Experiment 1 – OLTP & Basic Database Setup

1. What is OLTP?

OLTP (Online Transaction Processing) supports day-to-day transactional operations such as insert, update, and delete.

2. What is normalization?

Normalization organizes data to reduce redundancy and improve consistency.

3. What is a primary key?

A primary key uniquely identifies each record in a table.

4. What is a foreign key?

A foreign key creates a relationship between two tables.

5. What is a transaction?

A transaction is a logical unit of work that must execute completely or not at all.

- **Experiment 2 – Dimension & Fact Tables**

1. What is a dimension table?

A dimension table stores descriptive attributes like time, product, customer, etc.

2. What is a fact table?

A fact table stores measurable numerical data such as sales, quantity, revenue.

3. What is a surrogate key?

A system-generated unique key used in dimension tables.

4. Why do we need a star schema?

Star schema provides fast query performance and simple join structures.

5. What is granularity?

Granularity defines the level of detail stored in a fact table.

🔹 Experiment 3 – OLAP Operations

1. What is OLAP?

OLAP (Online Analytical Processing) supports multi-dimensional data analysis.

2. What is roll-up?

Roll-up increases the level of aggregation (e.g., day → month → year).

3. What is drill-down?

Drill-down moves from summarized data to more detailed data.

4. What is slicing?

Slice selects a single value of a dimension to create a sub-cube.

5. What is dicing?

Dice selects multiple dimension values to create a smaller sub-cube.

🔹 Experiment 4 – Decision Tree (J48 in WEKA)

1. What is a decision tree?

A classification model where nodes represent tests and leaves represent class labels.

2. What is J48?

J48 is WEKA's implementation of the C4.5 decision tree algorithm.

3. What is entropy?

Entropy measures impurity or randomness in data.

4. What is information gain?

Information gain measures reduction in entropy after splitting on an attribute.

5. What is pruning?

Pruning removes unnecessary branches to reduce overfitting.

🔹 Experiment 5 – Naïve Bayes Algorithm

1. What is Naïve Bayes?

A probabilistic classifier based on Bayes' theorem with independence assumptions.

2. Why is it called "naïve"?

Because it assumes all features are independent from each other.

3. What is prior probability?

The probability of an event before observing data.

4. What is posterior probability?

The probability of an event after considering evidence.

5. What are common applications of Naïve Bayes?

Spam detection, text classification, sentiment analysis.

- **Experiment 6 – K-Means Clustering**

1. What type of learning is K-Means?

Unsupervised learning.

2. What does K represent?

The number of clusters to be formed.

3. What is a centroid?

The mean position of all the points in a cluster.

4. What is the goal of K-means?

To minimize the distance between data points and their cluster centroid.

5. Does K-means require labeled data?

No, it works on unlabeled datasets.

- **Experiment 7 – Agglomerative Clustering**

1. What is hierarchical clustering?

A clustering method that builds a hierarchy of clusters.

2. What is agglomerative clustering?

A bottom-up method where each object starts as its own cluster and merges step-by-step.

3. What is a dendrogram?

A tree diagram showing how clusters merge at each level.

4. What distance metric is used commonly?

Euclidean distance.

5. What is linkage?

A rule to measure distance between clusters (single, complete, average).

- **Experiment 8 – Apriori Algorithm**

1. What is Apriori algorithm used for?

Finding frequent itemsets and generating association rules.

2. What is support?

The frequency of an itemset in the dataset.

3. What is confidence?

The likelihood that rule **A → B** is true.

4. What is the Apriori property?

If an itemset is frequent, all its subsets must be frequent.

5. Applications of Apriori?

Market basket analysis, recommendation systems, retail analytics.

1. Data Warehouse Architecture

A data warehouse architecture consists of **Source Layer**, **ETL Layer**, **Data Warehouse Storage**, and **Front-End Tools**.
It integrates data from multiple sources, transforms it, stores it in a central repository, and provides analysis through OLAP/reporting tools.

---

2. Data Warehouse

A data warehouse is a **centralized repository** used for decision-making.
It stores **historical**, **subject-oriented**, **non-volatile**, and **integrated** data optimized for analysis, not transactions.

---

3. OLTP vs OLAP

OLTP	OLAP
Handles daily transactions	Handles analytical queries
Real-time, operational	Historical, decision-making
Highly normalized tables	Denormalized/star schema
Fast inserts/updates	Fast read/aggregation

---

4. OLAP Operations

- **Roll-up:** Less detail (month → year)
- **Drill-down:** More detail (year → month)
- **Slice:** Select one dimension value
- **Dice:** Select multiple dimension values
- **Pivot:** Rotate view of dimensions

---

5. Data Mining

Data mining is the process of discovering **patterns, knowledge, and insights** from large datasets using machine learning and statistical techniques.

---

6. Issues in Data Mining

Includes noisy data, missing values, scalability, privacy issues, algorithm complexity, and data diversity.

---

7. Challenges in Data Mining

Handling **huge volumes**, **high-dimensional data**, **distributed data**, **real-time mining**, and **security concerns**.

---

8. Data Preprocessing

Preprocessing improves data quality using:

- Cleaning
- Integration
- Transformation
- Reduction
  to prepare raw data for mining.

---

9. Data Integration, Transformation, Cleaning & Reduction

- **Integration:** Combine data from multiple sources
- **Cleaning:** Fix missing, inconsistent, noisy data
- **Transformation:** Normalize, aggregate, generalize data
- **Reduction:** Reduce dataset size while keeping information (PCA, sampling)

---

10. Naïve Bayes, Decision Tree – Induction Algorithm

- **Naïve Bayes:** Probabilistic model based on Bayes theorem assuming independence.
- **Decision Tree (ID3/C4.5):** Builds tree using **entropy** and **information gain** to choose best attribute at each node.

---

11. Data Warehouse Schema

- **Star Schema:** One fact table + multiple dimension tables
- **Snowflake Schema:** Normalized dimensions
- **Fact Constellation:** Multiple fact tables share dimensions.

---

12. Accuracy & Error Measures

Accuracy = (Correct Predictions / Total Predictions).
Error metrics include **Precision**, **Recall**, **F1-score**, **Mean Absolute Error**, **Root Mean Square Error**.

---

13. Cross Validation

A technique to evaluate model performance.
In **k-fold cross-validation**, data is divided into k parts; model is trained on k−1 folds and tested on the remaining fold.

---

14. K-Means with Example

K-means divides data into **k clusters** by assigning each point to the nearest centroid.
Example: With k=2, points {1,2,3} cluster around centroid 2, while {8,9,10} cluster around centroid 9.

---

15. K-Means vs K-Medoids

K-Means	K-Medoids
Uses mean as centroid	Uses actual data point (medoid)
Sensitive to outliers	Robust to outliers
Faster	Slower
Works on Euclidean distance	Works with arbitrary metrics

---

16. Market Basket Analysis (MBA) with Examples

MBA finds items frequently bought together.
Examples:

- {Bread → Butter}
- {Milk → Cornflakes}
- {Shampoo → Conditioner}
- {Shoes → Socks}

---

17. Frequent Itemsets, Closed Itemsets

- **Frequent Itemset:** Items with support ≥ minimum threshold.
- **Closed Itemset:** A frequent itemset with **no superset having the same support**.

---

18. Apriori Algorithm

Apriori finds frequent itemsets using the principle:
If an itemset is frequent, all its subsets are also frequent.
It uses **support** and **confidence** to generate association rules.

---

19. Web Content Mining

Extracting useful information such as **text, images, videos** from web pages.

---

20. Web Structure Mining

Analyzing the **link structure** of the web using graph theory (nodes = pages, edges = links).

---

21. Difference: Web Content Mining vs Web Structure Mining

Content Mining	Structure Mining
Extracts page data	Analyzes links between pages
Works on text/media	Works on graph structure
NLP techniques	Graph algorithms

---

22. PageRank (with Example)

PageRank measures the importance of web pages based on incoming links.
Example: If Page A is linked by 10 popular pages, its PageRank is high.
Formula uses **inbound links + link weight** to calculate importance.

---

23. Association Rule Concept

Rules like **A → B** show relationships between items.
Evaluated using **support**, **confidence**, and **lift**.

---

24. Clustering Applications & Challenges

Applications: Customer segmentation, image analysis, anomaly detection.
Challenges: Choosing correct number of clusters, high-dimensional data, noise, scalability.

Basic Data Mining Viva Questions

(Your teacher can easily ask these)

---

1. What is Data Mining?

It is the process of discovering patterns and useful knowledge from large datasets.

---

2. What is the difference between Data Mining and Machine Learning?

Data mining focuses on extracting patterns; ML focuses on creating models that learn and predict.

---

3. What is Knowledge Discovery in Databases (KDD)?

A complete process of cleaning data, selecting data, mining patterns, and evaluating results.

---

4. What are the major tasks of Data Mining?

Classification, clustering, association rule mining, prediction, and anomaly detection.

---

5. What is a dataset?

A collection of data records consisting of attributes and values.

---

6. What are attributes in data mining?

Features/fields that describe the data (e.g., age, salary, marks).

---

7. What is supervised learning?

Learning using **labeled data** (e.g., classification).

---

8. What is unsupervised learning?

Learning from **unlabeled data** (e.g., clustering).

---

9. Give two examples of supervised algorithms.

Naïve Bayes, Decision Tree.

---

10. Give two examples of unsupervised algorithms.

K-means, Hierarchical clustering.

---

11. What is classification?

Predicting a categorical output (example: pass/fail, spam/not spam).

---

12. What is clustering?

Grouping similar data points into clusters without labels.

---

13. What is association rule mining?

Finding relationships like **A → B** in transactional data.

---

14. What is support?

How frequently an itemset appears in the dataset.

---

15. What is confidence?

The probability that rule **A → B** is correct.

---

16. What is lift?

Measure of how strong the association between A and B is, compared to random chance.

---

17. What is the Apriori algorithm?

An algorithm that finds frequent itemsets using the Apriori property.

---

18. What is data preprocessing?

Cleaning, transforming, integrating, and reducing data before mining.

---

19. Why is data cleaning important?

Because raw data has missing values, noise, and inconsistencies.

---

20. What is normalization?

Scaling numerical values to a similar range.

---

21. What is a decision tree?

A tree structure used to classify data using attribute-based rules.

---

22. What is entropy?

A measure of impurity or randomness in data.

---

23. What is information gain?

The reduction in entropy after splitting data by an attribute.

---

24. What is Naïve Bayes used for?

Spam detection, classification, sentiment analysis.

---

25. What is a centroid in K-means?

The mean point representing the center of a cluster.

---

26. How does K-means work?

Assign points to closest centroid → update centroid → repeat until stable.

---

27. What is hierarchical clustering?

Clustering that builds a tree (dendrogram) using merging/splitting.

---

28. What is an outlier?

A data point that is very different from others.

---

29. What is dimensionality reduction?

Reducing the number of attributes (e.g., using PCA).

---

30. What is OLAP?

Online Analytical Processing—used for multi-dimensional analysis.

---

31. Difference between OLAP and OLTP?

OLTP = transactions, OLAP = analysis.

---

32. What is a data warehouse?

A central repository storing historical data for analysis.

---

33. What is feature selection?

Choosing the best attributes to improve model accuracy.

---

34. What is overfitting?

When the model learns noise instead of pattern and performs poorly on new data.

---

35. What is cross-validation?

Technique to evaluate accuracy by training/testing on different splits of data.

✅ How Each Experiment's Algorithm (Used in WEKA) Works — In Short

---

1️⃣ Decision Tree (J48 / C4.5)

- Calculates **entropy** of data.
- Computes **information gain** for each attribute.
- Chooses the attribute with highest gain for splitting.
- Repeats until pure leaf nodes are formed.
- Applies **pruning** to avoid overfitting.

2️⃣ Naïve Bayes (Experiment 5)

- Computes prior probability of each class.
- Computes likelihood of features assuming independence.
- Applies Bayes theorem to find posterior probability.
- Assigns class with highest probability.

3️⃣ K-Means Clustering (Experiment 6)

- Choose **k** random centroids.
- Assign each data point to nearest centroid.
- Recalculate new centroids as mean of assigned points.
- Repeat until centroids stop changing.

4️⃣ Agglomerative Hierarchical Clustering (Experiment 7)

- Start with each data point as its **own cluster**.
- Merge the two closest clusters based on distance.
- Continue merging until only one cluster remains or target cluster count is reached.
- Result is shown using a **dendrogram**.

5️⃣ Apriori Algorithm (Experiment 8)

- Find frequent individual items (support ≥ min support).
- Generate larger itemsets by joining smaller ones.
- Remove itemsets whose subsets are not frequent (Apriori property).
- Generate **association rules** based on confidence.
- Output rules like: *If A is bought, B is also likely bought.*

6️⃣ OLAP Operations (Experiment 3)

- **Roll-up:** Summarize (day → month → year).
- **Drill-down:** Go into more detail (year → month).
- **Slice:** Choose a single value of a dimension.
- **Dice:** Select multiple values of dimensions.
- **Pivot:** Rotate view of cube for better visualization.

7️⃣ Data Warehouse (Experiment 1 & 2)

- Collects data from many sources.
- ETL: **Extract → Transform → Load**.
- Stores in **fact tables** (measures) and **dimension tables** (descriptions).
- Used for OLAP analysis.