"Data Mining and Warehousing"

Unit - 1

Data Warehousing :-

Data warehousing is the process of constructing and using a data warehouse.
A Data warehouse is constructed by integrating data from multiple ~~hetro-~~ heterogeneous sources that support analytical reporting, structured and decision making.
Data warehousing involves data integration, data cleaning & data consolidations.

Need :-

- Consolidated Data :
  Collects data from multiple sources into one place, making it easier for analysis.
- Fosters Decision making :
  Provides timely and accurate insights for business decisions.
- Historical Analysis :
  Allows businesses to analyze trends over time.
- Improved Data Quality :
  Cleans and structures data, ensuring consistency.

Characteristics =>

- Subject – oriented
- Integrated
- Time – Varient
- Non – Volatile.

Functions =>

- Data Consolidation
- Data Analysis
- Data Mining
- OLAP

Types =>

There are 3 types of data warehouses :

① Enterprise Data Warehouse (EDW) =>

A Centralized warehouse containing all the organizational data for reporting across departments.

② Data Mart :

A Smaller, focused subset of the data warehouse, Catering to Specific departments or business units.

③ Virtual Data Warehouse :

A Warehouse built using Virtual Views and real-time Queries over the operational database rather than storing physical data.

* Delivery process :-

The data warehousing delivery process involves the following steps :

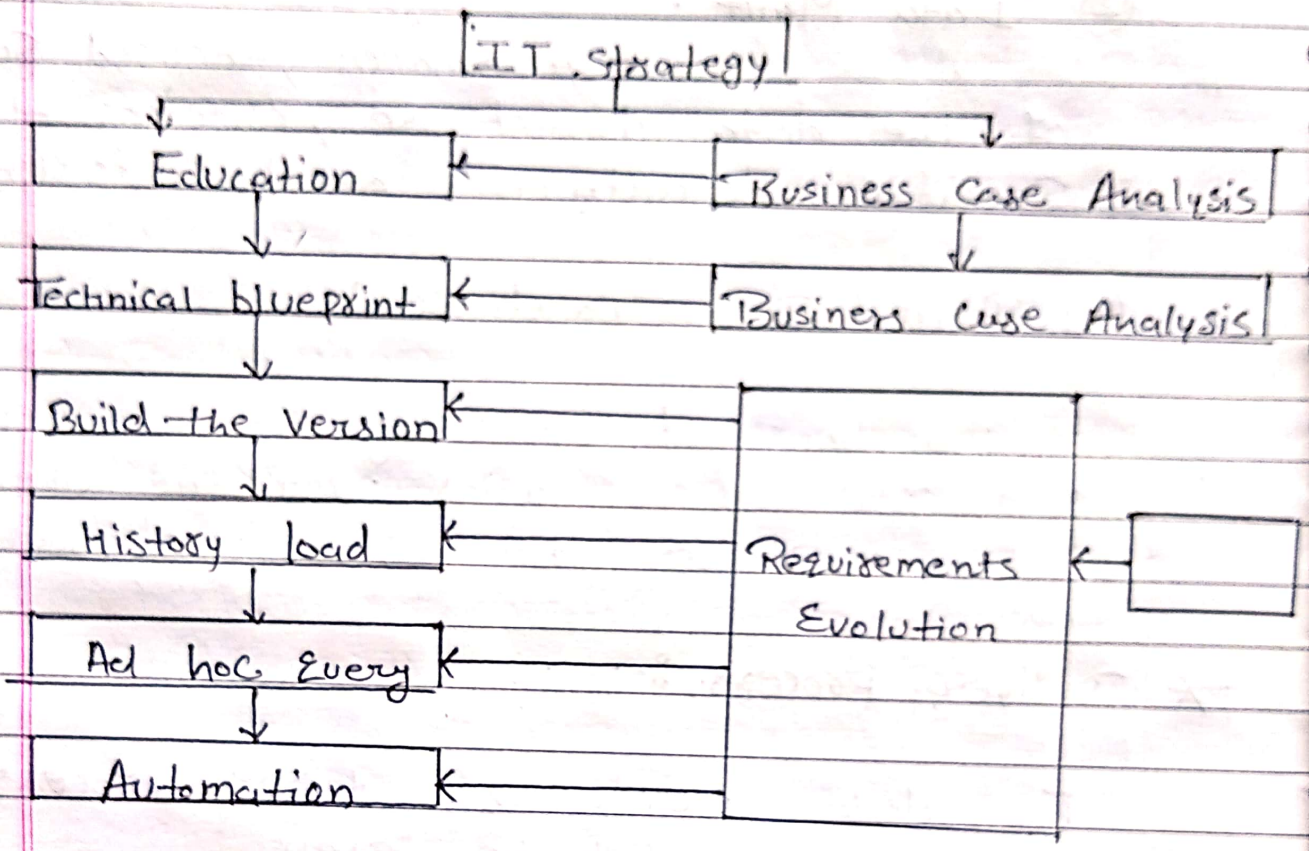
→ Extracting and loading data from different Source Systems.

→ Cleaning and transforming the data into a form Suitable for analysis.

→ Generating aggregations from predefined definitions within the data warehouse.

→ Determining aggregations to maintain Sys performance.

→ Backing up, restoring & archiving the data.



* Data Warehouse Architecture :-

Data Warehouse Architecture uses a structured framework to manage and store data effectively.

There are two common approaches to constructing a data warehouse :

-> Top - Down Approach

-> Bottom - Up Approach

Components of Data Warehouse Architecture ⇒

A data warehouse architecture consists of several key components that work together to store, manage and analyze data.

* External Source :

External sources are where data originates. These sources provide a variety of data types, such as structured data, semi-structured data or unstructured data.

* Staging Area :

The staging area is a temporary space where raw data from external sources is validated and prepared before entering the data warehouse. To handle this preparation effectively, ETL (Extract, Transform, Load) tools are used.

* Data Warehouse :

The data warehouse acts as the central repository for storing cleaned & organized data.

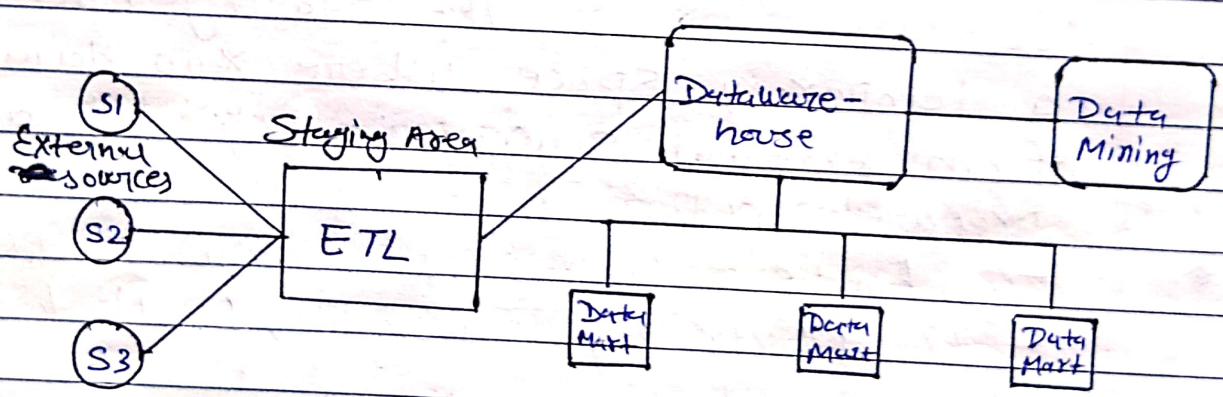
* Data Marts :

A data Mart is a subset of data warehouse that store data for a specific team or purpose, like sales or marketing.

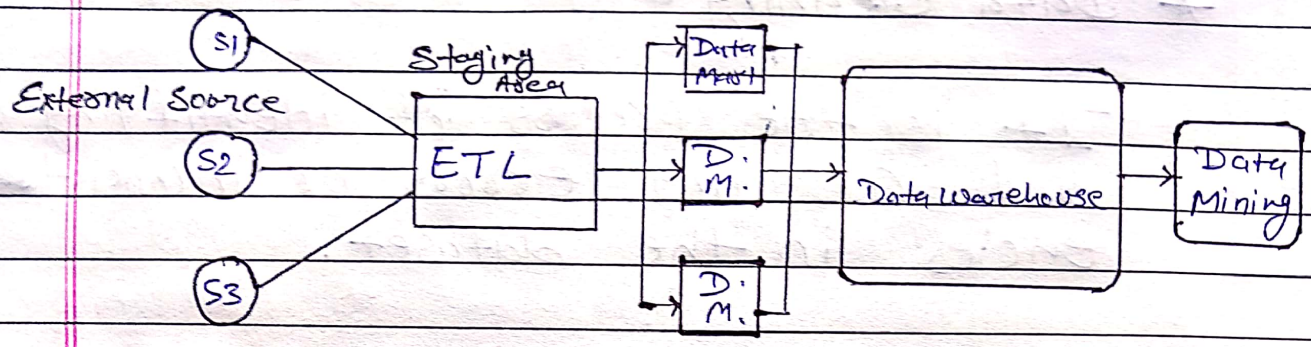
* Data Mining :

Data mining is the process of analyzing large datasets stored in the data warehouse to uncover meaningful patterns, trends, and insights.

(1) Top - Down Approach Archi. =>



(2) Bottom - UP =>



Data Pre-Processing :-

Data Pre-processing is the process of preparing raw data for analysis by cleaning and transforming it into a useable format.

→ Goal is to improve the quality of data.
→ Helps in handling missing values, removing duplicates and normalizing data.

→ Ensures the accuracy and consistency of the dataset.

Steps in Data Preprocessing :

Some key steps in Data preprocessing are Data Cleaning, Data Integration, Data Transformation and Data Reduction.

* Data Cleaning :

It is the process of identifying and correcting errors or inconsistencies in the dataset.

It involves handling missing values, removing duplicates and correcting incorrect or outlier data to ensure the dataset is accurate and reliable.

Clean data is essential for effective analysis, as it improves the quality of results and enhances the performance of data models.

- Mising Values =>

This occure when data is absent from a dataset. You can Either ignore the row with missing data or fill the gaps manually, with the attribute mean or by using the most probable value.

- Noisy Data =>

It refers to irrelevent or incorrect data that is difficult for machine to interpret, often caused by errors in data collection or entry.

It Can be handle in several ways:

- Binning Method
- Regression
- Clustering

- Removing Duplicates:

It involves identifying and Eliminating repeated data entries to ensure accuracy and Consistancy in the dataset.

* 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 and meanings. Techniques like record linkage and data fusion help in Combining data efficiently, ensuring Consistency & accuracy

Record linkage =>

R.L. is the process of identifying & matching records from diff. datasets that refer to the same entity, Even if they are represented differently.

It helps in Combining data from various sources by finding corresponding records based on common identifiers.

Data Fusion :-

DF involves Combining data from multiple sources to create a more comprehensive and accurate dataset.

It integrates info. that may be inconsistent or incomplete from different sources, ensuring a unified & richer dataset for analysis.

* Data Transformation :

It involves Converting data into a formate suitable for analysis. Common techniques include normalization, which scales data to a common range.

Standardization, which adjusts data to have zero mean and unit variance.

and discretization, which converts continuous data into discrete categories.

These techniques help prepare the data for more accurate analysis.

- Data Normalization
- Discretization
- Data Aggregation
- Concept Hierarchy Generation

* Data Reduction =>

It reduces the dataset's size while maintaining key info. This can be done through feature selection, which chooses the most relevent features, and feature extraction, which transform the data into a lower dimensional space while preserving imp. details. It uses various reduction techniques such as.

→ Dimensionality Reduction (eg PCA)

→ Numerosity Reduction

→ Data Compression

* Uses of Data Preprocessing :

→ Data Warehousing
→ Data Mining
→ M.L.

→ Data Science
→ Web Mining
→ Business Intelligence (BI)
→ Deep Learning purpose.

Advantages :

→ Improved data Quality
→ Better model performance
→ Efficient data Analysis
→ Enhanced Decision making

Disadvantages :

→ Time - Consuming
→ Resource - Intensive
→ Potential Data Loss
→ Complexity.

Data Warehousing Schemas :-

Schema :-

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.

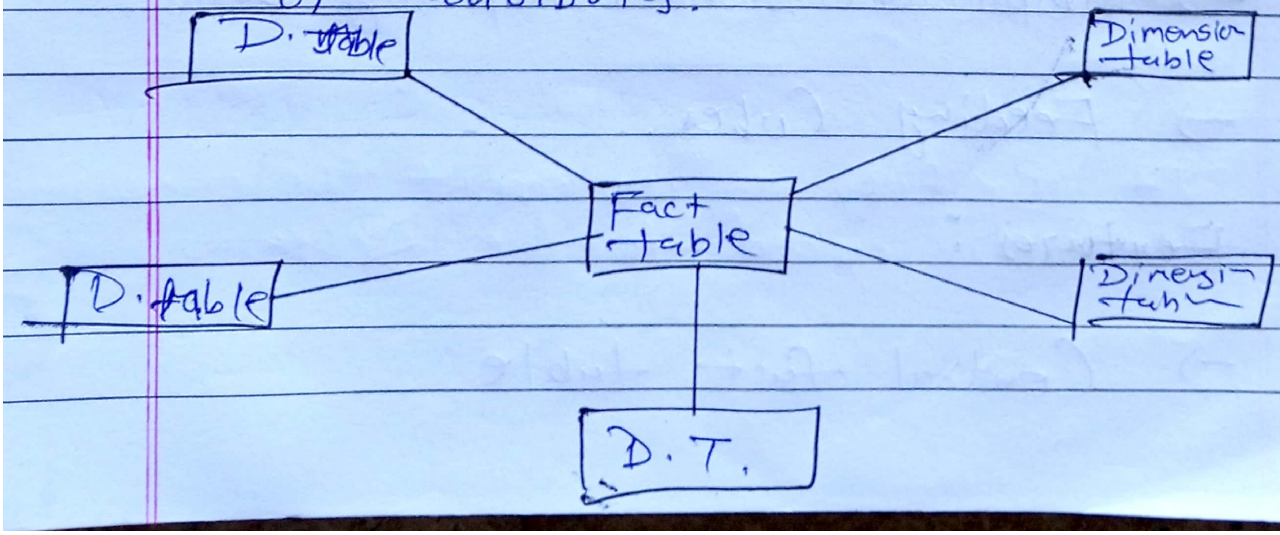
Much like a database, a data warehouse also requires to maintain a schema.

A database uses relational model, while a data warehouse uses Star, Snowflake and fact Constellation Schema.

(i) Star Schema =>

-> Each dimension in a Star schema is represented with only one dimension table.

-> This dimension table contains the set of attributes.



→ There is a fact table at the center. It contains the keys to each of four dimensions.

→ The fact table also contains the attributes.

→ This schema is widely used to develop or build a data warehouse & dimensional data marts.

→ A star schema having multiple dimensions is termed as centipede schemes.

→ It is easy to handle a star schema which have dimensions of few attributes.

Advantages:

→ Simpler queries

→ Simplified Business Reporting Logic

→ Feeding Cubes.

Features:

→ Central fact table

→ Dimension tables

→ Denormalized Structure

→ Simple Queries

→ Aggregated data

→ Fast Performance

→ Easy to understand

(ii) Snowflake Schema :-

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

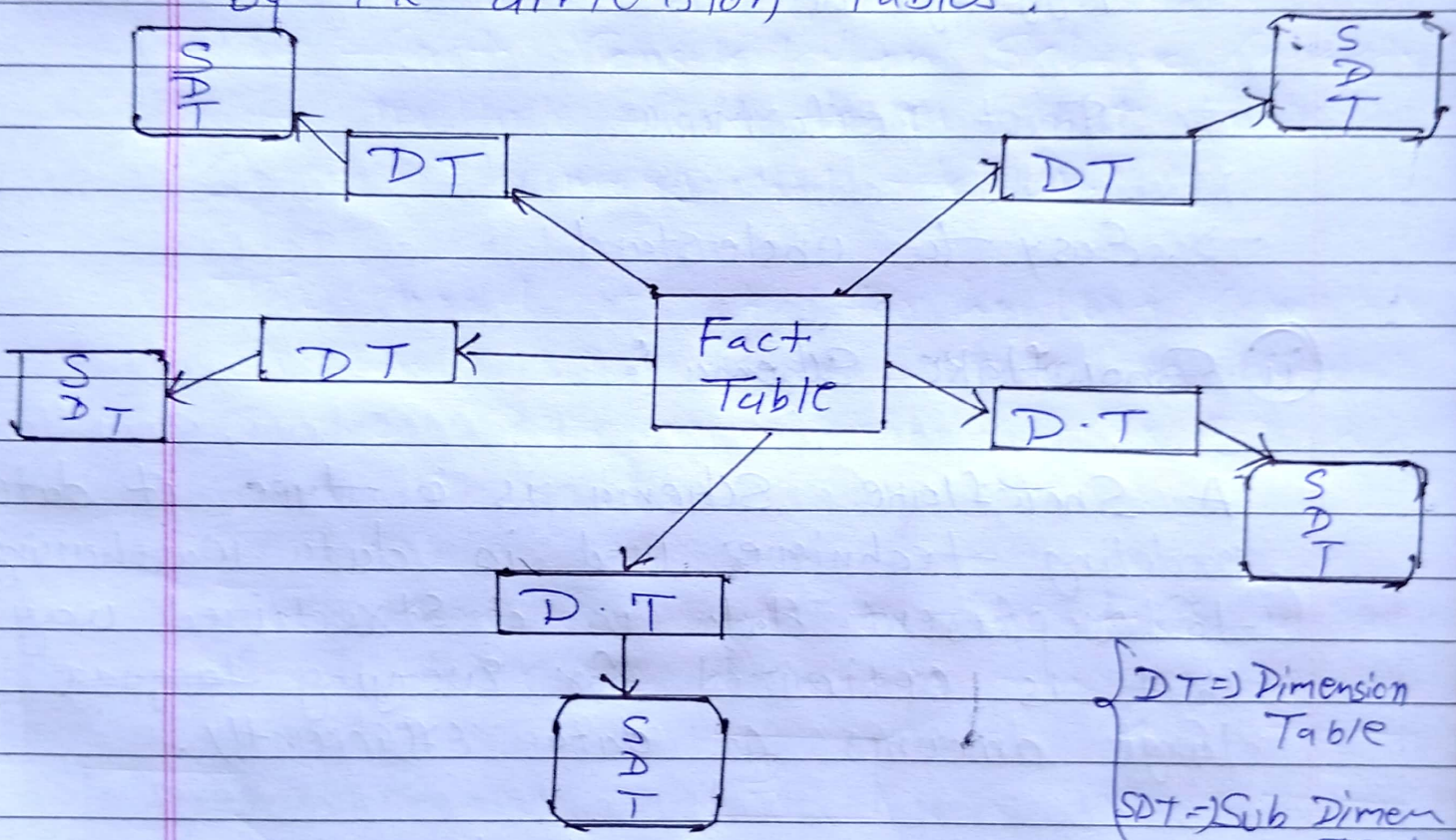
It is a varient of the star schema.

Here, the centeralized fact table is connected to multiple related tables.

The Snowflake effect affects only the dimension tables & does not affect the fact tables.

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

The fact table is still located at the center of the schema, surrounded by the dimension tables.



Characteristics :-

→ The Snowflake schema uses small disk space.

→ It is easy to implement the dimension that is added to the schema

→ There are multiple tables, so performance is reduced.

→ The dimension table consists of 100 or more sets of attributes.

Features :-

→ Normalization
→ Hierarchical structure
→ Multiple levels
→ Joins
→ Scalability

Advantages :-

→ It provides structured data which reduces the problem of data integrity

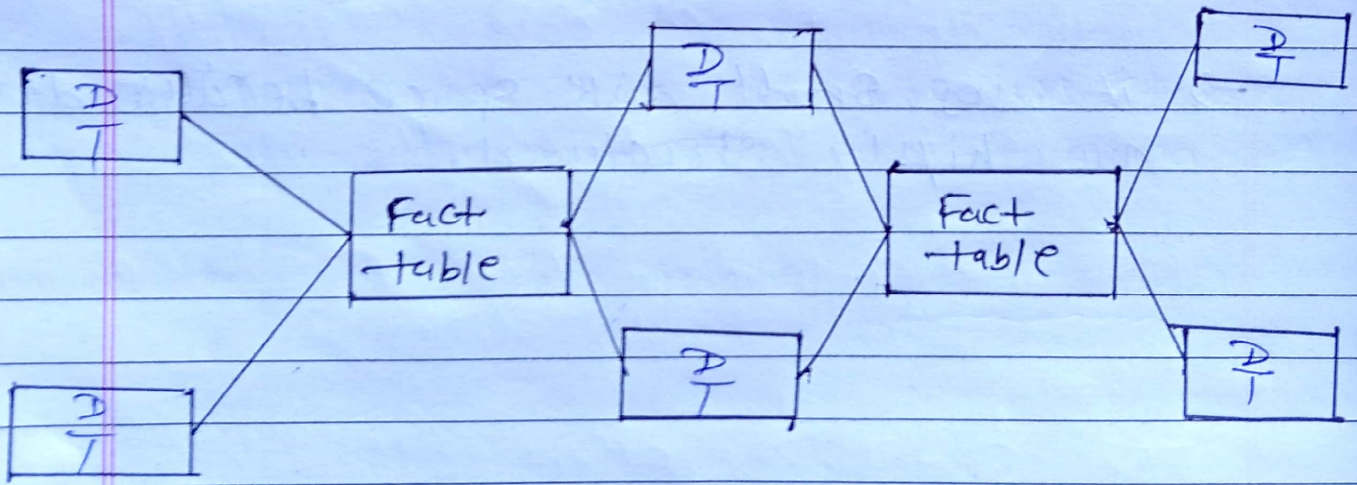
→ it uses small disk space because data are highly structured.

(iii) Fact Constellation :-

Fact Constellation Schema, also known as the Galaxy Schema, is an advanced data modeling technique used in designing data warehouse. Unlike simpler model like Star schema and Snowflake Schema the fact Constellation Schema Consists of multiple fact tables that share Common dimensional tables.

This model is ideal for handling Complex Systems and large-scale analytical queries, offering flexibility for BI & data mining.

The core Components of the FCS include Fact table & dimension tables.



Benefits =>

→ Enhanced Query Performance

→ Flexiblity in Reporting & Analysis

→ Improved Scalability

→ Simplified Data Management

→ Enhanced Data Consistency

→ Support for Complex Analytical Query

Challenges =>

→ increased Complexity in design

→ Performance issues with Complex Queries

→ Difficulty in maintaining Consistency

→ Data Redundancy & Storage Overhead

Feature	Star Schema	Snowflake Schema
structure	Central fact table connected to dimension tables	Fact table connected to normalized dimension tables.
Data normalization	Denormalized dimension tables	Normalized dimension tables.
performance	Faster query Execution due to fewer joins	Slower query Perfor due to multiple joins
Design Complexity	Simple & easy to understand	Complex design with multiple levels of relationship
Space Usage	Uses more storage due to denormalization	Use less storage due to normalization
Data Redundancy	Higher data redundancy	lower data redundancy
Foreign Keys	fewer foreign keys	More foreign keys
Learning Curve	Easier to learn & implement	More Complex to learn & implement

Pattern Warehousing :—

A pattern Warehouse is a specialized repo. that stores patterns identified through data mining processes in a persistent manner, Unlike traditional data warehouses, which focuses on storing vast amounts of raw data for historical analysis, a pattern Warehouse Consolidates refined and relevant patterns that carry Significant Sementic info, enabling quicker & more insightful decision making.

Key features =>

- Stored Patterns for Later use
- Context – Based Org.
- Fast & Smart Searching
- Works with Data Mining

Compare =>

→ Data Warehouse stores large amounts of raw data while pattern War stores summaries or patterns from that data.

→ DW good for reports & past data analysis while PW better for spotting trends & making decisions quickly.

Application →

→ Business

→ Healthcare

→ Finance

→ Social Media

Conclusion →

Pattern warehousing is like moving from just storing everything to storing only the useful patterns.

This helps business & org make faster, smarter decisions based on past behaviors & trends.

Unit => 2

OLAP :-

Online analytical processing (OLAP) is soft. technology you can use to analyze business data from different points of view.

Org's collect & store data from multiple data sources, such as website, app., smart meters, and internal system.

OLAP Combines & groups this data into categories to provide actionable insights for strategic planning.

for eg: a retailer stores data about all the products it sells, such as Color, Size, cost, and location. The retailer also collects Customer purchase data, such as the name of the items ordered & total sales value, in a different system. OLAP Combines the datasets to answer qus. such as which Color products are more popular or how product placement impacts sales.

OLAP Archi =>

Online analytical Processing (OLAP) systems store multidimensional data by representing info. in more than two dimensions,

or Categories. Two dimensional data involves Columns and rows, but multidimensional data has multiple characteristics. for Eg, multidim data for product Sales might Consist of the following dimensions:

→ product Type
→ Location
→ Time

Data Enginers build a multi OLAP Sys. that Consists of the following Elements:

* Data Warehouse
* ETL tools
* OLAP Server
* OLAP database
* OLAP Cubes
* OLAP Analytic tools

How does OLAP work?

An OLAP Sys. works by Collecting, organizing, aggregating & analyzing data using the following steps:

① The OLAP Server collects data from multiple data sources, including relational databases & data warehouses.

② Then, the ETL tools clean, aggregate, precalculate & store data in an OLAP cube according to the no of dimension specified.

③ Business analysts use OLAP tools to query & generate reports from the multi-dim data in the OLAP cube.

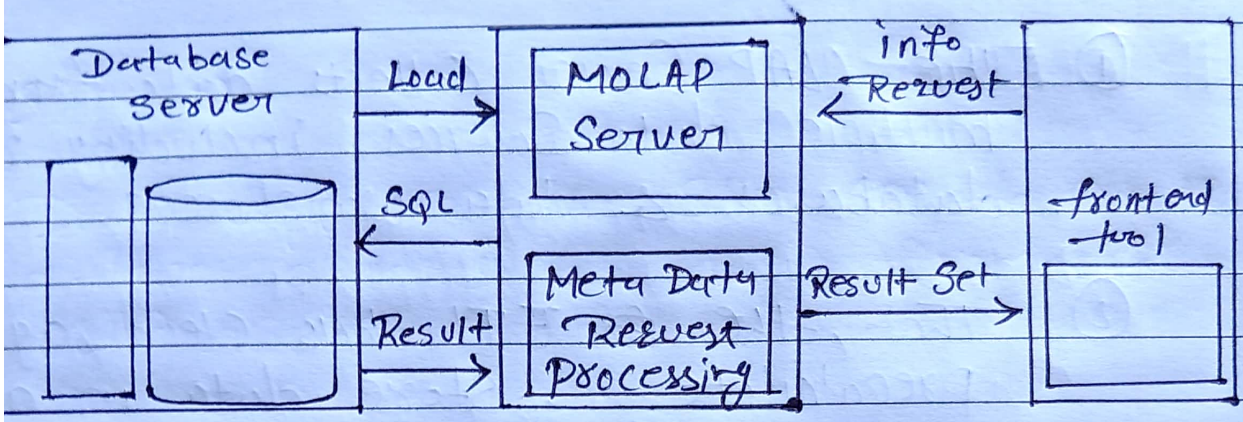
Types of OLAP :-

OLAP Sys. operate in 3 main ways :

* Multidimensional OLAP (MOLAP) :

MOLAP involves creating a data cube that represents multi-dim data from a data warehouse.

The MOLAP Sys. stores precalculated data in the hypercube. Data engineers use MOLAP because this type of OLAP technology provides fast analysis.



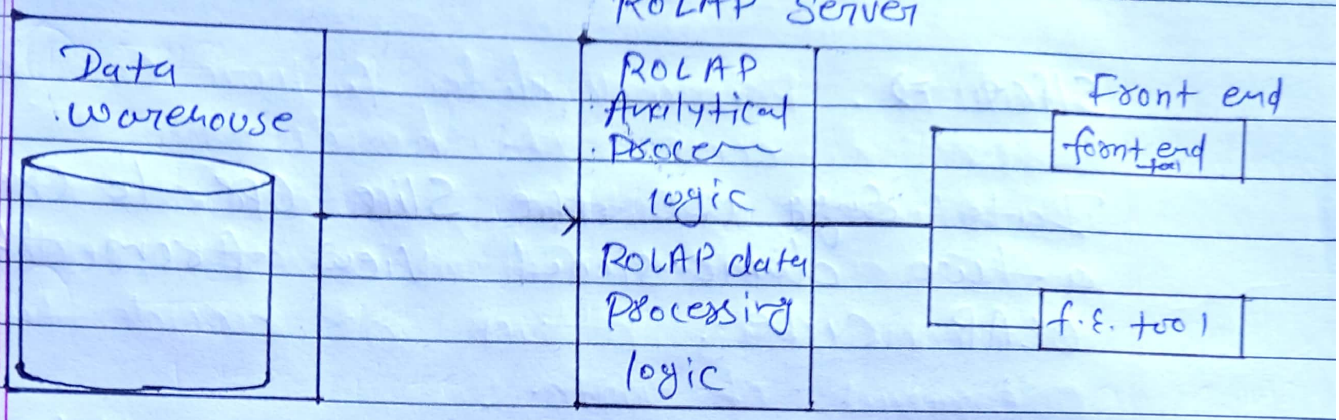
MOLAP

* Relational OLAP (ROLAP) :-

Instead of using a data cube, ROLAP allows data engineers to perform multi-dim data analysis on a relational database.

In other words, data engineers use SQL queries to search for and retrive specific info based on the required dimensions.

ROLAP is suitable for analyzing extensive & detailed data. However, ROLAP has slow query performance compared to MOLAP.



* Hybrid OLAP (HOLAP) :-

Hybrid online analytical processing Combines MOLAP & ROLAP to provide the best of both archi. HOLAP allows data Engg. to quickly retrieve analytical results from a data cube & extract detailed info from relational database.

OLAP Operations :-

Business analysts perform several basic analytical opr with a mul~ OLAP Cube.

Roll up =>

In roll up, the OLAP Sys. summarizes the data for specific attributes.

Drill down =>

Drill down is the opposite of the roll up opr.

Slice =>

Data Engg use the Slice Opr to create a two dimensional view from the OLAP cube.

for Eg, a MOLAP cube sorts data according to products, cities & month

By Slicing the cube, data Engg can create a spreadsheet-like table Consisting of products & cities for a specific month.

Dice =>

Data Engg use the dice opr to create a smaller subcube from an OLAP Cube. They determine the required dimensions & build a smaller cube from the orginial hypercube.

Pivot =>

The pivot opr. involves rotating the OLAP cube along one of its dimensions to get a diff perspective on the multi-dim data model.

for eg, a three-dimensional OLAP cube has the following dimensions on the respective axes:

- X-axis — product
- y-axis — location
- Z-axis — time

Upon a pivot, the OLAP cube has the following configuration:

- X-axis - location
- y-axis - time
- z-axis - product

Data Warehouse Backup & Recovery:

Backup & recovery define the process of backing up records in the method of a loss & setting up systems that enable that data recovery because of data loss. Backing up data needed copying & archiving computer info., so that it is applicable in case of data deletion or corruption.

The goal of the backup is to make a copy of data that can be renewed in the event of a primary data

failure. Primary data failure can be the outcome of hardware or software failure, data corruption, or a human caused event, including a malicious attack or accidental removal of info.

There are various types of backup which are as follows —

- Complete backup →

The entire database is backed up simultaneously. This includes all data files, control file, & journal files.

- Cold backup →

It is a backup that is taken while the database is completely shut-down.

- Hot backup →

The term hot is used because the database engine is up & running.

Recovery is the phase of reconstructing a database after some element of a database has been hidden.

The recovery model of a current database is inherited from the model database when the new database is generated.

The model for a database can be changed after the database has been created.

→ full recovery model

→ Bulk-logged recovery model

→ Simple recovery model

Unit => 4

"Supervised Learning"

Supervised learning is a subcategory of machine learning and AI that involves training algo's on labeled datasets to classify data or predict outcomes accurately.

In SL the machine is trained on a set of labeled data, which means that the i/p data is paired with the desired o/p. The machine then ~~Predict~~ learns to predict the o/p for new i/p data.

Types =>

SL is classified into two categories of algo:

① Regression:

Used to predict Continuous Values, such as house prices or stock prices. Common regression algo's include linear regression, polynomial regression, SVM ——, Decision Tree ——, & Random forest ——

(2) Classification :

Used to predict Categorical values, such as whether an Email is Spam or not. Common classification algo's include logistic regression, SVM, Decision Trees, Random forests & Naive Bayes.

Ques What is the imp of Data Marts in DW ?

A data mart is a specialized subset of datawarehouse, designed to focus on a specific functional area or department within an organization.

It provides a simplified & targeted view of data, address specific reporting & analytical needs.

Data marts are small in scale & scope, typically holding relevent data for a specific group of users, such as Sales, marketing or finance.

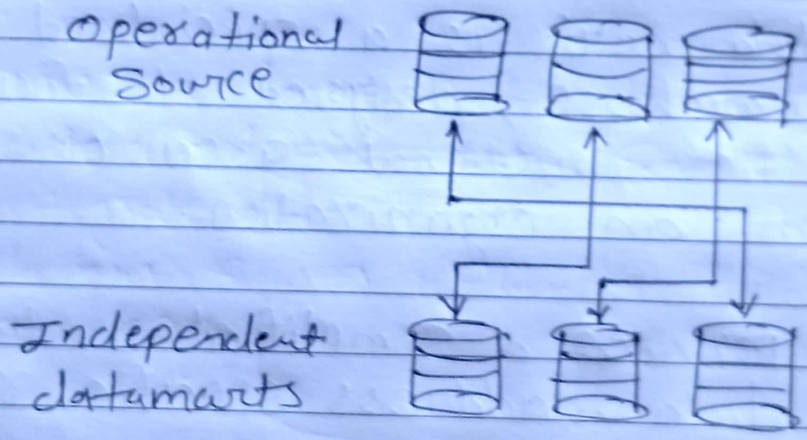
Types =>

There are 3 Common types of data marts :

① Independent Data Mart :

Created and maintained separately from the data warehouse, it satisfies the particular needs of a specific business unit or department.

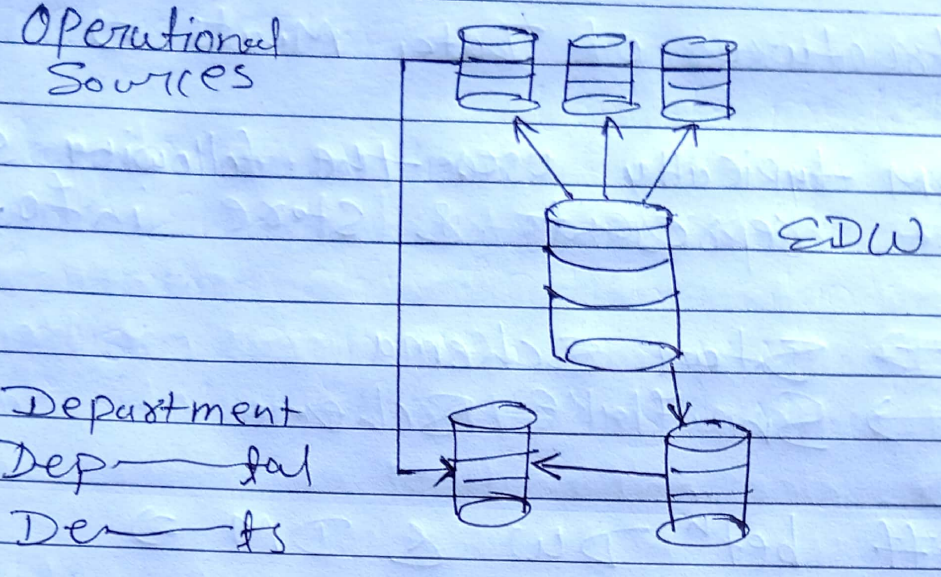
Ind~ D. M. offer flexibility & agility but may result in data redundancy & inconsistency.



② Depender D M =>

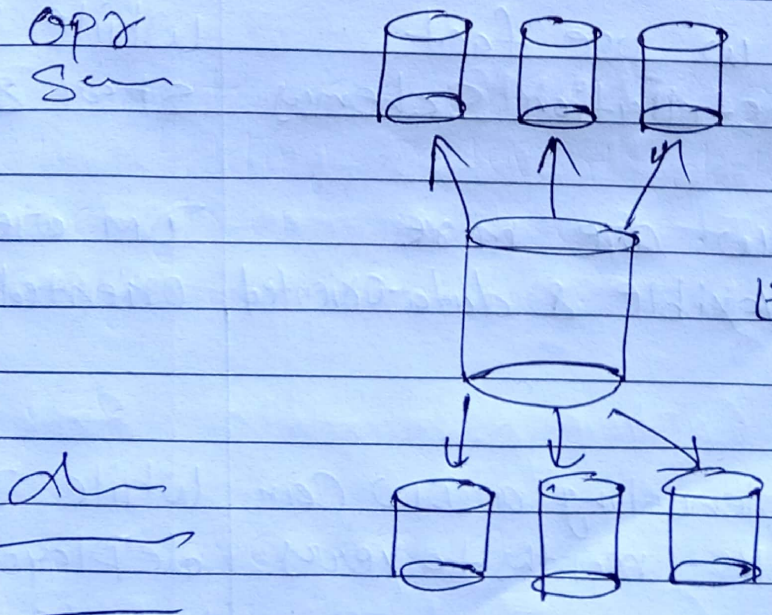
Generated directly from a data warehouse, it arranges data to meet the needs of a specific industry.

Dependent data marts benefit from the data integration, quality and consistency provided by the DW , preventing data duplication.



② Hybrid D—M :—

Combines Components of both indep & dep data marts. It uses the Centralized data Wa~ for core data integration & Consistency while incorporating additional data Sources specific to a business unit or departm.



Structures of Data Marts =>

DM typically uses the following structs to represent & store info. :

→ Star schema
→ Snowflake Schema

Diff bet^n DW & DM =>

Feature	D.W.	D.M.
Scope	DW are large, ~~center~~ centeralized sys that integrate data from various sources within an org.	While DM are smaller, decentralized sys. focused on specific business areas.
Data Integration	D W use fact Constellation Schemas	While Dm use star & Snowflake
flexiblity	DW are more flexible & data-oriented	DM are project-oriented & less flexible
Implementation time	implementing a DW can take months to years.	While DM can be deployed in few months.

Advantages ⇒

→ Optimized performance

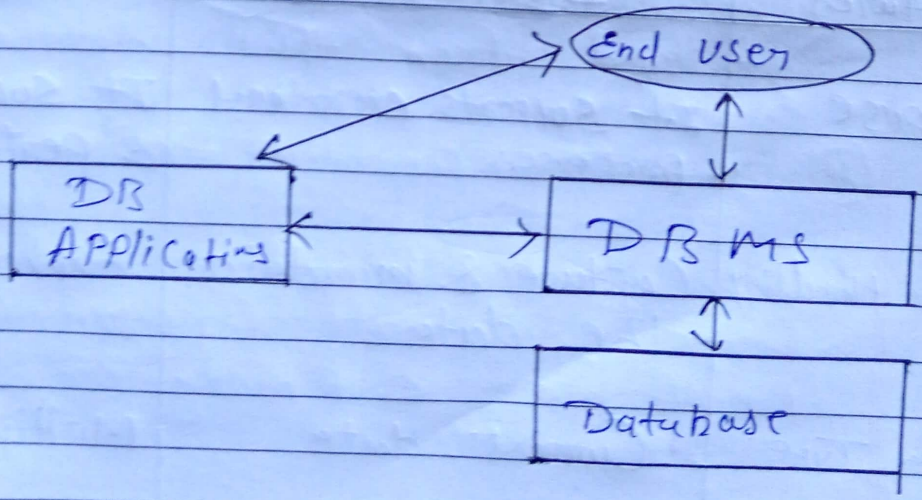
→ Faster Query Response

→ Empower Business users

DBMS vs DW ⇒

DBMS :

DBMS is used in the traditional way of storing & retrieving data. The major task of a database system is to perform query processing. These sys are generally referred to as online transaction processing systems These sys are used in the day-to-day operations of any organization.

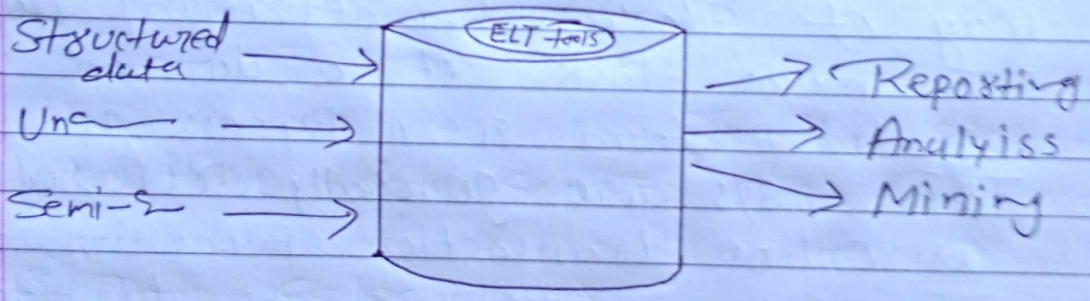


Data Warehouse :-

DW is the place where huge amount of data is stored.

It is meant for users or knowledge workers in the role of data analysis & decision making.

These systems are referred as Online analytical processing



Features	DBMS	DW
purpose	It supports operational processes	It supports analysis & performance reporting
Data Handling	Capture & maintain the data	Explore the data
Data Type	Current data	Multiple years of history

Data Verification	D.V. occurs when Entry is done	D V occurs after the fact.
Data Size	100 MB to GB	100 GB to TB
Data Model	ER based	Star/Snowflake
Orientation	App. oriented	Subject ori~
Data Specificity	primitive & highly detailed	Summarized & Consolidated
Storage Structure	Flat relational	Multidime~

Knowledge discovery :

Knowledge discovery refers to the complete process of uncovering valuable knowledge from large datasets. It start with selection of relevent data, followed by preprocessing to clean & organize it, transformation to prepare it for analysis data mining to uncover patterns & relationships, & concludes with the evolution & interpretation of results,

Ultimately producing valuable knowledge or insights.

KD is widely utilized in fields like ML, pattern recognition, statistics, AI & Data Visualization.

Here is the list of steps involved in the KD process :-

(1) Data Cleaning =>

In this, step, the noise & inconsistent data is removed.

(2) Data Integration =>

In this step, multiple data sources are combined.

(3) Data Selection =>

In this step, data relevant to the analysis task are retrieved from the database.

(4) Data Transformation =>

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

⑤ Data Mining =>

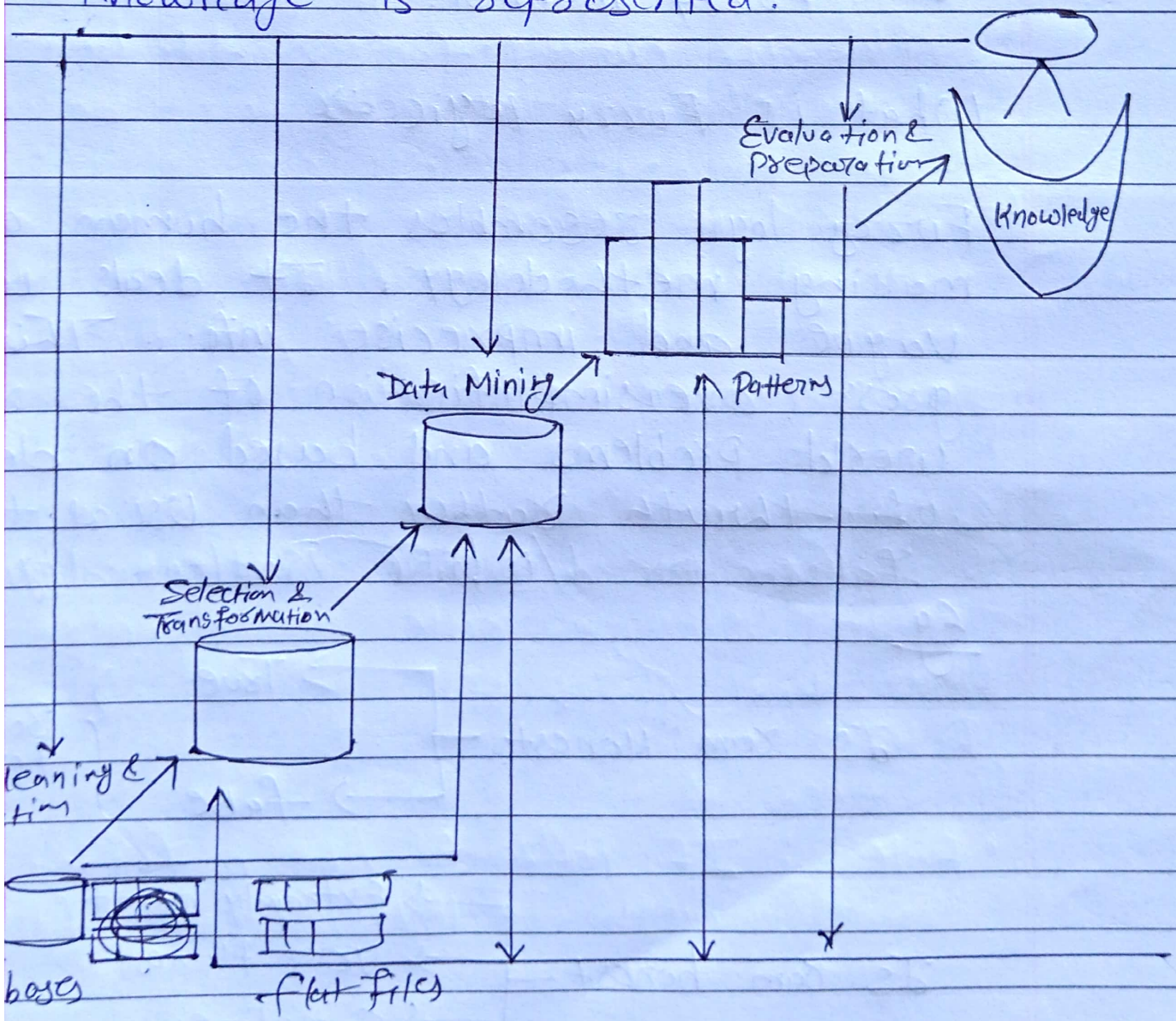
In this step, intelligent methods are applied in order to extract data pattern.

⑥ pattern Evaluation =>

In this step, data patterns are evaluated.

⑦ Knowledge presentation =>

In this step, knowledge is represented.



Fuzzy logic & Sets =>

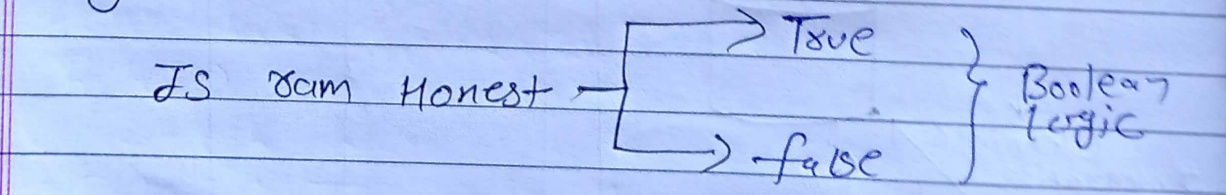
Fuzzy logic :

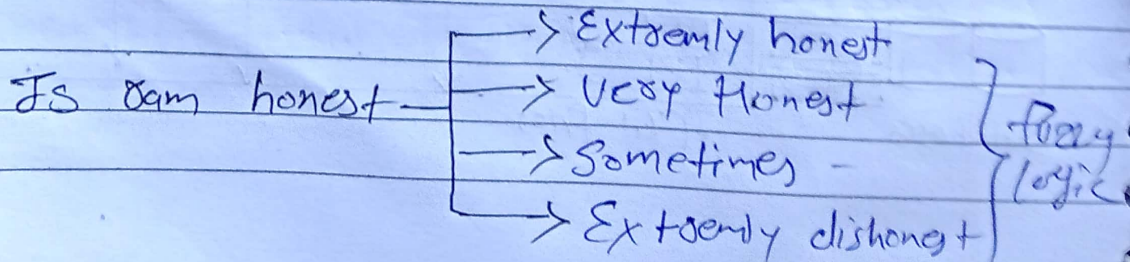
The word Fuzzy refers to things which are not clear or are vague. Any event, process, or function that is changing continuously can not always be defined as either true or false, which means that we need to define such activities in a Fuzzy manner.

What is Fuzzy logic ⇒

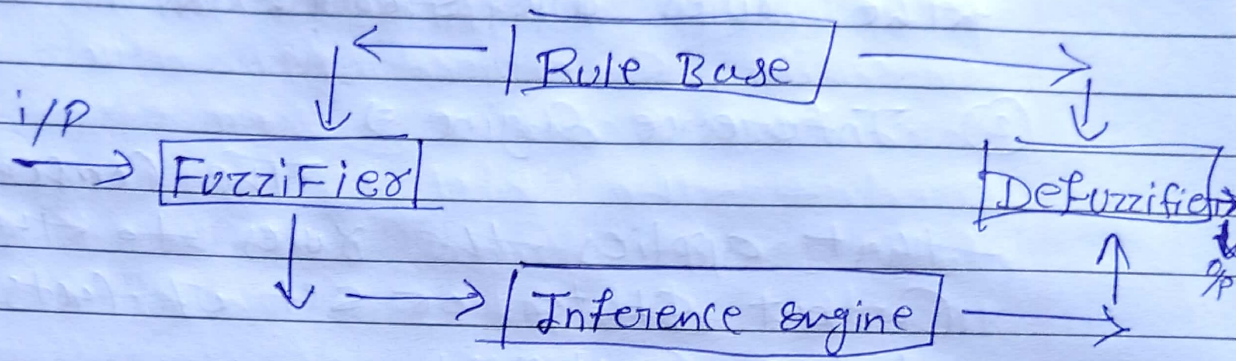
Fuzzy logic resembles the human decision making methodology. It deals with vague and imprecise info. This is gross oversimplification of the real world problems and based on degrees of thruth rather than usual true/false or 1/0 like Boolean logic.

Eg





Architecture :



Rule based Algo. =)

A rule based algo is a sys that uses a set of specific rules to reach decisions or conclusions.

These rules typically follow an "IF Condition THEN action" structure.

for eg, a rule might state, "if the temp is above 30°C THEN issue a heat warning."

This transperency makes such algo easy to understand & implement, especially in field like AI & data mining.

How rule based algo work :

① Knowledge Base =>

This is where all the rules are stored.

② Inference Engine =>

The Component that applies the rules to the Current State or set of facts to derive Conclusions.

③ Working memory =>

Holds the current facts & data being processed by the algo.

④ User interface =>

Allows users to interact with the sys, inputting data & receiving outputs.

Eg :-

Consider a Simple medical diagnosis System using a rule based algo :

Rules =>

- Rule 1 : IF patient has a fever AND Cough THEN consider flu.

- Rule 2 : IF patient has a sore throat AND fever THEN Consider Strep throat.

Applications =>

-> Expert Sys
-> Decision Support Sys
-> Control Sys

Advantages =>

-> Transparency
-> Consistency

Disadvantages =>

-> Rigid
-> Maintainance

KNN :-

K-Nearest Neighbors (KNN) is a simple, versatile, Supervised ML algo used for classification & regression that predict the label or value of a data point based on the 'K' NN in the dataset.

Overview of KNN =>

-> S.L. algo applicable in both clas & reg tas

-> KNN classifies a data point based on the majority class among its KNN or predicts a value using the avg of the K neighbors values.

-> KNN does not build a model upfront; it stores the training data & performs calculations at the time of prediction, making it a non-parametric method.

How KNN works =>

① Select K :

Choose the no of neighbors (K) to consider.

② Calculate distance :

Compute the distance bet^n i/p data point & all points in the training dataset using distance metrics such as :

- Euclidean Distance
- Manhatten ——
- Minkowski ——

③ Find nearest neighbors :

Sort the distance from the query point to every other point & select the top K closest points.

④ predict o/p :

- For classification, assign the class label that is most common among the K neighbors.
- For rege——, return the avg value of the K neighbors.

How to choose right K value :

Choosing the optimal K is critical ; techniques include :

→ Cross - validation
→ Elbow Method

Advantages =>

- Simplicity
- flexiblity
- Adaptability

Disadvantages =>

- Computationally intensive
- Sensitive to outliers
- Curse of dimensionality

Applications =>

- Recommendation Sys
- Image Classification
- Medical Diagnosis.

Data Types Uses in Clustering ?

Main Data Types in clustering =>

1. Numerical Data =>

Num or Quantiative data Consists of Continuous values that can be measured & ordered.

Clustering algo's like K-mean & Gaussian mixture models (GMM) typically work well with num data.

(2) Categorical data :

Categorical or qualitative data refers to variables that represent distinct categories or groups.

Algo's like K-modes can be used for Clustering Cater data.

Types of Clustering algo's & data uses =>

* Centroid-Based Clustering primarily for num
* Density — Suitable for both
* Distribution — primarily numerical data
* Hierarchical — works with both numerical depending on linkage method
* Fuzzy — Mostly numerical data, but can be used with categorical data when membership probabilities apply.

Apriori algo :-

The Apriori algo, proposed by Rakesh Agrawal and Ramakrishnan Srikant in 1994, is widely Employed in data mining to uncover patterns of items that frequently occur together in transactional databases.

It's particularly effective for app. like market basket analysis, where businesses aim to understand Consumer buying behaviour & optimize product placement.

Key Concepts :

① Itemsets =>

An itemset is a collection of one or more items.

② Frequent itemsets =>

These are itemsets that appears in the dataset with a frequency above a set threshold known as support.

(3) Support ⇒

This measures how often an itemset occures in the dataset

$$\text{Support}(X) = \frac{\text{no of transactions containing } X}{\text{Total no of transactions}}$$

(4) Confidence ⇒

Confidence measures the likelihood that an item y is purchased when item X is purchased.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

(5) Lift ⇒

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Confidence}(X \rightarrow Y)}{\text{Support}(Y)}$$

How its work :

* Set parameters
* Find Frequent 1-itemsets
* Generate Candidate itemsets
* prune Candidates
* Repeat for Higher-Order itemsets
* Generate association rules

Applications :

Market Basket Analysis
Recommen sys
Inventory mon
Fraud detec

Advantage =)

The algo is Easy to implement & understand, providing interpretable results in the form of association rules

limitations =)

The Apriori algo can be Computationally Expensive, Especially with large datasets, due to the multiple database scans required.

BIRCH :-

BIRCH is a Scalable and efficient clustering technique that enables effective data grouping in applications where datasets are too large to fit into memory.

It is designed to Summarize extensive datasets into smaller clusters, which can subsequently be clustered further

by other algo's, enhancing performance & reducing processing time.

Key Components =>

① Clustering Feature (CF):

The CF is a summary of a cluster that comprises three statistics:

- N => The no of points in the cluster.
- LS => The linear sum of the points.
- SS => The squared sum of the points.

② CF Tree :

The CF tree is hierarchical tree structure where each node represents a cluster.

It enables efficient storage of clustering features, allowing the algo to summarize large datasets without needing to store all individual data points

Advantages :

Scalability
Memory Efficiency
Incremental Clustering

Limitations of BIRCH :

Sensitivity to parameters

Assumption of Spherical clusters

Distance Metric limitations.

Data Mining Tasks :-

Data Mining Task Can be Classify into two Categories :

1. Descriptive Data Mining :

This Category focuses on finding human - interpretable patterns describing the data.

It aims to Summurize, Visualize and Understand the inherent characteristics of dataset.

Common task within this Category include :

- Classification => predicting if an Email is Spam or not based on specific features.

- Clustering ⇒

Clustering is the task of grouping Similar data objects together based on their attributes.

- Association Rule learning ⇒

This technique identify interesting relationships bet" variables in large datasets., often used in market basket analysis.

- Summarization ⇒

It Can Summarize key patterns, trends or statistics in a more compact form for decision making.

- Outlier decision ⇒

This task identifies data points that deviate significantly from the majority of the data.

(2) Predictive Data Mining :

predictive tasks aim to develop a model that predicts future outcomes based on existing data.
This tasks typically use historical data

to drive predictions. Key types :

- Regression Analysis =>

This technique is used to predict a quantitative response based on the relationships bet^n independent variables and the dependent variable

A common app is predicting sales based on advertising spend.

- Time Series Analysis =>

Technique analyze trends over time for forcasting future values like stock market predictions.

- Forecasting =>

Similar to regr^n but often used for broader prediction contexts, forcasting helps predict unknown values within a dataset & it is commonly used in various industries for demand or supply forecasting.

DBSCAN :-

DBSCAN is an unsupervised ML algo used for clustering tasks. It groups together points that are closely packed in the feature space while marking points in low-density regions as noise.

DBSCAN is particularly useful in scenarios where the no of clusters is not pre-defined & is capable of discovering clusters of arbitary shapes.

Key Concepts

- Core points =>
  Points that have at least a minimum no of other points within a specified distance (ε).

- Border Points =>
  Points that are within ε distance of a core point but do not have enough neighboos to be considered core points themselves.

- Noise Points =>
  Points that are neither core nor border points; they don't belong to any cluster.

How its work =>

(1) Parameter Selection :

Set two parameters, epsilon (ε) for the neighborhood radius & MinPts for the mini. no of points required to form a dense region.

(2) Cluster formation :

→ for each point in the dataset, DBSCAN retrieves the neighborhood within ε distance.

→ If enough neighbors (MinPts) are found, a new cluster is initiated; otherwise, the point is labeled as noise

→ The algo recursively Expands the cluster by adding all density-connected points to the cluster until no new points can be added

Advantages :

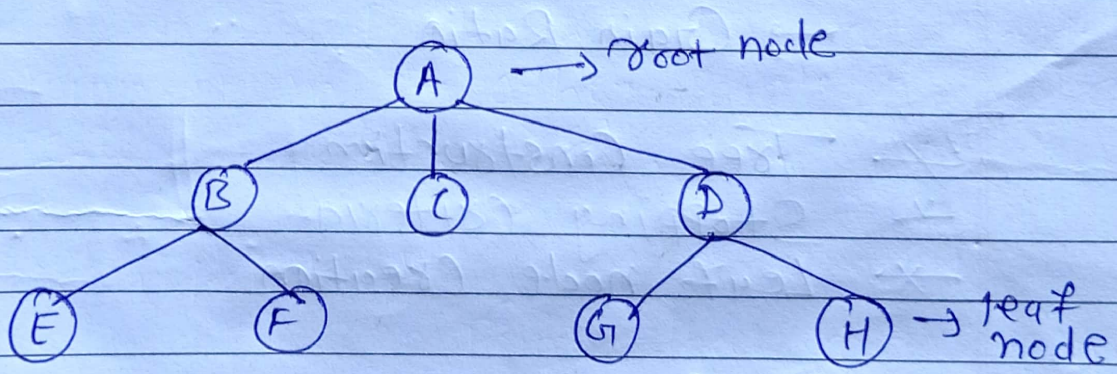
* No need for pre-defined no of clusters
* Flexibility in cluster shape
* Noise Handling

Disadv :

* Sensitive to parameters
* Difficulty with Varying Densities
* Performance on High Dimensions

Decision Tree induction algo :-

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.



The benefits of having a decision tree are as follows —

* It does not require any domain knowledge
* It is easy to comprehend
* The learning & classification steps of a decision tree are simple & fast

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algo know as ID3 (Iterative Dichotomiser).

In this algo, there is no backtracking; the trees are constructed in a top down recursive divide-and-conquer manner.

How its work =)

* Data Preparation
* Attribute Selection
  - → Info Gain
  - → Gini Index
  - → Gain Ratio
* Tree Construction
* Stopping Criteria
* Leaf node Creation

pruning :-

pruning is important to prevent overfitting, where the model becomes too complex & performs poorly on unseen data.

There are two main types of pruning:

pre-pruning => Stopes the tree growth early based on certain criteria

post ——— => first creating the full tree & then removing branches that have little significance to improving accuracy on validation data.

~~Numpy & Pandas~~

Advantages =>

Easy to Understand
Handles various data
Flexible Applications

Disadvantages =>

prone to overfitting
Sensitive to Data Changes
Bias Toward Certain Attributes.

Neural Network based algo ?

Neural Network Classification involves using various neural network archi. to automatically categorize i/p s into specific classes.

This technique has proven successful in domains such as image recognition, NLP & medical diagnosis, where complex data needs sophasticated handling.

Common neural network archi. =>

* CNN
* RNN
* FNN
* GAN etc

FP-Growth :-

The Frequent pattern Growth (FP-Growth) algo is widely used in data mining to discover frequent patterns or itemsets in transactional dataset, such as market basket analysis. It improves upon the Apriori algo by minimizing the no of database scans and eliminating the need for Candidate generation, which can be Computationally Expensive.

Key Concepts :

- Frequent Itemsets =>

Groups of items that appear together frequently in transactions.

- Support =>

The frequency of an itemset in the dataset. It is defined as the proportion of transactions that contain the itemset.

- FP - Tree =>

A Compressed tree structure that retains the essential info of the dataset, allowing for efficient frequent pattern mining.

How its works :

Step 1 : Build the FP - Tree

→ Scan the data
→ Sort items
→ Construct the Tree

Step 2 : Mine the FP-Tree for frequent patterns

→ start from the bottom of the tree & identify frequent itemsets up to the root

→ for each item in the tree, create a Conditional pattern base, which consists of the paths leading to that item.

→ Construct a new Conditional FP-tree from the Conditional pattern base to find associated frequent pattern

Advantages :

* Efficiency
* No Candidate Generation
* Scalability

Limitation :

* Complex Implementation
* Memory Usage

Naive Bayes :-

Naive Bayes is a family of probabilistic classifiers that apply Bayes' Theorem, assuming strong independence bet^n the features.

This means that the presence of one feature does not affect the presence of another, simplifying the computation of probabilities.

The algo computes the posterior prob. of a class given a set of features & selects the class with the highest probability. The formula is :

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

where,

$P(Y|X)$ is the posterior prob.,
$P(X|Y)$ is the likelihood,
$P(Y)$ is the prior prob.,
$P(X)$ is the evidence

Types :

→ Gaussian Naive Bayes

→ Multinominal Naive Bayes

→ Bernoulli ——— —

→ Complement ———

Advantages :

→ Simplicity
→ Speed
→ Performance

Disa— :

→ Independence Assumption
→ Zero Probability Problem

Application :

→ Text Classification
→ Medical Diagnosis
→ Recommendation Sys
→ Weather prediction

Category	OLAP	OLTP (Online Transaction Processing)
Definition	It is well known as online database query management sys.	It is well known as an online database modifying sys.
Data source	Consists of historical data from various DB	Consists of only operational current data.
Method used	It makes use of a data warehouse.	It makes use of a standard DBMS.
Application	~~It makes us~~ is subject oriented. Used for data mining, Analytics, Decisim making etc	It is app. oriented. Used for business tasks.
Normalized	In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized.
Nature of Audience	The process is focused on the Customer	——— the market
Operations	Only read & rarely write opr.	Both read & write opr.
DB design	Design with a focus on the subject	Design that is focused on the app.