

Unit-1 (DWDM)

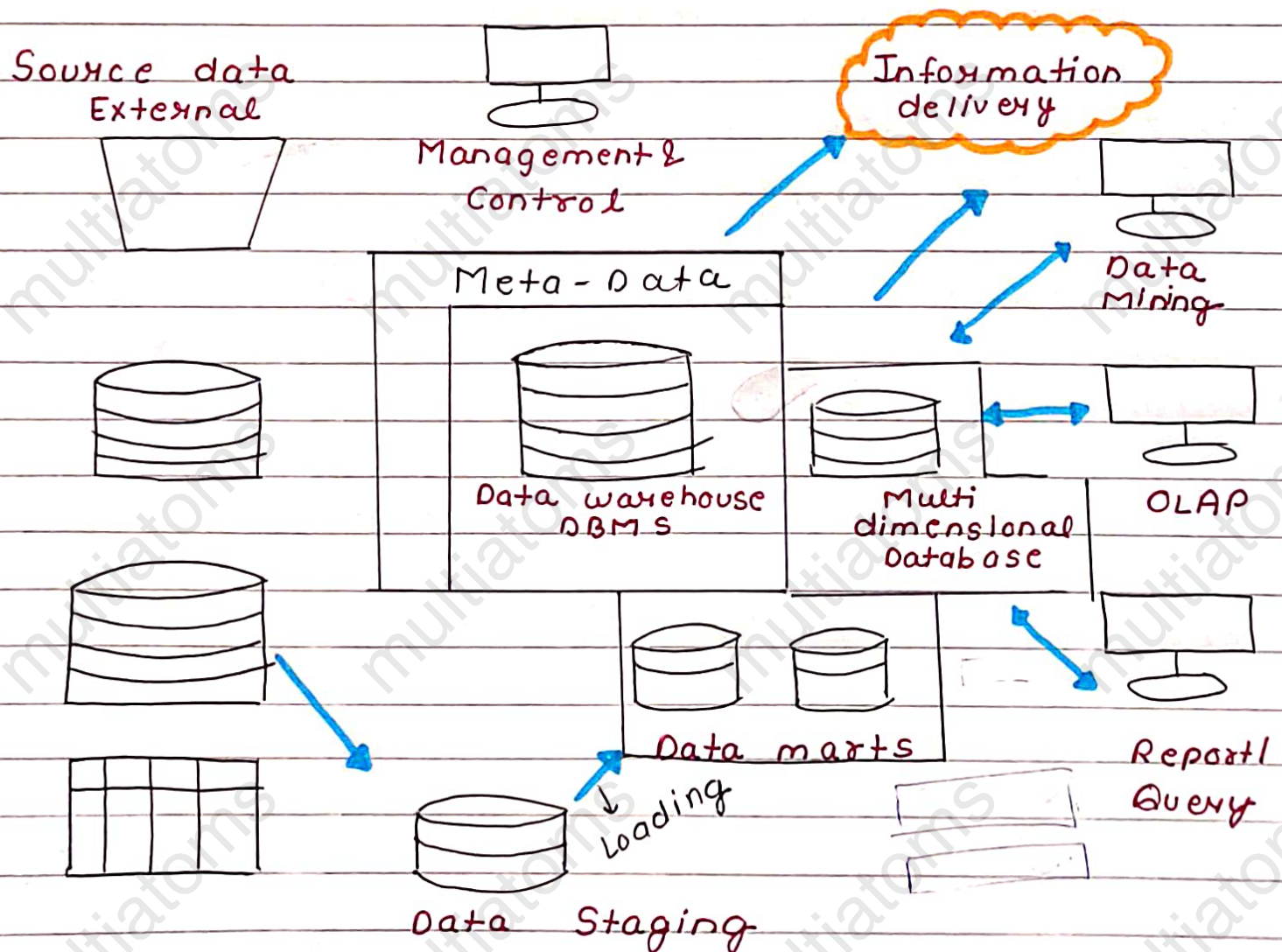
- Overview of Data Warehouse [AKTU-21-22, 22-23, 23-24]
- Data Warehousing Component
- Building a Data Warehouse [AKTU-21-22]
- Mapping the Data Warehouse to a Multi processor Architecture [AKTU-21-22]
- Difference between Database System and data warehouse
- Multi-dimensional Data model [AKTU-22-23, 23-24]
- Data Cubes [AKTU-22-23]
- Star, Snow Flakes [AKTU-22-23, 23-24]
- Fact Constellations [AKTU-22-23, 23-24]

SUBSCRIBE 

Data Warehouse :- [AKTU-21-22, 22-23, 23-24]

A data warehouse is like a big Storage Room where company keeps all its important information safely in an organized way. It collects data from different places, arranges it properly and makes it easy to find when needed.

Data Warehouse Components



1. Source Data

→ This is where the raw data comes from

→ It can be external sources (like → online database files) or internal sources (like → company records, sales data).

2. Data Staging (Temporary Storage & Processing)

→ Before storing data in the warehouse, it is collected and processed in a temporary storage area called data staging

3. Data Warehouse DBMS (main storage)

→ The Data Warehouse is the main place where all the cleaned and processed data is stored.

4. Data Marts

→ These are smaller sections of the Data Warehouse, created for specific departments (like Sales, Marketing or HR).

5. Multi-Dimensional Database

→ This is a special type of storage that helps analyze data from different

Perspectives (like Sales Per region, Per Year)

→ It helps in quick and efficient analysis and reporting

6. Information Delivery (How Users Access Data)

- Data Mining → Finding useful pattern in data

- OLAP (Online Analytical Processing) → Analyzing data from different angles.

- Report / Query → Generating reports and answering specific questions using stored data

7. Management & Control

→ This ensures the data warehouse is running smoothly, data is secure, and users get the right information.

Ques: Discuss in detail about the Building a Data Warehouse? [AKTU-21-22]

Sol^m Building a Data Warehouse is like constructing a large library where

all Company data is stored, organized and made easy to access for decision making.

Steps to Build a Data Warehouse

1. Identifying the Business Needs

→ Understand what the company wants from the data warehouse

2. Collecting Data

→ Data comes from different places like database, online systems, sales records

3. Data Staging (Cleaning & Processing)

→ The collected data is often messy with errors and duplicates

→ ETL Process (Extract, Transform, Load) is used to clean and organize the data before storing it.

4. Storing Data in the Data Warehouse

→ After cleaning, the data is stored in the data warehouse which act as

the main Storage System.

5. Creating Data Marts

→ Data is divided into smaller Data Marts based on different business areas like Sales, Marketing, Finance etc.

6. Using multi-Dimensional Database

→ Data is structured in multiple ways for fast analysis and reporting.

7. Query, Reporting and Data Mining

→ OLAP (Online Analytical Processing):- Viewing data from different angles.

→ Report & Query:- Getting specific answer

→ Data mining:- Find hidden patterns

8. Managing and Maintaining Data Warehouse.

→ Regular updates, Security and performance checks are done to keep the system running smoothly

Qw-2 Discuss about the Concept of Mapping the Data Warehouse to a Multiprocessor Architecture. [AKTU-21-22]

Sol^m A Data Warehouse stores and processes large volume of data, requiring high-speed processing for analysis and reporting. To handle this efficiently, a multiprocessor architecture is used where multiple processors work together to manage the workload.

- This improves performance, scalability and response time.

Why Use Multiprocessor Architecture for a Data Warehouse?

A single processor system may struggle with heavy data processing, leading to slow performance.

A multiprocessor system solves this

- Dividing tasks among multiple processors for faster execution.
- Handling multiple queries at the same time.

Types of Multiprocessor Architecture

1. Shared Memory Architecture

→ All processors share a common memory and database

2. Shared Disk Architecture

→ Each processor has its own memory but shares the same disk storage

3. Shared Nothing Architecture

→ Each processor has its own memory & storage, working independently

Difference between Database System and Data Warehouse

Aspect	Database - System	Data Warehouse
Purpose	Day-to-day Operation and Transaction	Analytical processing & decision-making
Data-type	Current & Real-time	Historical
Structure	Normalized (relational tables)	Denormalized (star schema)

Data-Source	Single Source	Multiple Sources
Complexity	Simple	Complex
Performance Focus	Fast read/write Operations	Fast data retrieval for analysis
Data-Volume	Smaller Volumes of current data	Large Volume of historical data
Processing Type	OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)

Qu-3 Explain Multi-Dimensional Data Model?

[AKTU-22-23, 23-24]

Sol^m A Multi-Dimensional Data Model is a way of organizing and storing data in a Data Warehouse so that it can be analyzed from different perspectives.

Think of it like a Rubik's Cube!

Just like you can look at Rubik's Cube from different angles (color, row, column), a Multi-Dimensional Data Model allows users to analyze data from different dimensions like → time, location, product, Sales etc.

Key Concepts of Multi-Dimensional Data Model

1. Facts (Measures)

→ These are the numerical values we analyze (ex → Total Sales, Profit, Revenue)

2. Dimensions

→ These are categories used to analyze facts.
like → time, region etc

3. Dimension Table

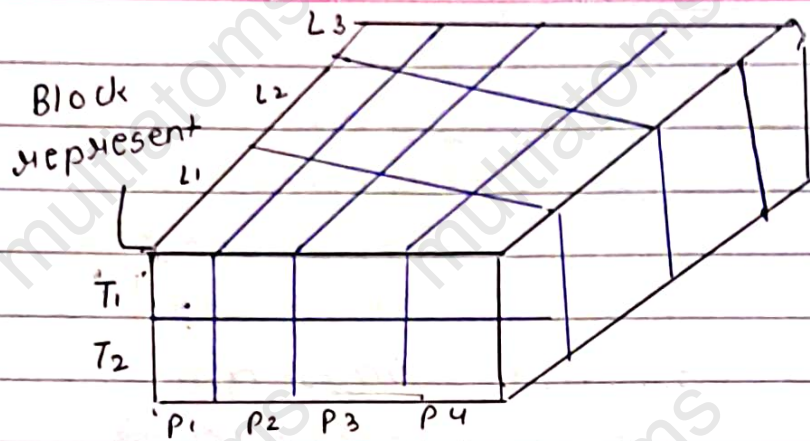
→ Contains descriptive information about each dimension

Types of Multidimensional Models:

1. Data Cubes :- [AKTU-22-23]

A data cube is a multi-dimensional representation of data. Think of it as a cube (3D box) where each dimension represents a different attribute or category of data.

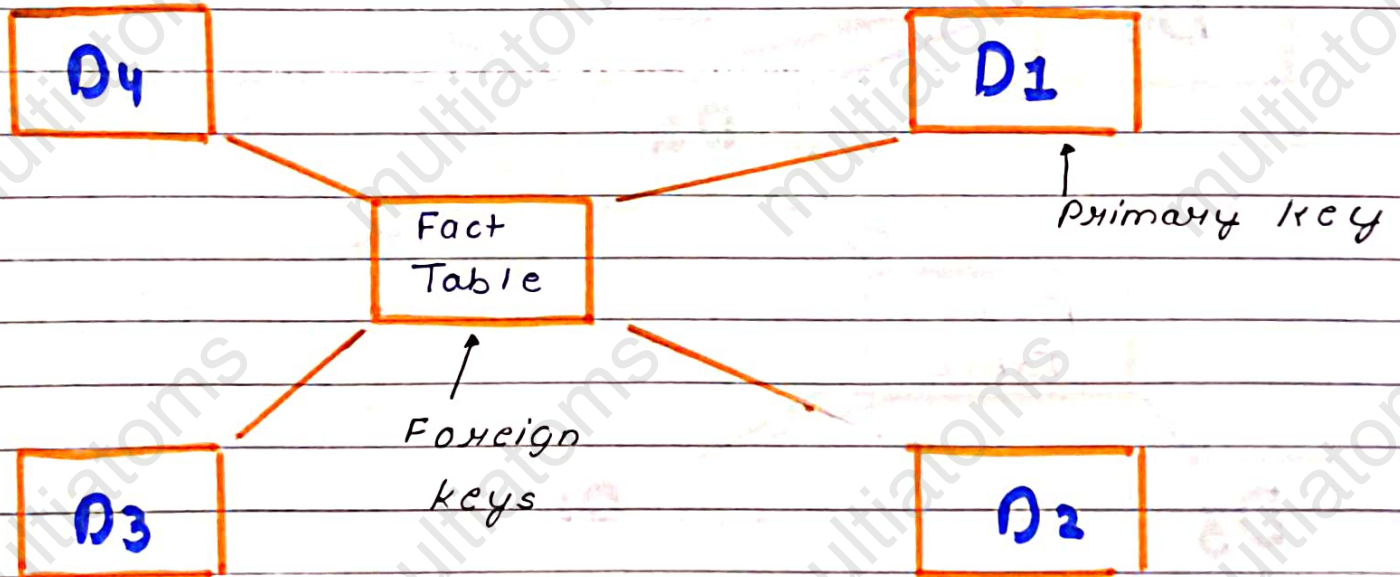
Dimensions → Time, Location, Product



$T \rightarrow$ Time
 $L \rightarrow$ Location
 $P \rightarrow$ Product

2. Star Schema :-

- Central Fact table (stores main data)
- Connected to multiple dimension tables
- Simple and Fast for queries

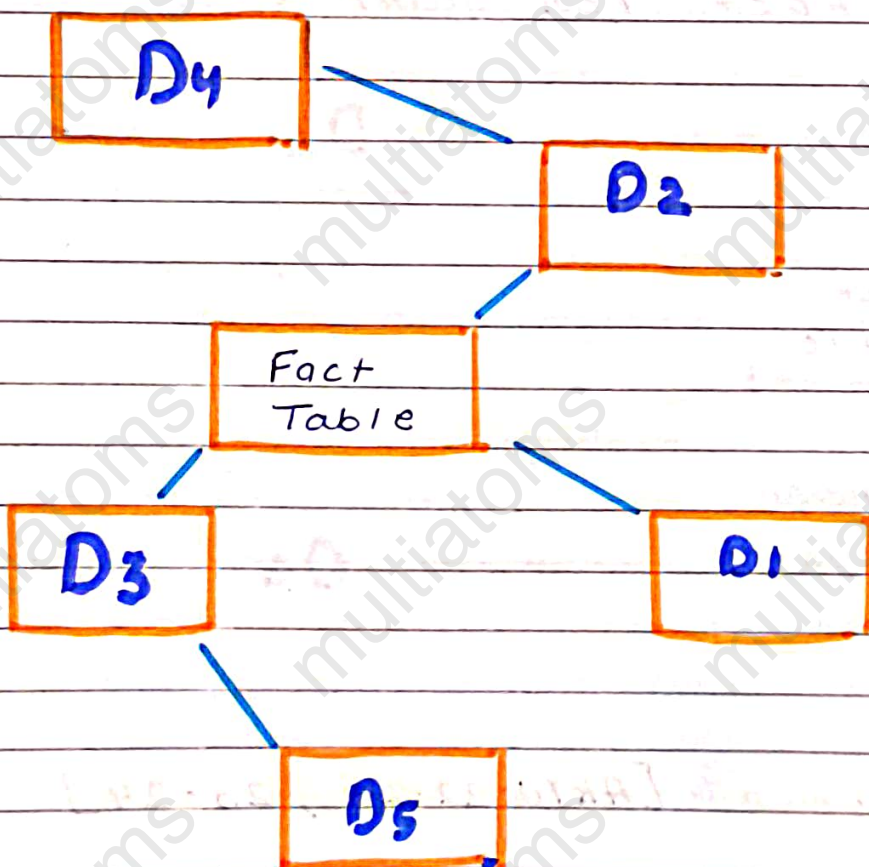


3. Snow Flake Schema :- [AKTU-22-23, 23-24]

→ The Snowflake Schema is a logical arrangement of tables in a data warehouse.

→ It gets its name becoz the diagram of the Schema looks like a Snowflake (due to its branching structure).

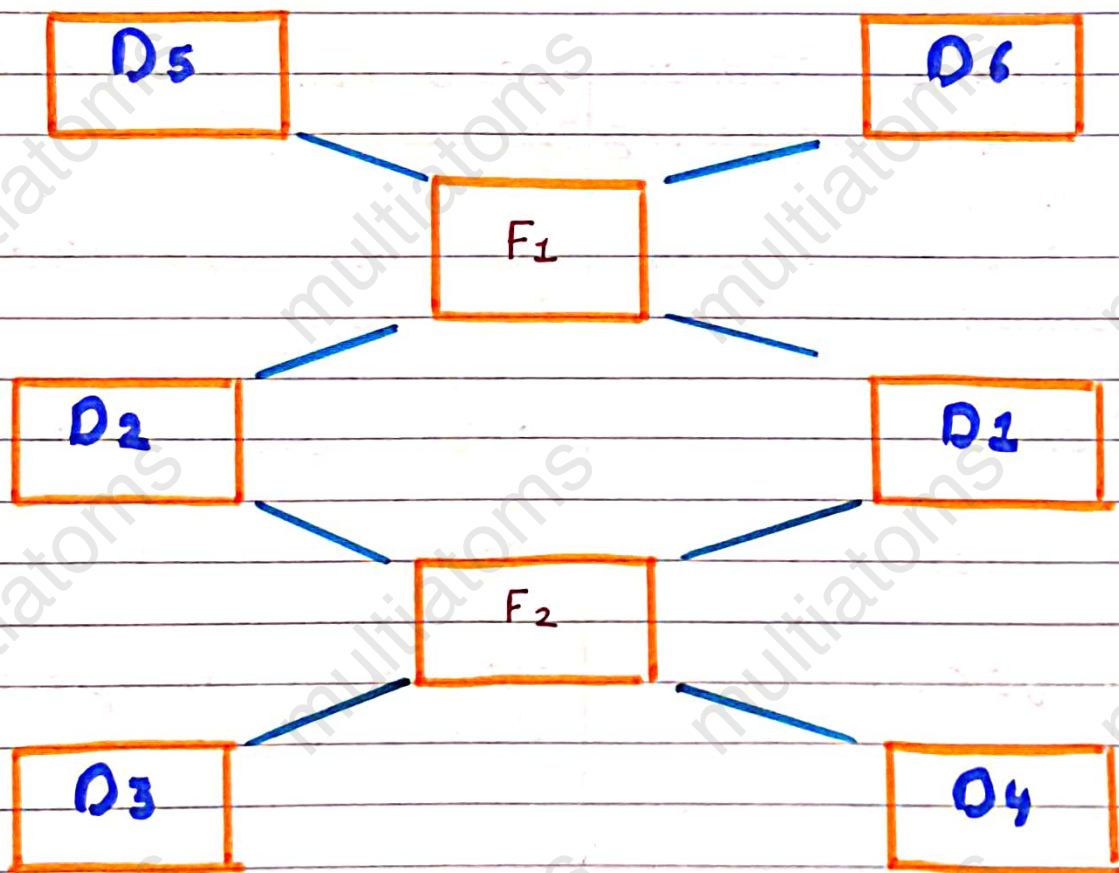
→ It reduce data redundancy by splitting dimension tables into smaller sub-dimension tables.



Fact Constellation Schema (Galaxy Schema)

[AKTU-22-23, 23-24]

- A Fact Constellation is a complex type of multidimensional data model
- It is also called Galaxy Schema because it contains multiple fact tables that share common dimension tables
- Complicated design & implementation



Qw-4 Write down the difference between Database System and Data Cubes.
[AKTU-22-23]

Feature	Database System	Data Cubes
Definition	A System that manage & stores Structured data.	A multidimensional schema used for analytical large datasets
Purpose	Used For Transaction Processing	Used For analytical Processing
Structure	Uses table with rows and Column	Uses a cube-like Structure
Operation	CRUD	Slice, Dice, Roll-up
Speed	Fast	Slow
Example	MySQL, Oracle	Microsoft analysis Services

Qw-5 Steps of knowledge discovery in Data? [AKTU-23-24]

Sol^m Knowledge Discovery in Data is the process of finding useful patterns and insights from large amount of data.

It involve multiple steps to extract knowledge that helps in decision-making

1. Data Selection

→ Choose the relevant data needed for analysis

2. Data Preprocessing

Clean the data by removing errors, duplicates and missing values

3. Data Transformation

Convert data into suitable format

4. Data Mining

Apply techniques like → classification, clustering or association rules

5. Pattern Evaluation

Identify the most useful and interesting pattern from the results.

6. Knowledge Presentation

Show the discovered knowledge using reports, graphs or dashboards for easy understanding

SUBSCRIBE 

JOIN TELEGRAM 

Unit-2

- Warehousing Strategy
- Warehouse management and Support Processes
- Warehouse Planning and Implementation
- Hardware & OS for DW [AKTU-21-22]
- Client / Server Computing Model
- Parallel Processors & Cluster System. [AKTU-21-22]
- Distributed DBMS [AKTU-22-23, 23-24]
- Warehousing Software [AKTU-22-23, 23-24]
- Warehouse Schema Design [AKTU-22-23, 23-24]
- Market Basket Analysis [AKTU-22-23, 23-24]

Warehousing Strategy :-

A Warehousing Strategy is a plan made by a Company to decide how and where to store goods (like - raw material) in a way that is efficient, Cost-effective and Fast.

Why is it important?

- To store Products safely
- To deliver goods faster to customers
- To reduce storage and transportation costs
- To avoid product damage or loss.

Types of Warehousing Strategies :-

1. Centralized Warehousing

- All goods are stored in one big warehouse
- Easy to manage

2. Decentralized Warehousing

- Multiple small warehouse in different areas
- Fastest delivery

3. Private Warehousing

→ A Company owns and operates its own warehouse

→ Used in Large businesses

→ Cost-effective for long-term use

4. Public Warehousing

→ Businesses rent space in shared warehouse (Pay-as-you-go)

→ Small businesses or seasonal demand

5. Contract Warehousing

→ A Company rents warehouse space for a fixed period (1-5 years)

→ Medium-sized businesses needing long-term but not permanent space.

6. Smart Warehousing (Automated)

→ Uses robots, AI and IoT for fast operations.

→ Amazon - robot-powered warehouse - pick and pack orders in minutes

Warehouse / management and Support Processes.

Warehouse management means taking care of everything inside a warehouse - like how goods are received, stored and sent out.

The goal is to make sure everything is stored properly, easy to find, and ready to send out when needed. It's all about keeping things smooth and efficient.

1. **Receiving** → Getting products from suppliers
2. **Storing** → Keeping goods in the right place safely
3. **Picking** → Choosing the right items when an order is received
4. **Packing** → Wrapping and getting product ready to send
5. **Shipping** → Sending goods to the customer or other locations

Support Processes in Warehouse :-

These are the extra tasks that help main warehouse activities run smoothly

1. Data Duplicacy :- Removing redundant data to optimize storage and improve performance

2. Backup and Recovery Processes :-

→ Implementing robust backup solutions for data protection

→ Regularly schedule backup with automated scripts to reduce downtime

3. Data management :-

→ Using software to track items

4. Maintenance :-

→ Fixing equipment and bugs

Warehouse Planning and Implementation

Warehouse Planning is like making a smart blueprint or game plan before building or organizing a warehouse

It's about deciding how to set up the Storage Space, what you need, and how everything will work

1. Define the Purpose and Goals :

Figure out why you need the warehouse and what you want to achieve

2. Data Source Identification :

Find out where the data will come from
Ex → database, Excel files

3. Define Data :-

- Decide which data should be stored
- Choose Useful data only (Remove unnecessary data).

4. Choose Architecture :

- Decide how the warehouse will be built

5. Tools & Technology :-

Choose software and hardware tools for Storage, Processing and Reporting

Warehouse Implementation :

Once planning is done, now it's time to build and use the warehouse

1. Data Extraction :-

Collect data from different sources {

2. Data Cleaning :-

Remove errors, duplicates or incomplete data

3. Data Transformation :-

Change data into a common format

4. Data Loading :-

Load the cleaned and transformed data into the warehouse

5. Data Indexing :-

Organize data to make searching and reporting faster.

6. Testing :-

Check if the data is correct and the system works properly

7. Deployment :-

Make the warehouse live for users to access.

8. Maintenance :-

Regular updates and fixing issues to keep it running smoothly

Hardware and Operating Systems for Data Warehousing :- [AKTU-21-22]

To run a Data Warehouse, we need proper hardware (the physical devices) and a good Operating System (the software)

Hardware for Data Warehousing :-

Data warehouses deal with huge amounts of data so strong hardware is needed

1. Processor (CPU)

- The brain of the system
- A fast multi-core processor is needed

to process data quickly

2. Memory (RAM)

- Used to temporarily hold data while it's being processed
- At least 64 GB or more is common in large warehouse

3. Storage (Hard ^{Drives} / SSDs)

- Stores all the data in the warehouse
- Usually, warehouse use a combination of both

4. Network :

- Need to transfer data quickly between systems
- High-speed LAN or Fiber networks are used

2. Operating System (O.S)

The OS manages hardware and run all warehouse software.

Linux, window server, unix

Linux	Fast, Secure, open-source
Window Server	User-friendly, support many data tools
UNIX	Stable and Powerful

Client Server Computing model :-

The client server computing model is a Foundational architecture widely used in data warehousing and data mining system

It involves a distributed Framework where tasks are split between a client and a server, communicating over a network

Client / Server model in data warehousing

A data warehouse is a centralized repository designed to store, manage and analyze large volume of historical data.

1. Client - Layer :-

This is typically the front-end interface where users interact with the system.

- Client send request such as SQL queries, reports.

- Ex → Desktop applications, web browsers.

2. Server - Layer :-

- The Server hosts the data warehouse itself, which includes the database management system like → Oracle, SQL Server.

- It processes client requests, retrieves data from the warehouse and perform operations like → joins, aggregation.

3. Communication :-

- Client and Server communicate via protocol like TCP/IP, ensuring seamless data exchange.

- Middleware may sit between the client and server to handle query optimization or load balancing.

Ques Describe about the Parallel Processes and Cluster System Under the Data Warehouse Process technology.

[AKTU-21-22]

Sol^m In the context of Data Warehouse Process technology, Parallel Processors and cluster system are critical innovations that enhance the performance, Scalability and efficiency of data warehousing operations

Parallel Processing in Data Warehousing

Parallel Processing refers to the use of multiple processing units (CPU or COMES) working simultaneously to execute tasks within a data warehouse

1. Parallelism Type :-

- Data Parallelism :- Data is partitioned across multiple processors, and each processor works on its subset independently
- Task Parallelism :- Different tasks are assigned to separate processors and execute concurrently

2. Architecture :

(a) Symmetric multiprocessing (SMP)

→ Multiple CPUs share the same memory and OS

→ Good for medium-sized data processing

b) Massively Parallel Processing (MPP)

→ Many processors (or nodes) work independently with their own memory and os

→ Used in large-scale data warehouses (like Google, Amazon).

2) Cluster Systems :-

→ A Cluster system is a group of computers connected together that works as a single system.

→ Think of it like group of friends solving a puzzle. Each one works on a part and then they put it all together

How it helps in Data Warehousing :-

→ Increases processing power and storage

→ Provide fault tolerance → if one system fails, others keep working

→ Helps in scalability → add more machines

Qw Explain Distributed DBMS Implementation?
[AKTU-22-23, 23-24]

Solⁿ A distributed DBMS manages a database that is fragmented or replicated across multiple servers connected via network.

- It stores data across multiple locations while making it appear as a single database to users

Key Components

1. Data Distribution

Data is stored in different locations (servers or systems)

Fragmented :- Split by rows or columns (e.g. Sales by region)

Replicates :- Copy of the same data at multiple sites

2. Transparency

User don't need to know where the data is.

DDDBMS handles :-

Location Transparency (where data is stored)

Replication Transparency (copies of data)

Fragmentation Transparency (how it's divided)

3. Transaction Management

Make sure data changes are correct and consistent

→ many users access it once.

→ A part of the system fails.

Points :-

1. A Distributed DBMS manages data stored across multiple locations

2. It gives the illusion of a single system to users

3. It supports fast, scalable and reliable data access and mining

Ques Define Warehousing Software?

[AKTU-22-23, 23-24]

Sol^m Data Warehousing Software is a tool or platform used to...

→ Collect, store and organize data from different sources

→ Clean, transform, and prepare the data for analysis

Software	Description
Amazon Redshift	Cloud-based warehouse from AWS. Fast and scalable
Google Big Query	Serverless data warehouse on Google cloud.

[Warehouse Schema Design

Already Cover in UNIT-1

Market Basket Analysis (MBA):

[AKTU-22-23, 23-24]

• Market Basket Analysis is a data mining technique used to find patterns in who + people buy together

• It is a data mining technique used to identify relationship between products purchased together in transactions

• These rules are typically expressed as "if $\{A\}$, then $\{B\}$ " (if bread is bought the butter is like bought too)

What Does It Do?

- It studies the buying habits of customers
- It finds pattern in what people usually buy
- Helps stores make better offers and combos

There are 3- Components

$$(A) \rightarrow \begin{array}{l} \text{Tot} = 200 \\ m = 70 \\ = \frac{70}{200} \end{array}$$

1. Support

It is been calculated with the no of transaction divided by the total no of transaction made

2. Confidence

It is been calculated for whether the product sales are popular on individual sales or through combined sales

3. Lift

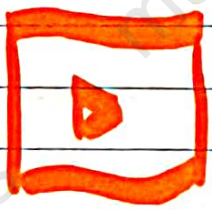
Lift is Calculated for knowing the ratio for the sales

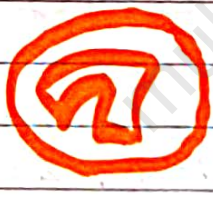
Real-world Use Cases:

Retail :- Recommend products (ex - Amazon, Flipkart)

Fast-Food :- Combo meals (burger + fries + drink)

MBA helps businesses understand customer behaviour and increase sales through smart recommendation and promotions

SUBSCRIBE 

JOIN TELEGRAM 

Unit-3 (DWDM)

- Data Mining & Functionalities [AKTU-22-23]
- Motivation
- Data Processing, Data Pre-Processing
- Data cleaning :- Missing Values, Noisy Data [AKTU-21-22]
- Binning, Clustering, Regression, Computer and Human Inspection [AKTU-23-24]
- Inconsistent data
- Data Integration & Transformation
- Data Reduction - data cube Aggregation [AKTU-22-23, 23-24]
- Dimensionality Reduction [AKTU-23-24]
- Data Compression, Numerosity Reduction [AKTU-21-22, 23-24]
- Discretization and Concept of hierarchy Generation
- Decision Tree [AKTU-22-23, 23-24]

Qu-1 Explain Data Mining and its functionalities?

[AKTU-22-23]

Sol^m Data Mining :- Data Mining is the process of discovering useful patterns, trends and information from a large amount of data.

Data mining turns raw data into useful knowledge, helping businesses grow smarter. It works hand-in-hand with data warehouses where all the data is stored neatly for analysis.

Functionalities of Data Mining :-

1. Classification :-

- Storing data into different categories
- It puts thing into categories or groups

Example :-

A bank use it to classify loan applications as "risky" or "safe".

2. Clustering :-

- It groups similar data items together without predefined labels.

Example:-

Grouping customers with similar buying habits

3. Association Rule Mining:-

→ Finding relationship between different items.

Example:-

~~market~~ If people buy bread, they often buy butter too

4. Prediction:-

→ It predicts future trends based on existing data

Example:-

Predicting what products a customer might buy next.

5. Outlier Detection (Anomaly Detection):

→ It finds data that doesn't fit the usual pattern

→ Spotting unusual data

6. Summarization

→ It gives a simple summary of

Overview of the data

Example

Showing average monthly sales

Motivation in Data Mining:-

Motivation in data mining means the reasons why we need data-mining - why it's important and useful in today's world.

1. Huge Amount of Data:-

- Every day, we collect a lot of data from websites, apps & banks.
- It's too much for handle manually

Motivation:-

We need smart methods to find useful info in this big data.

2. Hidden Patterns:-

- Useful patterns are not easy to see just by looking at the data.

Motivation:-

Discover hidden insights that can help make better decisions

3. Better Decision Making :

→ Companies and People want to make smarter choices

4. Competition and Business Growth :-

→ Data mining helps businesses understand customers, improve services, and increase profits

5. Automation:-

→ Data mining tool can analyze data automatically - fast and without human errors

Qw-2 Discuss the concept of Data Cleaning?
[AKTU-21-22]

Solⁿ Data Cleaning is the process of fixing or removing incorrect, incomplete or irrelevant data from a dataset

Qw-3 What are the needs of data Mining? [AKTU-23-24]

Solⁿ • To extract useful patterns from huge amounts of raw data

Stored in database.

2. To support decision-making in businesses by predicting trends, detecting fraud.

Data Processing :-

Data Processing is the step-by-step process of collecting, organizing and converting raw data into meaningful information that can be used for decision-making.

1. Extraction :- Collect data from database or external sources.
2. Transformation :- Clean, filter and convert data into a consistent format.
3. Loading :- Store processed data in the warehouse for easy access.
4. OLAP (Online Analytical Processing) :- Organizes data for fast querying.

Data Pre-Processing

Data Pre-Processing is the first and most important step before

doing any data mining. It's about preparing raw data so it's clean, complete and ready to be analyzed

1. Data cleaning :-

- Fixing dirty data
- Removes or corrects → missing values, Duplicates, Errors or types

2. Data Integration :-

- Combining data from different sources

3. Data Transformation :-

- Converting data into a proper format

4. Data Reduction :-

- Making data simpler.

i) Missing Values ?

These are blank or empty spots in your data - places where information is not available.

Name	Age	City
Riya	25	Delhi
Arun		Mumbai
Meena	30	

cii) Noisy Data?

Noisy data means data that has random errors, inaccuracies, or outliers (values that don't make sense)

Product	Price ₹
A	500
B	520
C	490
D	9500

Qn-4 Describe in detail about any two of the following

(i) Binning

(ii) Clustering

(iii) Regression [AKTU-23-24]

Solⁿ Binning :-

Binning is a method to smooth out noisy data by grouping values into smaller intervals ("bins"). It helps reduce minor data errors.

Binning is like sorting data into buckets or groups to make it easier to work.

data: 4, 8, 15, 21, 21, 24, 25, 28, 34

1:- Equal Partitioned Bin

- Divide the entire range of data into equal-sized intervals (bins)
- Then ~~replace~~ To reduce noise and To simplify the data

Bin:-1 4, 8, 15

Bin:-2 21, 21, 24

Bin:-3 25, 28, 34

2:- Bin Mean

Once you divide your data into bins,

you can replace the original values in each bin with the mean (average) value of that bin.

Bin-1 9, 9, 9

Bin-1 4, 8, 5

Bin-2 22, 22, 22

Bin-2 21, 21, 24

Bin-3 29, 29, 29

Bin-3 25, 28, 34

3.1 Bin Boundaries

- Bin Boundaries are the lowest and highest value in each bin (group) of data.
- In binning, after dividing data into bins, one way to smooth the data is to replace each value in a bin with the closest boundary value - either the lower boundary or the upper boundary.

Bin-1 4, 4, 15

Bin-1 4, 8, 15

Bin-2 21, 21, 24

Bin-2 21, 21, 24

Bin-3 25, 25, 34

Bin-3 25, 28, 34

(ii) Clustering :-

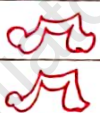
Clustering is like grouping similar things together without being told what the groups are

→ clustering groups similar data points together.

Songs



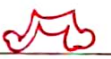
lofi-music



motivational songs



Rock-song



Pop-music

Types of Clustering :-

(2020)

1. k-means clustering → Pick a number of clusters (like 3) and group data into them based on similarity

2. Hierarchical clustering → Builds a tree of clusters, starting

with small groups and combining them

Use of Clustering :-

1. Organizes Data
2. Find Hidden Patterns
3. Target Customers
4. Handle Large Data

Regression :-

Regression is used to Predict missing or incorrect values using other related data.

It Create a mathematical relationship between Variables

like → house Price, Salary, weight

$$y = 2x$$

Dependent Variable :- What we want to Predict (ex - house Price)

Independent Variable :- Factor affecting the Prediction (ex - house-size, location)

Types of Regression

1. Simple Linear Regression:

Predicts a Value using only one input feature.

2. Multiple Linear Regression :-

Predict a Value using multiple input feature

Human Inspection :-

Human Inspection is when people manually look at data to check, understand or make decisions. It's like you tasting soup yourself to decide if it's yummy or not.

Computer Inspection :-

Computer Inspection is when a computer automatically checks and analyze data using special program or algorithm. It's like having a super-smart robot assistant that looks through a huge pile of data to find mistakes, patterns.

Inconsistent Data :-

Inconsistent data means data that doesn't match or follow the same format or rules across records.

It's happen when

- The same thing is written in different ways
- There are conflicts or contradictions in the data

Customer Name	Country
John Smith	USA
John Smith	United States
John Smith	U.S.A

Data Integration?

- Data Integration means combining data from different sources into one unified view
- Data Integration brings all this data together into a single system

Why is Data Integration Important?

- Helps create a complete picture of the business
- Make data consistent and usable
- Save time & efforts during analysis

Data Transformation :-

Data Transformation is the process of changing the data format, structure, or values to make it clean and useful.

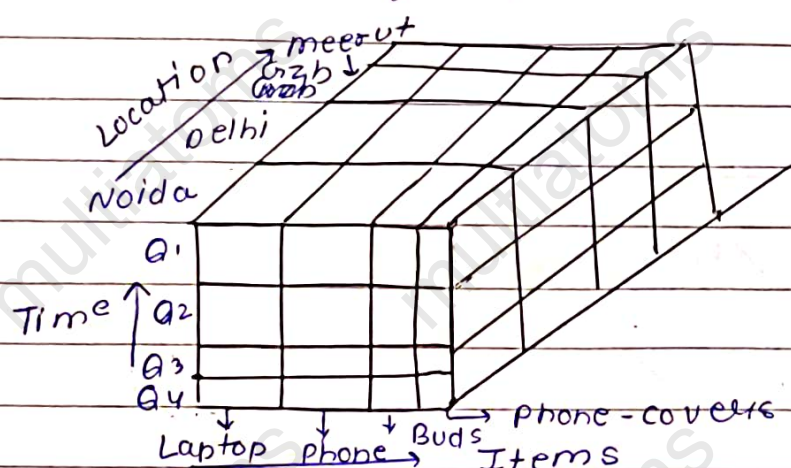
This is done after integration and before analysis or mining.

Types of Data Transformation :

1. Smoothing $\rightarrow$ Removing noise (errors)
2. Normalization $\rightarrow$ Scaling data into a small range (like 0 to 1)
3. Aggregation $\rightarrow$ Summarizing data (like total sales per month)
4. Attribute Construction $\rightarrow$ Creating new useful features from existing data

Ques What do you understand by Data Cube Aggregation and Dimensionality Reduction in Data Mining? [AKTU-23-24, 21-22]

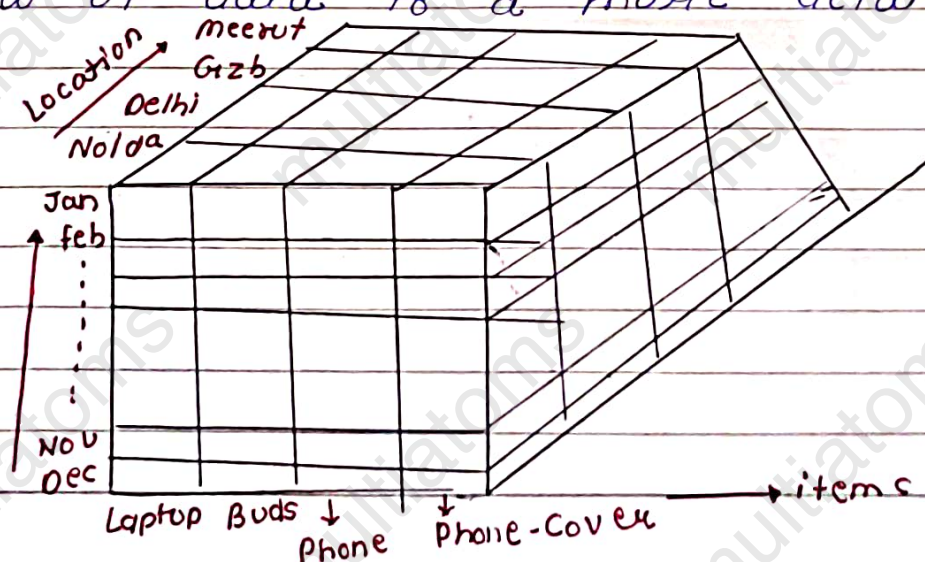
A data cube organizes data along multiple dimensions (time, location, product) to enable efficient analysis and querying. Aggregation refers to the process of summarizing data at different levels.



1. Operation on Aggregation

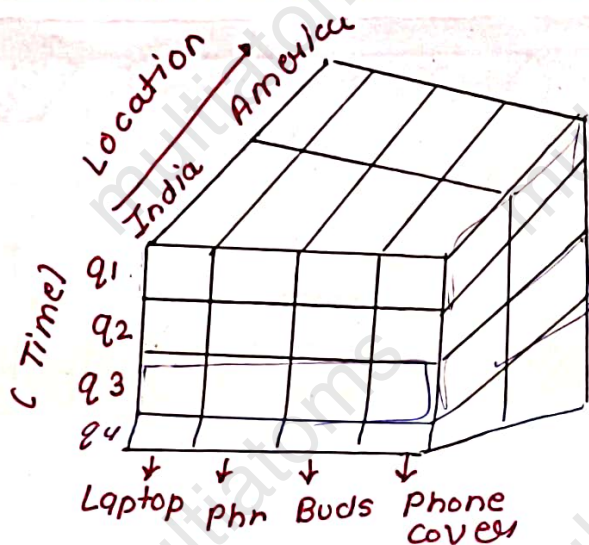
1. Drill-Down

The Drill-Down operation in data-mining and data cube analysis is a technique used to navigate from a summarized, high-level view of data to a more detailed



2. Roll-up :

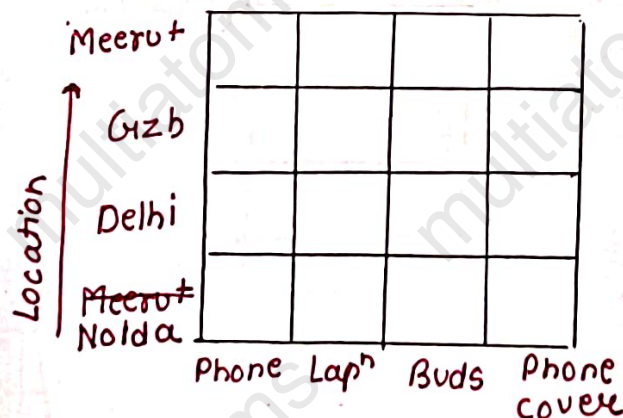
Roll-up operation in data-mining is a technique used to move from a detailed, granular level of data to a more summarized, high-level within a data cube.



3. Slice :

Pick on

get a 2D View

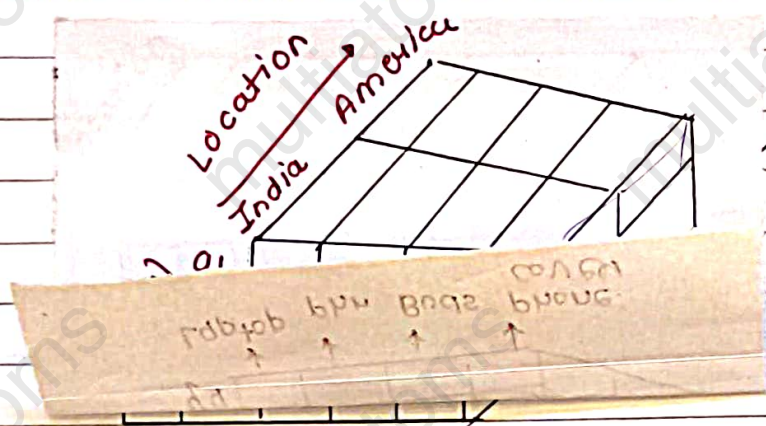


4. Dice :

Pick a range or multiple values from

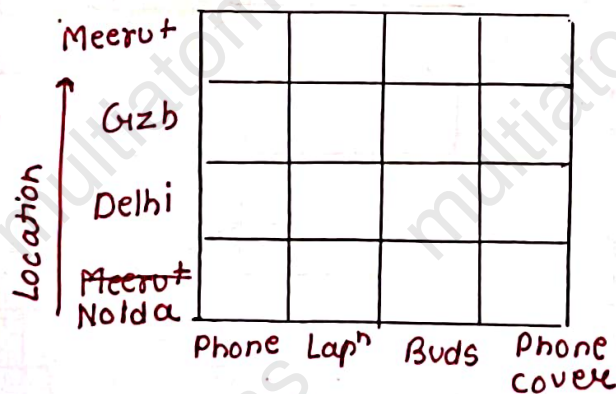
2. Roll-up :-

Roll-up operation in data-mining is a technique used to move from a detailed, granular level of data to a more summarized, high-level within a data cube.



3. Slice :-

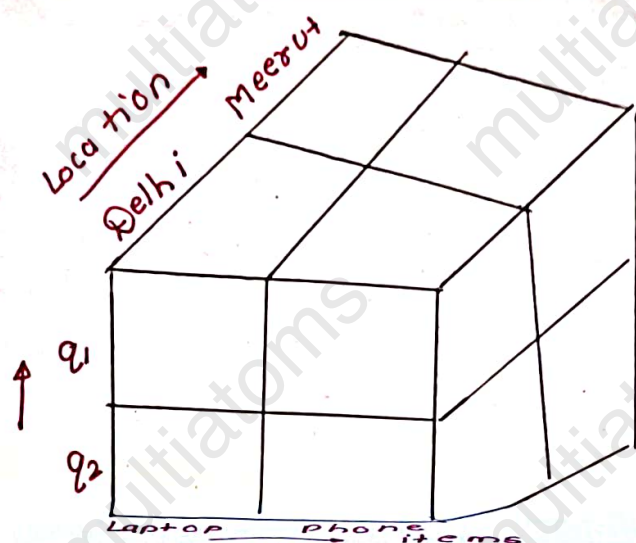
Pick one value for one dimension and get a 2D view.



4. Dice :-

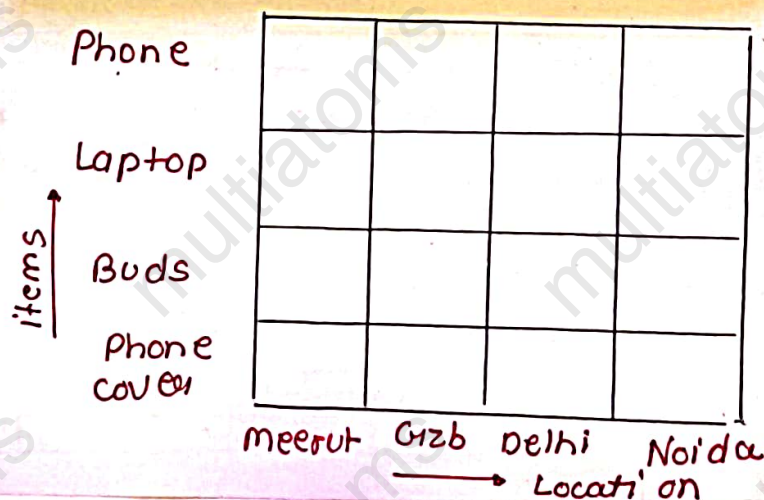
Pick a range or multiple values from

multiple dimensions



5. Pivot CR

Reorient the cube to look at it from different perspectives.



Dimensionality Reduction :-

Dimensionality Reduction is the process of reducing the number of input variables in a data set while preserving essential information. It helps in simplifying models, improving efficiency.

Goals of Dimensionality Reduction:

1. Remove noise and irrelevant data
2. Speed up data mining algorithm
3. Improve accuracy

Data Compression:

Data Compression is the process of reducing the size of data to save storage space, speed up transfer and improve computational efficiency.

Lossy Compression

Some data is lost
Compression Ratio is High

Lossless Compression

No loss of data
Compression Ratio is Low.

Qn- What is Numerosity Reduction?

Soln [AKTU-23-24, 21-22]

Numerosity reduction is a data reduction technique used in data mining to reduce the volume of data while maintaining the integrity and quality of the original dataset.

numerosity reduction focuses on representing large dataset with a smaller equivalent version.

Discretization

Discretization is the process of converting continuous data or attributes into a finite set of intervals.

This reduces data complexity and make it easier for certain algorithm to process.



10+	1 bl -
20+	✓
80+	✓✓

Concept Hierarchy Generation :-

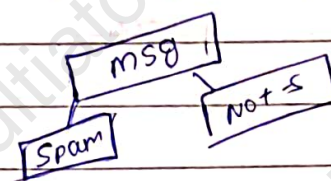
- It organizes data into a hierarchy of concepts, moving from specific to general categories. It create level of abstraction enabling analysis.
- Allows data to be analyzed at different level
- Enables the discovery of more general and meaningful patterns.

Ques Discuss Decision-tree Classifier in detail? [AKTU-23-24, 22-23]

Solⁿ

Decision tree based classifiers are a popular and into method used in data mining to classify data into categories

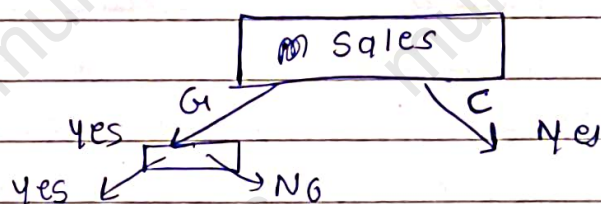
They work by breaking down complex decision into a series of simple, step-by-step, much like a flowchart



What is decision tree?

A decision tree is a visual model that represents decisions and their possible outcomes. Imagine you're trying to decide whether to go outside based on the weather

Is it raining?



→ If Yes, Stay inside

→ If No, is it sunny?

→ If Yes go outside

→ If no, maybe stay inside

Key- Algorithms :-

1. ID3 (Iterative Dichotomiser 3)

- ID3 is one of the earliest algorithm for building decision tree.

- ID3 uses something called Information gain to decide which feature (like age, gender) to split the data on.

2. C4.5

- C4.5 is upgraded version of ID3. It fixes some of ID3's problem and can handle more types of data.

- Handle continuous data, handle missing data

3. CART (Classification & Regression Trees)

- It like a multi-tool it can build decision trees Classification & Regression (how much will sale ?)

- Instead of Information gain it uses something called Gini impurity for Classification or Variance reduction for regression.

SUBSCRIBE 

UNIT- 4 (PART-01)

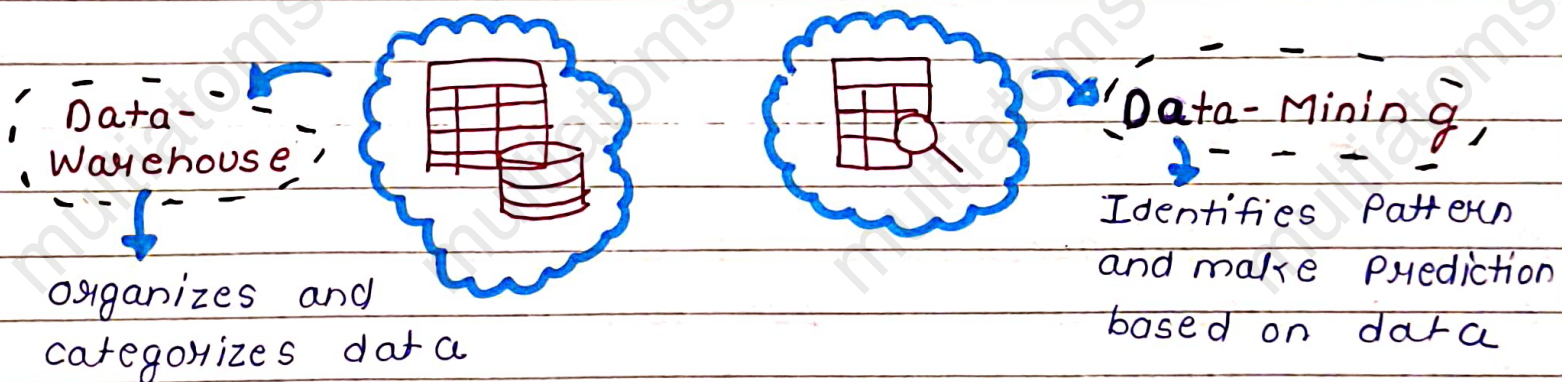
Classification

- Definition [AKTU-22-23]
- Data Generalization
- Analytical Characterization
- Analysis of attribute relevance, mining class comparisons
- Statistical measures in large database [AKTU-21-22]
- Statistical based algorithm
- Distance-Based algorithm [AKTU-21-22, 22-23]
- Decision-Tree-based algorithm [AKTU-21-22]

Classification :-

- Classification means organizing data into categories or groups based on certain characteristics
- Classification is a technique used to predict what group something belongs to based on past data

How to use Classification in data processing?



Data Generalization :-

- Data Generalization is the process of summarizing detailed data into higher-level, more general categories
- It helps simplify large dataset by replacing low-level details with broader concepts.

Why is Data Generalization Useful?

1. Simplifies Data Analysis - Instead of dealing with thousands of individual records, we work with summarized groups.
2. Improve Performance - Queries run faster on generalized data than on more detailed data.
3. Enhance Privacy - Hiding exact values protects sensitive information.

Analytical Characterization :-

Analytical Characterization is a step in data mining where we analyze and summarize data to find out :-

- Which information is important
- Which information is not useful and can be removed

Attribute Relevance :-

Attribute Relevance analysis is the process of evaluating and selecting the

most meaningful attributes for a given data mining task. It determine which attributes are significant for Pattern discovery, Classification or Prediction

1. Improve model accuracy
2. Reduce Computational Complexity
3. Eliminate noise and redundancy

Mining Class Comparisons

Class Comparison mining involves analyzing and contrasting different classes or categories in a dataset to identify significant differences between them

Comparing two or more groups of data to find out how they are different or what makes them unique.

Statistical measures in large Databases :- [AKTU-21-22]

In the context of large Databases, especially for data warehouse and data mining, statistical measures are essential for summarizing, analyzing. These measures help in understanding the distribution, relationship and trends within massive datasets

• These are really useful when we want to summarize data in a big table or data cube

1. Measure of Central Tendency (tells us where the "center" of the data is)

Measure of Central tendency such as mean, median, mode and mid-range

1. Mean

Sum of ^{values} ~~over~~ ~~all~~ divided by count

$$\text{Mean} = \frac{x_1 + x_2 + \dots + x_n}{N}$$

2. Median

The middle value in a sorted list.

• If there are an odd no of values $\rightarrow$ it's middle one

Ex:- 1, 4, 6, 7, 12, 14, 18

median:- 7

• If there are an even no $\rightarrow$ it's middle two / ones

Ex:- 1, 4, 6, 7, 8, 12, 14, 16

median:- 7.5

3. Mode

Most Frequent Value (s)

• If a value repeat the most $\rightarrow$ it's the mode

• A dataset can have one, two or even multiple modes

2. Measures of Dispersion (tell us how "spread out" the data is)

These show how much the values differ from each other

1. Range

The difference between the biggest and smallest values.

Formula:- Large Value - Smallest Value

3. Standard deviation & Variance :

Variance :- Variance measures how much the data points in a dataset deviate from the mean (average).

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Standard Deviation :- Standard deviation is the square root of the Variance providing a measure of spread

in the same units as the Original data

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Statistical Based Algorithm

In DWM statistical algorithm helps us discover Patterns, trends and relationship in large datasets. These algorithms use the Principles of Statistics and Probability to analyze data and make Predictions or decisions

1. Regression Analysis

A method to Predict a value (like $\rightarrow$ Price, Salary, score) based on one or more inputs

Types:-

1. Linear Regression:- Finds a straight-line relationship between variables.

2. Multiple Regression:- Uses more than one input variable.

2. Classification (Statistical - models)

These are used to categorize data into groups (like $\rightarrow$ spam or not-spam)

1. Logistic Regression:

Predicts binary outcomes (Yes/No)

2. Naive Bayes Classifier

Uses Probability to classify data based on Past Occurrences.

3. Bayesian - Network (Directed - acyclic - graph)

A graph-based model that shows Probabilistic relationship between Variables.

Rain $\rightarrow$ Dog bark $\rightarrow$ Cat hide

4. Time-Series Analysis

Analyzing data that's collected over time (like $\rightarrow$ sales per month), To Forecast Future values based on Past trends.

5. Hypothesis Testing

A statistical method used to determine if a claim or assumption about data to be true.

Ques- Discuss Various issues regarding classification & Prediction: [AKTU-22-23]

Solⁿ Classification:- Classification is about assigning data into predefined categories (like spam or not spam)

Prediction:- Prediction is about forecasting a numeric or continuous value (like Predicting a house price)

1. Data Quality:-

Missing values, noisy data, or inconsistent formatting can lead to inaccurate models

2. Imbalanced data:-

When one class significantly outnumbers the others

(98% of customers don't churn, 2% do)

3. Feature Selection and Engineering

Irrelevant or redundant features reduce performance

Using domain knowledge to select important features

4. Scalability

Data in the real world is huge (millions or billions of rows)
models may not fit in memory

5. Dynamic Data:

Pattern in data can change over time
A model trained on customer behavior from 2022 may not work well in 2025.

Distance Based Algorithm

Distance-based algorithm use mathematical distance (or similarity) between data points to analyze, group or classify them.

These algorithm works by calculating the distance between data points (like numbers, customer profiles) to decide if they belong together or not.

4. K-Nearest Neighbors (KNN) [AKTU-22-23]

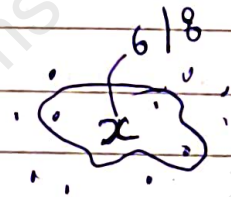
To predict a value or class, it finds the k -closest data points and makes decisions based on their majority class or average.

How KNN Works:

1. Choose k (no of neighbour you want to check)
2. Measure distance between the new point and all other points
3. Pick k -closest neighbours
4. Vote or average based on their values

Query :- $X = (DNDM = 6, CS = 8)$ at $k = 3$

	DNDM	CS	Result
1	4	3	Fail
2	6	7	Pass
3	7	8	Pass
4	5	5	Fail
5	8	8	Pass



$$\text{Euclidean distance} = \sqrt{|x_{01} - x_{A1}|^2 + |x_{02} - x_{A2}|^2}$$

$\downarrow 6$ $\downarrow 8$

$$\text{I :- } \sqrt{(6-4)^2 + (8-3)^2} = \sqrt{29} = 5.38$$

$$\text{II :- } \sqrt{(6-6)^2 + (8-7)^2} = \boxed{1}$$

$$\text{III :- } \sqrt{(6-7)^2 + (8-8)^2} = \boxed{1}$$

$$\text{VI :- } \sqrt{(8-8)^2 + (8-8)^2} = \boxed{0}$$

$$\text{IV :- } \sqrt{(6-8)^2 + (8-8)^2} = \boxed{2} \quad \text{V :- } \sqrt{(6-5)^2 + (8-5)^2} = \sqrt{1^2 + 9} = \sqrt{10}$$

Pros of KNN

- Super simple to understand and use
- No need to train a complex model

Cons of KNN

- Slow for large datasets
- Doesn't work well with noisy data

2. K-Means Clustering

K-means clustering is a machine learning algorithm that groups similar things together into K-clusters.

[Decision-Tree-Based-Algorithm
Covered in UNIT-3 (ID3, C4.5, CART)

Ques Compare and Contrast between Distance-Based Algorithms and Decision tree Based Algorithms? [AKTU-21-22]

Feature / Criteria	Distance-Based Algorithm	Decision Tree-Based Algorithms
Basic-Idea	Classify based on distance to other data points	Classify using if-then rule in tree-like structure
Example	KNN, k-means	ID3
Time	Fast	Slower
Prediction Time	Slow	Fast
Memory Usage	High	Low
Handle missing data	Not well	Better - can handle missing values
Working	Compare a test point to nearby data using a distance metric	Splits data based on attribute value using measures
Use-case Example	Image-recognition	medical diagnosis

SUBSCRIBE 

UNIT-4 (PART-2) DWDM

- Introduction
- Similarity and Distance Measures
- Hierarchical and Partitional Algorithm [AKTU-21-22, 23-24]
- CURE and Chameleon [AKTU-22-23]
- DBSCAN, Optics [AKTU-23-24]
- STING, CLIQUE [AKTU-22-23]
- Statistical Approach
- Association Rule
- Large Items Sets
- Parallel & Distributed Algorithm
- Neural Network Approach [AKTU-22-23]

Clustering :-

In the world of Computer and data Clustering does the same thing - it groups similar data point together. ~~without~~

Clustering is like organizing a messy room by grouping similar thing together

Similarity & Distance Measures

In Clustering, we group similar data points together. But how do we decide if two things are similar? That's where Similarity and distance measures come in!

What is Similarity & Distance?

- Similarity :- How alike two data points are
- Distance :- How different two data points are

Why is this important for Clustering?

- If two points are very similar they should be in the same cluster
- If two points are very different they should be in different clusters

Common Distance Measures

1. Euclidean Distance

- Measures the shortest distance between two points

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

2. Manhattan Distance

- Measure distance along grid like path

$$\text{Distance} = |x_2 - x_1| + |y_2 - y_1|$$

3. Cosine Similarity

- Measures how similar two vectors are based on their direction not magnitude

$$\text{Similarity} = \frac{A \cdot B}{|A| |B|}$$

Ques What do you understand by Hierarchical and Partitioned Algorithm under the Clustering in Data Mining and Warehousing?
[AKTU-21-22, 23-24]

Sol^m Hierarchical Clustering:-

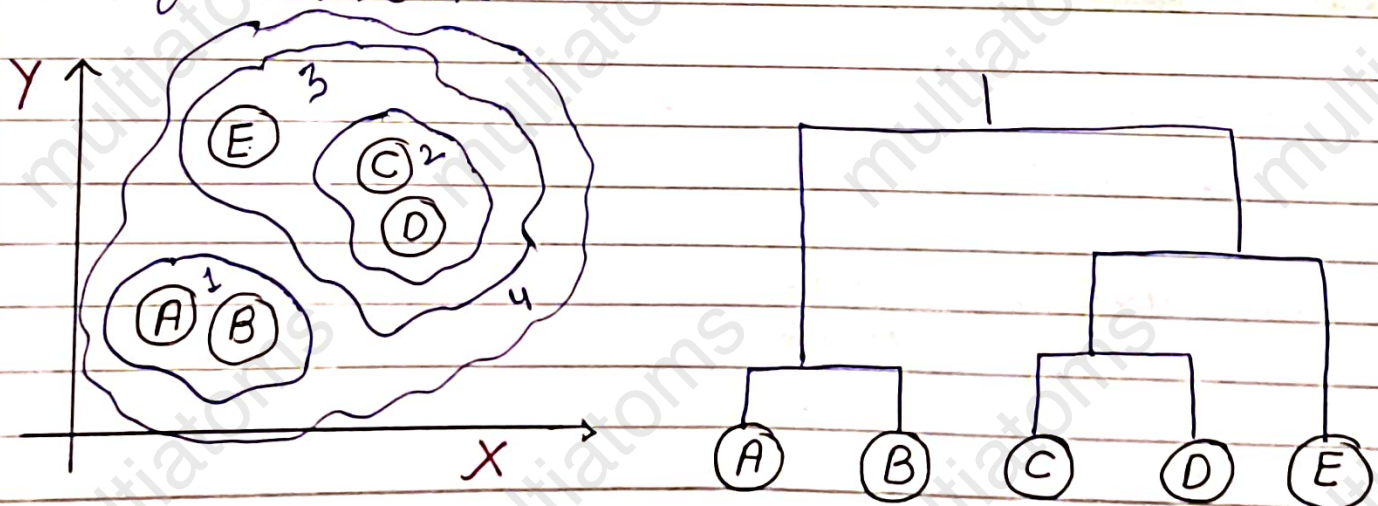
Hierarchical Clustering is a technique used to group similar data points together based on their similarity creating a hierarchy or tree-like structure.

Two-Types

1. Agglomerative

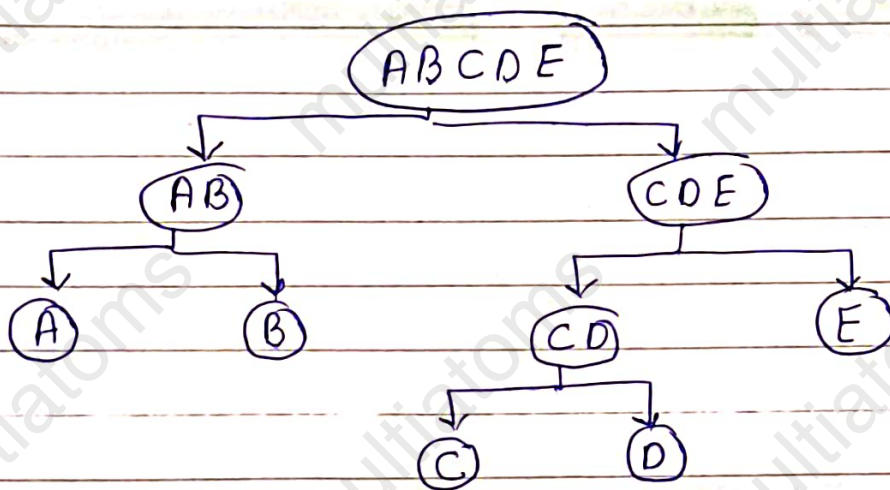
→ Bottom-Up

→ Start with individual points and merge them



2. Divisive

- Start with all data points in one big cluster
- Split the cluster into small groups based on differences
- keep splitting until you have many small clusters or decide to stop



Partitioned Algorithm



These algorithm divide the data into a fixed no. of clusters (say, k clusters) right from the start. Then try to group data points so that points in the same cluster are as similar as possible, and points in different clusters are as different as possible.

1. k-mean clustering

The most popular partitioned algorithm. It uses the average of points to define cluster center.

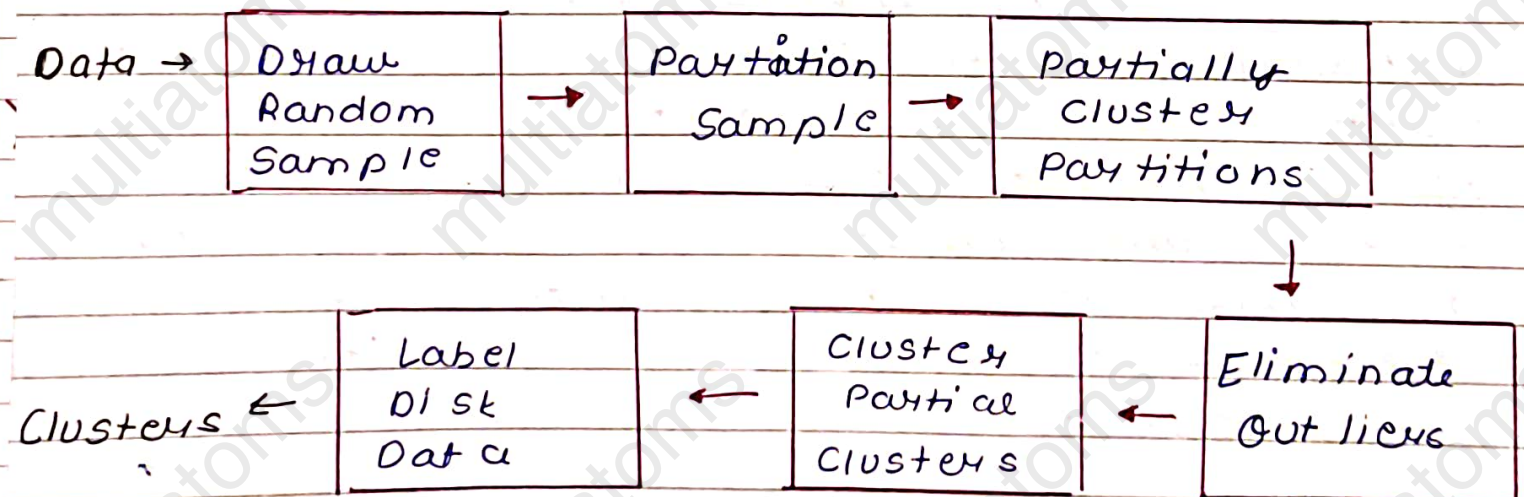
[AKTU-22-23]

2. CURE (Clustering Using Representatives)

CURE is a hierarchical clustering algorithm designed to handle large datasets and find clusters of non-spherical shape.

• Works very nicely with outliers

• Efficient for working with large dataset



2. Chameleon

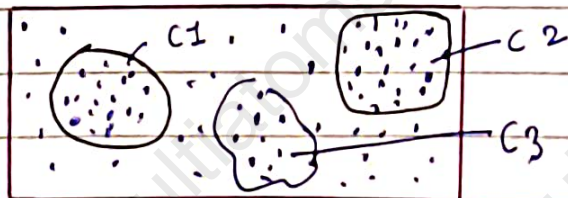
Chameleon is a hierarchical clustering algorithm that uses dynamic modeling to measure cluster similarity.

It was developed to address weaknesses in algorithm like CURE and ROCK.

Chameleon excels at finding high quality clusters of varying shapes, size and densities even in the presence of noise.

Ques Discuss DBSCAN Clustering algorithm with example? [AKTU-23-24]

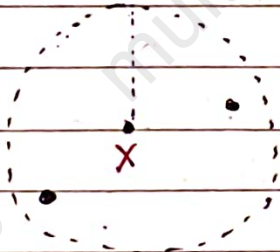
Soln DBSCAN is a density-based clustering algorithm that groups data points based on their density, meaning it finds areas where points are tightly packed together (high-density regions) and separates them from low-density regions.



Key - Concepts

1. ϵ (Epsilon)

The maximum distance (radius) within which two points are considered neighbours.



2. minPts (Minimum Points)

The minimum no. of points required within the eps radius to classify a point as a core point.

3. Core point

A point with at least minPts points in its eps neighborhood. Core points are in the dense interior of a cluster.



4. Border point

A point within ϵ ps of a core point but with fewer than minpts neighbours.

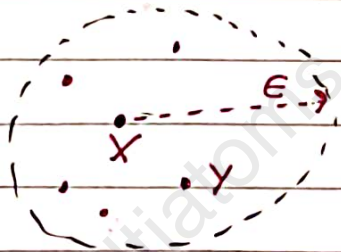
5. Noise point

A point that is neither a core nor border point

6. Density Reachable

A point is density Reachable From a

core point if it's within ϵ ps or connected through a chain of core points



Algorithm

Step-1 Pick a random point that hasn't been processed

Step-2 Points which are density reachable with Minpoint $\notin$ From Points, retrieve them

Step-3 The cluster shall be formed if the point is a Core Point

Step-4 If it is a border point and in that case it shall move to next point of dataset

Step-5 This process shall be continue until all points are processed

Advantage

- Handle non-spherical cluster
- Works well for datasets with consistent density

Disadvantage

- Struggle with variable density
- Less effective in high-dimensional data
- Computationally expensive

2. OPTICS

- Ordering points to identify the Clustering structure
- It's smarter version of DBSCAN
- It solves DBSCAN Problem when choosing the right distance (ϵ) is difficult.

Algorithm

Step-1 Initialize

- Start with an empty ordered list to store points
- Mark all points in the dataset as unprocessed

Step-2 Pick a Point

- Randomly select an unprocessed point from the dataset
- Make it as processed and add it to ordered list

Step-3 Check if it's a Core Point

- Find all neighbors of the point with distance ϵ
- If it's not a Core Point, move to

the next unprocessed point.

4. Expand the Neighborhood

→ For a core point, add all its unprocessed neighbors to a priority queue (called seed list)

→ Sort the neighbors by their reachability distance to the core point.

5. Process the Seed List

→ Pick the point with the smallest reachability distance from the seed list

→ Check if the new point is core point

→ Repeat this step until the seed list is empty

6. Repeat For All Points

→ If there are still unprocessed points in the data set, pick another unprocessed point and repeat step 2-5

→ Continue until all points are processed

Grid Based Method

STING, CLIQUE

Ques Discuss STING is different from CLIQUE?
[AKTU-22-23]

Solⁿ STING (Statistical Information Grid)

STING is a method used to group similar data together. It uses a grid system, kind of like a checkerboard to divide up the space where the data lives.

- Instead of looking at each data point one by one, STING looks at the area around the data, which is split into small square or rectangular cells.
- These cells are arranged in different levels: big cells at the top level and smaller ones when you go deeper.

These facts help STING quickly find and group data and do analysis without checking every single data point.

	first-layer
	⋮
	⋮
	⋮
	i-layer

CLIQUE Algorithm

CLIQUE is a clustering method that works well with high-dimensional data.

- It divides the data space into a grid (like cutting a cake into small cubes).
- It then checks how many data points fall into each small grid.
- Instead of checking every possible high-dimensional area, CLIQUE only looks at combinations that already seemed dense. This saves time and makes the algorithm very efficient.

Feature	STING	CLIQUE
Method	Grid-based, uses statistics	Grid-based, but subspace clustering
Good for	Simplex	Complex
Uses	Hierarchical grid structure	Dense units in subspaces
Output	Approximate clusters	more detailed subspace clusters

What is the model-Based method?

This method assumes that the data is generated by a model - usually a statistical model - and it tries to find the best model that fits the data. Once the model is chosen, data is grouped into clusters based on how well it fits the model.

Gaussian-mixture-model (GMM)

A common statistical model used here is the Gaussian mixture model. It assumes "Each cluster is like a bell-shaped curve and the data is a mix of several such curves".

The algorithm tries to find:

- How many bell curves are there?
- Where are they located?
- How wide or narrow are they?

This is done using Probability and algorithm like Expectation - Maximization (EM)

Association Rules:- Introduction

Association rules are patterns that show how items or events are related in a dataset. They are often used to find relationships between things that happen together frequently.

What are Large Itemsets? $A \rightarrow B$

A Large itemset is a combination of items that appear in many transactions more often than a user-defined threshold called minimum support.

Large itemsets are the starting point for finding association rules.

Parallel- Algorithm

- Run multiple parts of the algorithm at the same time on one machine with many processors.
- The goal is to finish tasks faster by splitting the work.

Distributed- Algorithm

- Run on multiple machines usually connected over a network

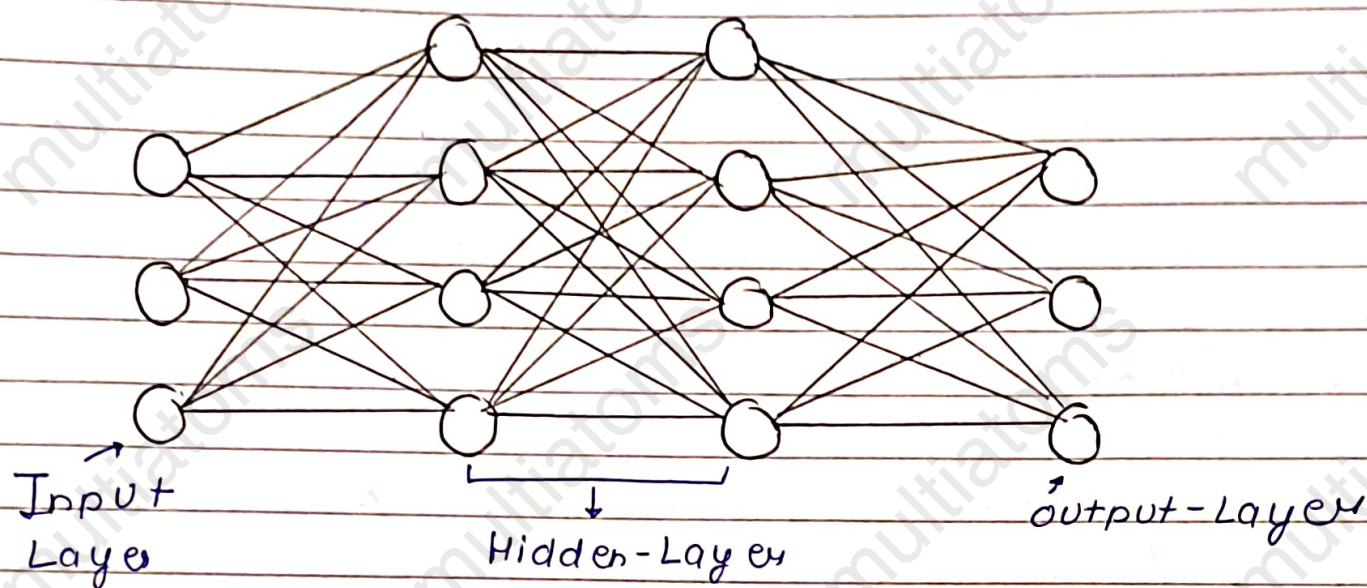
- Each machine works on part of the problem and they communicate with each other to complete the task.

Ques Explain Neural Network approach ?
[AKTU-22-23]

Solⁿ A Neural Network is like a digital brain made of interconnected nodes organized in layers. It's used to find hidden patterns, cluster data, classify data.

A Neural Network has 3 main parts

1. Input Layer :- Take the input
2. Hidden Layer :- Processes the data
3. Output Layer :- Gives the result



Advantages

- Handle Complex Patterns
- Flexible
- Adapts to noise

Disadvantages

- Slow to Train
- Needs Tuning
- Resource-Intensive

SUBSCRIBE 

JOIN TELEGRAM 

Unit-5 DWDM

- Aggregation [AKTU-21-22]
- Historical Information
- Query Facility
- OLAP Function and Tools [AKTU-21-22, 22-23, 23-24]
- OLAP Servers [AKTU-21-22]
- ROLAP, MOLAP, HOLAP [AKTU-21-22, 22-23, 23-24]
- Security and Backup & Recovery
- Tuning data Warehouse [AKTU-22-23]
- Types of Warehousing application [AKTU-22-23]
- Web Mining, Spatial Mining and Temporal Mining [AKTU-21-22, 22-23, 23-24]

SUBSCRIBE 

What is Aggregation? [AKTU-21-22]

Aggregation in data mining means collecting and summarizing data so we can understand it better and use it for analysis.

When we have a lot of data from different sources, it's hard to make sense of it all. So we group and summarize the data to get useful information.

How Does Data Aggregation Work?

Step 1:- Raw data is often too detailed or messy to be useful by itself.

Step 2:- So we combine similar piece of data to make it easier to understand.

Step 3:- We use simple calculations like Sum, Count, Average.

Step 4:- Aggregated data can be shown in report and dashboards to help people make quick and smart decision.

Historical Information?

Historical Information means old data collected over time - like past sales, customer behavior, or Product Performance

In a data warehouse, historical data is stored so businesses can look back and see:

- What happened?
- When did it happen?
- How did things change over time?

Query Facility?

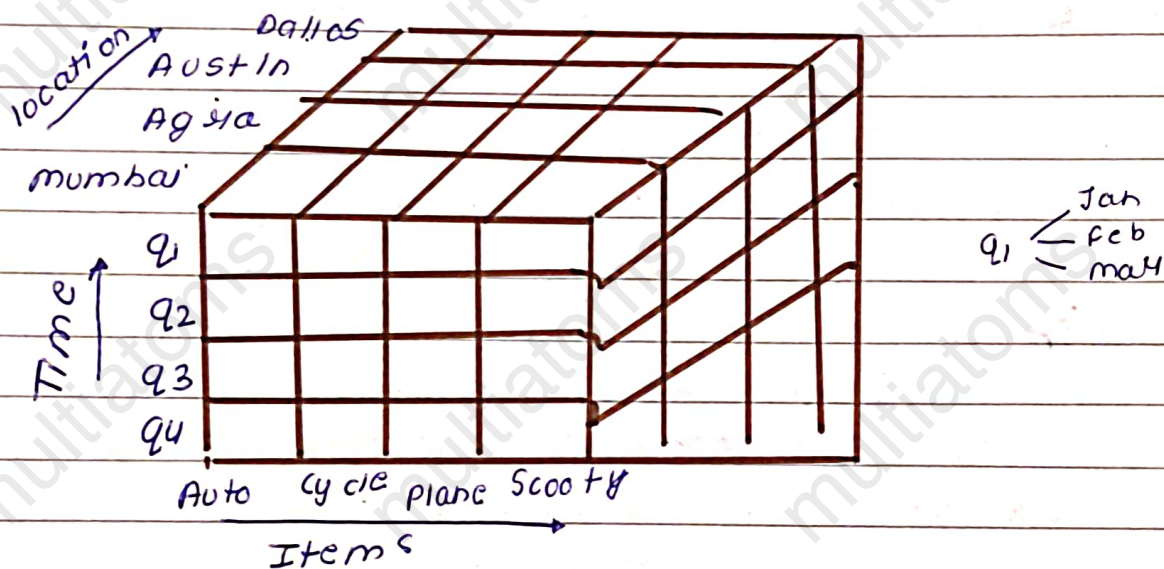
A query facility is a tool or system that lets users ask questions (queries) to retrieve, analyze or summarize data from a data warehouse or database. It's like a search engine for your data, helping you get specific information quickly.

What is OLAP?

OLAP (Online Analytical Processing) is a technology used in data warehousing to analyze large amounts of data quickly. It allows users to explore and

Summarize data from different angles to
Find insights, trends or patterns

OLAP is especially useful for handling multi-dimensional data, where data is organized into dimensions (ex → time, location, product) and measures (ex → sales, profit)

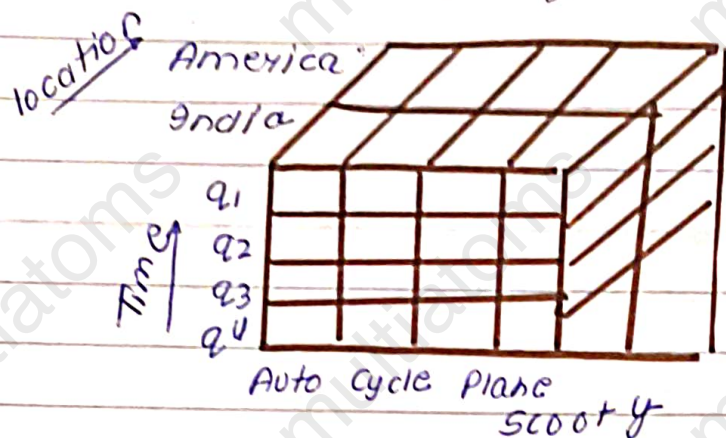


OLAP Functions [AKTU-23-24, 21-22, 22-23]

OLAP provides specific function to analyze data interactively. These functions work with aggregated data are accessed through query facilities

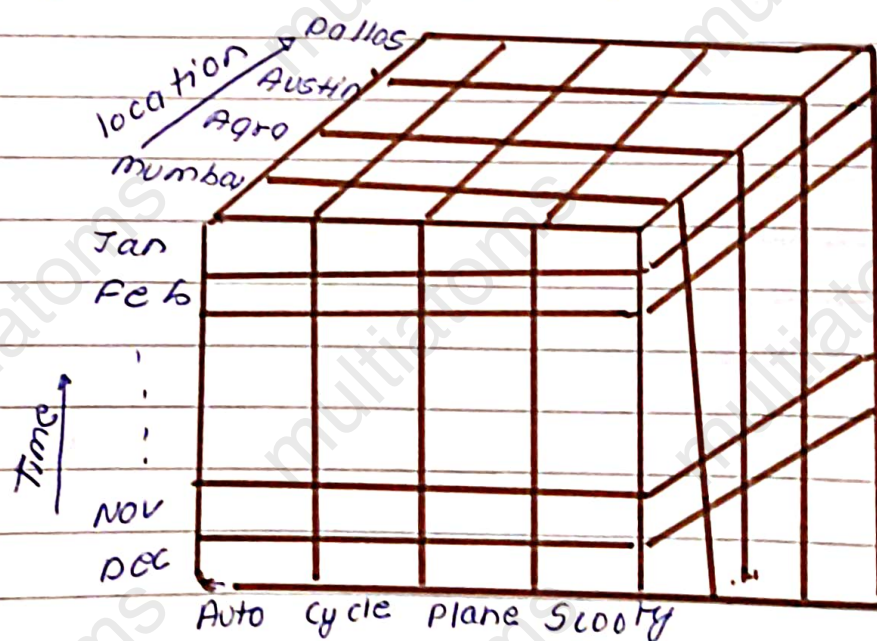
1. Roll-Up :-

- Combine data to a higher level of summary
- Ex: → Summarize daily sales into monthly sales



2. Drill-Down

- Goes from summarized data to detailed data
- Ex: - From monthly sales, drill down to see daily sales for a specific month



3. Slice

- Selects a specific subset of data by fixing one dimension
- Ex:- Look at sales for only one region (ex "Show sales in New York")

	Dallas			
↑	Austin			
	Agra			
	Mumbai			
		Cycle	Auto	Plane
				Scooty
		Items		

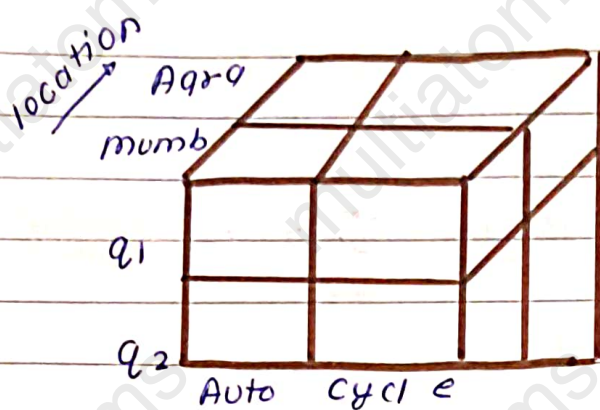
4. Pivot (Rotate)

- Change the view of data by swapping dimensions

	Auto			
↑	Cycle			
	Plane			
	Scooty			
		Dallas	Austin	Agra
				Mum
		Items		

5. Dice

- Select a smaller range of data across multiple dimensions



OLAP Tools :- [AKTU-21-22]

OLAP tools are software that make it easy to perform these functions, often with user-friendly interfaces.

1. Microsoft Power BI
2. Tableau
3. Oracle OLAP
4. SAP Business Objects

OLAP Server :- [AKTU-21-22]

An OLAP Server is a specialized software system that powers OLAP in a data warehouse. It organizes and processes large amount of data to support fast multidimensional analysis.

The OLAP Server act as a "brain" that handle queries from users or

OLAP tools and delivers results quickly

Different types of OLAP Server [AKTU-23-24, 22-23]

1. ROLAP (Relational OLAP)

→ ROLAP stores and processes data using relational databases (like SQL, Oracle) which organize data into tables.

→ ROLAP Servers are placed b/w relational backend server & client frontend tools

• Queries are written in SQL to perform OLAP Functions.

Advantages

→ Handle large datasets

→ Uses existing database system, so no need for new storage

Disadvantages

→ Slower for complex multidimensional queries because it

→ Performance depend on database Optimization.

2. MOLAP (Multidimensional OLAP)

- MOLAP stores data in a multidimensional cube structure, which is optimized for OLAP analysis. A cube organizes data by dimensions and measures.
- OLAP functions (slice, dice, pivot) are performed directly on the cube, making queries very fast.

Advantage

- Very Fast for multidimensional queries due to pre-aggregated data.
- Great for complex analysis.

Disadvantage

- Limited Scalability.
- Cube building takes time and storage.

3. Hybrid OLAP (HOLAP)

- HOLAP is combination of MOLAP & ROLAP. It balances speed and flexibility. It uses both relational databases and multidimensional cubes.
- Detailed data stays in relational databases.

Advantages

- Fast for summarized queries
- Flexible for detailed queries

Disadvantages

- More complex to set up and manage
- Performance depends on how well the cube and databases are integrated

Data Mining Interface

A data mining interface is a user-friendly tool or system that allows users to interact with data mining processes.

It helps users explore, analyze and extract pattern, trends from large data sets without needing deep technical expertise.

Role of Data Mining Interface :-

1. Data Selection
2. Data Preprocessing
3. Algorithm Selection
4. Result Visualization
5. Evaluation

Security :-

Security refers to protecting the data, systems and processes in a data warehouse environment from unauthorized access, theft and damage.

Key - Security Concept :-

1. Access - Control :- Restrict who can access the data warehouse tools
2. Data - encryption :- Protect data by encoding it so only authorized user can read
3. Network - Security Protect the server from external threats

Backup :-

Backup is the process of creating copies of the data to prevent data loss to hardware failure, human errors

Key - Backup Practices :-

1. Full Backups : Copy the entire data warehouse including raw data, aggregated data and metadata.

2. Incremental Backups : Copy only the data that has changed since the last backup.

3. Backup Storage : Store backup in secure, off-site locations.

Tuning Data Warehouse

A Data Warehouse is like a big storage room where a company keeps all its data and its important data in one organized place.

Tuning means making the data warehouse faster and more efficient like cleaning and organizing a messy room so you can find things quickly.

How Do You Tune a Data Warehouse?

1. Speed Up Slow queries

- look at which queries are too long
- Try to rewrite them in a simpler way

2. Use Indexes

- Indexes are like a table of content in a book - they help you find info faster

3. Partition Data

- This means splitting big tables into smaller pieces

4. Avoid Unused Data

- Don't pull in data you don't need

Ques Name any Five Challenges in Data Warehouse Testing? [AKTU-22-23]

Solⁿ Here are five common challenges

1. Huge Data Volumes :- Data Warehouse often hold millions or billions of records. Testing all that data takes a lot of time and resources.

2. Data Quality Issues :- Data comes from many sources, and inconsistencies can cause errors during testing.

3. **Complex Transformation**: Data is often transformed before storage. Testing these transformations to ensure they work correctly is tricky.

4. **Performance Testing**: Ensuring the warehouse responds quickly to queries, especially with large datasets, is challenging and requires careful testing.

5. **Data Integration**: Combining data from different systems can lead to mismatches or errors that are hard to detect or fix.

Ques Explain Warehousing application and recent trends? [ARTU-22-23]

Soln A warehouse application is a software system designed to manage to optimize the operations of a warehouse or data warehouse.

In the context of data it refers to tools or platform that handle the storage, organization, processing and analysis of large volume of data in a

data warehouse .

- It gathers data from different places
- Helps users quickly .

→ Amazon Redshift, Snowflake, Google Bigquery
Oracle Autonomous

Recent Trends in Warehousing Application

1. Cloud-Based Data Warehouses

- Instead of storing data on physical server in a company office, data warehouses are now hosted in the cloud

2. Real-Time Data Processing

- Warehousing application now process and analyze data as it arrives instead of waiting for batch updates

3. AI and Machine Learning Integration

- Warehousing applications are adding AI and machine learning to automatically find patterns or optimize queries.

Types of Warehousing Applications

1. Enterprise data warehouse (EDW)

→ A Central Warehouse for all data across a company

2. Cloud - Data Warehouse

→ Hosted on cloud platform like Amazon Redshift etc

3. Virtual Data Warehouse

→ Data stays in source system but appear combined through a virtual layer

Ques Discuss about the Web Mining, Spatial Mining and Temporal Mining under the Data Visualization? [AKTU-21-22, 23-24 J22-23]

Sol^m Web-mining :-

→ Web-mining is about extracting useful information from the web - such as websites, search engines or social-media.

How It Relates to Visualization

→ Web mining is about extracting useful

→ Helps visualize Online behavior

→ Useful Visuals:- click heatmaps, Flow-diagram

2. Spatial mining :

→ Spatial mining deals with geographic or location based data like maps, GPS coordinates or satellite images

How It Relates to Visualization

→ Shows where events happen and uncovers location-based patterns

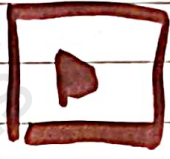
→ Finding relationship like how close stores are to customers or how weather affects sales in different areas

3. Temporal mining

→ Temporal mining focuses on data that changes over time - like logs, transactions. It's like looking at a calendar to see when things happen or repeat such as sales.

How It Relates to Visualization:

- Track when something happens and reveals trends or changes
- Useful visuals - like - line-graph, timelines

SUBSCRIBE 

JOIN TELEGRAM 