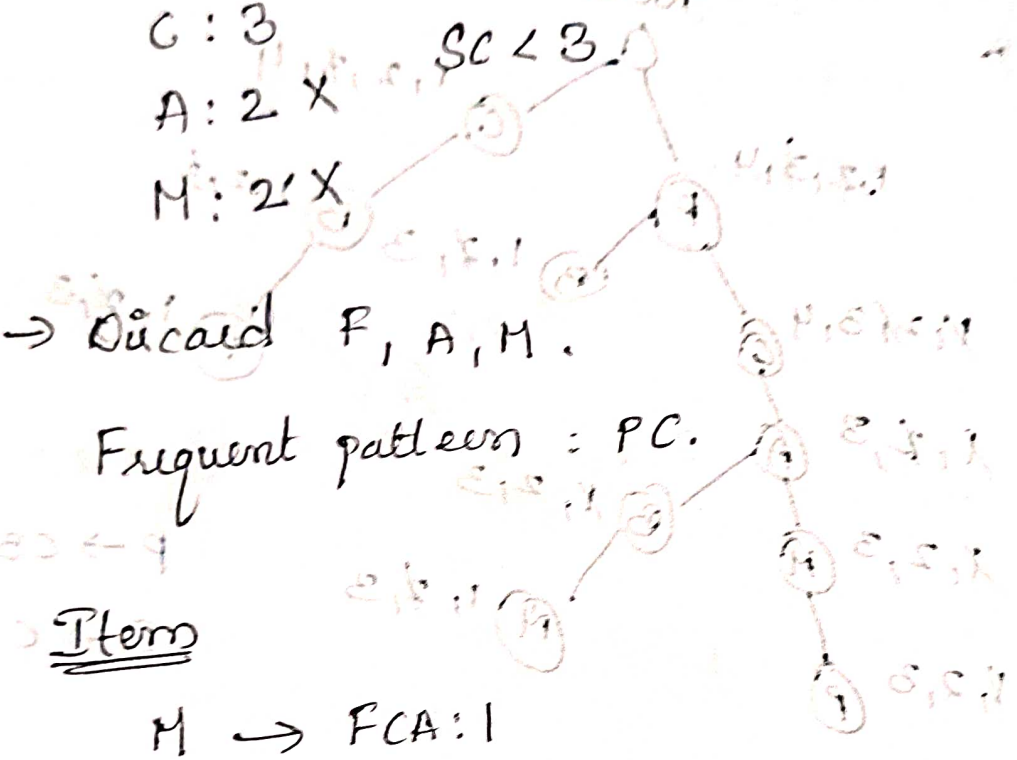
Item

P → FCAM:1
CB:1
FCAM:1

F:2 ✗
C:3
A:2 ✗
M:2 ✗

PC:3

SC < 3



→ Discard F, A, M.

Frequent pattern : PC.

Item

M → FCA:1
FCAB:1
FCA:1
F:3
C:3
A:3
B:1 ✗ SC < 1

F:4 ; C:4, A:4

FP = MF, MC, MA.

Item

A →
FC:1 F:3, C:3
FC:1 AF, AC → FP
FC:1

Item

C → F:1
F:1 F:3
F:1
FP = CF:3

Item

B → FCA:1
F:1
C:1

F:2, C:2, A:1

SC < 3

Discard F, C, A.

Frequent Patterns are

PC, MF, MC, MA, AF, AG, EF.

Apriori v/s FP

Apriori	FP
1) It generates exponential no. of CK.	It doesn't generate any candidate sets.
2) It performs self-join op on the generated candidate set	It doesn't perform self-join op.

Compact Representation of frequent Item Set

→ The no. of frequent itemsets generated by the 'Apriori Algo' can often be very large, so it is beneficial to identify a smaller representative set from which every frequent itemset can be derived.

↔ One such approach is using "Maximal frequent itemset".

Maximal Frequent Itemset

It is freq. itemset for which

None of its immediate superset are frequent. (or)

Immediate supersets are infrequent

$SC = 2$

Ex:-

TId	Items
1	a, b, d
2	a, c
3	a, b, c
4	a, d.

a - 4
b - 2
c - 2
d - 2

ab	ac	ad	bc	bd	cd
2	2	2	1	1	1

abc	abd	acd	bcd
1	1	0	0

a → ab, ac, ad.

2 — 2 2 2

'a', supersets all are frequents

So it is not Maximal.

b → ab, bc, bd

2 — 2 1 1

atleast 1 superset ab is fequent so Not Maximal

c → ac, bc, cd Not Maximal.
    2   1   1

atleast 1 superset ac is frequent so
                    Not Maximal.

d - ad, bd, cd Not Maximal.
    2   1   1.

ab — abc, abd
     1   1

both immediate supersets of ab are infrequent / not frequent so
ab is Maximal.

ac — abc, acd (Maximal)
     1   1

ad — abd, acd (Maximal)
     1   1

abd, acd are infrequent.

Maximal frequent Itemset :-

→ An itemset X is Maximal frequent Itemset only if

* X is frequent &
* No immediate superset of X is frequent. (or)
* None of its immediate superset are frequent

Closed frequent Itemset :-

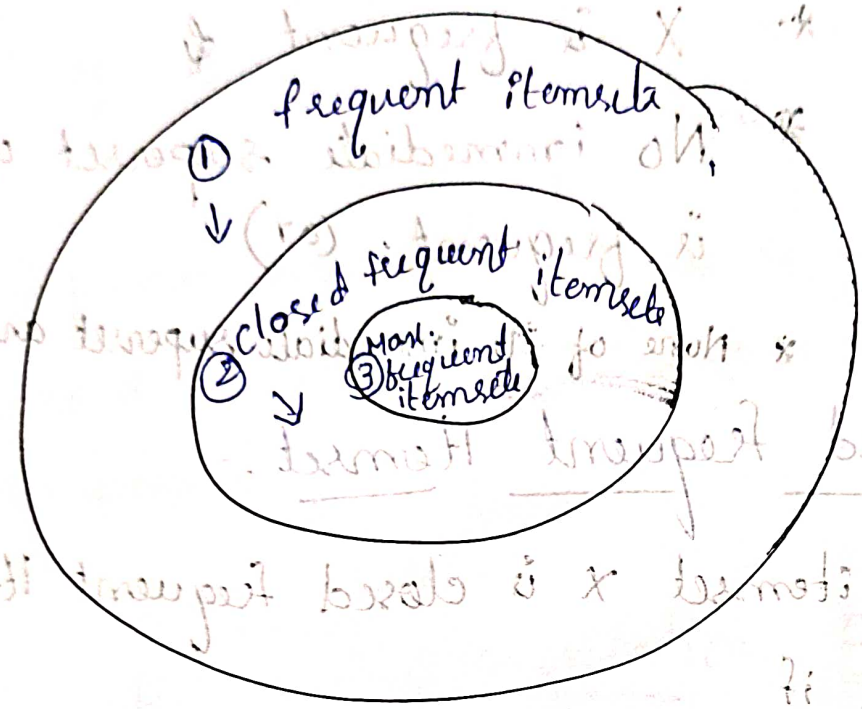
→ An itemset X is closed frequent Itemset only if

* X is frequent &
* No immediate superset of X has same support as X. (original item)

→ These closed frequent & maximal frequent itemset are useful when huge amt of data is used in association rule mining..

→ when the computation is very expensive & there is no interest to find additional

subsets that can be avoided by frequent itemset with max. length.



Ex:-

TID	list of items
T1	A, B, C, D
T2	A, B, C, D
T3	A, B, C
T4	B, C, D
T5	C, D

Min support count = 3

Find frequent, closed, Maximal Itemsets

Item	count
A	3
B	4
C	5
D	4

A, B, C, D are frequent because
SC (A, B, C, D) ≥ 3.

2 Items (Table 2)

Items	count
A, B	3
A, C	3
A, D	2 X SC < 3.
B, C	4
B, D	3
C, D	4

closed frequent items

item
A(3) (=) AB(3), AC(3), AD(2)
↓
count of A, subsets & their count.

A(3) ≠ 3, 3, 2 (NO)

So A is not closed.

→ A(3) Min. Support count 3.

A is not Maximal.

In A's immediate superset, itemset are present with min. SC i.e., 3 (frequent)

→ Item B

B(4) → AB(3), BC(4), BD(3)

B(4) = 3, 4, 3

B (sup count) is 4 its immediate superset support count (BC=4) is same as B SC.

B is not closed. so closed.

→ Immediate superset of B, Itemsets are present with min SC = 3 (frequent)

So B is not Maximal.

Item C

C(5) → AC(3), BC(4), CD(4)

C(5) = 3, 4, 4

So C is closed.

* SC of AC & BC are (3, 4) so C is not maximal.

Item D

D(4) → AD(2), BD(3), CD(4)

D(4) ⊇ 3, 4 not greater than 3, 4

So not closed.

SC of BD, CD are (3, 4)

SC ≥ 3, (yes)

So Not Maximal.

3 Itemsets (for Table form 3 itemsets)

{A, B, C} {A, B, D} {A, C, D}

{B, C, D}

Table 3

Items	count	
A, B, C	3	
A, B, D	2	X (AD) Not frequent. SC < 3.
A, C, D	2	X
B, C, D	3	

for item

AB superset are

AB (3) → ABC (3), ABD (2)

3 ≠ 3, 2 NO

Atleast one SC is equal = AB support count so closed.

* AB Not closed.

SC (ABC) ≥ 3 yes

Not Maximal.

Item AC

AC (3) → ABC (3), ACD (2)

3 ≠ 3 NO

SC (3) = SC (AC)

Not closed.

Not closed.

SC (ABC) ≥ 3 yes

Not Maximal.

Item CD

CD (4) → BCD (3)

4 ≠ 3 yes.

CD is closed.

BCD(3) ≥ 3 (support count)

CD: is not Maximal.

Now, for Table 3 itemsets form 4 itemset.

Items	count
A, B, C, D	2

In Table 3 Item 1

ABC(3) → ABCD(2)

↘ superset of ABC

3 ≥ 2 yes

ABC is closed

Superset ABCD(2) ≥ SC(3) .. NO /or infrequent

ABC is Maximal

BCD

BCD(3) → ABCD(2)

3 ≥ 2 yes

BCD is closed

ABCD(2) > SC(3) NO

ABCD is Maximal.

ABCD is not frequent

1 Itemset

Item	freq.	closed	Maximal
A(3)	✓	X	X
B(4)	✓	X	X
C(5)	✓	✓	X
D(4)	✓	X	X

2 Itemset:

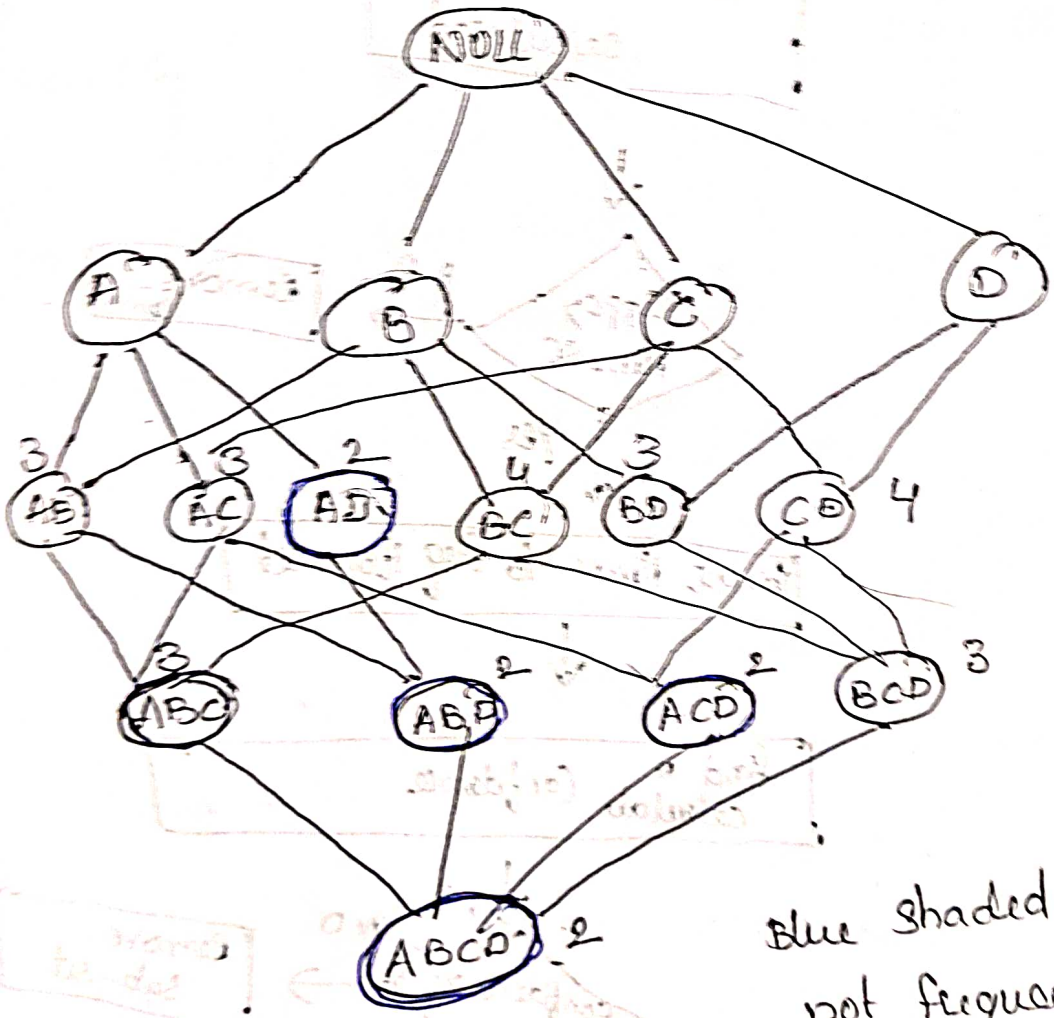
Item	freq	closed	Maximal
AB(3)	✓	X	X
AC(3)	✓	X	X
BC(4)	✓	✓	X
BD(3)	✓	X	X
CD(4)	✓	✓	X

3 - Itemset

Item	Freq	closed	Maximal
ABC(3)	✓	✓	✓
BCD(4)	✓	✓	✓

4 - Itemset

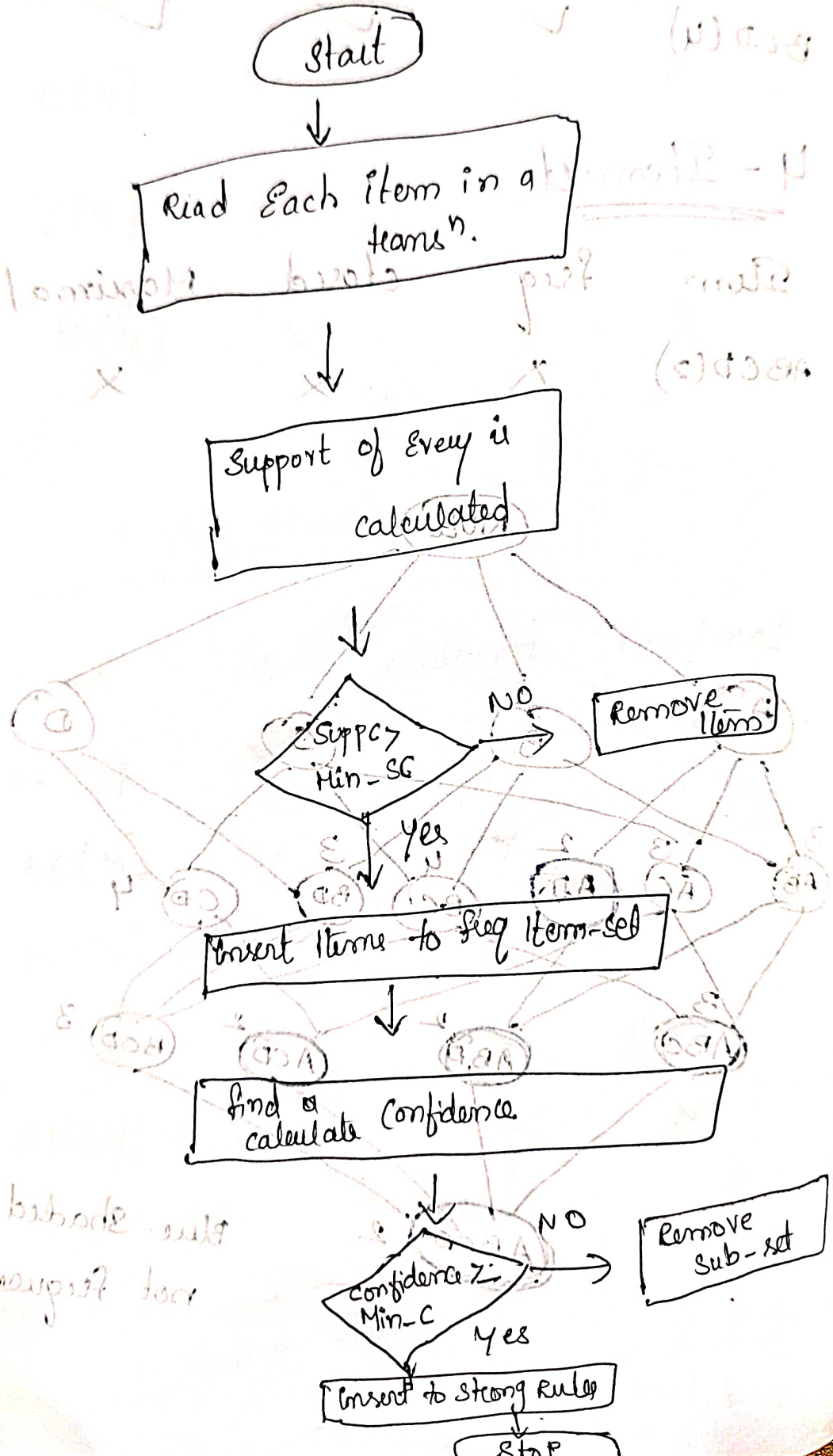
Item	Freq	closed	Maximal
ABCD(2)	X	X	X



Blue shaded are not frequent.

Flow chart of Apriori Algo

Frequent Itemset Mining Algo.



Apriori VS FP

	Apriori Algo	FP-Growth Algo.
space	Requires More space	Requires less space.
Time	Consume More time	Consume less time
Memory	uses Most of the Memory	uses less Memory
Accuracy	less accurate	More accurate
No. of Scans	Scan Each Candidate in the Itemset	Two scan only.
	self Joins	only sc.
	Join & prune	

The Partition Algorithm

→ This clustering method classifies the info into multiple groups based on the characteristics & similarity of the data.

→ It's the data analyst to specify the no. of clusters that has to be generated for the clustering methods.

→ In the partitioning method where D.B (D), that contains multiple(N) objects then the partitioning method constructs user-specified (K) partitions of the data in which each partitions of the data in which each partition represents a cluster & a ~~partition~~ ~~represents~~ particular region.

→ There are many algorithms that come under partitioning method

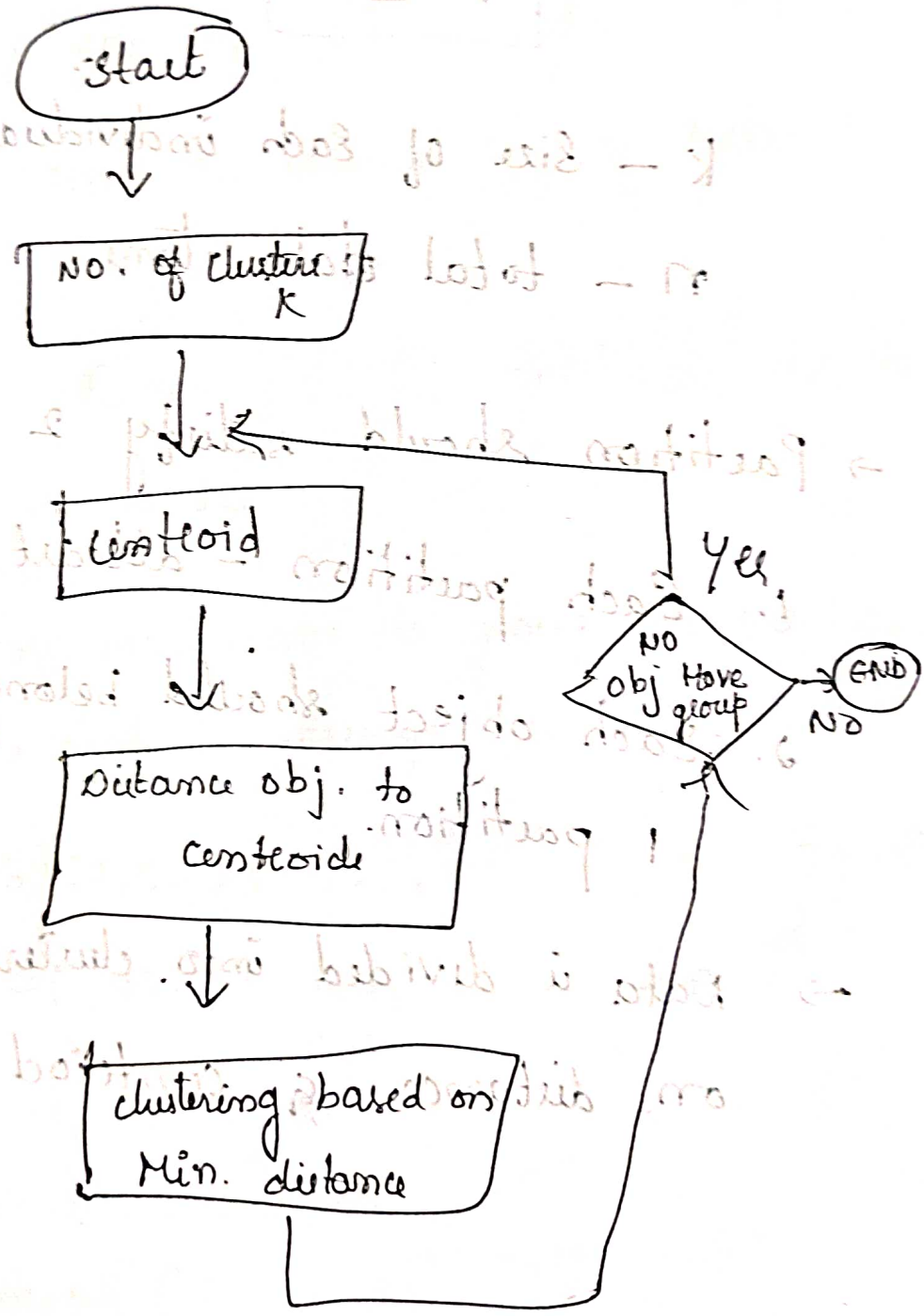
some of the popular ones are

K-mean,

PAM (K-Mediods)

CLARA algo (clustering large applications) etc.

K-Mean (A centroid based technique)



N data ple / objects / set / data ple $\xrightarrow{\text{divided}}$ K partitions or paths clusters.

Each partition represents one cluster

$$k \leq n$$

k – Size of each individual cluster

n – total data items

→ Partition should satisfy 2 rules.

1. Each partition – atleast 1 obj
2. Each object should belong to only 1 partition.

→ Data is divided into clusters based on distance & centroid value.

Example (1)

Steps: 1 Randomly assign k object from the dataset (D) as cluster centers C.



Step 2: Take the Mean value.

Step 3:- find the nearest no. of Mean & put in cluster.

Step 4 :- Repeat (1), (2) until we get the same Mean.

(1) Suppose we want to group the visitors to a website using their age as follows

16, 16, 17, 20, 20, 21, 21, 22, 23, 29,

36, 41, 42, 43, 44, 45, 61, 62, 66.

Initial cluster

K = 2 (2 clusters)

Centroid (C1) = 16 [16]

" (C2) = 22 [22]

→ C1, C2 are chosen randomly from the dataset (centroids)

Iteration - 1

K1 = {16, 16, 17} centroid

Nearest values to C1 [16]

K2 = {20, 20, 21, 21, 22, 23, 29, 31, 41, 42, 43, 44, 45, 61, 62, 66}

Mean of C1 = 16.33 $\left[\frac{16+16+17}{3}\right]$

Mean of C2 = 37.25

Iteration - 2

New Mean of C1, C2 are 16.33, 37.25

Nearest values to 16.33.

K1 = {16, 16, 17, 20, 20, 21, 21, 22, 23}

K2 = {29, 36, 41, 42, 43, 44, 45, 61, 62, 66}

Nearest value to New mean of C2 = 37.25

Mean of c1 = 19.55

Mean of c2 = 46.90

Iteration-3

K1 = {16, 16, 17, 20, 20, 21, 21, 22, 23, 29}

Neaest values of new Mean of c1 = 19.55

K2 = {36, 41, 42, 43, 44, 45, 61, 62, 66}

Neaest values of 46.90.

Means of c1 = 20.50

c2 = 48.89

Iteration-4

K1 = {16, 16, 17, 20, 20, 21, 21, 22, 23, 29}

K2 = {36, 41, 42, 43, 44, 45, 61, 62, 66}.

Mean of c1 = 20.50

c2 = 48.89.

K - cluster, c = centeiod

→ No change b/w Iteration 3 & 4, so we stop. ∴ we get the clusters (16-29) & (36-66) as 2 clusters using K-Mean algo.

Eg 2 arrange them in descending order

Datapts = K = {2, 4, 6, 9, 12, 16, 20, 24, 28}

No. of clusters = 2

c1 = 4
c2 = 12
} Pick any datapt as centeroids

cluster K1 = {2, 4, 6}

Values near to centeroid c1 = 4.

K2 = {9, 12, 16, 20, 24, 26}

Means of K1 = $\frac{2+4+6}{3}$ = 4

K2 = 18

Nearest values for c1 = 4

c2 = 18

K1 = {2, 4, 6, 9}

K2 = {12, 16, 20, 24, 26}

Mean of K1 = 5.25 = 5

K2 = 19.6 = 20

Nearest values for K1, K2 = 5, 20

K1 = {2, 4, 6, 9, 12} = 6.6 = 7

K2 = {16, 20, 24, 26} = 21.5 = 22

New cluster values c1 = 7

c2 = 22

Nearest values for 7 & 22 are

K1 = {2, 4, 6, 9, 12} = 6.6 = 7

K2 = {16, 20, 24, 26} = 21.5 = 22.

Acc. to step 3, we got same mean in

Iteration 2 & 3 i.e., 7 & 22.

final clusters are

$K1 = \{2, 4, 6, 9, 12\}$

$K2 = \{16, 20, 24, 26\}$.

Ex 3

Ht (x)	wt (y)
170	60
160	55
180	75
150	50
175	85
190	85

* considering 2 objects

Randomly pick any set from data & assign to centroid K1 & K2.

K1	170, 60
K2	160, 55

Now, 1st row belongs to K1, &

2nd " " K2

→ start Clustering from 3rd row.

→ Now, remaining values should be divided based on Euclidean distance

$$E.D = \sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$$

x_0 = value from table

x_c = centroid value.

for Row (3)

ED from C1 to (3)

$$K1 = \sqrt{(180-170)^2 + (75-60)^2}$$

$$= 18.02$$

from C2 to 3

$$K2 = \sqrt{(180-160)^2 + (75-55)^2}$$

$$= 28.28$$

Pick smaller distance among K1 & K2

18.02 < 28

K1 < K2

D3 ∈ K1

data pt 3 belongs to cluster K1.

Now, before we calculate for data pt 4 calculate new centroid:

data pt 3 belong to K1 so calculate new centroid for K1

Consider K1

just do addition for ED K1.

$$K1 = \left(\frac{180+170}{2}, \frac{75+60}{2}\right)$$ New centroids

$$= 175, 67.5$$

C1	175	67.5
C2	160	55

New centroid C1 = 175, 67.5

Distance for dt pt 4 using new centroids

$$K1 = \sqrt{(175-150)^2 + (67.5-50)^2}$$

$$= 30.51$$

$$K2 = \sqrt{(150-160)^2 + (50-55)^2}$$

$$= 11.18$$

K2 is Min so

d4 ∈ K2

datapt 4 ∈ cluster 2

Now calculate new centroid for K2

$$= \left(\frac{150+160}{2} , \frac{50+55}{2} \right)$$

155, 52.5

New centroids

C1	175	67.5
C2	155	52.5

Calulate ED for d5 using new centroids

$$K1 = \sqrt{(175-175)^2 + (65-67.5)^2}$$

$$= 2.5$$

$$K2 = \sqrt{(175-155)^2 + (65-67.5)^2}$$

$$= 23.58.$$

Min is 2.5

d5 ∈ K1

Change centroid for K1:

$$\left(\frac{175+175}{2}, \frac{65+67.5}{2}\right)$$

$$= 175, 66.25$$

New centroids

c1	175	66.25
c2	155	52.5

for d6:

$$EDC1 = \sqrt{(190-175)^2 + (85-66.25)^2}$$

$$= 24.01$$

$$C2 = \sqrt{(190-155)^2 + (85-52.5)^2}$$

$$= 47.76$$

Min is 24.01 so

d6 ∈ K1

Now, calculate new centroids for K1 since d6 ∈ K1.

→ If d6 ∈ K2, then we need to calculate new centroids for K2.

$$C1 = \left(\frac{190+175}{2}, \frac{85+66.25}{2} \right)$$

$$= \frac{182.5}{2}, 75.6$$

C1	182.5	75.6
C2	155	52.5

$$K1 = \{1, 3, 5, 6\}$$

$$K2 \quad \{2, 4\}$$

cluster 1 = { 170, 60
180, 75
175, 65
190, 85 }

cluster 2 = { 160, 55
150, 50 }

Eg 4

Ht (x)	wt (y)
185	72
170	56
168	60
179	68
182	72
188	77
180	71
183	84
180	67
180	76
177	76

Ans

cluster 1 = {1, 4, 5, 6, 7, 8, 9, 10, 11, 12}

cluster 2 = {2, 3}.

UNIT - IV

Classification: Problem Definition, General approaches to solving a classification problem, Evaluation of classifiers, classification techniques, Decision Trees - Decision tree Construction, Methods for Expressing attribute test conditions, Measures for selecting the Best split, Algo for decision tree Induction, Naive Bayes classifier, Bayesian Belief N/w, K-Nearest neighbours classification Algo & characteristics.

Supervised data

→ Supervised learning, has the supervisor as a teacher.

→ when we teach or train the machine using data i.e., well labeled, means some data is already tagged with correct ans.

→ After that, the machine is provided with a new set of examples, so that the machine is provided with new set of data (Ex) so that the supervised learning algo. analyses the set of training Ex. & produces a correct outcome from labeled data.

Ex:. fruit ex
shape round
color red
classifier → apple.

→ Train the machine with diff kind of fruits. one by one with features.

→ After training the data, given a new fruit & asked to identify.

→ Since the machine has already learned the things from previous data & this time has to use it

classifications:-

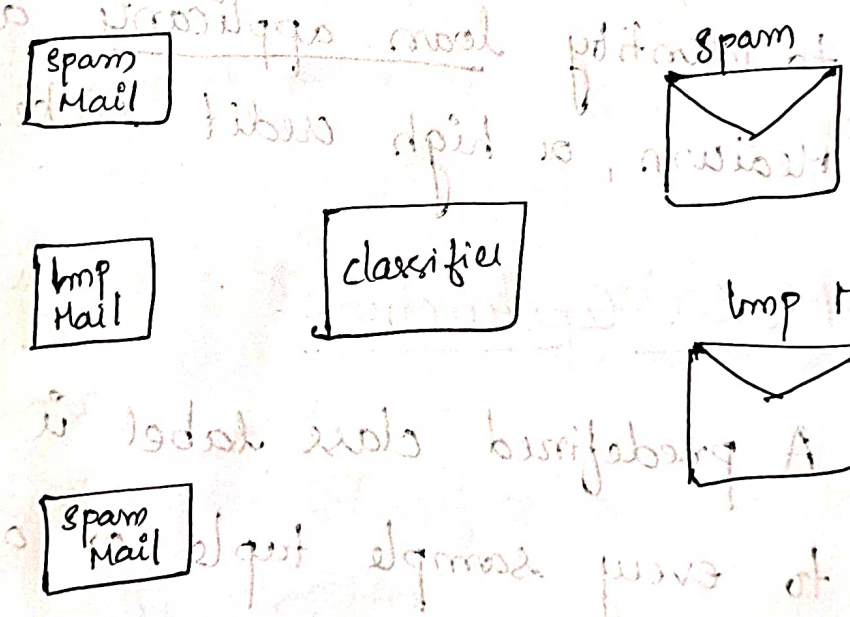
→ finding a Good model that is used to predict the class of objects whose class label is unknown.

(or)

→ Categorization of new data with the help of current past data.

Ex:- Grouping of the patients based on Medical records.

Mail Ex



Eg2 For the products in the o/p, we give feedback like 5*, 2*. These feedbacks has to be classified to +ve feedback or -ve.

Classification has 2 steps

* Model Construction
* Model usage.

→ classification is a DM fn. that assigns items in a collection to target categories to (or) classes.

→ the goal of classification is to accurately predict the target class for each case in the data.

Ex:- classification model could be used to identify loan applicants as Low, Medium, or high credit risks.

Model Construction:

→ A predefined class label is assigned to every sample tuple or object.

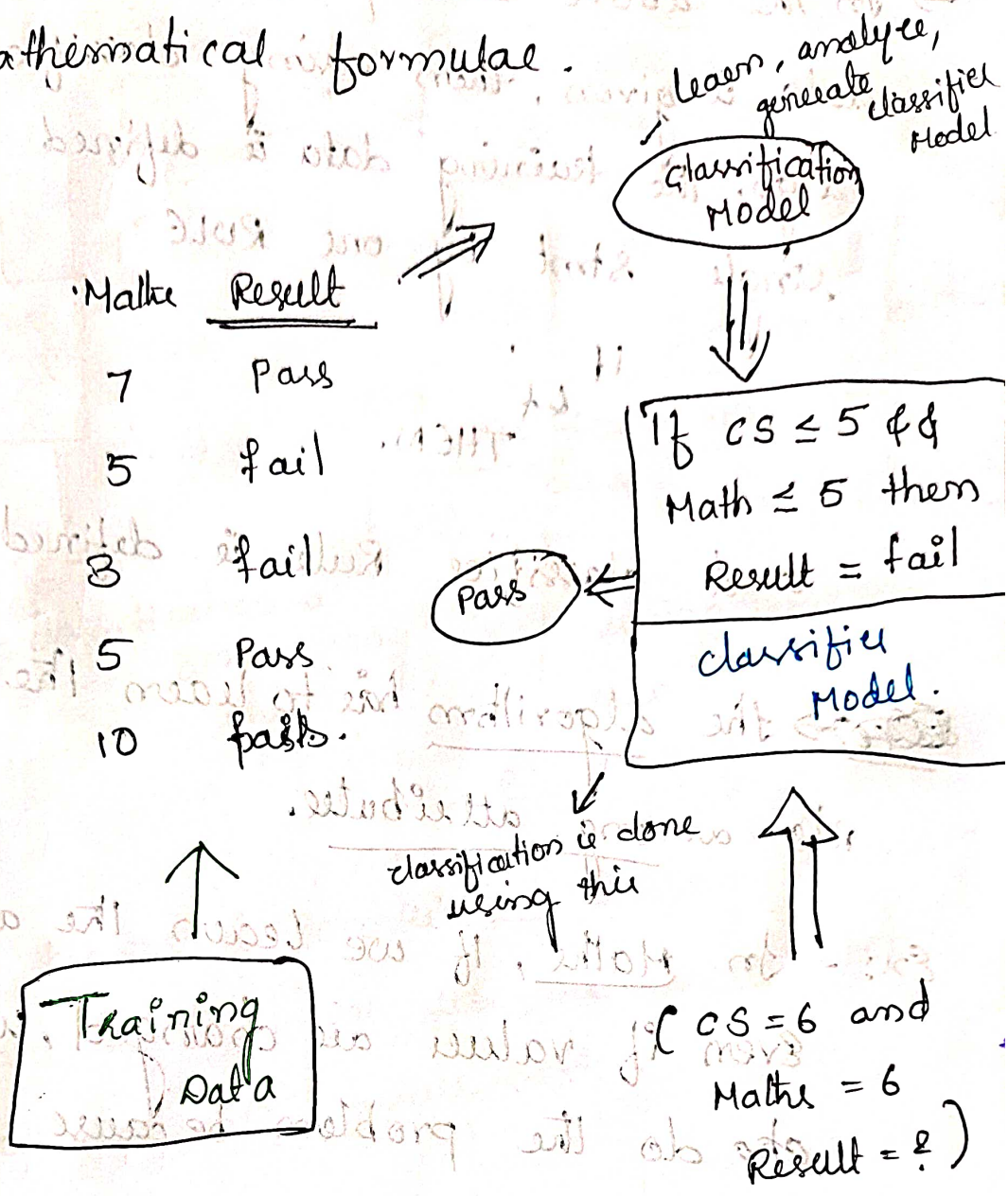
→ These tuples or subsets data are known as "training data set".

→ The constructed model, which is based on training set is represented as

classification rule, decision trees or mathematical formulae.

Ex:-

CS	Maths	Result
6	7	Pass
5	5	fail
2	8	fail
7	5	Pass
9	10	pass.



→ Step-1 is the learning step or learning phase.

→ In this step classification algorithms build the classifier.

→ The classifier is build from training set made up of DB tuples & their associated class labels.

→ In the above example, assume Training data is given. Then using classification algo. the training data is defined in single stmt by one RULE

IF && THEN.

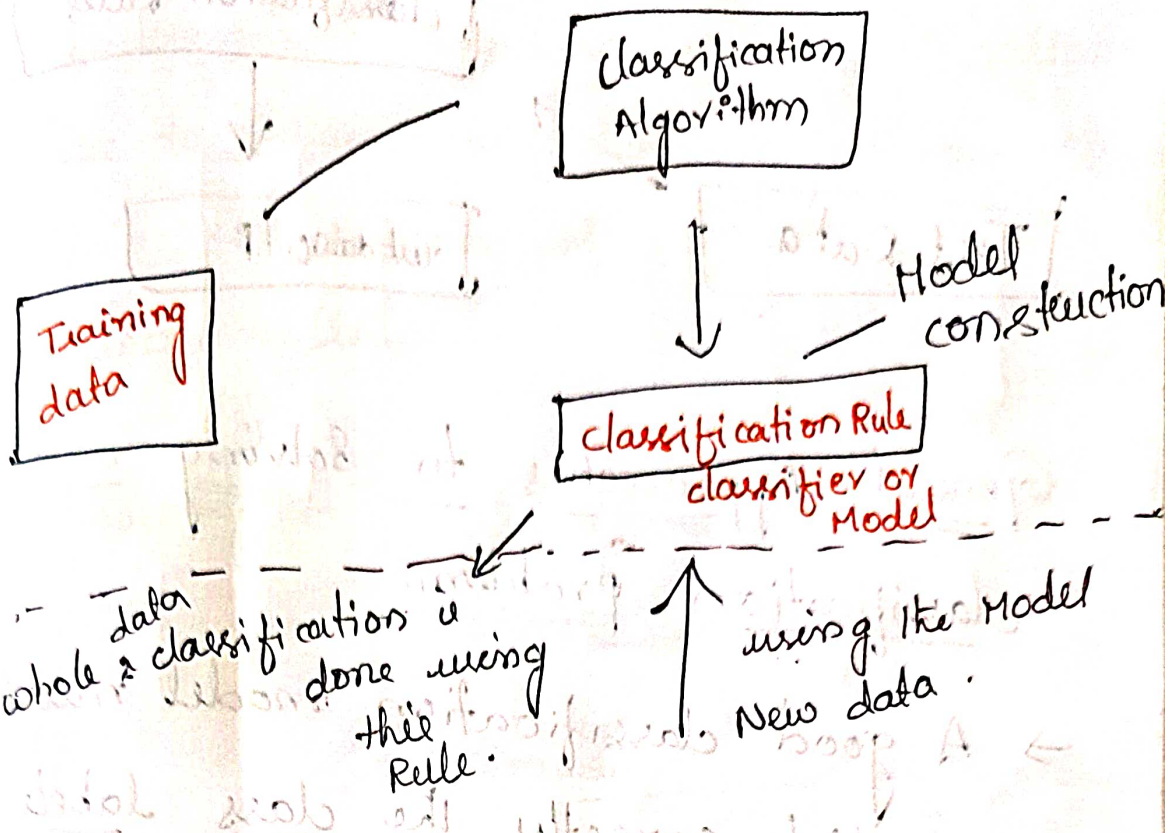
and classifier Rule is defined.

→ the algorithm has to learn the patterns & rls among attributes.

Ex:- In Maths, If we learn the algorithm even if values are changed, we can also do the problem because we learnt the Model.

→ In the same way, classification will learn, analyse & generate the classifier model.

Ex 2 :- In univ, DB can be classified into type above avg & below avg.



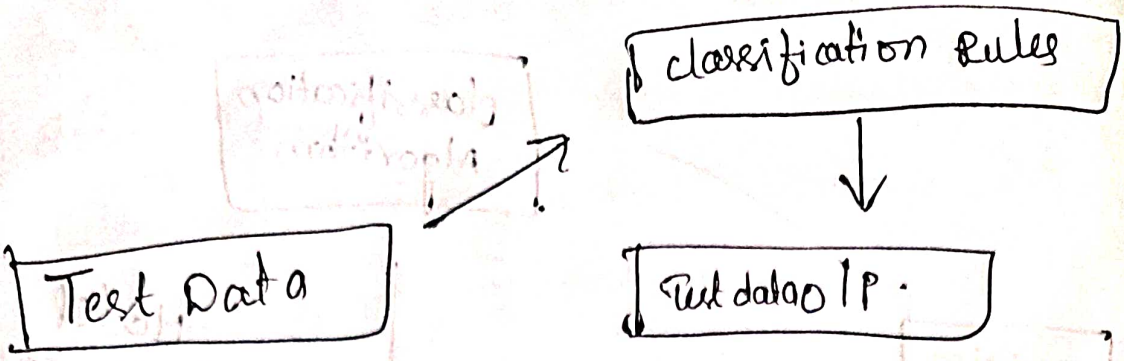
→ A Model is constructed using training data & then we use the model for new data or future data.

Model usage :-

→ In this step, classifier is used for classification.

→ Here the test data is used to estimate the accuracy of classification rule.

→ The classification Rule can be applied to the new data tuples if the accuracy is considered acceptable.



General Approaches to Solving a classification problem.

→ A good classification model must predict correctly the class labels of records it has never seen before

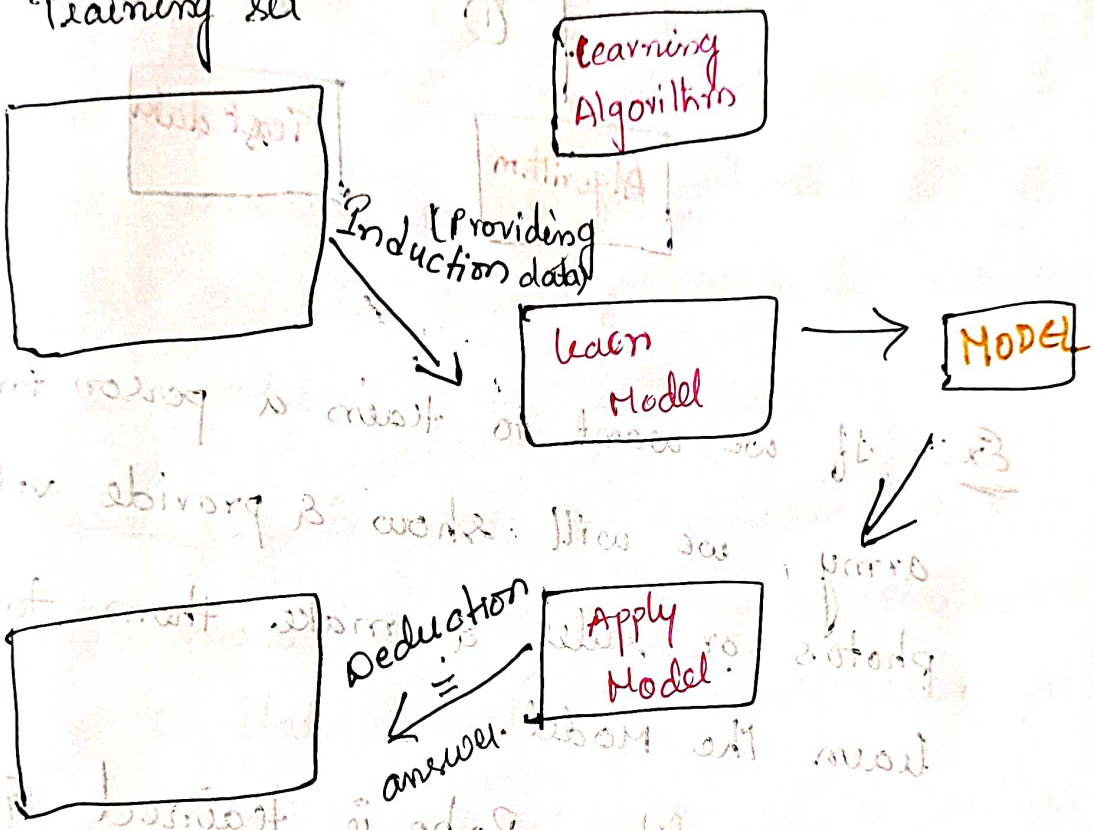
→ A General Strategy to solving a classification problem is that

* first, the i/p data is divided into two disjoint sets, known as the Training set & test set, respectively

Training Set: will be used for building a classification model.

→ the created model is later applied to the test set to predict the class label of each test record.

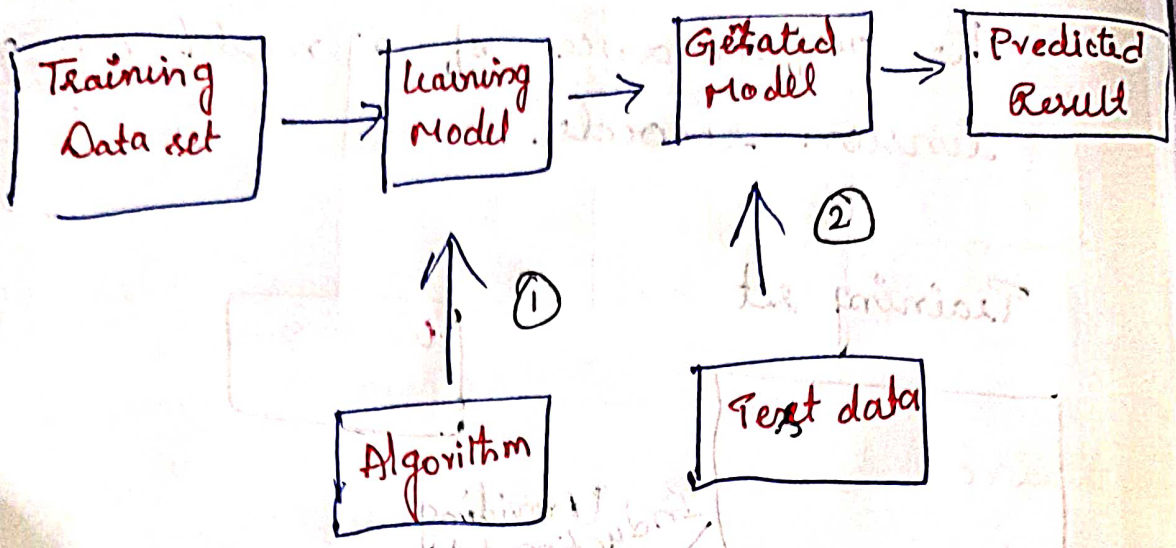
→ The strategy of dividing the data into independent training & test sets allows us to obtain an unbiased estimate of the performance of a model on previously unseen records.



deduction = answer, assumption.

→ Evaluation of the performance of the classification model is based upon the no. of test records predicted correctly & wrongly by the model.

→ The counts are tabulated in a table known as a confusion matrix.



Ex:- If we want to train a person for a army, we will show & provide videos photos or rules & make them to learn the model.

→ In ROBO film, Robo is trained & made ROBO to learn model using some algorithms.

→ After that a new model is generated for that Test data is given to check accuracy of the model & checks the results for accuracy.

Evaluation of classifiers:-

→ Once the model is created, they need to be tested to findout how effectiveness of the model based on metrics & dataset.

→ As the model are tuned to the training set during the process of training the ML model, if their performance is evaluated on the training data the o/p of the model will be wrong.

→ Therefore in order to estimate the error, the model is required to be evaluated on the data which it hasn't seen yet.

→ Evaluate the model performance on test data.

→ for performance measures, A confusion Matrix is used to understand the performance.

→ A confusion Matrix is a table that is often used to describe the performance of

a classification model or a classifier.

	Actually +ve (1)	Actually -ve (0)
Predicted +ve (1)	TP	FP
Predicted -ve (0)	FN	TN

TP - True +ve
FP - False +ve
FN - False -ve
TN - True -ve

Ex: Assume Model is evaluated using 180 Samples

⇒ Binary classification of the task. (That means 0 & 1 (actual & predicted).

180	Actual: Positive	Actual: Negative
Predicted +ve	60 ✓	30 X
Predicted -ve	40 X	50 ✓

TP = +ve
TN = -ve

⇒ Here we are checking created model using 180 test data. we need to check how model is classifying the data to respective class.

→ Ex:- Consider 180 corona patients list. classifier has to classify +ve to that +ve class & -ve class. (Prediction).

→ out of total 180, 100 are +ve samples, 60 are correctly classified as +ve & the remaining 40 are Mis-classified as -ve.

→ out of total 80 -ve samples, 50 are classified as -ve while the remaining 30 are wrongly predicted as +ve

→ Means, 30 patients are belongs to class +ve but predicted as -ve.

→ 40 patients are belongs to class -ve but wrongly predicted as +ve.

* 60 are correctly classified.

* 30 samples are ~~belor~~ tested -ve belongs to -ve class but wrongly predicted as +ve (by the Model)

$$\text{Accuracy} = \frac{(TP + TN)}{TP + TN + FP + FN}$$

Accuracy — overall, how often is the classifier correct

$$= \frac{60 + 150}{180}$$

$$= 0.61$$

Misclassification Rate — How often it's wrong.

$$\frac{FP + FN}{\text{Total}}$$

$$= \frac{30 + 40}{180}$$

$$= 0.38$$

calculate precision & Recall.

Equivalent = $\boxed{1 - \text{Accuracy}}$

Also known AS "ERROR RATE".

→ In certain cases, calculating accuracy is not a good measure.

→ for imbalanced dataset, accuracy is not a good measures for assessing model performance.

Ex:-

N=1000	Actual: cancer	Actual: No cancer	
Predicted: cancer	0 (TP)	0 (FP)	0
Predicted: No cancer	50 (FN)	950 (TN)	1000
	50	950	

Here, the above confusion matrix for cancer dataset is imbalance, because TP & FP are 0. remaining TN, FN are 950 & 50 respectively.

→ All data in classes are not balanced.

→ Out of 1000, 950 are non-cancers 50 are cancers.

Classification Techniques

→ There are many classification techniques which have originated from the ML field.

→ These are need to be scalable for large data sets. Some of the classification techniques are :

* Decision Trees.
* Naive - Bayes classifier
* Bayesian Belief N/w.
* K - Nearest Neighbour.

→ Classification can be performed on structured or unstructured data.

→ classification is a technique where we categorize data into a given no. of classes.

→ The main goal of a classification problem is to identify the category/

class to which new data will fall under.

Few terms w.r.t classification.

classifier :- An algo. that maps the i/p data to a specific category.

classification Model :- A CM tries to draw some conclusion from the i/p values given for training. It will predict the class labels / categories for the new data.

features :- A feature is an individual measurable property of a phenomenon being observed.

Binary classification :- classification task with 2 possible outcomes

Eg :- Gender (M/F).

Multi-class classification :- classification with more than 2 classes. In MCC each sample is assigned to one & one

target label.

Eg:- An animal can be cat or dog but not both at the same time.

Multi-label classification:- classification task where each sample is mapped to a set of target ~~class~~ labels. (more than one class).

Eg:- A news article can be about sport, a person, & loc at the same time.

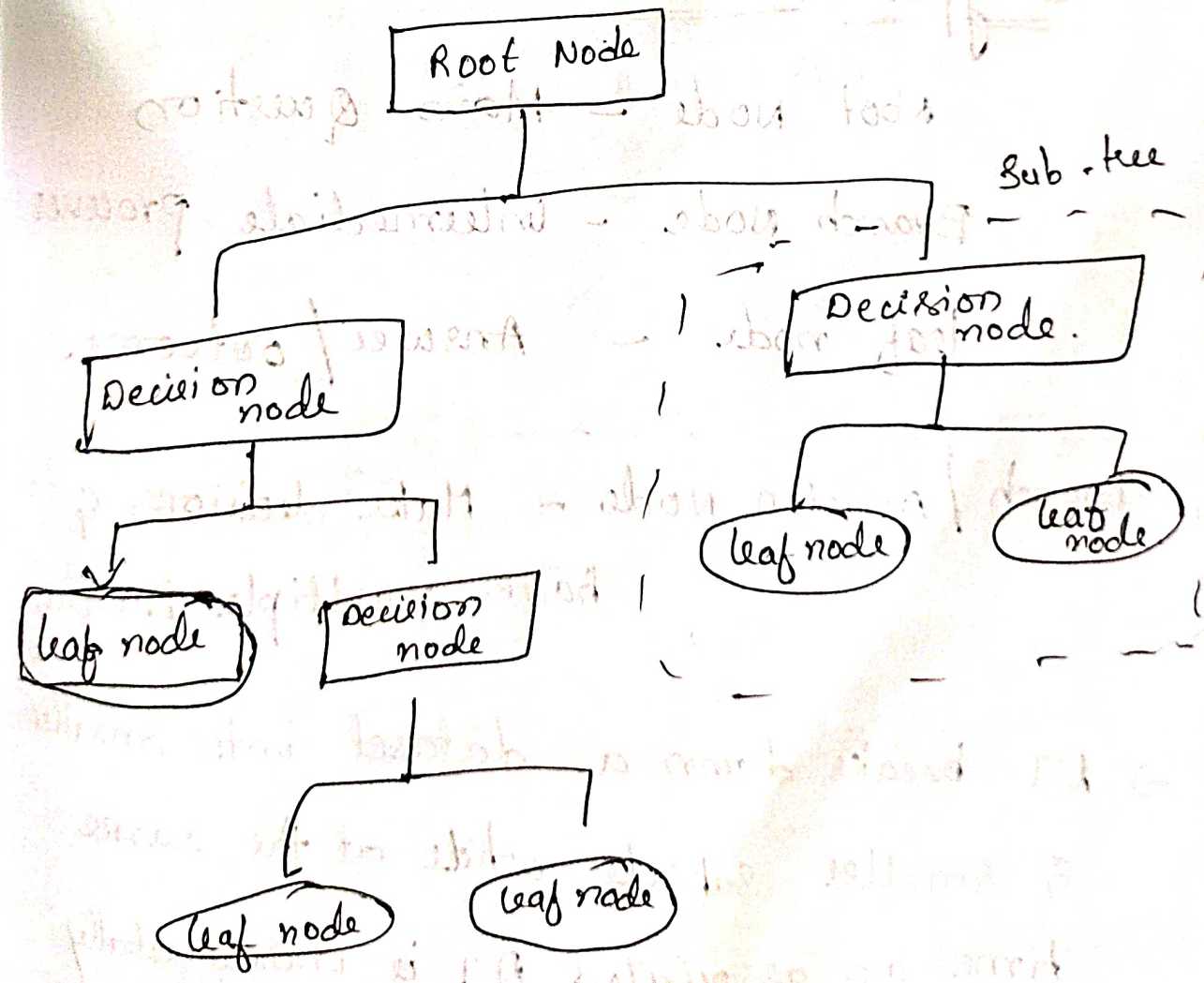
⇒ The following are the steps involved in building classification model.

* Initialize the classifier to be used
* Train the classifier, All the classifiers should fit the given model.
* Predict the target, given an unlabeled observation x, the predict(x returns the predicted label y.
* Evaluate the classifier Model.

Decision Trees:-

→ DT is powerful tool for classification & prediction.

A DT is a flowchart like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, & each leaf node (terminal node) holds a class label.



Top of the node – Root node 

Branch node – decision node 

leaf node – final outcome 

→ Each internal node holds a feature

depth of the tree – ht. of the tree.

→ It support in taking decisions.

→ It defines the rules visually in form of tree.

3 types of nodes

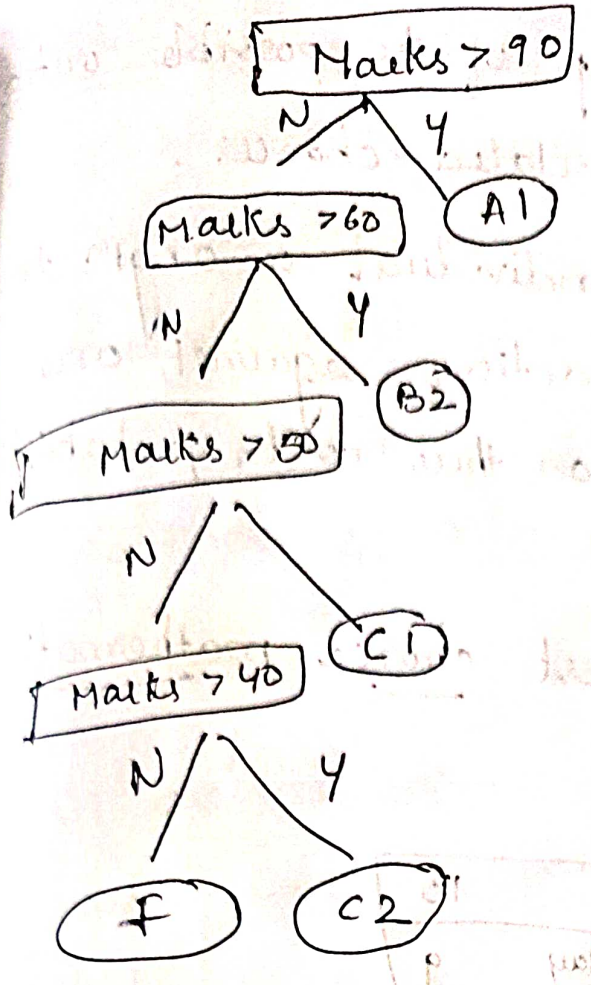
Root Node – Main Question

Branch Node – Intermediate processes

leaf node – Answer / outcome.

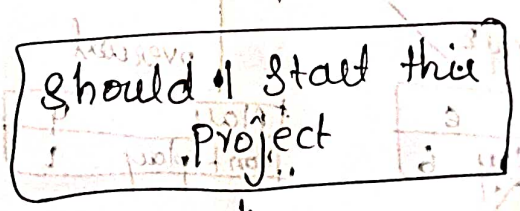
Branch / Decision Node – Make decisions & have multiple branches

→ DT breaks down a dataset into smaller & smaller subsets while at the same time an associated DT is incrementally developed.



Decision tree for grade calculation

Project Mgmt Decision Helper



Yes

Is it imp for the company

NO

Ast not to be involved

too risk

Yes

Yes

Consider not to be involved

NO

Start the proj.

NO

Is it imp for my career

Yes

Start the project.

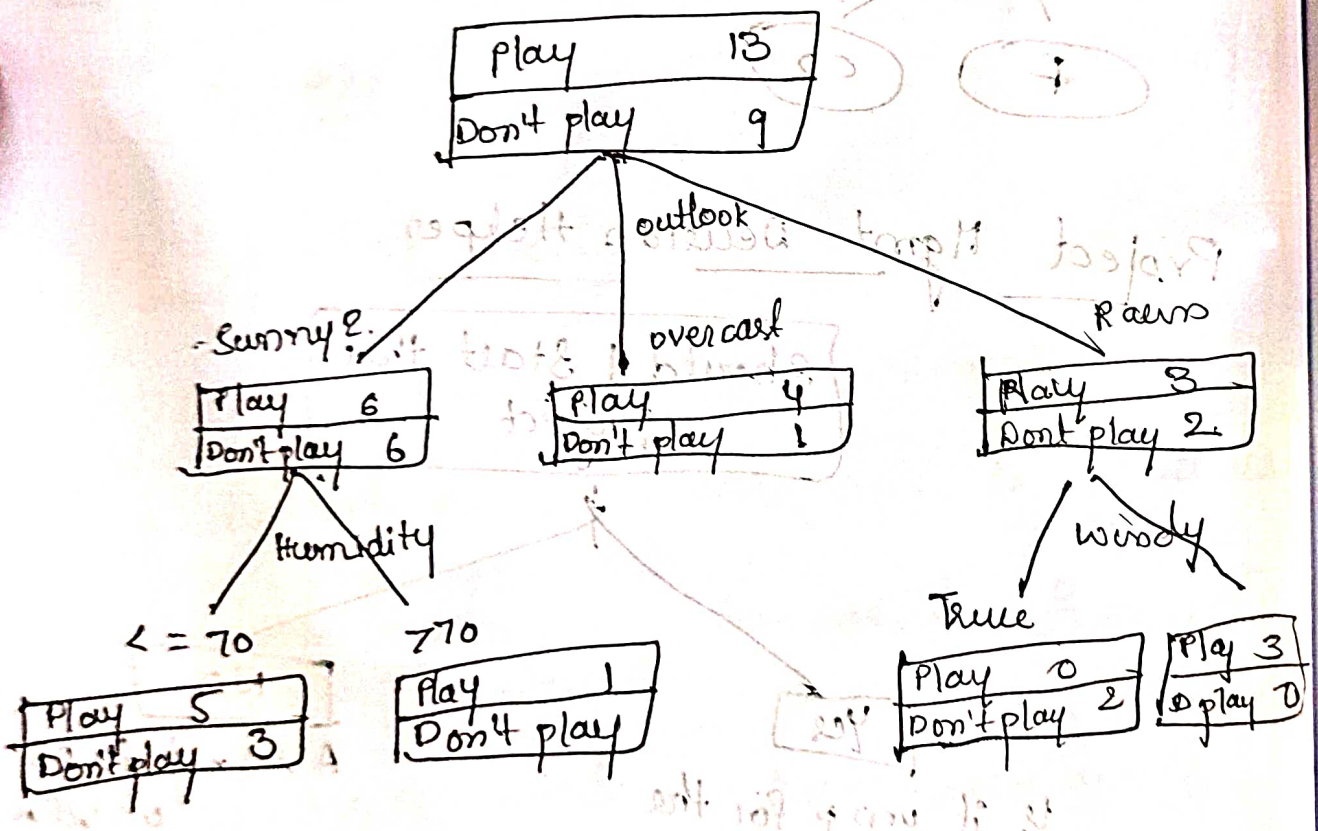
NO

Don't start

→ A DT is a map of the possible outcomes of a series of related choices.

→ It allows an individual or orgtn to weigh possible actions against one another based on their cost, probabilities & benefits.

→ It predicts best choice mathematically



→ A decision tree typically starts with a single node, which branches into possible outcomes.

Adv

→ DT generate understandable rules

→ DT perform classification without requiring much compilation.

→ DT are cabable of handling both continous & categorical variables.

Dis Adv

→ DT are less approciate for estimation tasks where the goal is to predict the value of a continous attribute.

→ DT can be computational expensive to train. The process of growing a decision tree is Expensive. At each node, Each candidate splitting field must be sorted before its best split can be found.

→ The final result ~~of~~ is a tree with decision nodes & leaf nodes.

* Terms Related to decision Tree.

Entropy (quality).

→ It is a measure of the randomness in the info. being processed.

→ The higher the entropy, the harder it's to draw any conclusion from that info.

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

or

$$E(S)$$

Information Gain :-

→ It is the amount of info. gained about a variable.

→ It is considered as the difference b/w entropy of parent node & weighted avg. Entropy of child nodes.

$$IG(S, A) = H(S) - H(S, A)$$

(or)

$$IG(S, A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

* There are 2 algorithm to build a decision tree.

1. CART (Classification & Regression Trees)

2. ID3 (Iterative Dichotomiser 3)

Constructing Decision Tree using ID3

NO.	outlook	Temperature	Humidity	Wind	play
1	Sunny	Hot	High	weak	NO.
2.	Sunny	Hot	High	Strong	NO
3.	overcast	Hot	High	weak	yes
4.	Rain	Mild	High	weak	yes
5.	Rain	cool	Normal	weak	yes
6.	Rain	cool	Normal	strong	NO
7.	overcast	cool	Normal	strong	yes
8.	Sunny	Mild	High	weak	NO
9.	Sunny	cool	Normal	weak	yes
10	Rain	Mild	Normal	weak	yes
11.	Sunny	Mild	Normal	strong	yes
12.	overcast	Mild	High	strong	yes
13.	overcast	Hot	Normal	weak	yes
14.	Rain	Mild	High	strong	NO.

Target class = play (Yes / No)

or

class attribute

Step 1:-

. find the parent node for our decision tree.

→ for that follow the steps

find Entropy of class variable or for Entire dataset

→ In weather data set, there are 14 instances.

→ In class attribute play, there are

9 - yes

5 - NO.

→ Based on it calculate Entropy.

$$E(S) = -\left[\left(\frac{9}{14}\right)\log\left(\frac{9}{14}\right) + \frac{5}{14}\log\left(\frac{5}{14}\right)\right]$$

$\frac{9}{14}$ → 9 yes out of 14

$\frac{5}{14}$ → 5 NO out of 14

S → whole dataset

→ there are 4 attributes. we need to calculate entropy for each individual attribute. Based on ~~Entrophy~~ IG value (Max), Root node is selected.

→ attribute having Max ~~Entrophy~~ IG will become Root Node.

Attribute Outlook

outlook has 3 ~~attribute~~ value

Sunny, overcast, Rain.

	Play		tot
	Yes	NO	
Sunny	2	3	5
overcast	4	0	4
Rain	3	2	5
			14

→ for Sunny, in class attribute play there are 3 yes & 2 NO

$E(S_{sunny}) =$

$E(3,2) = -\left[\frac{3}{5}\log\left(\frac{3}{5}\right) + \frac{2}{5}\log\left(\frac{2}{5}\right)\right]$

3 yes / 5 total for sunny; 2 NO.

$= 0.97.$

$E(\text{overcast}) =$

$E(4,0) = -\left[\frac{4}{4}\log\left(\frac{4}{4}\right) + 0\right]$

$= 0.$

$E(\text{Rainy}) =$

$E(2,3) = -\left[\frac{2}{5}\log\left(\frac{2}{5}\right) + \frac{3}{5}\log\left(\frac{3}{5}\right)\right]$

2 \ total Rainy days

$= 0.97.$

Total ~~Entropy~~ Information Gain of outlook w.r.t to:

Entropy whole data is

$E(S, \text{outlook}) = \frac{E(S)}{}$

*use logarithm calculator Rapidtables.com & Expand.

~~$E(S, \text{outlook}) = E(S) + E(O) + E(R)$~~

Total Information Gain of attribute outlook w.r.t whole dataset.

$$IG(S, \text{outlook}) = E(S) - \sum_{v \in (\text{Sunny, overcast, rain})} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

$E(S) = 0.97$

$$= 0.94 - \frac{5}{14} E(S_{\text{sunny}}) + \frac{4}{14} \text{Entropy}(S_{\text{overcast}})$$

5 are sunny out of 14

$$+ \frac{5}{14} \text{Entropy}(S_{\text{rain}})$$

$$= 0.97 - \frac{5}{14} * 0.97 + 0. + 0.35 * 0.97$$

$E(S_{\text{sunny}}) = 0.97$

$E(S_{\text{overcast}}) = 0$

$E(S_{\text{rain}}) = 0.97.$

$= 0.97 - 0.9$

$= 0.247.$

Calculate Entropy for attribute Temp

Values are Hot, Mild, cold

	Yes	NO	Tot
Hot	2	2	4
Mild	4	2	6
cold	3	1	4

$E_{Hot} = E(2,2) =$ (1 Yes, 1 NO)

$$-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$= 1$$

$$E_{Mild} = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$$

$$= 0.913$$

$$E_{cold} = E(3,1)$$

$$= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2 1/4$$

$$= 0.8113$$

$$IG(S, Temp) = E(S) - \sum_{v \in (H,M,C)} \frac{|S_v|}{|S|} E(S)$$

$E(S)$ = Entropy of class attribute

$= 0.94$

IG for attribute Temp w.r.t whole dataset.

$$= 0.94 - \frac{4}{14} * 1.0 - \frac{6}{14} * 0.91 - \frac{4}{14} * 0.81$$

- $\frac{4}{14}$: 4 hot out of 14
- $\frac{6}{14}$: 6 mild
- $\frac{4}{14}$: 4 cold out of 14. in Temp attribute column

$= 0.02$

Entropy for attribute = Humidity

Value = High, Normal.

	Yes	NO	Tot
High	3	4	7
Normal	6	1	7
			14

$E(S_{High}) = E(3,4)$

$$-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$$

$$= 0.98$$

$E(S_{Normal}) = E(6,1)$

$$-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$

$$= 0.5$$

$IG(S, Humidity) =$

$$E(S) - \sum_{VE(High, Normal)} \frac{|S_v|}{|S|} E(S)$$

$$= 0.94 - \frac{7}{14} * 0.98 - \frac{7}{14} * 0.59$$

$$= 0.15$$

Attribute Wind

Values = Strong, weak

	yes	no	tot
strong	3	3	6
weak	6	2	8
			14

$E(S_{strong}) = E(3,3)$

$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$

$= 1$

$E(S_{weak}) = E(6,2)$

$= -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$

$= 0.8113$

$IG(S, Wind) = E(S) - \sum_{v \in (S,w)} \frac{|S_v|}{|S|} E(S)$

$= 0.94 - \frac{6}{14} * 1.0 - \frac{8}{14} * 0.8113$

$= 0.047$

IG (S, outlook) = 0.246

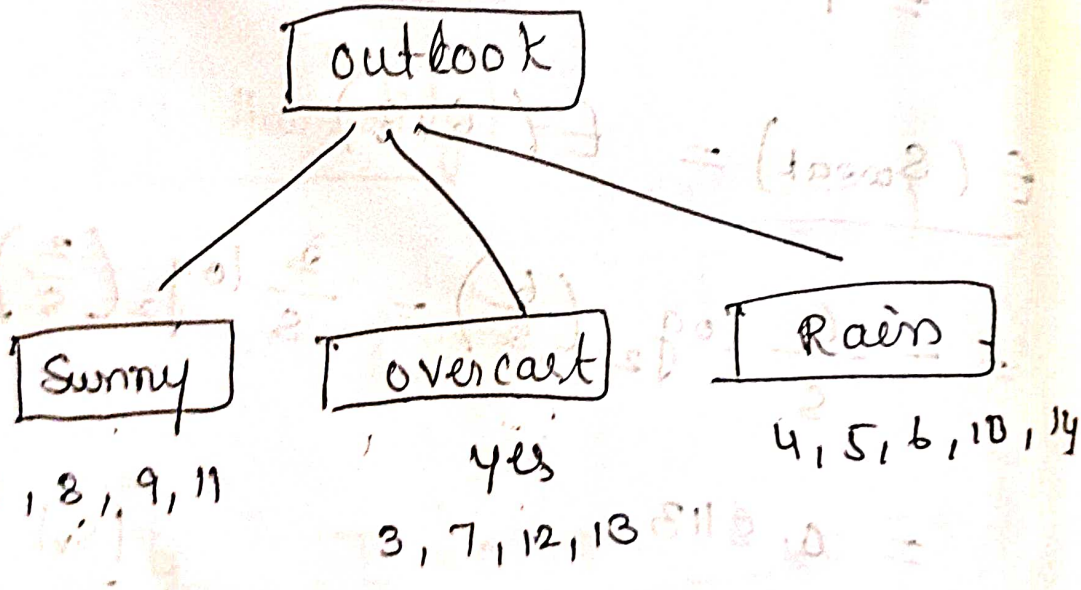
IG (S, Temp) = 0.02

IG (S, Humidity) = 0.15

IG (S, wind) = 0.047

Max IG value = Root Node

outlook = 0.24 Max out of all attributes



→ for overcast, all value in class attribute play are 'yes' so can directly write yes.

→ for sunny, 2 are yes } in class
3 are NO } attribute

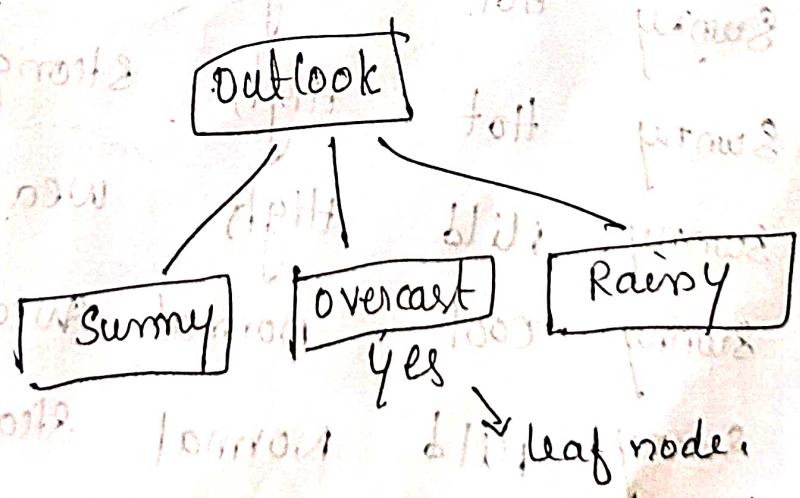
→ So here for sunny node, we need to expand.

→ In ~~overcast~~ Rainy also, 3 yes } class
2 NO } attribute

→ Rainy node has to be expand.

overcast = leaf node

→ That means, if outlook is overcast they can play.



→ The next step is to find the next node for sunny and rainy in our decision tree.

→ we have to determine attribute (Temp. Humidity, wind) which has higher IG.

for parent Entropy E (Sunny)

E (Sunny) = 0.97 (calculated previously)

→ we need to find next node for Sunny so, considering Entropy of Sunny.

→ Calculate IG of Temp. w.r.t Sunny.

outlook	temp	Humidity	wind	play
Sunny	Hot	High	weak	No
Sunny	Hot	High	Steong	No
Sunny	Mild	High	weak	No
Sunny	cool	Normal	weak	yes
Sunny	Mild	Normal	Steong	yes.

→ We are Expanding only Branch Node Sunny so, consider only Sunny & their respective temp values.

values of Temp w.r.t Sunny are

Hot Mild cool.

	Yes	NO	Tot
Hot	0	2	2
Mild	1	1	2
cool	1	0	1
			5

$E(Hot) = E(0,2)$

$= 0$

$E(Mild) = -\frac{1}{2}\log_2 1/2 - \frac{1}{2}\log_2 1/2$

1 - Yes, 1 - NO

1 Yes
2 tot.

$= 0.4$

$E(cool) = E(1,0)$

$= 0$

$IG(Sunny, Temp) =$

$\frac{2}{5} * 0 + \frac{1}{5} * 0.4 + \frac{2}{5} * 0$

2 hot in temp out of 5 sunny

1 Mild in temp column w.r.t sunny

$= 0.57$

$I G(\text{Sunny, Temperature}) = 0.971 - 0.4$

$= 0.571.$

Values of Humidity w.r.t to Sunny

	Yes	NO	tot
High	0	3	3
Normal	2	0	2

$E(S_{High}) = E(0,3)$

$= 0$

$E(S_{Normal}) = E(2,0)$

$= 0.$

$IG = E(S) - \frac{|S_v|}{|S|} E(S)$

$E(S_{High})$

$= 0.97 - \frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} E(S_{Normal})$

3 High out 5 sunny, 2 Normal in Humidity attribute w.r.t Sunny

$= 0.97 - \frac{3}{5} * 0 + \frac{2}{5} * 0$

$= 0.97$

values of Wind w.r.t sunny

	Yes	NO	tot
weak	1	2	3
strong	1	1	2
			5

$E(S_{weak}) = E(1,2) =$

$-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$

$= 0.9183$

$E(S_{strong}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

$= 1$

$IG(S_{sunny}, Wind) = E(S) - \sum_{v \in (S,w)} \frac{|S_v|}{|S|} E(S_v)$

$= 0.97 - \frac{2}{5} * 1 + \frac{3}{5} * 0.918$

$= 0.02$

IG(Sunny, Temp) = 0.570

IG(Sunny, Humidity) = 0.97

IG(Sunny, wind) = 0.0192

Max Gain = 0.97

Humidity

→ So, Humidity is the node that comes under Sunny.

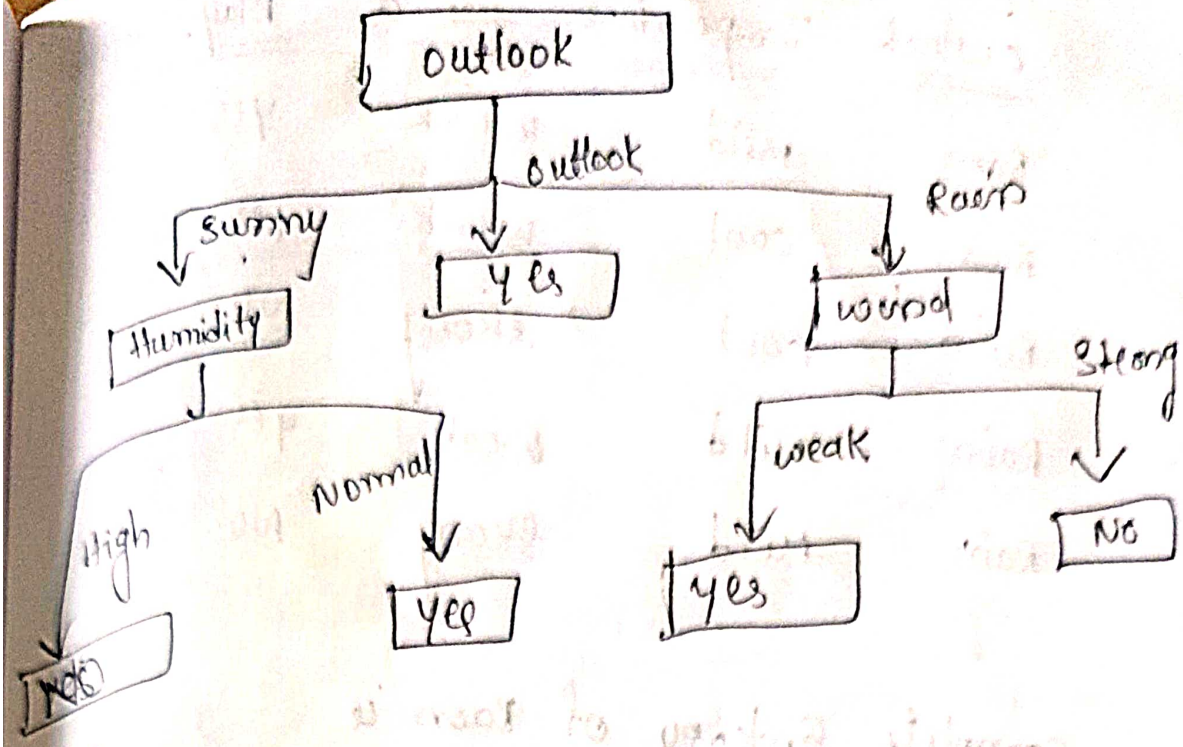
Humidity w.r.t Sunny

	yes	NO	tot
High	0	3	3
Normal	2	0	2

→ from above table, play will occur

if humidity is Normal

if " " High play is NO.



Rules

→ If when outlook = Sunny

Humidity = High

it's a pure class of category = "No"

If when outlook = Sunny

Humidity = Normal

Category = "High"

→ Now find best attribute for splitting Rain.

Remaining attributes are Temperature and wind.

outlook	Temperature	wind	Play.
Rain	Mild	weak	Yes
Rain	cool	weak	Yes
Rain	cool	strong	NO
Rain	Mild	weak	Yes
Rain	Mild	strong	NO

complete Entropy of Rain is

$E(S) = E(3, 2)$

yes NO total 5

$E(S) = 0.97$ (calculated in initial steps).

first attribute Temp.

Rain w.r.t Temp.

Mild cool (only for Rain.

Entropy for $E(\text{Temp}_{\text{cool}})$

Rain

	Yes	NO	Tot
Mild	2	1	3
cool	1	1	2
			5

$E(\text{Temp cool}) =$

$$-\tfrac{1}{2}\log\left(\tfrac{1}{2}\right) - \tfrac{1}{2}\log\left(\tfrac{1}{2}\right)$$

$$= 1$$

$E(\text{Temp}_{\text{Mild}}) = E(2,1)$

$$= -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\frac{1}{3}$$

$$= 0.91$$

$$IG(\text{Rain, Temp}) = E(S) - \sum_{v \in H,C} \frac{|S_v|}{|S|} E(S$$

$$= 0.97 - \frac{2}{5} * 1 + \frac{3}{5} * 0.91$$

$$= 0.97 - 0.9508$$

$$= 0.024$$

Entropy of Wind w.r.t Rain

	Yes	NO	
weak	3	0	3
strong	0	2	2
			5

$$\epsilon(\text{Wind}_{weak}) = \epsilon(3,0)$$

$$= 0$$

$$\epsilon(\text{Wind}_{strong}) = \epsilon(0,2)$$

$$= 0.$$

$$IG(\text{Rain}, \text{wind}) = \epsilon(S) - \sum \frac{|S_v|}{|S|} \epsilon(S_v)$$

$$= 0.97 - 0.$$

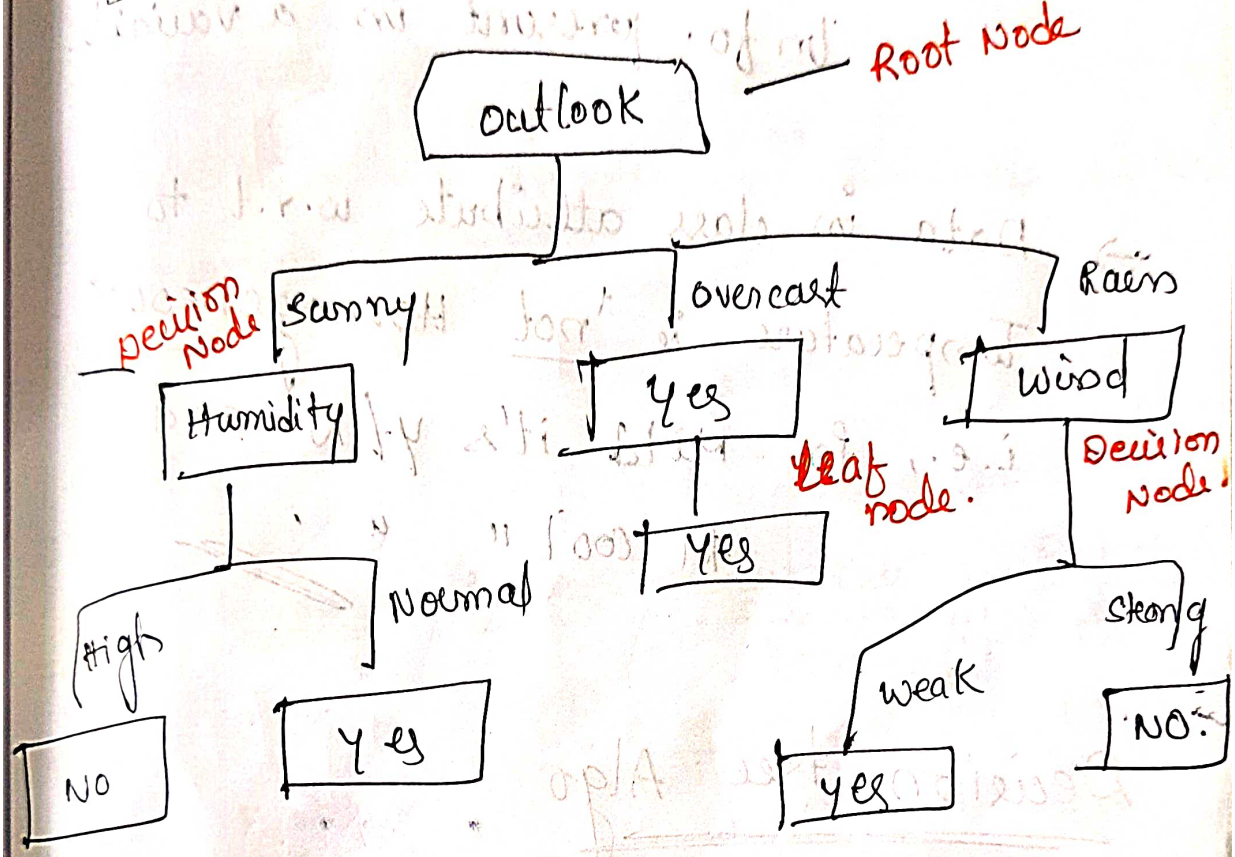
$$= 0.971.$$

$IG(Rain, Temp) = 0.02$

$IG(Rain, wind) = 0.971$

Max is wind $= 0.97$.

So, the decision tree built with wind



Rules

If outlook = Rain
wind = weak then class = yes
can play

If outlook = Rain
wind = Strong then class = No

→ for Temperature Entropy is less / Min
So, data in temperature is 'Min' impure
or disorder.

Entropy Measure — "amount of
Info. present in a variable.

→ Data in class attribute w.r.t to
Temperature is 'not Homogeneous'
i.e., for Mild it's Y/N and
for 'cool' " "

Decision tree Algo

→ This algo. is based on 3 parameter

① D

② Attribute - list &

③ Attribute - selection - Method.

→ attribute - list is a set of attributes defining the tuples.

→ Attribute - selection - Method specifies a process of choosing the attribute that best discriminates the given tuples acc. to class.

→ ASM process applies an Attribute selection measure.

Ex-2

age	Income	Region	Credit Rating	Buy_Mobile
<20	High	USA	Low	NO
<20	High	USA	High	NO
21..50	High	USA	Low	yes
>50	Medium	USA	Low	yes
>50	Low	PK	Low	yes
>50	low	PK	High	NO
21..50	Low	PK	High	yes
<20	Medium	USA	~~High~~ low	NO
<20	Low	PK	Low	yes
>50	Medium	PK	Low	yes
<20	Medium	PK	~~Low~~ High	yes
21..50	Medium	USA	High	yes
21..50	High	PK	Low	yes
>50	Medium	USA	High	NO.

Entropy :-

→ Measured b/w 0 & 1

High Entropy – data have more variance with each others

Low Entropy – data have less variance with each others.

P = Total yes = 9

N – Total No = 5.

Entropy of class attribute

$$= -\frac{9}{14} \times \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \times \log_2\left(\frac{5}{14}\right)$$

$$= 0.94.$$

Entropy for age

Values are

<20 | 21..50 | >50

<20 = 2 yes 3 No.

Entropy = 0.97

21..50 = 4 yes 0 NO

Entropy = 0

>50 = 3 yes 2 NO

Entropy (>50) = 0.970.

IG (age, class attribe)

IG (<20, S) = 0.248

IG (>50, S) = 0.248.

IG Income = 0.029

IG Credit Rating = 0.048

IG of Region = 0.15.

Max Gain is 0.24

Root Node = age

Age	Income	Region	CR	Buy Mobile
<20	High	USA	L	NO
<20	H	USA	H	NO
<20	M	USA	L	NO
<20	L	PK	L	yes
<20	M	PK	H	yes

Entropy for Income

High 0 yes 2 NO

$E = 0$

Medium = 1 yes 1 NO

$E = 1$

Low 1 yes 0 NO

$E = 0$

Entropy for Region

Region

USA 0 yes 3 NO

$E = 0$

PK 2 yes 0 NO

$E = 0$.

for Credit Rating

credit

low 1 yes 2 NO

$E = 0$

High | yes | NO

$G = 0$.

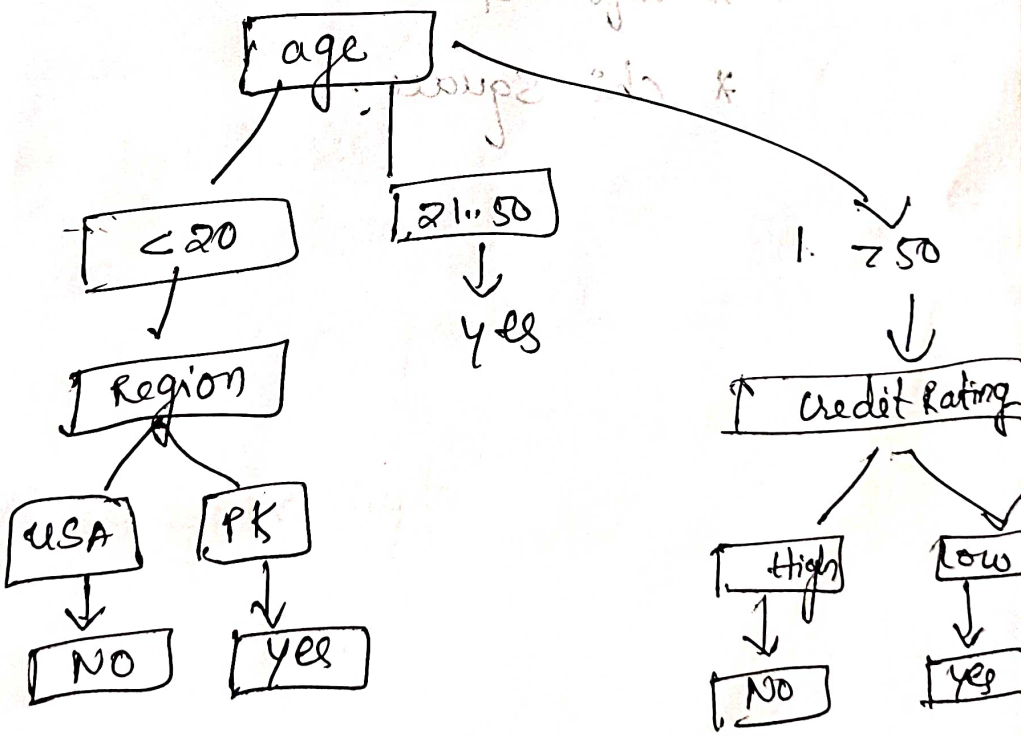
$IG(age, Income) = 0.57$

$IG(age, Region) = 0.97$

$IG(age, Credit\ Rating) = 0.02$

$Max = 0.97$

branch = Region



Similary calculate for >50 node.

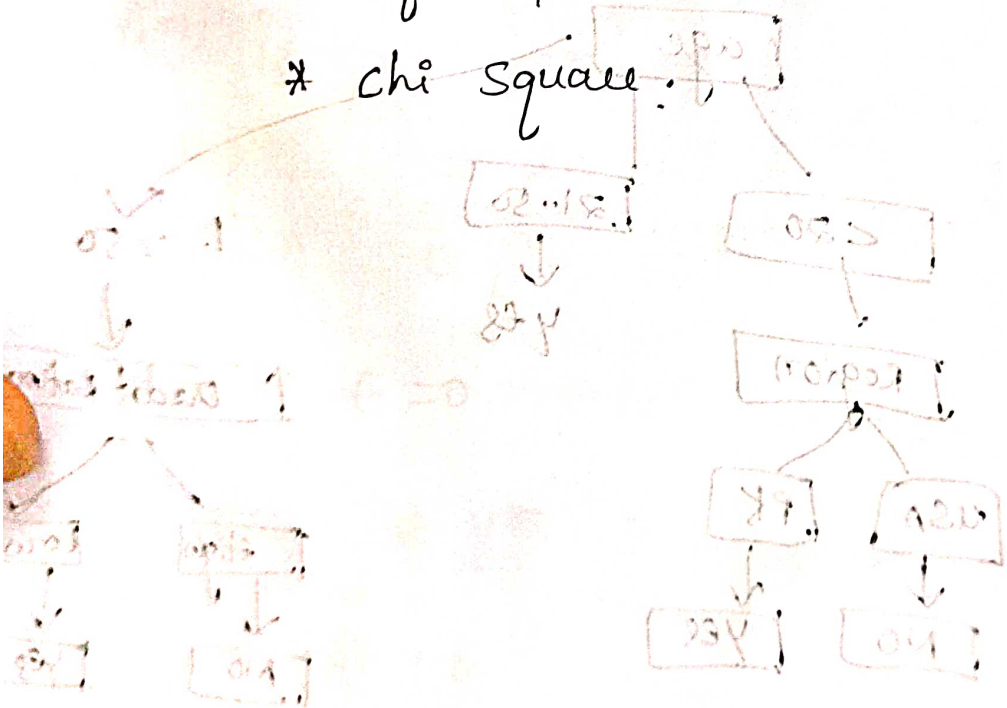
Measures for Splitting / selecting the Best split

→ Splitting is the process of dividing a node into multiple sub-nodes to create relatively pure node.

2 types

for categorical target variable

* Gini Impurity
* Info. Gain
* chi Square.



Naive - Bayes classifiers

→ It's mainly used in text classification that includes a high-dimensional training dataset.

→ It predicts the probabilities of class memberships are called Bayesian classifiers.

$$P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right) P(c)}{P(x)}.$$

C - Hypothesis
i.e.; data tuple x belong to a particular class 'C'.

X = Data tuple

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) P(A)}{P(B)}.$$

$P\left(\frac{c}{x}\right)$ = The probability of 'c' being true Given that x is true.

or

$P\left(\frac{A}{B}\right)$

$P\left(\frac{x}{c}\right)$ = the probability of "B" being true given that "A" is true.

or

$P\left(\frac{B}{A}\right)$

$P(C)$ = the probability of "A" Being true.

or

$P(A)$

$P(x)$ = probability of B being true.

or

$P(B)$

Naive Bayes on <u>whealther Nominal.</u>

Total no. of instance in whealthes nomenal = 14

out of 14, yes = 9

$P(yes)$

Probability (yes) = $\frac{9}{14}$

$P(NO) = \frac{5}{14}$

→ The posterior probability can be calculated by first constructing a frequency table for each attribute against the target.

Individual Attribute Probabilities:-

	Play		
outlook	yes	NO	Total
sunny	2/9	3/5	5/14
overcast	4/9	0/5	4/14
Raeiny	2/9	3/5	5/14

$$P\left(\frac{x}{c}\right) = P\left(\frac{\text{sunny}}{\text{yes}}\right) = \frac{2}{9}$$

$$= 0.23$$

$$P(x) = P(\text{sunny})$$

$$= \frac{5}{14} = 0.36.$$

$$P(c) = P(\text{yes}) = \frac{9}{14} = 0.64.$$

Posterior Prob $P\left(\frac{c}{x}\right) = \frac{P\left(\frac{x}{c}\right)P(c)}{P(x)}$.

$$= 0.33 \times 0.64 \div 0.36$$

$$= 0.597$$

$$P\left(\frac{x}{c}\right) = P\left(\frac{\text{sunny}}{\text{NO}}\right)$$

$$= \frac{2}{5} = 0.4$$

$$P(x) = P(\text{sunny}) = \frac{5}{14} = 0.36$$

$$P(c) = P(\text{NO}) = \frac{5}{14} = 0.36$$

$$P\left(\frac{c}{x}\right) = P\left(\frac{\text{NO}}{\text{sunny}}\right) = 0.40 \times \frac{0.36}{0.36}$$

$= 0.40$

when x value is sunny P(playing) is NO

		yes	NO
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		yes	NO
Temp	Hot	2/9	2/5
	Mild	4/9	2/5
	cool	3/9	1/5

		yes	NO	Totdl
windy	False	6/9	2/5	8/14
	True	3/9	3/5	6/14

Q If outlook = Rainy Temp = cool, Humidity = High, windy = True play = ?

If 15th row value are those, to which class it will belong to.?

$P\left(\frac{yes}{x}\right)$

x = Rainy, cool, High, True

$$P\left(\frac{yes}{x}\right) = P\left(\frac{Rainy}{yes}\right) \times P\left(\frac{cool}{yes}\right) \times P\left(\frac{High}{yes}\right) \times P\left(\frac{True}{yes}\right) \times P(yes)$$

$$= \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$$

$$= 0.00529$$

$$P\left(\frac{NO}{x}\right) = P\left(\frac{Rainy}{NO}\right) \times P\left(\frac{cool}{NO}\right) \times P\left(\frac{High}{NO}\right) \times P\left(\frac{True}{NO}\right) \times P(NO)$$

$$= 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14$$

$$= 0.02057.$$

from above value : 0.02057

Max so class is "NO".

→ Normalize above values to in Range 1.

$$P\left(\frac{yes}{x}\right) = \frac{0.00529}{0.02057 + 0.00529}$$

$$= 0.2$$

$$P\left(\frac{No}{x}\right) = \frac{0.02057}{0.02057 + 0.00529}$$

$$= 0.8$$

Max = 0.8 ∈ NO class.

So, 15th row ∈ NO class.

Bayesian Belief N/w

→ BBN are type of probabilistic Graphical Model that can be used to build models from data or Expert opinion.

→ They can be used for a wide range of tasks: including prediction, anomaly detection, diagnostics, automated insight, reasoning time series prediction & decision making under uncertainty.

→ BBN also called Bayes Net, Belief N/w & Casual N/w.

→ A Bayesian n/w is a graph which is made up of Nodes & directed links b/w them.

Ex:-

→ Suppose you are trying to determine if a patient has asthama.

→ you observe the following symptoms.

(1) the patient has a cough.

(2) The patient has a fever.

(3) the patient has difficulty breathing.

→ you would like to determine how likely the patient is infected with asthama, given that the patient has a cough, a fever & difficulty in breathing.

→ We are not 100% sure that the patient has asthama because of these symptoms.

→ We are dealing with Uncertainity (doubts & chances)

→ Now, you order patient an X-ray & observe that, the patient has a wicker, flum in the lungs.

→ U, belief that the patient is infected with asthama is now higher.

→ what you observed affected your belief that the patient is infected with asthama.

Reasoning with Uncertainity =>

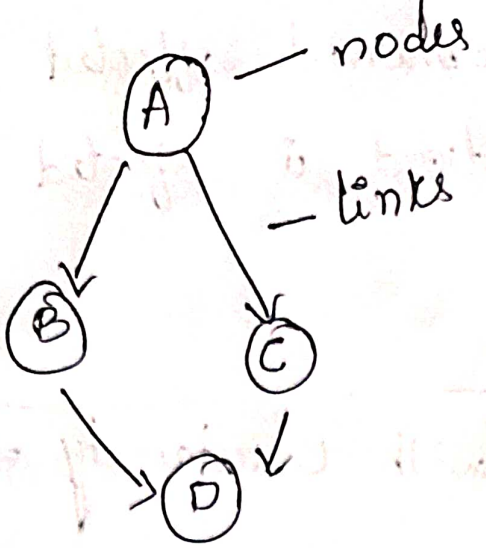
Methodology => Bayesian N/w.

BBN

→ It is n/w based framework for representing & analyzing models involving uncertainity.

→ BN calculates uncertainity using mathematically in a simple way.

→ A Bayesian n/w is always directed acylic Graph.



asyclic.

4 types of nodes.

1) Conditional node
2) Evidence "
3) Hypothesis "
4) Independent "

Conditional Node :- A node having parents.

B has parent A

C " " A

D " " B, C.

B, C, D are cond. Nodes.

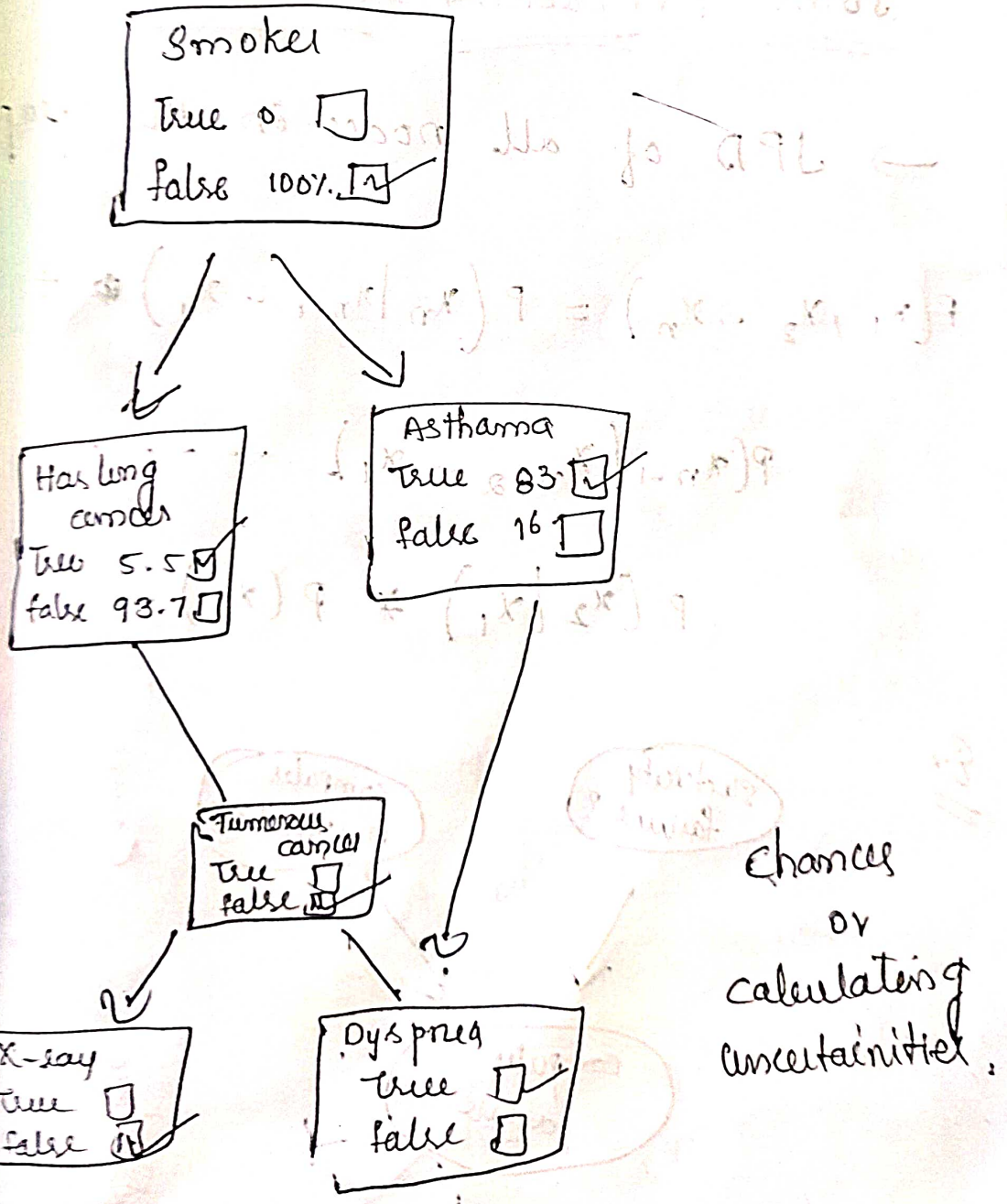
Evidence nodes → If there is a value for node then called EN.

Hypothesis Node → No children.

D → doesn't have children.

Independent node → No parent

A. → No parent.



Chances or calculating uncertainities.

Conditional Probability Distribution

→ Effect of parent node on particular node (child node).

$$P(x_i \mid x_{i-1} \cdots x_1) = P(x_i \mid \text{Parent}(x_i))$$

↓ child node.

Joint Probability Distribution

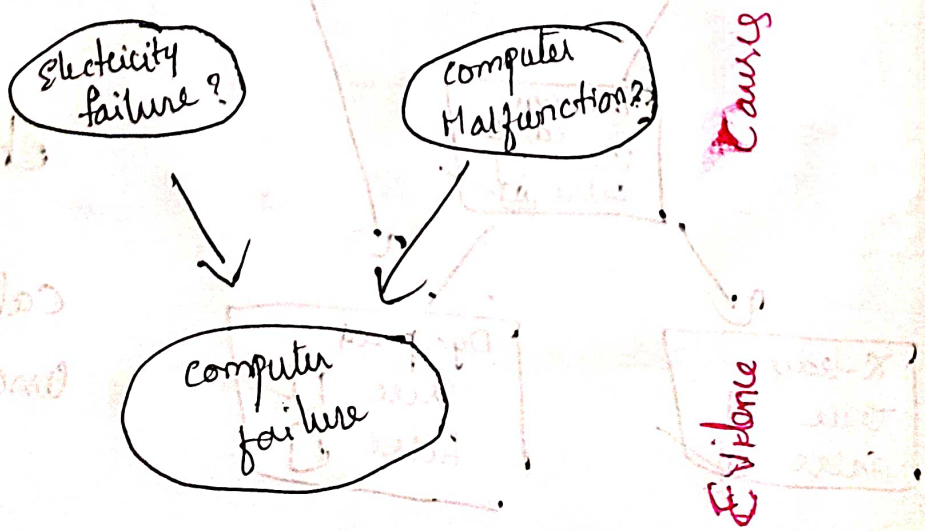
→ JPD of all nodes in the Graph.

$$P(x_1, x_2 \ldots x_n) = P(x_n \mid x_{n-1} \ldots x_1) * \ldots$$

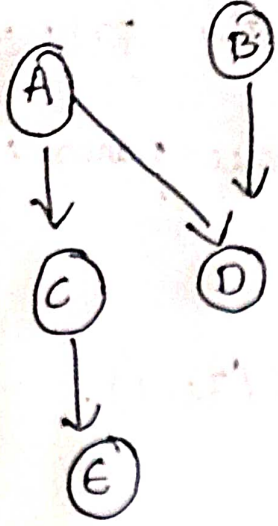
$$P(x_{n-1} \mid x_{n-2} \ldots x_1) \ldots$$

$$P(x_2 \mid x_1) * P(x_1)$$

Ex



Ex:



A, B – Independent

C – Parent is A

D – Parent is B, A.

E – Parent is C.

→ Suppose the following Graph/net has following probabilities (values will be given in problem for True). Calculate.

P(A)	C. (True)	for C (false)
T	0.8	1 − 0.8 = 0.2
F	0.4	1 − 0.4 = 0.6.

changes in A will effect the probability of C (because A is a parent of C).

Parent/Independent Nodes

P (A = T) = 0.3 P (A = False) = 1 − 0.3 = 0.7

P (B = T) = 0.6 P (B = false) = 0.4

→ A & B will effect the prob. of D because 'A & B' are parents of D.

A	B	P(D) (True)	calculate when D(false) D = false
T	T	0.7	1-0.7 = 0.3
T	F	0.8	1-0.8 = 0.2
F	T	0.1	1-0.9 = 0.9
F	F	0.2	1-0.2 = 0.8

⇒ C will effect the E. because C is the parent of E.

C	P(E) when T	Calculate for false
T	0.7	1-0.7 = 0.3
F	0.2	1-0.2 = 0.8.

→ for E, there is only one parent. So, there will be only two values T, F.

→ If there are 2 parents, possible values will be four.

Computate

(1) Prob (D = T)

(3) Prob (D = F, C = T)

(5) Prob (A = T | C = T)

Parents of D = A, B.

(1) P(D = T)

for D = True, A, B possible values are

T T, T F, F T, F F.

Prob of parents when T.

P(D=T | A=T, B=T) * P(A=T) P(B=T) +

P(D=T | A=T, B=F) * P(A=T) P(B=F) +

P(D=T | A=F, B=T) * P(A=F) P(B=T) +

P(D=T | A=F, B=F) * P(A=F) P(B=F).

= 0.7 x 0.3 x 0.6 + 0.8 x 0.3 x 0.4 +

0.1 x 0.7 x 0.6 + 0.2 x 0.7 x 0.4

= 0.32.

(2) $P(D=F, C=T)$

Parents of $D = A, B$.

Parents of $C = A$. (common is A)

Question D = false)

Parents of D. Parent of C

$P(D=F \mid A=T, B=T)\ P(A=T)\ P(B=T)\ P(A=T) +$

(other table (0.3)) 0.3 0.6 0.3

$P(D=F \mid A=T, B=F)\ P(A=T)\ P(B=T)\ P(A=T) +$

$P(D=F \mid A=F, B=T)\ P(A=F)\ P(B=T)\ P(A=F)$

$P(D=F \mid A=F, B=F)\ P(A=F)\ P(B=F)\ P(A=F)$

$= 0.3 \times 0.8 \times 0.6 \times 0.3 + 0.2 \times 0.8 \times 0.3 \times 0.4 + 0.9 \times 0.4 \times 0.7 \times 0.6 + 0.8 \times 0.4 \times 0.7 \times 0.4$

$= 0.3032$

(3) $Prob(A = True \mid C = T)$

A — no parent

C — A is parent.

$P(A=T \mid C=T)$

first, calculate for $C=T$ true

$= P(C=T \mid A=T)\, P(A=T) +$

$P[C=T \mid A=F]\, P(A=F)$

$= 0.8 \times 0.3 + 0.4 \times 0.7$

$= 0.52.$

Now calculate $P(A=T \mid C=T)$

when $A=T$, How it will effect C.

$= P(C=T \mid A=T)\, P(A=T) / P(C=T)$

↓ when $A=T \mid C=T$.

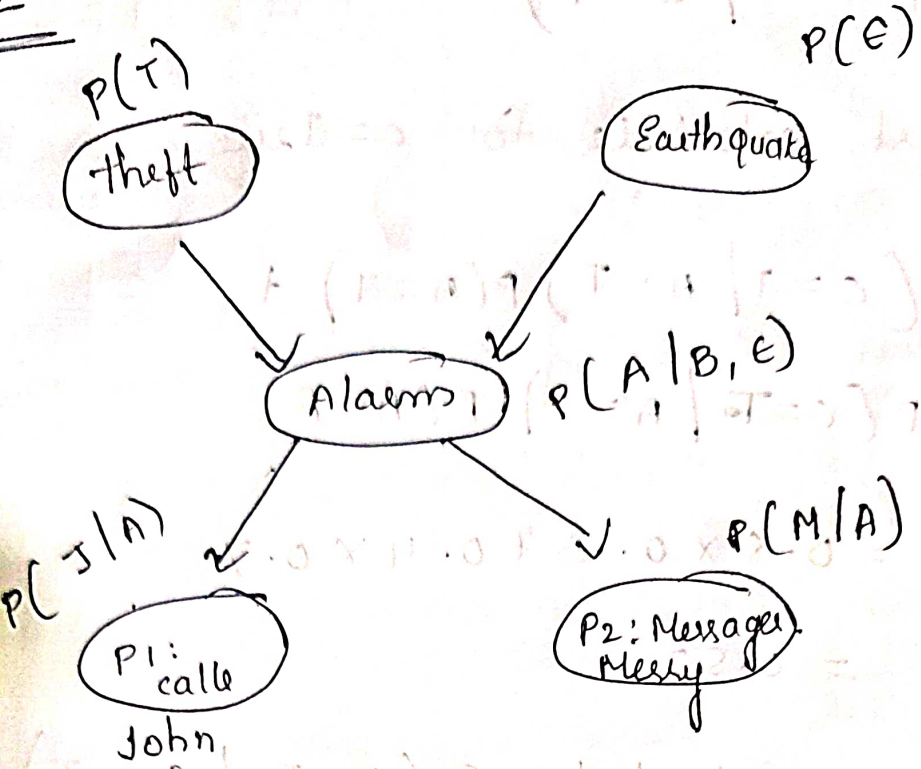
$= 0.8 \times 0.3 / 0.52$

$= 0.46$ //.

(4) Prob (~~$A=T \mid D=T$~~) =

~~for~~ D ~~$=$~~ A, B ~~are~~ parents

Ex 2



Alarm System Example (Prob. Stmt)

→ Your house has a alarm system against burglary. You live in a seismically active area & the alarm detects burglary but also sometimes it responds to minor earthquake.

→ You have 2 neighbours, Mary & John, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.

→ John always calls when he hears the alarm, but sometimes considers telephone ringing as alarm & calls you. Mary likes loud music & the mixer, the alarm, sometimes.

→ we want to represent the probability distribution of events — Burglary, Earthquake, Alarm, Mary calls, Jhon calls.

Probability values will be given

P(B)
0.001

- Prob. of Burgulry, it means that
  0.001 there is a theft
  0.999 there is no theft.

P(E)
0.002

- Prob of getting Earthquake.

A	P(M/A)
T	0.70
F	0.01

Prob. of Mary calls when there is a Alarm. (T)

Prob. of Mary calls when there is no Alarm (F)

A	P(J/A)
T	0.90
F	0.05

Prob. of Jhon calls when there be heavy alarm (T)

← conditional probabilities

B or Theft	E	P(A/B, E)	Probability of
T	T	0.95	there is a Earthquake, Burglary
T	F	0.94	Theft, NO Earthquake
F	T	0.29	NO Theft, Earthquake
F	F	0.001	NO Theft, NO Earthq

Q1 :- what is the prob. that the alarm rings, but neither a burglary nor an Earthquake has occured & both John & Mary calls you.

→ In the above que, all 5 variables are Mentioned

P. ($J \wedge M \wedge A \wedge \sim B \wedge \sim E$) – As per given que.

Prob. of John calls and Mary calls when there is a alarm but there no theft & no Earthquake (As per question)

Parent of John – Alarm

" Mary – "

Alarm – Theft, Earthquake

No parents for Theft & Earthquake.

$P(J \wedge M \wedge A \wedge \sim B \wedge \sim E)$ (Joint Prob)

$= P(J|A)\, P(M|A)\, P(A|\sim B, \sim E)\, P(\sim B)\, P(\sim E)$

$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998$

$= 0.000628$

→ there is no dependency b/w J, M, A

So use Joint prob & Multiply.

these events are not dependent

(2) what is the prob. that John calls you.

conditional prob.

$$P(J) = P(J|A)P(A) + P(J|\sim A)P(\sim A)$$

parent

calculate Prob. of Alarm.

Joint prob.

$$P(A) = P(A|B,\epsilon) \times P(B,\epsilon) + P(A|B,\sim\epsilon) \times P(B,\sim\epsilon) + P(A|\sim B,\epsilon) P(\sim B,\epsilon) + P(A|\sim B,\sim\epsilon) P(\sim B,\sim\epsilon)$$

$$= 0.95 * 0.001 * 0.002 + 0.94 \times 0.001 \times 0.998 + 0.29 \times 0.999 \times 0.002 + 0.001 \times 0.999 \times 0.998$$

$$= 0.00252.$$

$$P(\sim A) = P(\sim A|B,\epsilon) \times P(B,\epsilon) + P(\sim A|B,\sim\epsilon) \times P(B,\sim\epsilon) + P(\sim A|\sim B,\epsilon) \times P(\sim B,\epsilon) + P(\sim A|\sim B,\sim\epsilon) \times P(\sim B,\sim\epsilon)$$

$= 0.05 \times 0.001 \times 0.002 + 0.06 \times 0.001 \vee$

$0.998 + 0.71 \times 0.999 \times 0.002 +$

$0.999 \times 0.999 \times 0.998$

$= 0.99748$

$P(J) = 0.90 \times 0.00252 +$

0.05×0.99748

$= 0.052142$

UNIT - V

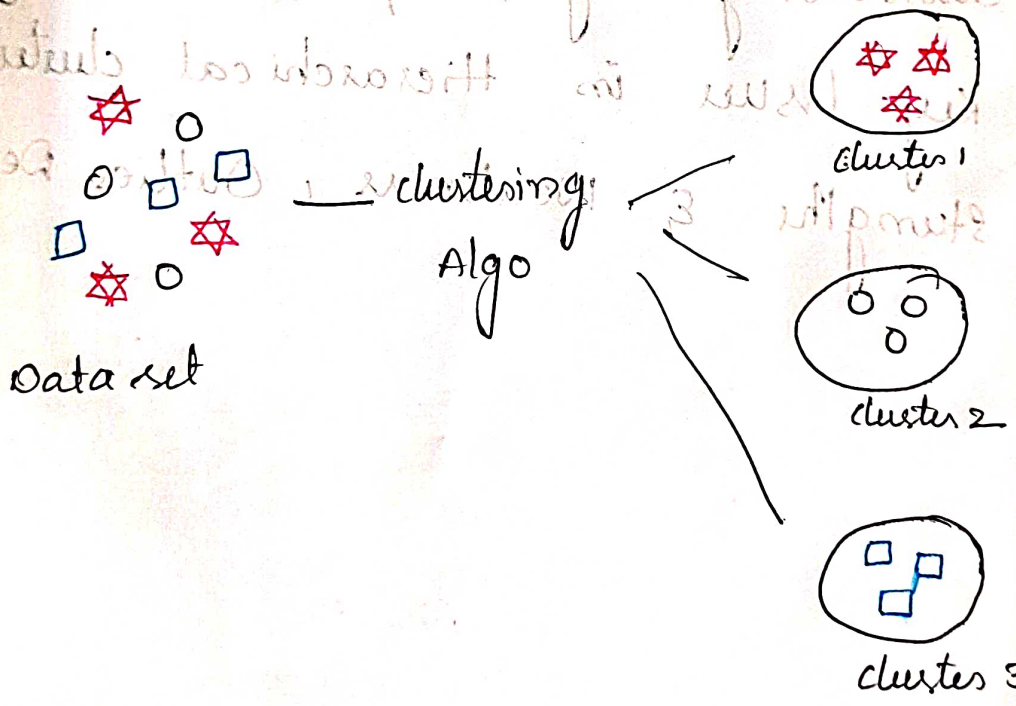
clustering :- Problem Definition, clustering overview, Evaluation of clustering Algo, Partitioning clustering - K-Means Algo, K-Means Additional issues, PAM Algo, Hierarchical clustering - Agglomerative Methods & divise Methods Basic Agglomerative Hierarchical clustering Algo, specific techniques, Key issues in Hierarchical clustering Strengths & Weakness, Outlier Detection.

Clustering

→ It's an unsupervised ML-based algo.

→ clustering helps to split data into several subsets.

→ Each of these subsets contains data "alike" to each other, & these subsets are called clusters.



→ No predefined labels

→ large dataset is divided into smaller subsets called "clusters"

→ cluster contains "alike" type of objects.

PAM Algo

K - Mediods algo. also called as

PAM

PAM - Partition around Medoids

→ PAM searches for K representative obj. in a dataset (K Medoids) & then assigns each objects to the closet mediod in order to create cluster.

K - Mediod
↓
No. of clusters

↳ centeally located objects

PAM - the Method partitions our data into K clusters.

→ K Mediods clustering & PAM Method are interchangeble terms.

distance = cost

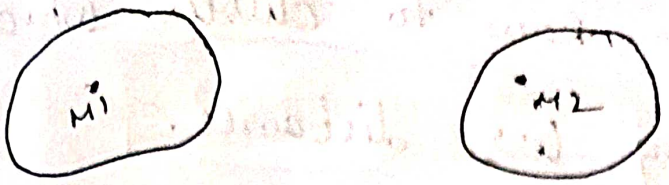
clustering process (objective): Reduce the overall cost.

Sample data :

SNO.	a	b
1	18	21
2	18	23
3	15	28
4	19	29
5	20	28
6	11	24
7	16	21
8	19	23
9	16	23
10	15	24.

→ Initially, Select 2 Medoid for clusters (Randomly)

a	b	
20	28	— cluster Mediod 1
15	24	— " " 2



cluster cluster

→ calculate the distance b/w each of the 10 rows from Mediode.

using Manhattan Distance. (can use any distance formula.

$$|x_1 - x_2| + |y_1 - y_2|$$

distance = cost.

S.NO.	a	b	1st cluster Mediod (20,28)	2nd cluster Mediod (15,24)
	13	21	$\|13-20\| + \|21-28\| = 14$	$\|13-15\| + \|21-24\| = 5$
	18	23	7	4
	15	28	5	4
	19	29	2	9
	20	28	0	9
	11	24	13	4
	16	21	11	4
	19	23	6	5
	16	23	9	2
	15	24.	9	0.

→ Assign Row to cluster which is having less distance.

SNO	a	b	1st cluster	2nd cluster	Assign to cluster
1	13	21	14	5	C2
2	18	23	7	4	C2
3	15	28	5	4	C2
4	19	29	2	9	C1
5	20	28	1	9	C1
6	11	24	13	4	C2
7	16	21	11	4	C2
8	19	23	6	5	C2
9	16	23	9	2	C2
10	15	24	9	0	C2

→ As per PAM, the cost of assigning Row1 to cluster2 is 5.

Means, the Manhattan distance b/w 1st Row and 2nd Mediod is 15!

cost for Making first cluster.

= 2 + 0. (considers distance of Rows ∈ c1)

= 2

cost for Making Second cluster

= 5 + 4 + 4 + 4 + 4 + 5 + 2 + 0

= 28.

Total cost for 2 clusters

1st cluster + 2nd cluster

= 2 + 28 = 30.

Next step: Calculate distance / cost of these 10 Rows with every possible combination of Mediode

let's Randomly pick: (Second Mediod)

Row 5 & Row 8

New clusters.

a	b
20	28
19	23

a	b	1st cluster	2nd cluster	Assigned to cluster
13	21	$\|13-20\|+\|21-28\| = 14$	$\|13-19\|+\|21-23\| = 8$	C2
18	23	7	1	C2
15	28	5	4	C1
19	29	2	6	C1
20	28	0	6	C1
11	24	7	9	C2
16	21	1	5	C2
19	23	6	0	C2
16	23	9	3	C2
15	24	9	5	C2

Cost for making 1st cluster = 7

" " 2nd " = 31

Total cost for making 2 clusters

$= 7 + 31 = 38$

→ Total cost for making two clusters using initial medoids is 30. &

for cost for making 2 clusters using 2nd medoids is 38.

→ Based on PAM Method, 1st choice of mediods are better than the 2nd choice of 'a'.

→ calculate distance/cost of these 10 rows with every possible combination (change the mediods & calculate)

→ calculate total cost associated with every mediod & finally choose one combination that results in one Min. total cost.

Again changing Mediods & calculate distance using New Mediods.

a	b
20	28
16	23.

calculate for every Row

Ex Row1 :-

$|13-20| + |21-28| = 14$

$|13-16| + |21-23| = 5$

Apply calculate for all rows.

→ for newly choosen mediods distance/ cost for making first cluster

$= 5+2+0 = 7.$

→ Cost for making 2nd cluster.

$5+2+6+2+3+0+2 = 20.$

Total cost for making 2 cluster are

$20+7 = 27.$

→ finally 3rd Mediods are taking less cost for Making cluster. So we swap Mediods to

20 28

16 23 and continues for all future instance of that dataset.

K Means Vs K Mediods

* In K-Means centroids updates whenever any row joins a cluster.
* Possible drawbacks: presence of a outliers could impact the clustering Result

⇒ K Medoids with lowest cost remains the same throughout the clustering process.

PAM

1. Given K.
2. Randomly Pick K instances as initial Mediods
3. Assign each instance to the nearest Medoid x.
4. Calculate the objective fn:
   * the sum of dissimilarities of all instances to their nearest mediods.
5. Randomly select an instance Y.
6. Swap x by Y if the swap reduces the objective fn.
7. Repeat (3-6) untill no change.

Hierarchical clustering

→ HC is an algorithm that groups similar objects into groups called 'clusters'

→ Groups the data into tree of clusters → Dendrogram

↓

has sequence of all merge & split.

or

shows hierarchical Rls b/w obj.

→ There are two types of HC:

① Agglomerative (bottom-up)

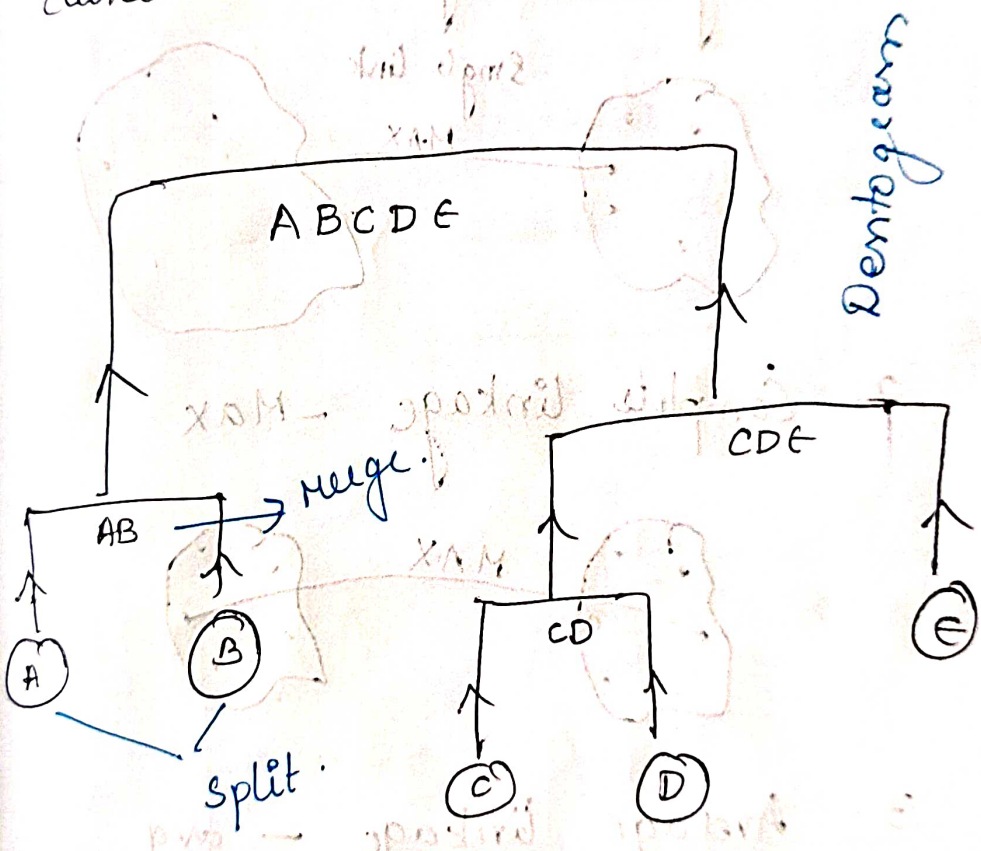
② Divise (top-down).

Agglomerative

Steps

1. calculate similarity of one cluster w.r.t all other clusters

2. Consider every data pt. as individual data
3. Merge the clusters with highest similarity
4. Recalculate similarity for each cluster
5. Repeat step ③ & ④ until single cluster is obtain.



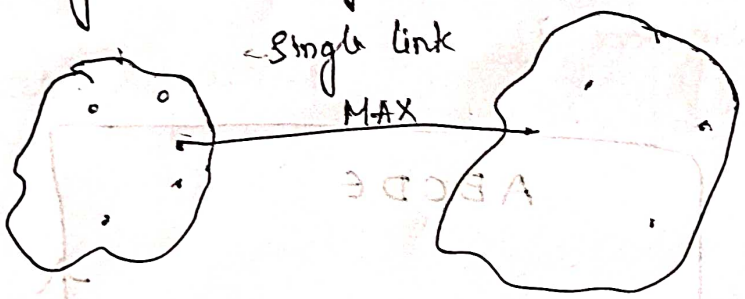
→ 1st iter, b/w A & B are calculated if distance is low merge it & make it into one cluster AB

→ Same for C, D & E.

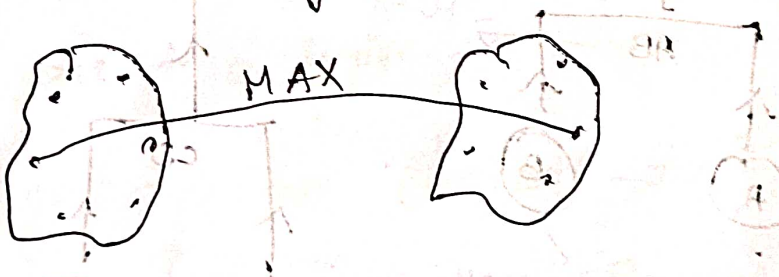
→ In Real world, flipkart, amazon will consider every customer individu and cluster/ group them as one customer base.

→ Agglomerative will be operated in 3 modes.

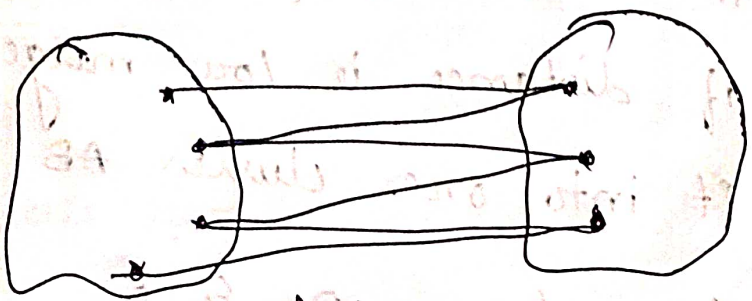
1. Single linkage – Min (similarity)



2. Complete linkage – Max



3. Average linkage – avg



Single linkage : Complete linkage

MAX(d(x1, x2))

Sample NO	x	y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Step 1:- Compute the distance Matrix.

→ use Euclidean distance blw each & every pt to calculate distance.

$$d(A,B) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

$$d(P_1,P_2) = \sqrt{(0.22-0.40)^2 + (0.38-0.53)^2}$$

$$= 0.23$$

$$d(P_1,P_3) = \sqrt{0.35-0.40)^2 + (0.32-0.53)^2}$$

$$= 0.22$$

$$d(P_2,P_3) = \sqrt{(0.35-0.22)^2 + (0.32-0.38)^2}$$

$$= 0.14$$

→ Clustering starts by computing a distance b/w every pair of pts & we want to cluster.

→ A distance matrix will be symmetric because the distance b/w x & y is the same as the distance b/w y & x. & will have zeros on the diagonal (because every item is distance zero from itself).

* below is an example of a distance Matrix.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

These values are calculated using Euclidean distance if a table is given.

→ Pick which is having smallest / Min or smallest distance.

→ from above Matrix, 2 is the Min. distance. (3,5) clusters 3 & 5 are Merged.

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0



→ Create New distance Matrix, Remove 3,5 entries & replace it by an entry (35,T)

→ Since we are using complete linkage clustering, the distance b/w "35" & every other item is the MAX of of the distance b/w this item

	35	1	2	4
35	0			
1	[11]	0		
2	[10]	9	0	
4	[9]	6	5	0

$$d(1, "35") = d(1,3)\ d(1,5)$$

$$= MAX\left(d(1,3), d(1,5)\right)$$

$$= MAX\ (3, 11)$$

$$= 11$$

from Initial Matrix
$d(1,3) = d(3,1) = 3.$

So, $d(1, "35") = 11$

$$d(2, "35") = MAX\left(d(2,3), d(2,5)\right)$$

$$= MAX(7, 10)$$

$$d(2, "35") = 10.$$

$$d(4, "35") = MAX\left(d(4,3), d(4,5)\right)$$

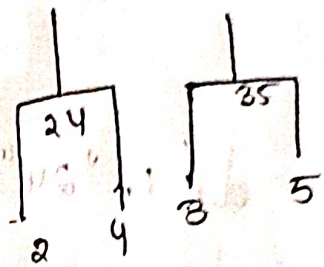
$$= MAX(9, 8)$$

$$= 9$$

→ Continue Step, Picking Min distance pt. & Merge.

→ from above Matrix 5, is Min, so Merge 2,4.

	35	1	2	4
35	0			
1	11	0		
2	10	9	0	
4	9	6	5	0



After Merging, the Matrix will be,

	35	24	1
35	0		
24	10	0	
1	11	9	0

d("24", "35")

→ Now the distance blw 35 & each item is now the MAX (2, "35") (4, "35")

$$d((24),35) = d(2,35)\ d(4,35)$$

$$= MAX(d(2,35), d(4,35))$$

$$= MAX(10, 9)$$

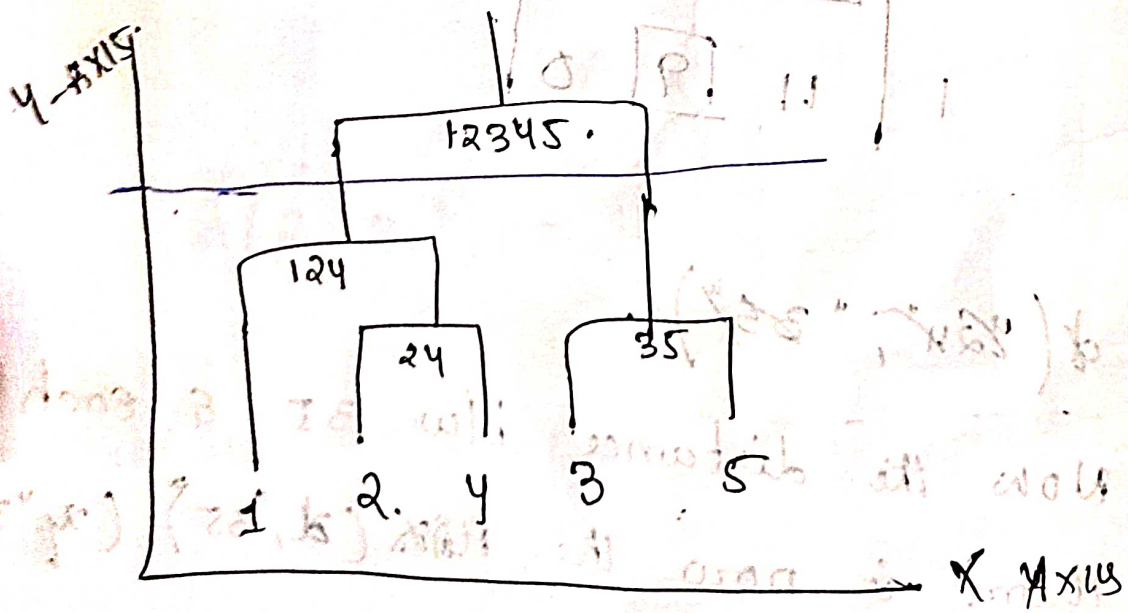
$$= 10$$

$$d((1),(24)) = MAX(d(1,2), d(1,4))$$

$$= MAX(9, 6)$$

$$= 9$$

→ from above Matrix, 9 is the "MIN" distance, so Merge (1,2,4).

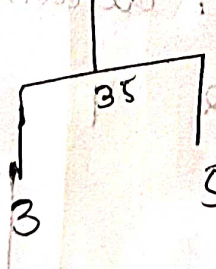


Single linkage:

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

→ Pick Min/smallest distance and merge them.

→ In above, 2 is Min distance so, Merge 5 3



clusters.

→ Create new distance Matrix & calculate distances.

	35	1	2	4
35	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

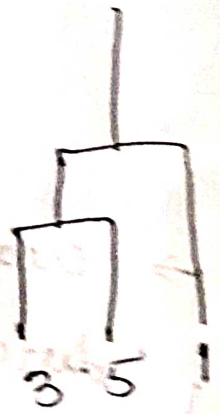
$$d(1,"35") = MIN(d(1,3), d(1,5))$$
$$= MIN(3, 11)$$
$$= 3.$$

$$d(2,"35") = MIN(d(2,3), d(2,5))$$
$$= MIN(7, 10)$$
$$= 7.$$

$$d(4,"35") = MIN(d(4,3), d(4,5))$$
$$= MIN(9, 8)$$
$$= 8.$$

→ Again Pick, Min distance and Merge.

	135	2	4
135	0		
2	7	0	
4	6	5	0



Dendrogram.

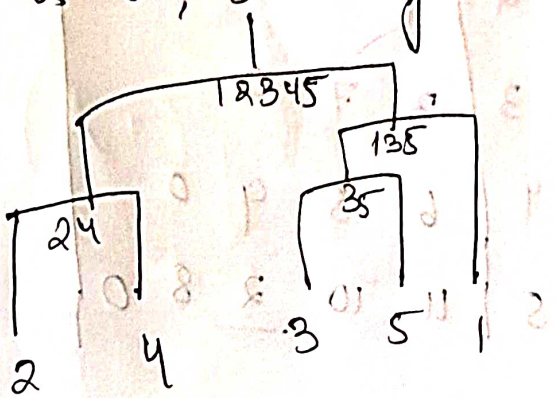
$$d(2,"135") = MIN(d(2,1), d(2,3), d(2,5))$$
$$= MIN(9, 7, 10) = 7$$

$d(.4,^4 135^4) = d(4,1)\ d(4,3)\ d(4,5)$

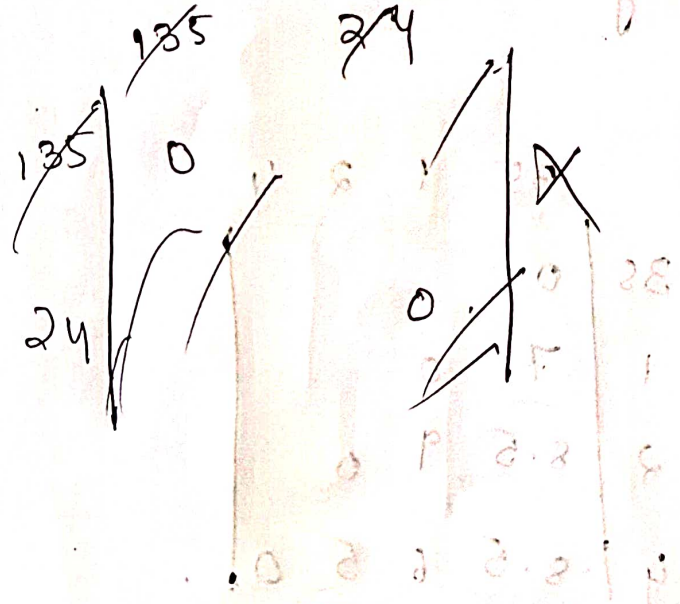
$= MIN(6, 9, 8)$

$= 6.$

→ MIN is 6, so Merge 2,4.



dendogram

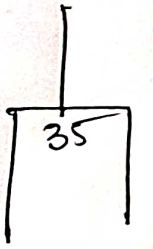


Average Linkage :-

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

MIN distance = 2

Merge 5, 3.



	35	1	2	4
35	0			
1	7	0		
2	8.5	9	0	
4	8.5	6	5	0

$$d(1, "35") = Avg\,(d(1,3), d(1,5))$$

$$= Avg\,(3, 11)$$

$$= 14/2$$

$$= 7$$

$d(2, "35") = d(2,3)\ d(2,5)$

$= Avg(7, 10)$

$= 8.5$

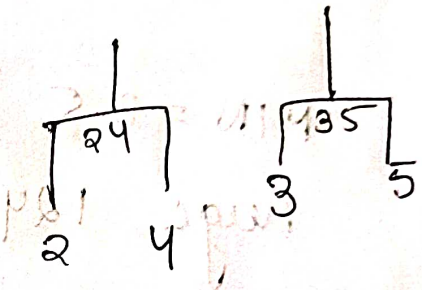
$d(4, "35") = Avg(d(4,3), d(4,5))$

$= Avg(9, 8)$

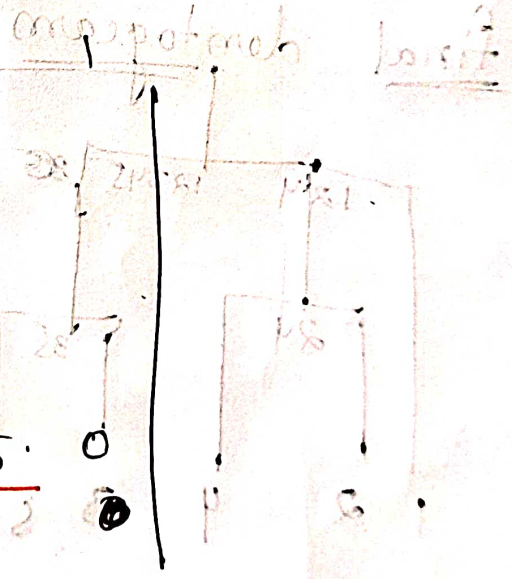
$= 8.5$

MIN = 5

merge = 2, 4



	35	24
35	0	
24	8.5	0
1	7	3.5



$d(24, "35") = Avg(d(2, "35"), d(4, "35"))$

$= Avg(8.5, 8.5)$

$= 8.5$

$d(1, "35") = 7$ (~~Already~~ calculated previously)

$d(9, "24") = Avg(d(1,2)\ d(1,4))$

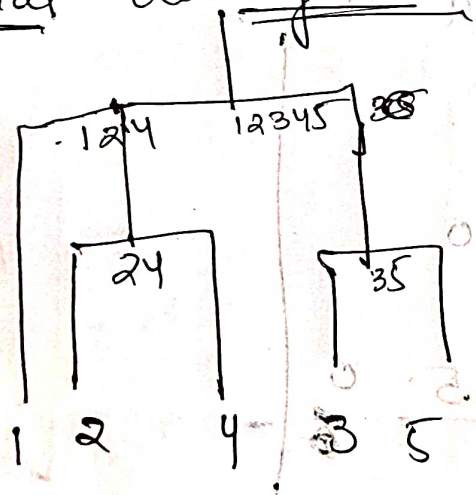
$= Avg(1, 6)$

$= 3.5$

MIN = 3.5

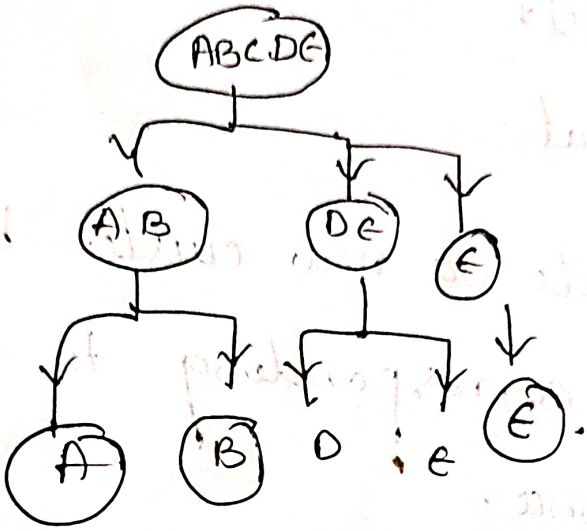
Merge 124

Final dentogram



Divisive Methods

→ In this data objects are grouped in top down manner.



→ Initially all objects are in one cluster.

→ Then the cluster is subdivided into smaller & smaller pcs untill each obj. forms a cluster on its own or untill it satisfies certain termination condition as the desired no. of clusters is obtained.

Divisive Algorithm

1. Compute a Min. Spanning tree (MST) for the given adjacency matrix.
2. Repeat.
3. Create a new cluster by breaking the corresponding to the largest distance.
4. Until only singleton clusters remains

Distance Matrix

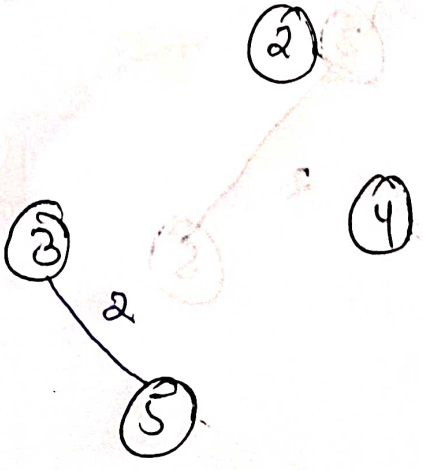
	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

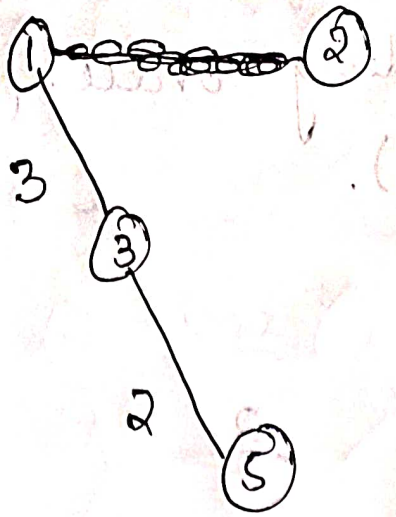
→ from adjacent Matrix, consider edges & cost in ascending order (start from lowest distance).

Edge	cost
5 - 3	2
3 - 1	3
4 - 2	5
4 - 1	6
3 - 2	7
5 - 4	8
4 - 3	9
2 - 1	9
5 - 2	10
5 - 1	11

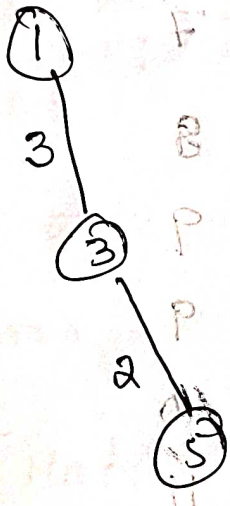
→ From above table, create MST by Prim's or Kruskal's Algo.

1st
5 - 3 - 2

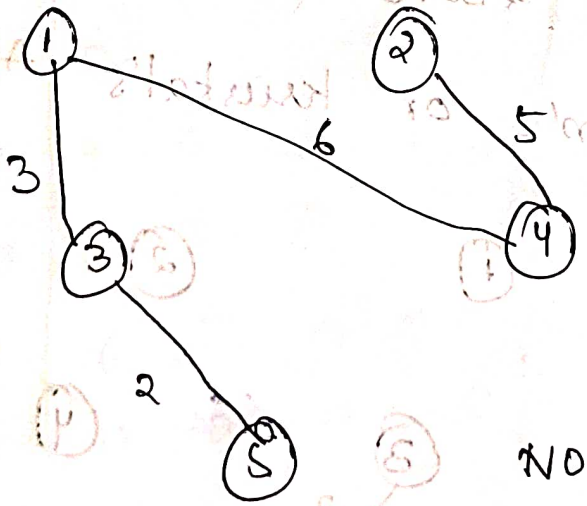




3 - 1 = 3



4 - 2 - 5



4 - 1 = 6

No. of edges = 4

No. of vertices = 5

→ Stop constructing a MSP. because

$$V - 1 = e$$

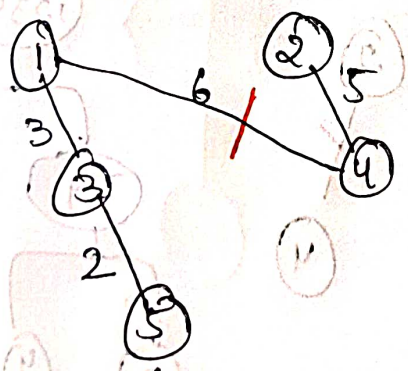
$$5 - 1 = 4$$

$$\text{Vertices} - 1 = \text{edges}$$

→ we need to construct upto $V-1 = e$, after that even if we construct a tree it will become cyclic (loop.

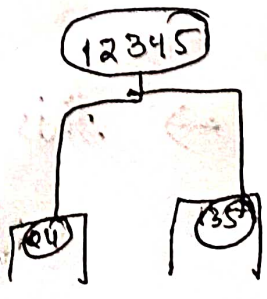
→ MSP should not be cyclic.

→ Cut the largest edge i.e., blw 1 & 4.



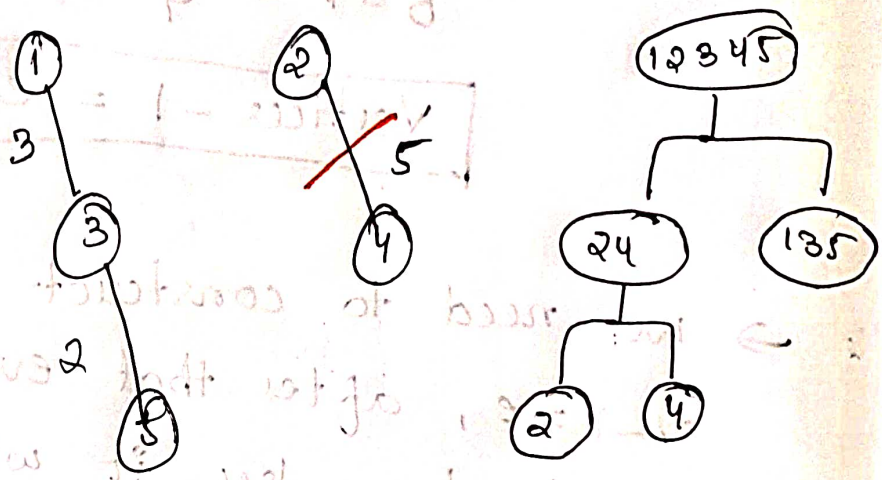
→ Cutting this edge results into two clusters i.e.,

{2, 4} {1, 3, 5}.



→ Again Remove the largest edge i.e.,

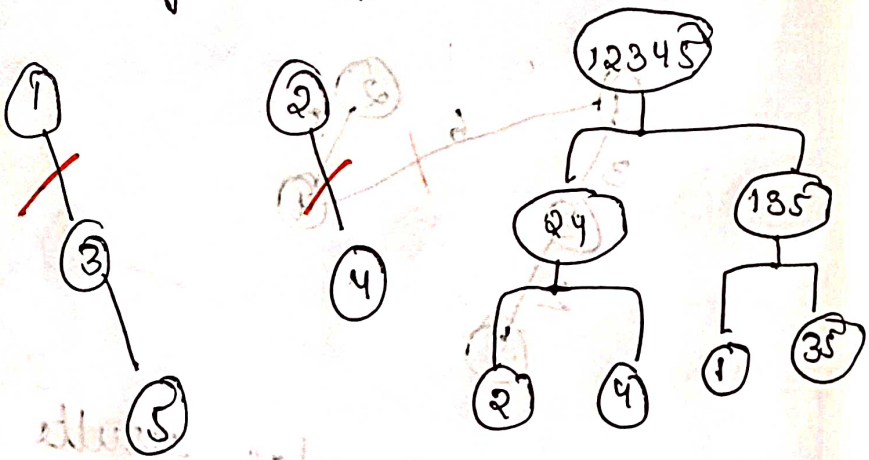
2, 4 – 5



→ This split creates 3 clusters

{2} {4} {135}

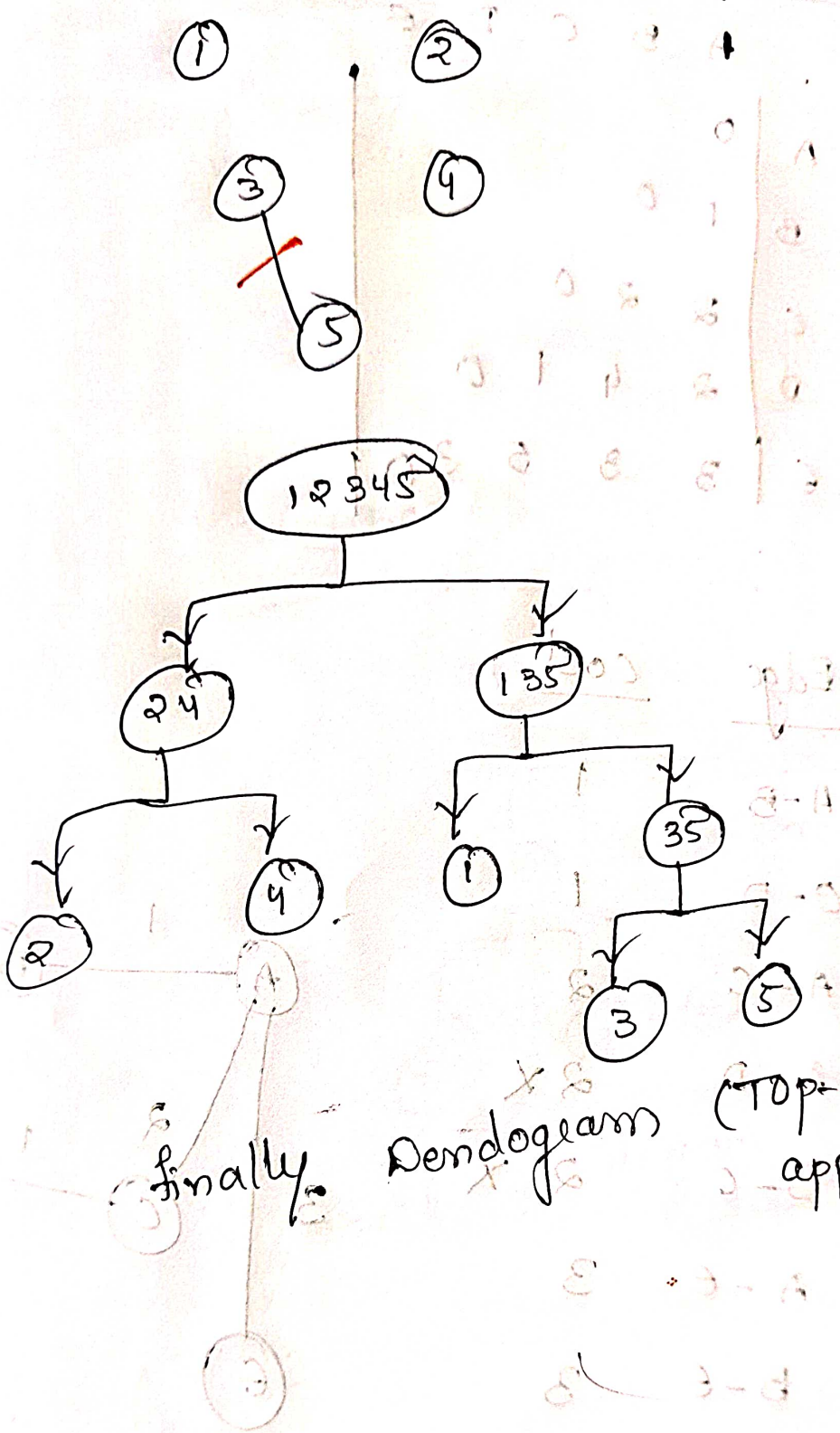
→ Remove larget edge i.e., 13



→ This split creates 4 clusters.

{2} {4} {1} {35}

→ finally, Remove last edge 3 5

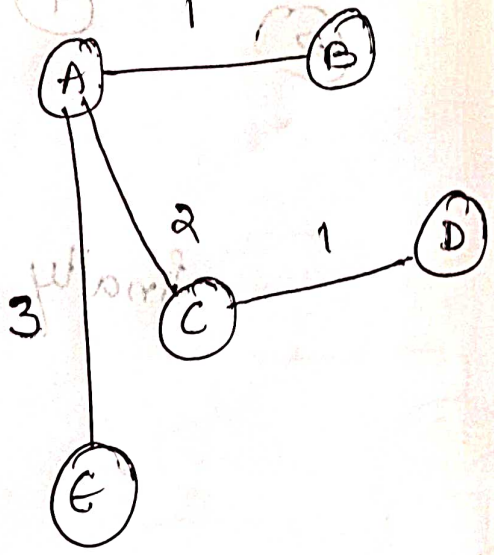


finally Dendogram (Top-down approach).

Ex-2

	A	B	C	D	E
A	0				
B	1	0			
C	2	2	0		
D	2	4	1	0	
E	3	3	5	3	0

Edge	cost	
A-B	1	
C-D	1	
A-C	2	
A-D	2	X
B-C	2	X
A-E	3	
B-E	3	
D-E	3	
B-D	4	
C-E	5	



AD, BC are not considered because they form loop / cyclic.

V reached

$V-1=e$

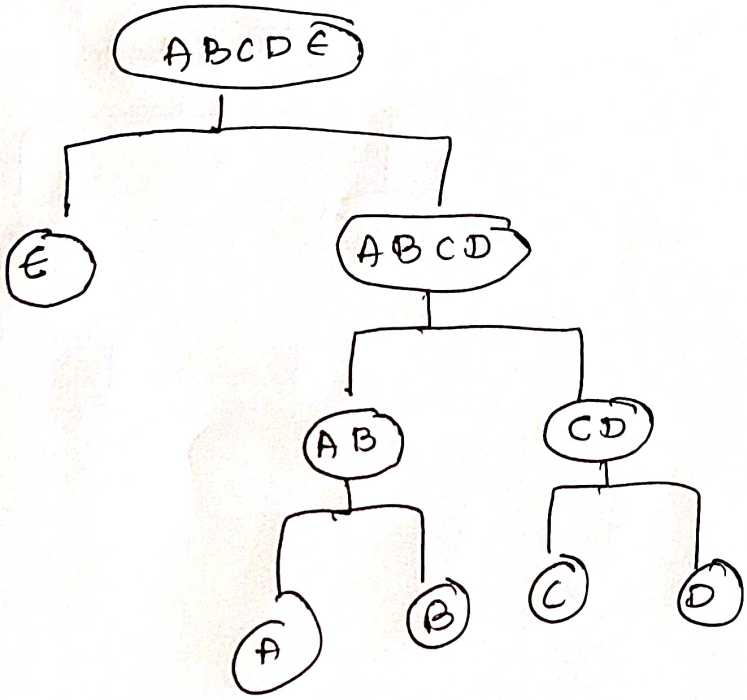
→ cutting edge one by one, till we reach individual clusters.

cutting edge AE - 3

AC - 2

AB - 1

CD - 1



final Dendogram