ASSIGNMENT
OF
DATABASE MANAGEMENT SYSTEM - II

1. What is OLAP? Explain its purpose in data analysis.

Ans: OLAP (Online Analytical Processing) is a category of software that enables users to interactively analyze multidimensional data from multiple perspectives.It is designed for complex queries, multi-dimensional analysis, and decision support.

Purpose in Data Analysis:

- To enable multi-dimensional analysis of data.
- To support decision-making through trends, patterns, and comparative analysis.
- To provide quick summaries and aggregations.

2. Describe four key OLAP operations (e.g., roll-up, drill-down, slice, dice) and give a short example for each.

Ans: Four keys of OLAP operations are -

i. Roll-up:Aggregates data to a higher level in a hierarchy. Exp: Monthly sales to Quarterly sales.

ii. Drill-down:Breaks down data to more detailed levels . exp: Quarterly sales to Monthly sales.

iii. Slice:Extracts a single layer of data based on one dimension .exp: Sales for Q1 2025 only .

iii. Dice: Extracts a subcube selecting specific values from multiple dimensions. Exp: Sales for Q1 & Q2 in Japan for Electronics.

3. Define a data warehouse. List and explain its four main characteristics.

Ans: A data warehouse is a centralized repository designed to store integrated data from multiple sources. It supports business intelligence (BI), reporting, and data analysis.

Four main characteristics:

1. Subject-Oriented: Organized around key subjects like customers, products.

2. Integrated: Data from multiple sources is combined and standardized.

3. Time-Variant: Historical data is stored to analyze trends over time.

4. Non-Volatile: Once data is loaded, it is stable, only read or appended.

4. Depict and describe the core components of a data warehouse (e.g., ETL, data storage, metadata, OLAP engine).

Ans:

i. ETL (Extract, Transform, Load)	Pulls data from sources, cleans and formats it, then loads into storage
ii. Data Storage	Centralized repository, often in relational or multidimensional format
iii. Metadata	Describes data's source, transformations, and usage
iv. OLAP Engine	Enables multidimensional analysis and querying

5. Differentiate star schema and snowflake schema with a brief comparison of their structure and advantages.

Ans:

Feature	Star Schema	Snowflake Schema
Structure	Fact table linked to de-normalized dimension tables.	Fact table linked to normalized dimension tables.

Query Speed	Faster (fewer joins).	Slightly slower (more joins).
Storage	Uses more space.	More space-efficient.
Ease of Use	Easier to understand.	More complex.
Use Case	Best for simpler reporting.	Better when dimensions have many levels.

6. What is a Decision Support System (DSS)? Describe its main purpose in organizational decision-making.

Ans: A decision support system (DSS) is a computer program used to improve a company's decision-making capabilities. It analyzes large amounts of data and presents an organization with the best possible options available.

A decision support application might gather and present the following typical information:

- Comparative sales figures between one week and the next.
- Projected revenue figures based on new product sales assumptions.
- The consequences of different decisions.

The purpose of a DSS is to gather, analyze and synthesize data to produce comprehensive information reports that an organization can use to assist in its decision-making process. Unlike tools that are limited to just data collection, DSSes also process that data to create detailed reports and projections.DSSes are an adaptable tool meant to meet the specific needs of the organization using it. Finance, healthcare and supply chain management industries, for example, all use DSSes to help in their decision-making processes. A DSS report can provide insights on topics like sales trends, revenue, budgeting, project management, inventory management, supply chain optimization and healthcare management.

7. Explain at least three types of DSS (e.g., data-driven, model-driven, knowledge-driven), providing an example for each.

Ans: Decision support systems are types of information systems that analyze data to support executives, managers and staff in business decision-making.
Here are three of the most common examples of decision support systems we might encounter in the workplace :

1. Data-driven DSS:A data-driven DSS gives users access to a large amount of internal and external data. This DSS will query a database using the web, an external server or a company's mainframe.
Software examples of a data-driven DSS include:

i.Geographic Information Systems (GIS)

ii.File drawer systems

iii.Executive information systems

iv.Computer-based databases with query systems

2. Model-driven DSS:A model-driven DSS allows a user to analyze and manipulate specific models of data, such as statistics, finances or scheduling. These decision support systems are specific to the type of model the user wants to interact with and typically offer less data than other DSS types.
Software examples of a model-driven DSS include:

i.Scheduling software

ii.Financial modeling

iii.Decision analysis modeling

iv.Optimization software

3.Knowledge-driven DSS:With a knowledge-driven DSS, a knowledge-management system monitors continually updated data about an organization to support decisions. The DSS uses diagnosis, prediction, interpretation and classification to recommend actions consistent with the business.
Software examples of a knowledge-driven DSS include:

i.Software that identifies new or current customers who might be interested in products

ii.Product selection software

8. Define Data Mining. Discuss two real-world applications and their impact.

Ans: Data is simply raw facts or figures, like numbers or text, which by themselves don't mean much. But when processed, they become useful information. Today, we collect huge amounts of data—from simple measurements to complex formats like images, videos, and web content. As the amount of data grows rapidly, data mining techniques help us find useful patterns and insights. For example, banks use data mining to study customer transactions and predict who might be interested in loans, credit cards, or insurance.

Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects

Intrusion Detection: Network intrusion refers to any unauthorized access or activity on a digital network, often aimed at stealing or misusing resources. Data mining plays a key role in detecting such intrusions by identifying unusual patterns, anomalies, and potential threats within large datasets. It helps classify and extract relevant data to support Intrusion Detection Systems (IDS), which monitor network traffic and raise alerts for suspicious activities.

- Detect security violations
- Misuse Detection

9. What are association rules? Define support, confidence, and lift with formulas or descriptions.

10. Describe a real-world scenario where you would use OLAP, then apply data mining techniques (association, classification, clustering) to the same dataset—outline the integrated workflow.

Ans: Scenario: An online retailer wants to improve customer retention.

Step 1: OLAP Analysis

- Use OLAP to examine:
    - Customer purchases over time (drill-down by month)
    - Regional sales trends (slice by location)
    - Product category profitability (dice by category + region)

Step 2: Data Mining Techniques

- Association: Identify commonly bundled items (e.g., "laptop + mouse")
- Classification: Predict if a customer is likely to churn (e.g., using Decision Trees)
- Clustering: Segment customers by behavior (high-spenders, occasional buyers)

Integrated Workflow:

1. Use OLAP to identify potential areas of concern (e.g., declining sales in a segment)

2. Apply data mining to uncover patterns, predict outcomes, or find customer groups
3. Combine both insights to tailor marketing or optimize supply chain.

11. Differentiate classification and clustering in terms of goals, methodology, and types of learning (supervised vs. unsupervised).

Ans: Difference between Classification and Clustering

S.No.	Classification	Clustering
1	It is an approach to classifying the input instances on the basis of related class labels.	It is used to set the instances on the basis of their resemblance without class labels.
2	Classification is a type of supervised learning method.	Clustering is a kind of unsupervised learning method.
3	It prefers a training dataset.	It does not prefer a training dataset.
4	Classification is more complex as compared to clustering.	Clustering is less complex as compared to the classification.
5	Here, we utilised the labels for training data.	Here, we don’t prefer the labels for training data.

12. Name two classification algorithms (e.g., Decision Trees, Naïve Bayes) and describe briefly how one of them works.

Ans: Decision Tree and Naive Bayes are two popular classification algorithms. Both are widely used in various applications such as spam filtering, fraud detection, and medical diagnosis. However, they are based on different theoretical foundations, and their performance varies depending on the nature of the data. How Decision Trees Work:

1. Splitting the Dataset: The algorithm selects a feature that best splits the dataset into distinct classes.
2. Tree Structure: The decision tree is built recursively. Each node represents a feature and the branches represent decisions based on that feature.
3. Termination: The tree continues splitting the dataset until it cannot be split further, either because the data is perfectly classified or the maximum depth of the tree is reached.

13. Name two clustering algorithms (e.g., K-Means, DBSCAN) and explain one key difference between them.

Ans: 1. K-Means Clustering : K-means is a centroid-based or partition-based clustering algorithm. This algorithm partitions all the points in the sample space into K groups of similarity. The similarity is usually measured using Euclidean Distance .The algorithm is as follows :

Algorithm:

- K centroids are randomly placed, one for each cluster.
- Distance of each point from each centroid is calculated

- Each data point is assigned to its closest centroid, forming a cluster.
- The position of K centroids are recalculated.

2. DBScan Clustering : DBScan is a density-based clustering algorithm. The key fact of this algorithm is that the neighbourhood of each point in a cluster which is within a given radius (R) must have a minimum number of points (M). This algorithm has proved extremely efficient in detecting outliers and handling noise.The algorithm is as follows :

Algorithm:

- The type of each point is determined. Each data point in our dataset may be either of the following :
- Core Point: A data point is a core point if there are at least M points in its neighborhood i.e., within the specified radius (R).
- Border Point: A data point is classified as a BORDER point if:
- Its neighborhood contains less than M data points, or
- It is reachable from some core point i.e., it is within R-distance from a core point.
- Outlier Point: An outlier is a point that is not a core point, and also, is not close enough to be reachable from a core point.
- The outlier points are eliminated.
- Core points that are neighbors are connected and put in the same cluster.
- The border points are assigned to each cluster.

14. Explain the Apriori algorithm for discovering frequent itemsets and generating association rules.

Ans: The Apriori algorithm is used to mine frequent itemsets and generate association rules. How It Works:

1. Set minimum support and confidence thresholds.
2. Identify itemsets that meet the minimum support (e.g., {milk, bread} occurs in 30% of transactions).
3. Use those itemsets to generate rules (like milk → bread) that meet the confidence threshold.
4. Repeat with larger itemsets (iteratively growing).

Optimization: Uses the Apriori Property: If an itemset is infrequent, all its supersets will be infrequent, allowing early pruning.

15. Explain how you evaluate a classification model, including metrics like precision, recall, accuracy, or F1-score.

16. Explain how you evaluate clustering results, mentioning at least one metric (e.g., silhouette score, SSE).

Ans: Clustering is a technique in Machine Learning that is used to group similar data points. While the algorithm performs its job, helping uncover the patterns and structures in the data, it is important to judge how well it functions. Several metrics have been designed to evaluate the performance of these clustering algorithms.

17. Outline the six phases of CRISP-DM. Why is the 'Data Preparation' phase often the most time-consuming?

Ans: The CRISP-DM process is organized into six major phases, each with specific tasks and deliverables. These phases are:

1. Business Understanding: Define objectives,success criteria.
2. Data Understanding: Collect and explore initial datasets.

3. Data Preparation: Clean,transform and integrate data.
4. Modeling: Apply algorithms and adjust parameters
5. Evaluation : Verify that the model meets business needs.
6. Deployment: Deliver model insights into operations.

Why is Data Preparation time-consuming?

- It involves data cleaning (removing duplicates, missing values)
- Requires formatting and transforming diverse datasets
- Often, 60–80% of time goes into this phase because real-world data is rarely analysis-ready

18. Compare OLAP, ROLAP, MOLAP, and HOLAP, focusing on their architectures, strengths, and typical use cases.

19. Discuss ethical considerations and challenges (e.g., data privacy, bias, interpretability) in applying data mining and machine learning in decision support environments.

When using data in Decision Support Systems, ethics and transparency matter greatly.

Major Challenges:

- Data Privacy: Handling personal data securely; e.g., using anonymization
- Bias and Fairness: Biased data can lead to unfair outcomes (e.g., hiring algorithms)

- Model Interpretability: Complex models like deep learning may lack explainability

Responsible Practice Includes:

- Transparency about how data is collected and used
- Auditing ML models for bias or discrimination
- Ensuring consent and compliance with data regulations (like GDPR)