

# 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 thing change over time?

## Query Facility ?

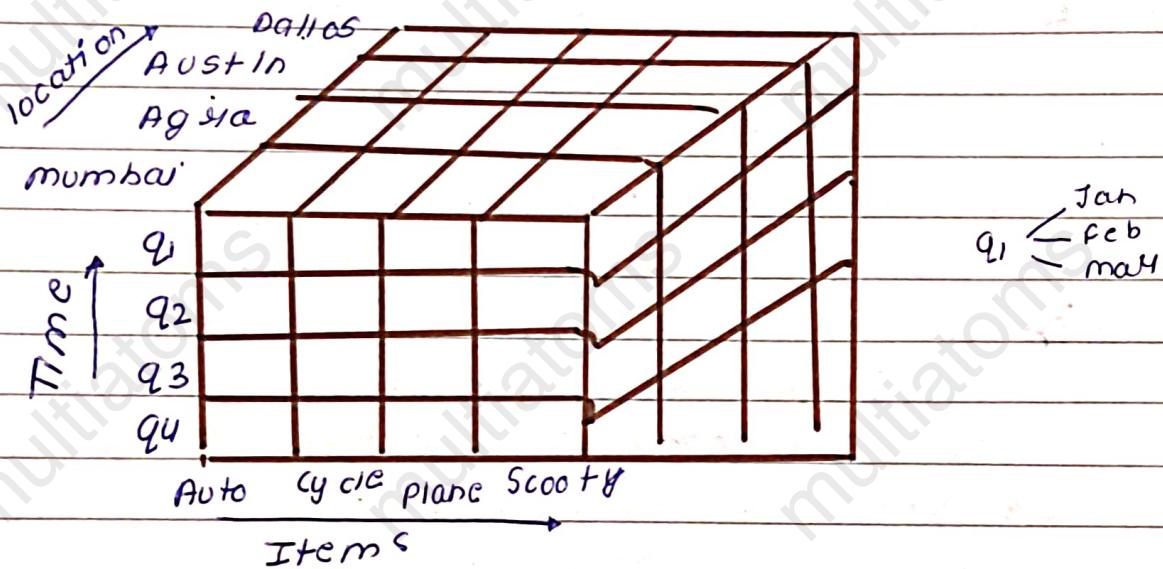
A query facility is a tool or system that lets users ask questions (queries) to retrieve, analyze or summarize data from a data warehouse or database. It's like a search engine for your data, helping you get specific information quickly

## What is OLAP?

OLAP (Online Analytical Processing) is a technology used in data warehousing to analyze large amount 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)

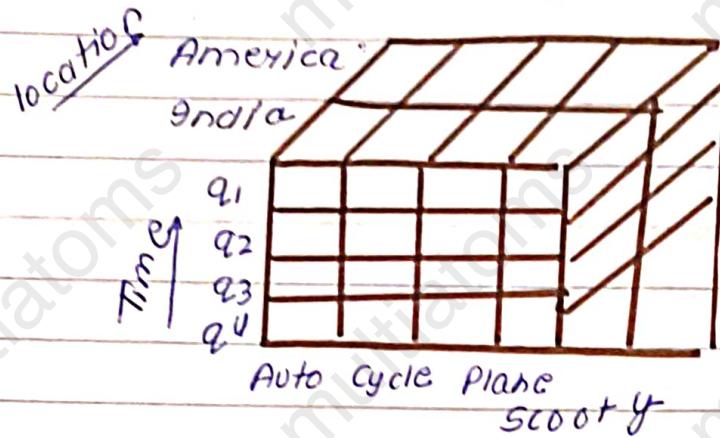


### OLAP Functions [AKTU-23-24, 21-22, 22-23]

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

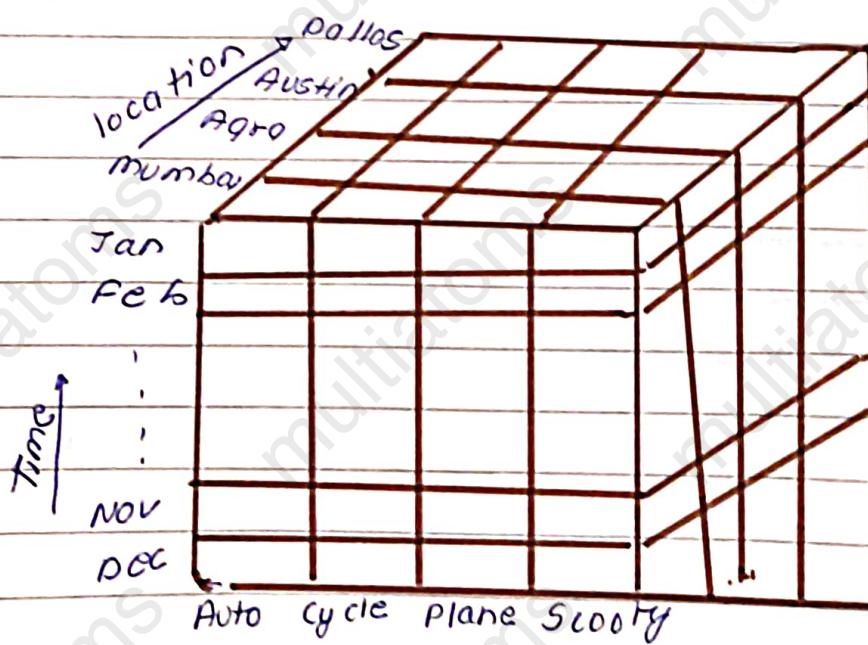
1. ROLUP :-

- 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

→ E.g. - Look at sales for only one region  
(ex "Show sales in New York")

Dollars				
Austin				
Agra				
Mumbai				

cycle Auto Plane Scooty

Items

### 4. Pivot (Rotate)

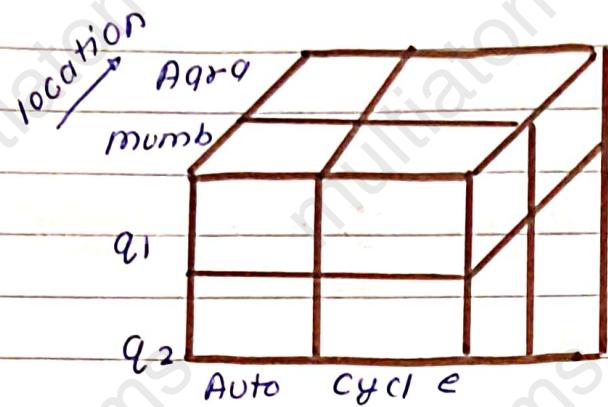
→ Change the view of data by swapping dimensions

Auto				
Cycle				
Plane				
Scooty				

Dollars      Austin      Agra      Mumbai

### 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 Servers [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 between relational backend servers & 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 depends 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 database

### 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 patterns, trends from large datasets 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 ~~down~~ 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 back up
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]

Sol There are five common challenges

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

2. Data Quality Issues :- Data comes from many sources, and inconsistency 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.

Qn Explain Warehousing application and recent trends? [AKTU-22-23]

Sol'n A Warehouse application is a software system designed to manage and optimize the operations of a warehouse or data warehouse.

In the context of data, it refers to tools or platforms that handle the storage, organization, processing and analysis of large volumes 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 servers 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

Qn discuss about the Web mining, Spatial mining and Temporal mining under the Data Visualization? [AKTU-21-22, 23-24 J 22-23]

### So, what is 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 visualis:- 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 @

# 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]

Qn-1 Explain Data mining and its functionalities?  
[AKTU-22-23]

Sol<sup>m</sup> 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 :-

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

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

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

**Ques-3 What are the needs of data mining?** [AKTU-23-24]

Sol" To extract useful patterns from huge amounts of your 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 your 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
Arijun		number
Meha	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	950 0

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:  $\rightarrow 4, 8, 15, 21, 21, 24, 25, 28, 34$

### 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:  $\rightarrow 4, 8, 15$

Bin:  $\rightarrow 21, 21, 24$

Bin:  $\rightarrow 25, 28, 34$

### 2:- Bin Mean

Once you divide your data into bins,

you can replace the original values in each bin with the mean coverage 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

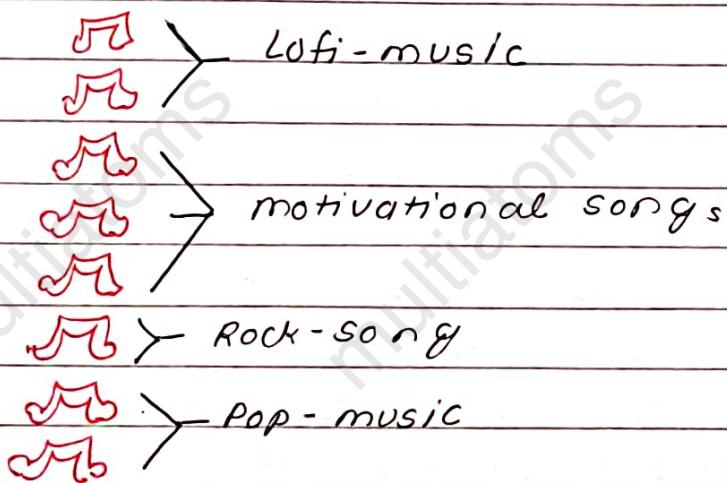
Bin-3:- 25, 25, 34

| 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



### Types of Clustering :-

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

$$y = 2^x$$

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.

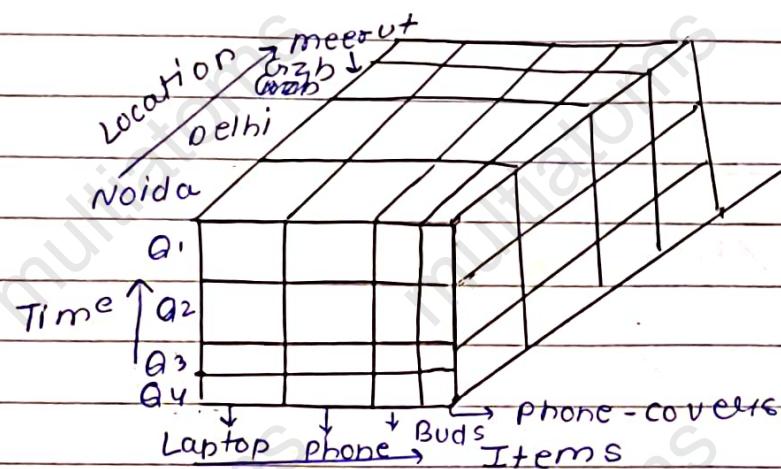
This is done after integration and before analysis or mining.

## Types of Data Transformation :

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

Qn What do you understand by Data cube Aggregation and dimensionality reduction in data mining? [AKTU-23-24, 21-22]

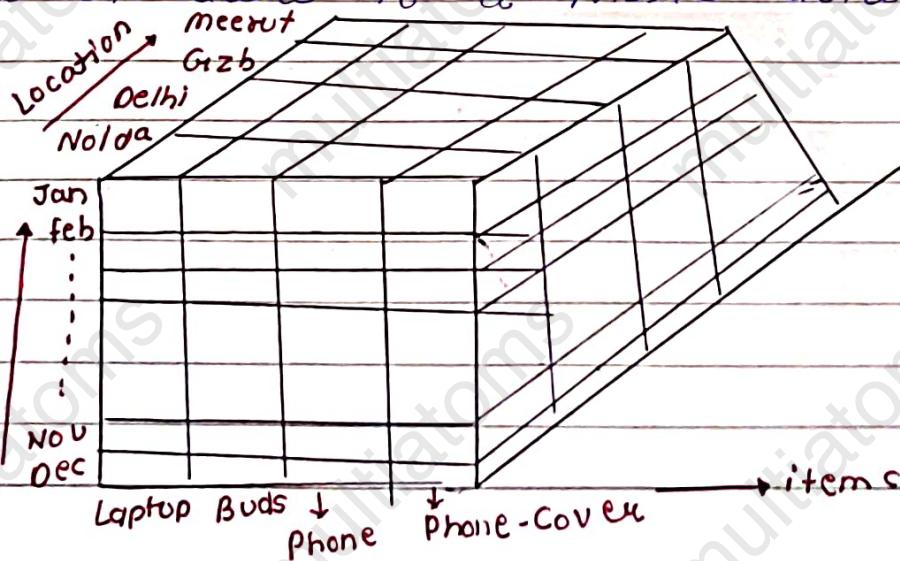
A data cube organizes data along multiple dimensions (time, location, product) to enable efficient analysis and querying. Aggregation refers to the process of summarizing data at different levels.



## 10. Operation on Aggregation

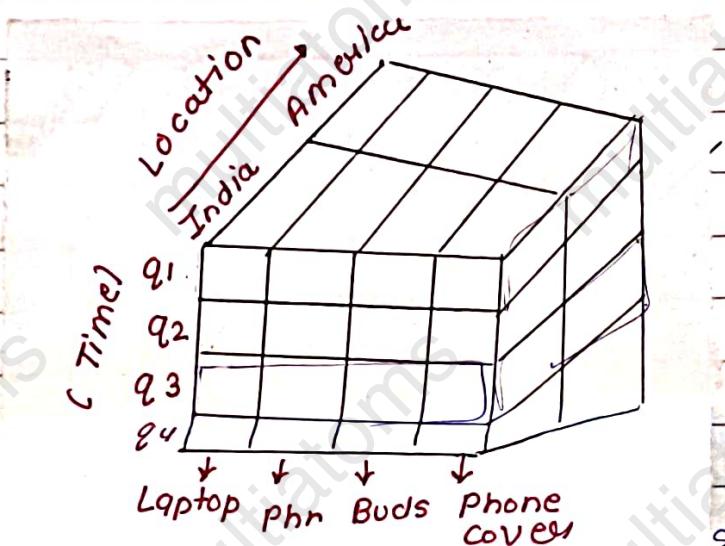
### 1. Drill-down

→ 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

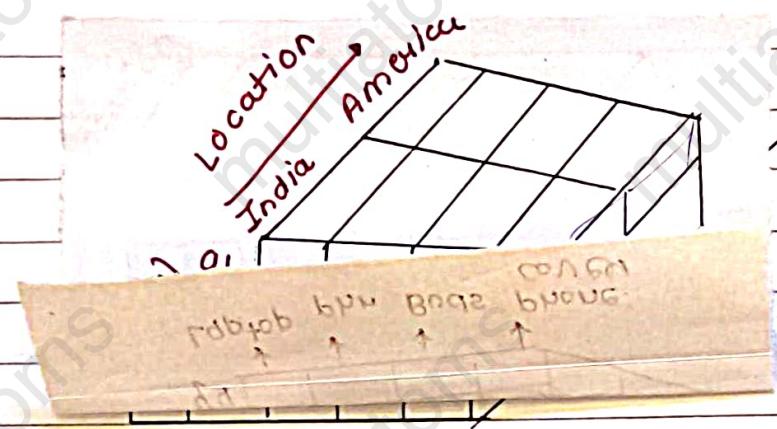
Location		Meerut	Gizb	Delhi	Metro Noida
		Phone	Lap <sup>n</sup>	Buds	Phone cover

## 4. Dice :

Pick a range on 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.



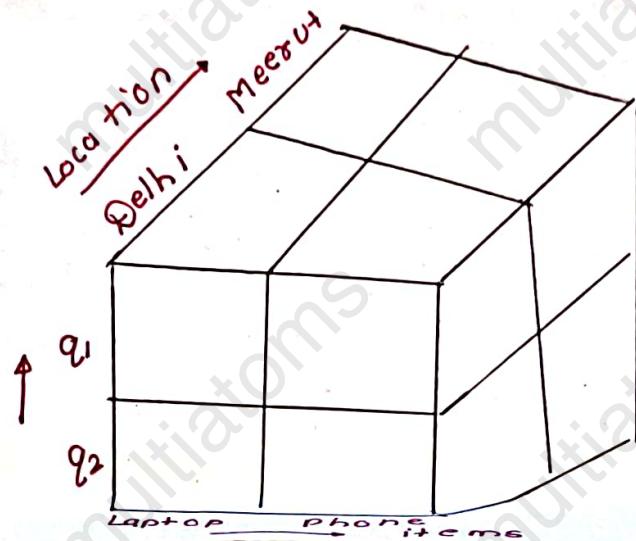
## 3. Slice :-

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


## 4. Dice :-

Pick a range or multiple values from

multiple dimensions



### 5. Pivot Cube

Reorient the cube to look at it from different perspectives.


meerut    Gizb    Delhi    Noida  
→ Location

### 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?

Sol [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+	161-
20+	-
80+	✓

### Concept Hierarchy Generation

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

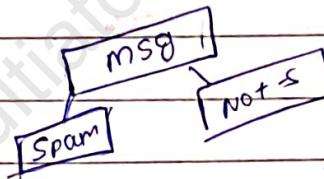
Qn Discuss Decision-tree classifier in detail? [AKTU-23-24, 22-23]

Soln

Decision tree based classifier are a popular and intu 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?



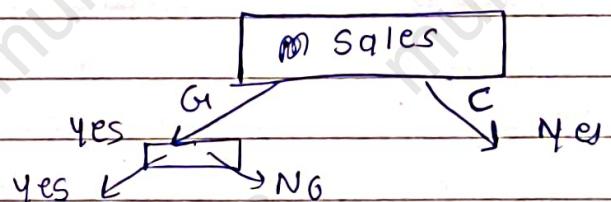
A decision tree is a visual model that represents decisions and their possible outcomes. Imagine you're trying to decide whether to go outside based on the weather

Is it raining?

→

→ If Yes, stay inside

→ If No, Is it sunny?



→ If Yes go outside

→ If no, may be 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 ▶**