(KOE-093)

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]

Soln 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 customer 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 montly 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]

Sol^n 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]

Sol^n • 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
Arjun		mumbai
Meera	30	

(ii) 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

Qu-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

I:- 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. Bin Boundaries

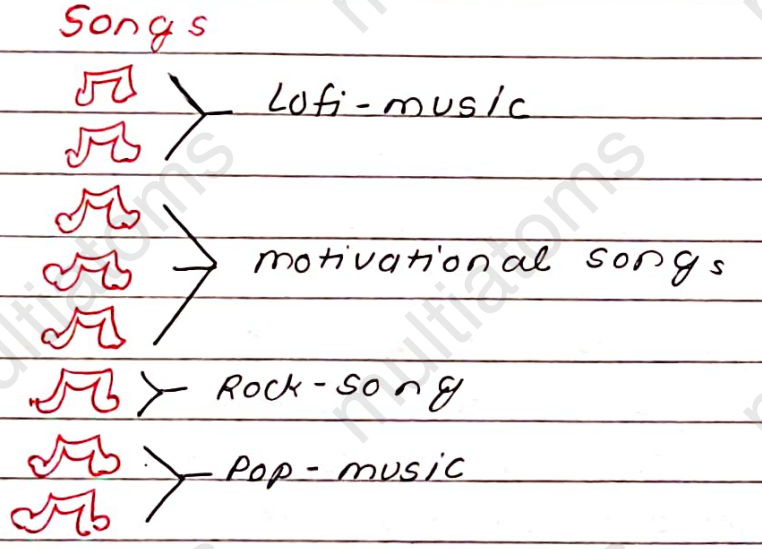
- Bin Boundaries are the lowest and highest value in each bin (group) of data
- In binning, after dividing data into bins, one way to smooth the data is to replace each value in a bin with the closest boundary value - either the lower boundary or the upper boundary.

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.



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 $y = 2x$

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.

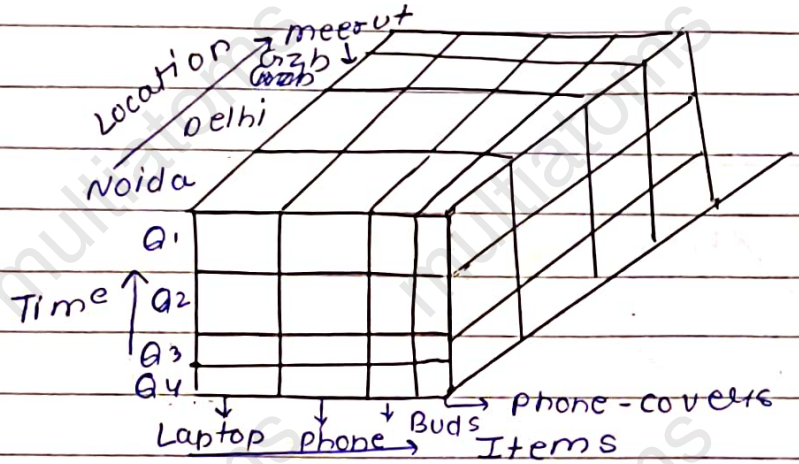
This is done after integration and before analysis or mining

Types of Data Transformation :

1. Smoothing → 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

Qu What do you understand by Data Cube Aggregation and Dimensionality reduction in Data mining? [AKTU- 23-24, 21-22]

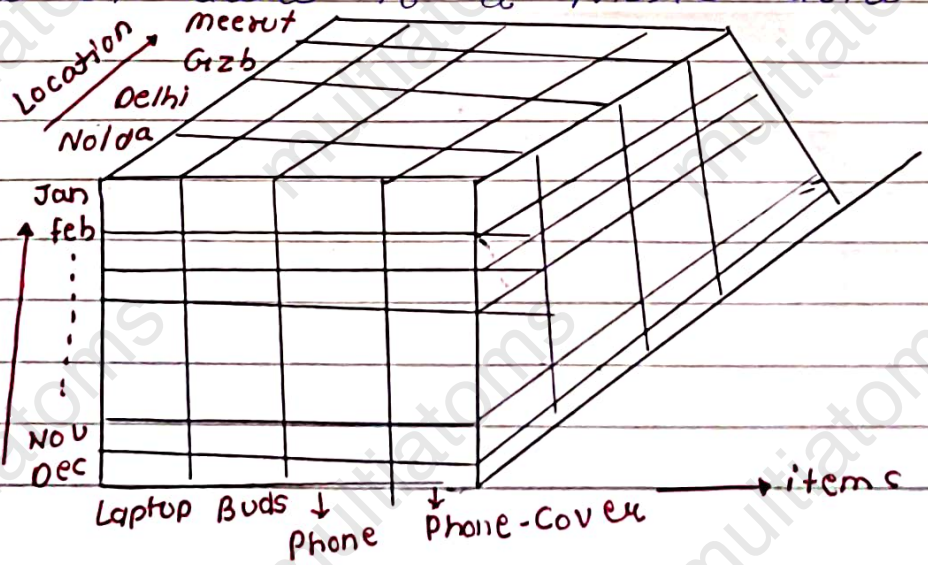
A data cube organizes data along multiple dimensions (time, location, product) to enable efficient analysis and querying. Aggregation refers to the process of summarizing data at different level



11. Operation on Aggregation

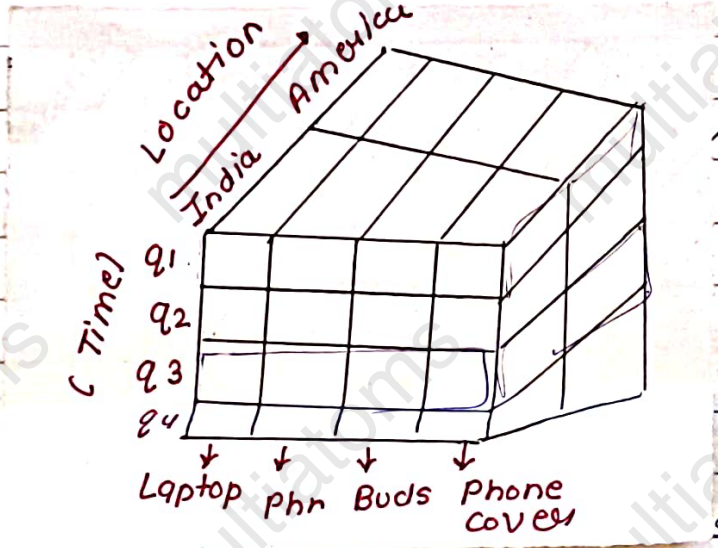
1. Drill-Down

→ 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



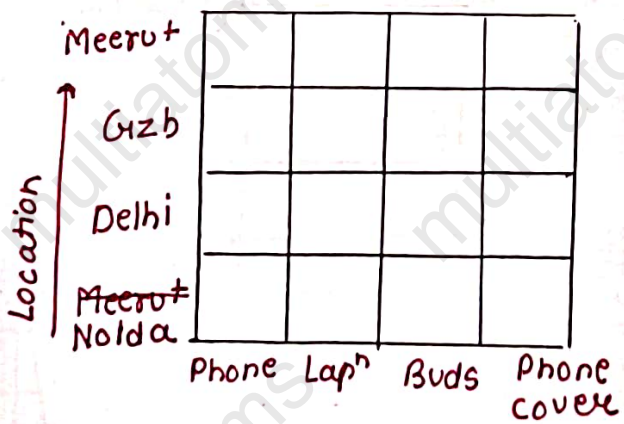
2. Roll-up :-

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 sion and get a 2D view

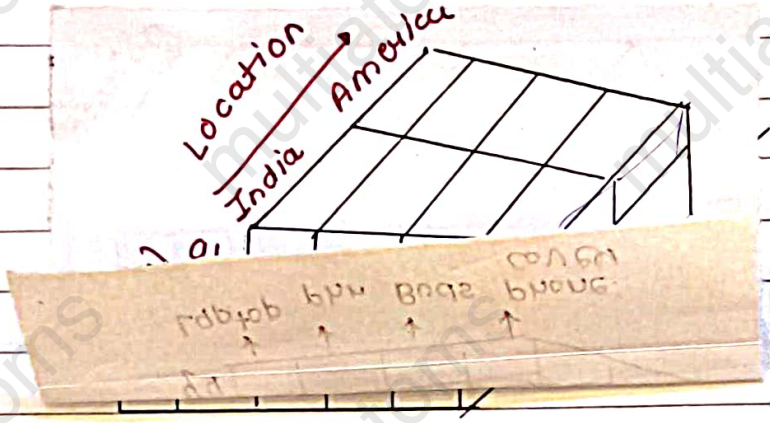


4. Dice :-

Pick a range or multiple values from

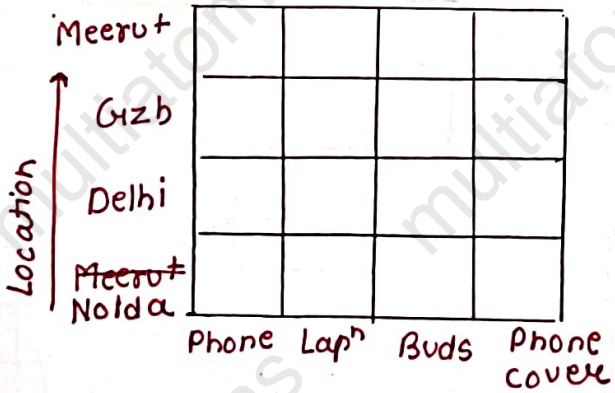
2. Roll-up :-

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 :-

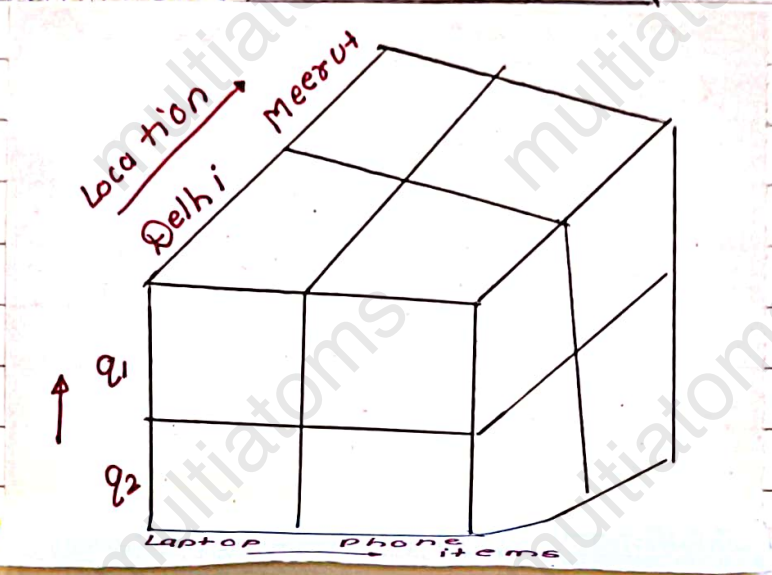
Pick one value for one demension and get a 2D View



4. Dice :-

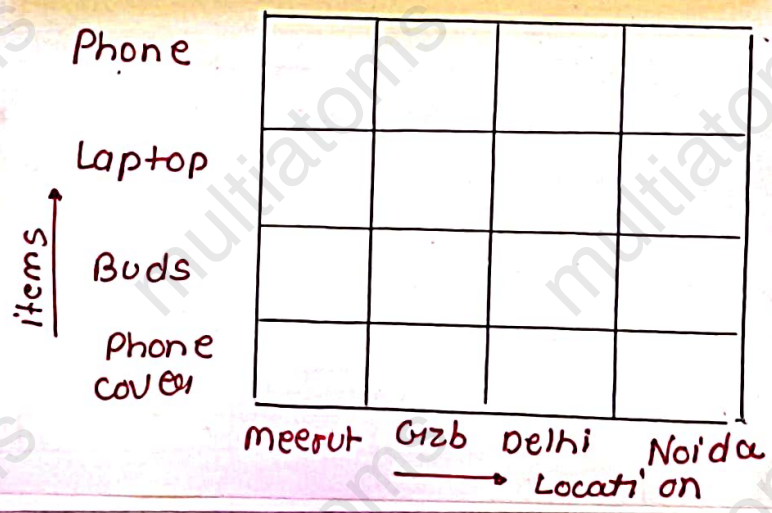
Pick a range or multiple values from

multiple dems/ons.



5. Pivot (Ro

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.

Qu- What is Numerosity Reduction ?
Sol^n [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 bi
20+	
80+	

Concept Hierarchy Generation :-

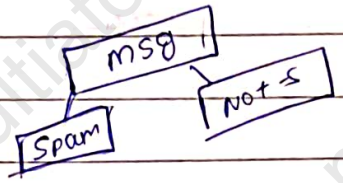
- It organizes data into a hierarchy of concepts, moving from specific to general categories. It create level of abstraction enabling analysis.
- Allows data to be analyzed at different level
- Enables the discovery of more general and meaningful patterns.

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

Soln

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

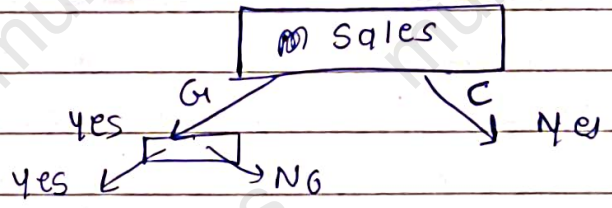
They work by breaking down complex decision into a series of simple, step-by step, much like a flowchart



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, 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 continous 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