DATA WAREHOUSING

What is data warehousing?

- **Data Warehousing**:
    - Centralized repository for storing large volumes of structured data.
    - Supports decision-making by integrating data from multiple sources.
    - Enables efficient querying and analysis.
    - Used for business intelligence, reporting, and trend analysis.

What is Metadata?

- **Metadata**:
    - Data about data, providing descriptive information.
    - Includes details like data source, structure, format, and usage.
    - Helps in organizing, locating, and understanding data.
    - Crucial for managing databases, data warehousing, and data mining processes.

What is need of data warehousing?

- **Need for Data Warehousing**:
    - Integrates data from diverse sources for consistency.
    - Facilitates business intelligence, reporting, and decision-making.
    - Improves query performance for large datasets.
    - Enables historical data analysis and trend forecasting.
    - Supports strategic planning and operational efficiency.

What do you mean by data extraction?

- **Data Extraction**:

- Process of retrieving data from various sources like databases, files, or APIs.
- First step in ETL (Extract, Transform, Load) processes.
- Ensures data availability for analysis and warehousing.
- Converts unstructured or semi-structured data into usable formats.

What do you mean by data quality?

- **Data Quality**:
  - Measure of the reliability, accuracy, and completeness of data.
  - Ensures data is consistent, valid, and relevant for analysis.
  - High-quality data supports better decision-making and business outcomes.
  - Key aspects: accuracy, consistency, completeness, timeliness, and integrity.

What is data pre-processing?

- **Data Pre-processing**:
  - Preparation of raw data for analysis by cleaning, transforming, and organizing it.
  - Steps include data cleaning, integration, reduction, and normalization.
  - Eliminates inconsistencies and errors, ensuring data quality.
  - Essential for effective data mining and machine learning.

What do you mean by data cleaning?

- **Data Cleaning**:
  - Process of detecting and correcting errors, inconsistencies, or inaccuracies in data.
  - Includes handling missing values, removing duplicates, and correcting invalid entries.
  - Ensures data is accurate, consistent, and ready for analysis or storage.

What is data reduction?**Data Reduction**:

- Process of minimizing the volume of data while retaining its essential information.
- Techniques include dimensionality reduction, data compression, and aggregation.
- Reduces storage requirements and improves processing efficiency.
- Essential for handling large datasets in data mining and analysis.

Why we use data Mining?

- **Why Use Data Mining**:
  - Extracts meaningful patterns and insights from large datasets.
  - Supports decision-making, trend prediction, and business strategies.
  - Helps identify customer behavior, fraud detection, and market segmentation.
  - Improves efficiency in various fields like healthcare, finance, and marketing.

Explain the differences between Knowledge discovery and data mining.

- **Knowledge Discovery**:
  - Broad process of extracting useful knowledge from data.
  - Involves multiple steps: data cleaning, integration, selection, transformation, mining, and interpretation.
- **Data Mining**:
  - Specific step within knowledge discovery.
  - Focuses on identifying patterns and relationships in data using algorithms.

SECTION-B

11. Write difference between operational database system & data warehouses.

- **Operational Database System**:
    - Supports daily transactions and routine operations.
    - Focuses on current, real-time data.
    - Highly normalized for efficiency in updates.
    - Example: Banking system for deposits/withdrawals.
- **Data Warehouses**:
    - Designed for analysis and decision-making.
    - Stores historical data from multiple sources.
    - Denormalized structure for faster querying.
    - Example: Sales analysis system for trends.

12. Explain data warehouse models & its development process.

- **Data Warehouse Models**:
    1. **Enterprise Warehouse**: Centralized data for the entire organization.
    2. **Data Mart**: Subset of a warehouse for specific departments.
    3. **Virtual Warehouse**: Provides views of data without physical storage.
- **Development Process**:
    1. Define business objectives.
    2. Identify data sources.
    3. Design warehouse schema (star, snowflake, or fact constellation).
    4. Extract, Transform, Load (ETL) data.
    5. Implement and test the warehouse.
    6. Optimize for performance and scalability.

13. Explain physical design process of data warehouse.

- **Physical Design Process of Data Warehouse**:
    1. **Define Storage Structures**: Allocate physical storage for tables, indexes, and metadata.
    2. **Partitioning**: Divide data into segments for efficient access.
    3. **Indexing**: Create indexes to optimize query performance.
    4. **Denormalization**: Use flattened structures for faster queries.
    5. **Materialized Views**: Precompute complex queries for speed.
    6. **ETL Implementation**: Plan extraction, transformation, and loading workflows.
    7. **Performance Optimization**: Fine-tune for load balancing and scalability.

14. What do you mean by data warehouse uses for information processing.

- **Data Warehouse Uses for Information Processing**:
    - **Query and Reporting**: Supports generating detailed and summary reports.
    - **OLAP (Online Analytical Processing)**: Enables multidimensional data analysis.
    - **Decision Support**: Provides historical and integrated data for strategic decisions.
    - **Trend Analysis**: Identifies patterns and forecasts future scenarios.
    - **Business Intelligence**: Enhances insights into business operations.

15. Describe architecture of data mining system.

Architecture of Data Mining System:

1. **Data Sources**: Includes databases, data warehouses, or flat files.
2. **Data Preprocessing**: Cleans, transforms, and integrates data.
3. **Data Mining Engine**: Core system for pattern recognition and analysis.

4. **Pattern Evaluation Module**: Interprets and validates mined patterns.
5. **Knowledge Base**: Stores domain knowledge to guide mining.
6. **User Interface**: Enables interaction for query input and result visualization.

16. Explain in detail about Data mining functionalities?

- **Data Mining Functionalities**:
    1. **Association Rule Mining**: Identifies relationships between items (e.g., market basket analysis).
    2. **Classification**: Categorizes data into predefined classes.
    3. **Clustering**: Groups similar data without predefined labels.
    4. **Regression**: Predicts continuous values based on input data.
    5. **Outlier Detection**: Identifies anomalies in data.
    6. **Trend Analysis**: Discovers patterns over time for forecasting.
    7. **Summarization**: Generates concise representations of data (e.g., reports).

17. Explain the following? a) Data Integration b) Data Transformation methods.

- **a) Data Integration**:
    - Combines data from multiple sources into a unified view.
    - Resolves data conflicts and redundancy.
    - Techniques: ETL process, schema mapping, and data cleaning.
- **b) Data Transformation Methods**:
    - **Smoothing**: Reduces noise (e.g., moving averages).
    - **Normalization**: Scales data to a standard range.
    - **Aggregation**: Summarizes data (e.g., monthly to yearly).
    - **Discretization**: Converts continuous data into categories.

SECTION-C

18. Explain project planning & management of data warehouse. Describe architecture of data warehouse.

Project Planning & Management of Data Warehouse:
Planning involves identifying business objectives, defining scope, and selecting tools. Key steps include analyzing data requirements, estimating costs, and forming a skilled team. Effective management ensures timely ETL processes, data validation, and user training. Milestones and feedback loops are essential for successful implementation.

Architecture of Data Warehouse:
A data warehouse comprises:

1. **Data Sources**: Operational databases, external files, and applications.
2. **ETL Process**: Extracts, transforms, and loads data into the warehouse.
3. **Warehouse Storage**: Centralized repository with star, snowflake, or fact constellation schemas.
4. **Metadata Repository**: Stores information about the data's structure and lineage.
5. **Query and Analysis Tools**: Enable reporting, OLAP, and data mining for decision-making.
6. **User Interface**: Provides access to data through dashboards and reports.

The architecture ensures seamless data flow and analytical efficiency.

19. Explain star, snowflakes and fact constellations schemes for multidimensional data models of data warehouse.

Star Schema:

- Central fact table linked to dimension tables.
- Simple and denormalized structure for fast querying.
- Example: Sales fact table connected to dimensions like time, product, and location.

Snowflake Schema:

- Extension of the star schema with normalized dimensions.
- Dimension tables are split into related sub-tables.

- Reduces redundancy but increases complexity.
- Example: Product dimension split into categories and subcategories.

Fact Constellation Schema:

- Multiple fact tables share dimension tables.
- Supports complex relationships and data analysis.
- Example: Sales and inventory fact tables share dimensions like time and product.

These schemas organize multidimensional data to enhance data retrieval and analysis in warehouses.

20. What is OLAP in data warehouse? Explain typical OLAP operations with diagram.

OLAP in Data Warehouse:
OLAP (Online Analytical Processing) enables multidimensional data analysis for decision-making. It allows users to analyze data interactively by slicing, dicing, drilling, and pivoting across dimensions like time, product, and region. OLAP tools support business intelligence by providing insights into trends and patterns.

Typical OLAP Operations:

1. **Slice**: Filters data on one dimension (e.g., sales in 2025).
2. **Dice**: Creates a sub-cube by selecting multiple dimensions.
3. **Drill Down/Up**: Navigates between detailed and summarized data (e.g., year → month → day).
4. **Pivot**: Rotates the data view to change dimensional perspectives.

Diagram:
A cube with three axes: Time, Product, and Region. Operations like slicing show a single time period, dicing selects a subset (e.g., 2025 sales in India for Electronics), and drill down reveals monthly sales.

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

KDD Process in Data Mining:
KDD (Knowledge Discovery in Databases) is a structured process for extracting meaningful patterns and knowledge from large datasets. It involves several key steps:

1. **Data Selection**: Identifying and retrieving relevant data from the source.
2. **Data Preprocessing**: Cleaning and preparing data by handling missing values and removing noise.
3. **Data Transformation**: Converting data into suitable formats for mining (e.g., normalization).
4. **Data Mining**: Applying algorithms to discover patterns and relationships.
5. **Pattern Evaluation**: Validating and interpreting the patterns for usefulness.
6. **Knowledge Representation**: Presenting the findings in a comprehensible format like reports or charts.

Diagram: A flowchart showing the sequence of steps:
Data Selection → Preprocessing → Transformation → Mining → Evaluation → Representation.

The KDD process transforms raw data into actionable knowledge for decision-making.

22. Explain data generalization & summarization based characterization.

Data Generalization:
Data generalization involves summarizing detailed data by replacing low-level attributes with higher-level concepts using concept hierarchies. For example, transactional dates can be generalized into months or years. This simplifies data for analysis by focusing on broader patterns.

Summarization Based Characterization:
Summarization provides an overview of a dataset by identifying general features or trends. It creates concise descriptions using techniques like aggregation, averages, or histograms. For instance, customer purchase data can be summarized into average spending per month or frequent purchase categories.

Both techniques aim to extract meaningful insights from complex datasets by reducing granularity and emphasizing key characteristics.