

Data Mining & Warehouse

Date
Page

1) Differentiate b/w OLAP and OLTP.

OLAP

1) OLAP stands for Online Analytical Processing.

2) Used to analysis and decision making.

3) Works with historical data

4) Handles large amount of data

5) Queries are complex

6) Data is updated less frequently.

7) Used by managers and analysts.

example - business sales analysis.

OLTP

1) OLTP stands for Online Transaction Processing.

2) Used for day-to-day transactions.

3) Works with current data

4) Handles small and detailed transactions.

5) Queries are simple and fast

6) Data is updated continuously.

7) used by ~~ex~~ Customers and employees.

example - ATM, banking, online Ticket booking.

X

X

2) what is data Mining? Also define the key issues in data mining.

Ans) Data Mining is the process to extracting useful information, hidden patterns and knowledge from large amount of data.

It helps organizations in analysis and decision making.

ex - online shopping websites analyze customer buying behavior using data Mining.

→ Key Issues in Data Mining

- 1) Data Quality - Incomplete or incorrect data can produce wrong results.
- 2) Data Integration - Combining data from diff. source is difficult.
- 3) Privacy and Security - personal and sensitive data must be protected.
- 4) Scalability - Handling very large amounts of data requires efficient methods.
- 5) Complex Data type - Data may be in the form of text, images, videos etc, which are difficult to analyze.
- 6) Performance and speed - Data Mining should provide fast and accurate result.
- 7) Interpretation of Result - Understanding and using the discovered patterns correctly can be challenging.

Date _____
Page _____

3) Definition among ROLAP, MOLAP, HOLAP Server,
stands for Relational OLAP

1) ROLAP - ROLAP store data in relational
databases using tables.

It can handle large amounts of
data but may be slower in processing.

Ex - SQL-based database systems.

2) MOLAP - stands for Multidimensional OLAP.

MOLAP stores data in Multidimensional Cubes.
It provides fast data analysis and quick
query response.

ex - Sale analysis using data cubes.

3) HOLAP - Stands for Hybrid OLAP.
It stores some data in relational
tables and some in multidimensional cubes.

ex - Large business system using both speed
and large storage.

X

X

✓ 4) Define data Model and its various types.

Ans) A data Model is a collection of concepts used to organize, store and manage data in a database.

It defines the structure of data and the relationship b/w different data items.

ex - A relational data model stores data in table with rows and columns.

→ Types of Data Model

1) Hierarchical DM - Data is organized in a tree-like structure with parent-child relationships.

2) Network DM - Data is connected through links and one record can have multiple relationships.

3) Relational DM - Data is stored in tables consisting of rows and columns.

4) Entity-Relationship (ER) Model - Represents data using entities, attributes and relationships.

5) Object-oriented Data Model - Data is stored in the form of objects similar to object-oriented programming.

✓

Q6) What is clustering? Classify various clustering methods.

Ans) Clustering is a data mining technique used to group similar data items into clusters.

Objects in the same cluster are similar, while objects in different clusters are different.

ex - Grouping customers into low, medium and high spending groups.

Types of Clustering Method

1) Partitioning Method - Divides data into diff. groups or clusters.

ex - K-Means clustering.

2) Hierarchical Method - Creates clusters in the form of a tree structure.

3) Density-Based Method - Forms clusters based on dense regions of data.

4) Grid-Based Method - Divides data into grids and forms clusters.

5) Model-Based Method - Uses mathematical models to create clusters.

7) Explain classification techniques in data Mining.
Also define its types.

Ans) Classification in data Mining technique used to assign data into predefined classes or categories.

It helps in predicting the category of new data based on existing data.

ex - Classifying emails as spam or not spam.

→ Type / Techniques of Classification

1) Decision Tree - Represents data in a tree structure using nodes and branches.

2) Naive Bayes classification - Uses probability to classify data.

3) K-Nearest Neighbor (KNN) - classifies data based on nearest neighboring data points.

4) Neural Network - Uses interconnected nodes similar to the human brain.

5) What is Machine Learning? Also differentiate b/w Machine Learning and data Mining.

Ans) Machine Learning is a branch of Artificial Intelligence that enables computers to learn from data and improve automatically without being explicitly programmed.

ex - Spam email detection system.

→ Diff. b/w ML & DM

ML

DM

- | | |
|--|--|
| 1) It helps computers learn from data automatically. | 1) It is used to find hidden patterns from data. |
| 2) Main goal is prediction and learning. | 2) Main goal is data analysis. |
| 3) Used to improve machine performance. | 3) Used for decision making. |
| 4) Works with algorithms and training models.
ex - Spam detection | 4) Works with large databases.
ex - Customer buying pattern analysis. |
| 5) It is a part of AI. | 5) It is a process of knowledge discovery. |
| 6) Focuses on future predictions. | 6) Focuses on extracting useful information. |

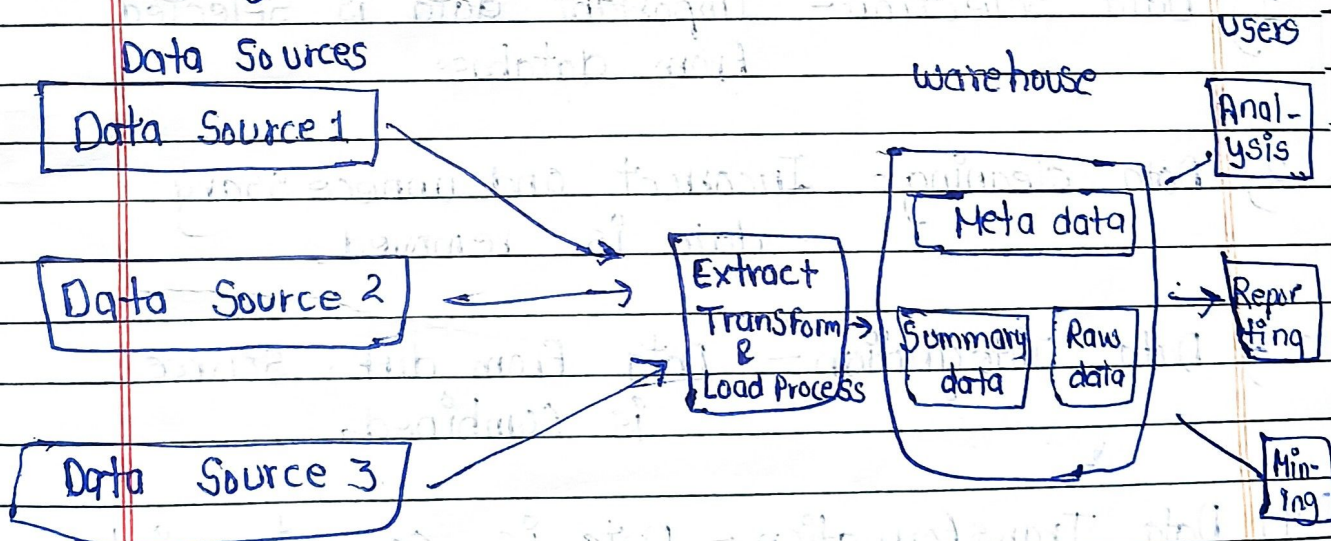
Date _____
Page _____

5) Support Vector Machine (SVM) - Separates classes using a boundary line.

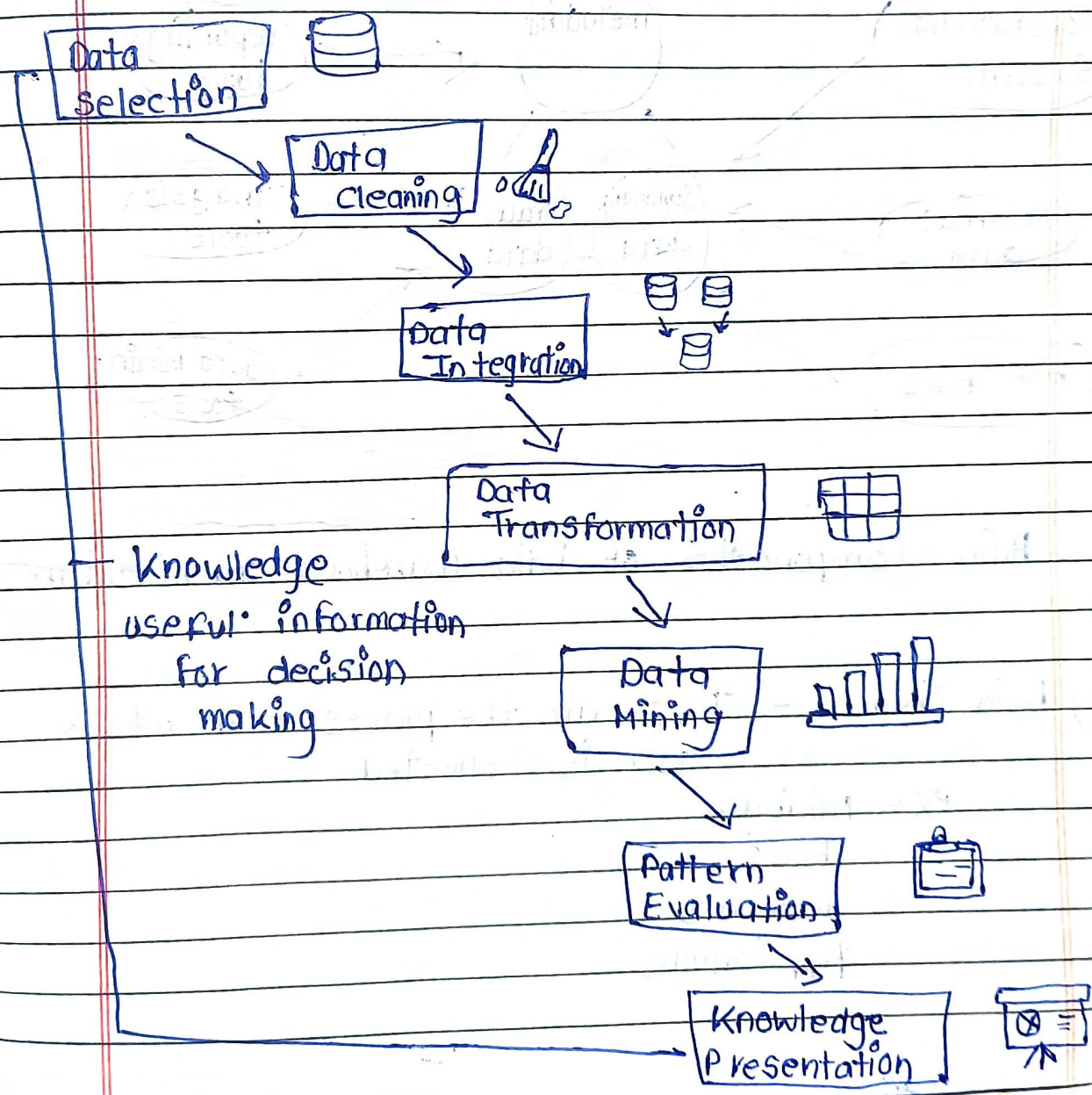
8) With a neat diagram explain the architecture of data warehouse.

Data Warehouse- A data warehouse is a central repository used to store large amounts of data collected from diff. sources. It is used for data analysis, reporting and decision making.

ex - A company stores sales data from all branches is one data warehouse for analysis

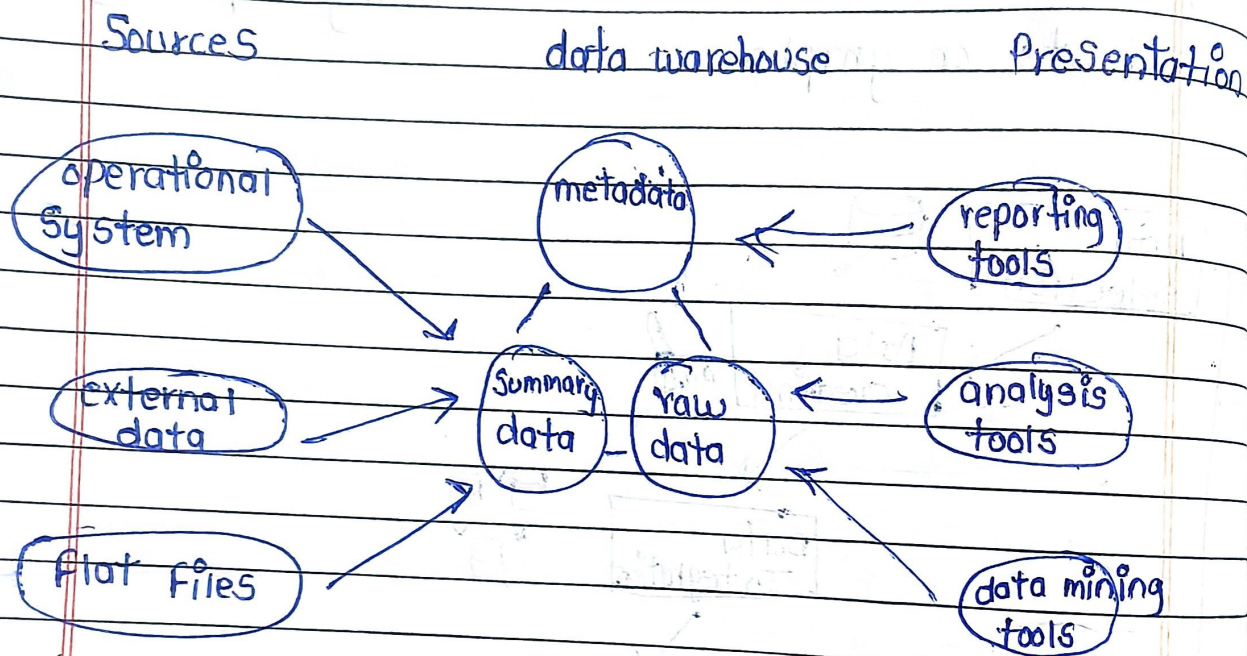


- 5) Data Mining - Useful patterns are extracted using techniques.
- 6) Pattern Evaluation - Important and meaningful patterns are identified.
- 7) Knowledge Presentation - result are presented using reports, charts or graphs.



10) Explain the architecture of data warehouse

Ans) A Data warehouse is a system used to store large amount of data from different sources for analysis and reporting.



Main Components of Data warehouse Architecture

1) Data Sources - These are the places from where data is collected.

ex - Database
Files
websites
Applications

2) ETL Process

ETL stands for

E → Extract → collect data from Sources

T → Transform → Convert data into proper Format

L → Load → Store data into warehouse.

3) Data Warehouse - It is the central storage where processed data is stored.

4) Data Marts - These are smaller parts of the data warehouse made for specific departments.

ex - Sales department

HR

Finance "

5) OLAP server - OLAP mean Online Analytical Processing.

It helps users - Analyze data quickly.

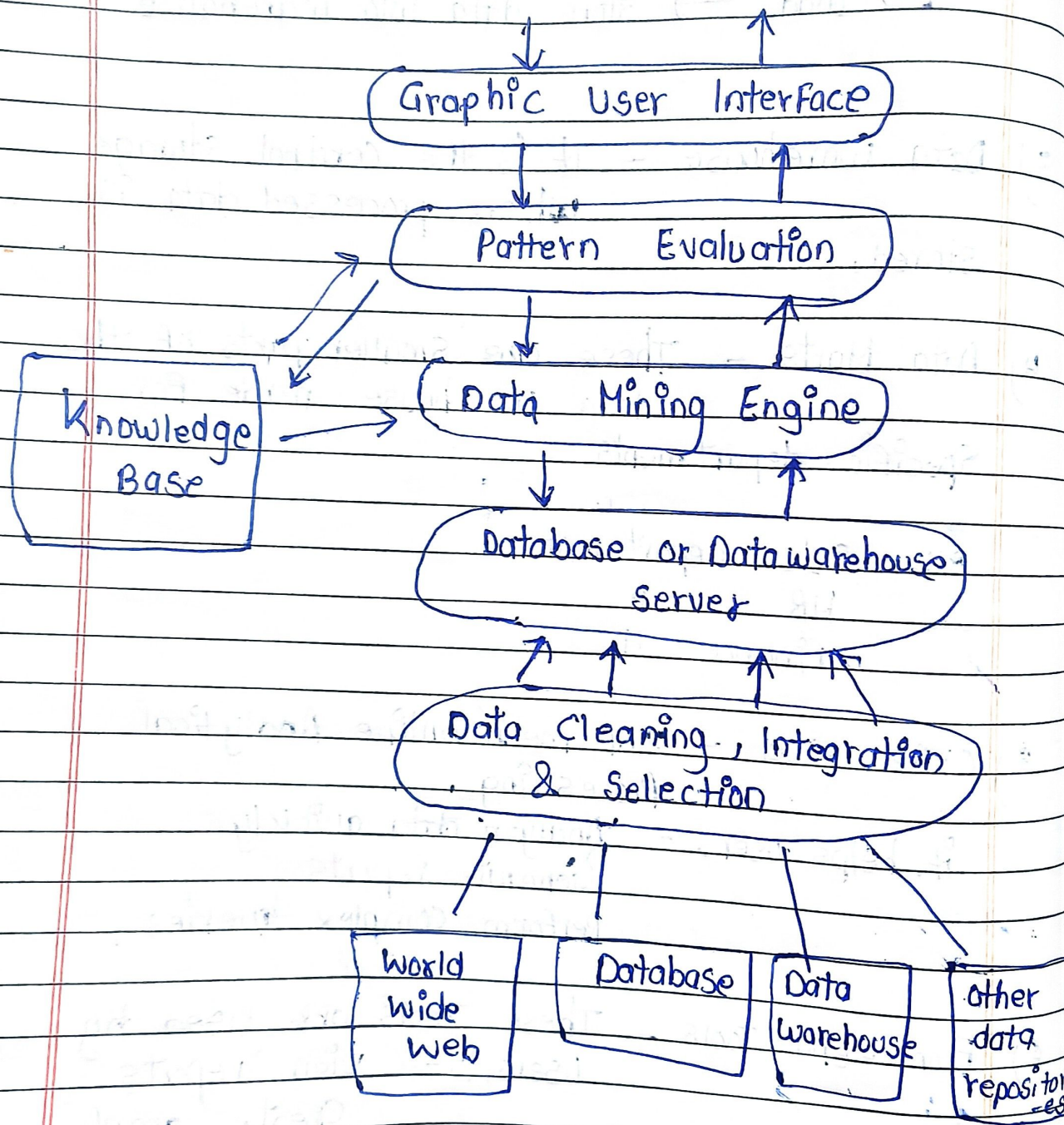
Generate reports

Perform complex queries.

6) Front-End Tools - These tools are used by users to - view reports
Create graph
Analyze information

11) Explain the architecture of data Mining with diagram.

Ans) Data Mining architecture shows how data is collected, processed, analyzed and converted into useful information.



Components of Data Mining Architecture

- 1) Data Sources - These are the places from where data is collected.
ex - database
 - data warehouse
 - Internet
 - files
- 2) Database / Data warehouse Server - This Server manages and stores large amounts of data.
- 3) Data cleaning and Integration - This step improves data quality.
 - removes errors
 - Removes duplicate data
4. Data Mining Engine - It is the main part of the system.
 - performs mining operations
 - finds patterns and relationships

Techniques used -

- classification
- clustering
- Association
- Prediction

- 5) Pattern Evaluation - This module checks which patterns are useful.
- 6) Knowledge base - It stores useful information and rules used during mining.
- 7) Graphical User Interface (GUI) - GUI allows users to interact with the system.

—————X—————X—————

12) What is Data warehouse?

Ans) A data warehouse is a large storage system used to collect and store data from different sources for reporting and analysis.

It stores :

- Historical data
- Organized data
- Business information

example - A shopping company stores

- Customer details
- Product sales
- Payment records
- Online orders

Types of Data Warehouse

1) Enterprise Data Warehouse (EDW) :-

- Stores complete organization data.
- Used by large companies

2) Data Mart - Small part of data warehouse. Used for a specific department.

ex - HR data Mart

Sale Data Mart

3) Virtual Data Warehouse - Data is viewed virtually without storing physically.

→ Advantages

- 1) stores large amount of data
- 2) Helps in decision making
- 3) Fast report generation
- 4) Improve business analysis
- 5) keep historical records.
- 6) Data From multiple Sources can be Combined.

X

X

Types of DM

Q3) Explain classification, prediction, Association

Ans) Classification - classification is a DM technique used to place data into predefined categories or classes.

It predicts the class label of data based on training data.

ex - Email $\rightarrow$ Spam or Not Spam

Student $\rightarrow$ Pass or Fail

Loan $\rightarrow$ Approved or Rejected.

$\rightarrow$ Steps of Classification

1. Collect training data
2. Build classification model
3. Test the model
4. Predict Class / Category.

$\rightarrow$ Types of Classification Algorithms

- Decision Tree
- Naive Bayes
- K-Nearest Neighbor (KNN)
- Support vector Machine (SVM)

$\rightarrow$ Advantages

- Easy to understand
- Fast

- Helps in risk analysis
- Used in medical and banking fields.

2) Prediction - Prediction is a technique used to estimate future values or outcomes based on existing data.

It predicts continuous or unknown values.

ex - Predicting weather
 Predicting stock market prices
 predicting student marks
 predicting future sales

→ Steps of prediction

1. Collect past data.
2. Analyze patterns.
3. Build prediction model
4. Predict Future result.

→ Methods used in Prediction

- Regression Analysis
- Neural Network
- Machine Learning Algorithms

→ Advantages of Prediction

- Helps in future planning.

- Reduces business risk
- Improves decision making
- Increases accuracy

✓
3) Association — Association is a data Mining technique used to Find relationships or Connections b/w items in a large database.

It shows which items are frequently bought or used together.

ex - In a Supermarket :

Customers who buy bread also often buy butter.

This relationship is called an association rule.

→ Types of Association rules

1) Single Dimensional Association rules —

This rule uses only one type of data or attribute.

ex - Milk → Bread

Only buying items are considered.

2) Multidimensional Association Rule -

This rule uses multiple attributes.

ex - Age = young $\rightarrow$ Buys Mobile

It includes age and buying behavior together.

3) Boolean Association Rule -

This rule shows whether an item is present or absent.

ex - Buys Tea $\rightarrow$ Buys Biscuit.

only yes/no condition is checked.

4) Quantitative Association Rule - This rule uses numeric value.

ex - ~~20~~ Age 20-30 $\rightarrow$ Buy 2 Cold Drinks

Numbers and quantities are included.

$\rightarrow$ Advantages -

- Helps in market analysis
- Improves product sales.
- Useful in recommendation system.

- Helps businesses understand Customer behavior.

14) What is Data Mart and define the various application of Data Mart.

Ans) A Data Mart is a small part of a data ~~mart~~ warehouse designed for a specific department or group in an organization.

It stores only department-related data instead of complete organizational data.

ex - Sales department data,
HR department data,
Finance " "

→ Advantages:

- Faster access to data
- Easy to manage
- Low cost
- Improves department analysis
- Better decision making

→ Application of DM

1) Banking Server - Used for : Fraud detection,
Loan approval
Risk analysis

2) Education Sector - Used for :

- Student performance analysis
- Predicting result
- Attendance analysis

3) Healthcare - used for :

- Disease prediction
- Medical diagnosis
- Patient record analysis

4) E-Commerce and Shopping - Used for :

- Product recommendations
- Customer behavior analysis
- Market analysis

5) Telecommunications - Used for :

- Network analysis
- Customer retention
- Fraud detection

6) Marketing - used for :

- Target advertising
- Customer segmentation
- Sales prediction

7) Weather Forecasting - Used for :

- climate analysis
- Rainfall prediction
- Temperature forecasting

X ————— X

Key features

- Stores integrated data
- Contains historical data
- Used for analysis and reporting
- Supports decision making.

✓ X X

Q) What is KDD process in data mining? Explain with diagram.

Ans) KDD stands for Knowledge Discovery in Databases.

It is the process of finding useful and hidden knowledge from large data.

→ Steps of KDD Process

- 1) Data Selection - Important data is selected from database.
- 2) Data cleaning - Incorrect and unnecessary data is removed.
- 3) Data Integration - Data from diff. sources is combined.
- 4) Data Transformation - Data is converted into suitable form.