Author: Madhurima Rawat

CT 1 Question Bank: Data Warehouse

Unit 1

Question 1: Explain the role of a data warehouse in Business Intelligence.

Solution: Centralized Data Storage

A **data warehouse** plays a crucial role in **Business Intelligence (BI)** by providing a centralized repository for structured data collected from multiple sources. It enables businesses to perform **complex queries, generate reports, and conduct data analysis** to gain meaningful insights. By storing **historical and current data**, a data warehouse helps in identifying trends, making predictions, and improving decision-making processes.

Question 2: What are the key benefits of using a data warehouse in an organization?

Solution: Key Benefits

A data warehouse provides several benefits, including:

- **Improved Data Quality:** Ensures consistency, accuracy, and completeness of data.
- **Faster Query Performance:** Optimized for analytical queries, reducing retrieval time.
- **Enhanced Decision Making:** Supports strategic and operational decisions.
- **Historical Data Storage:** Maintains historical records for trend analysis.
- **Data Integration:** Combines data from multiple sources into a single repository.
- **Scalability:** Can accommodate growing data needs efficiently.

Question 3: Describe the process of data integration in a data warehouse.

Solution: ETL Process

Data integration involves collecting, transforming, and loading data from different sources into a unified database. The process includes:

- **Extract:** Retrieve data from multiple sources (databases, files, applications).
- **Transform:** Cleanse, filter, and structure data for consistency and accuracy.

- **Load (ETL Process):** Store the transformed data into the data warehouse for analysis.
- **Data Refreshing:** Regularly updating data to reflect recent changes.

Question 4: How does a data warehouse support decision-making processes?

Solution: Decision-Making Support

A data warehouse supports decision-making by:

- Providing **timely and accurate information** for business strategy planning.
- Allowing **trend analysis and forecasting** using historical data.
- Enabling **multi-dimensional analysis** through OLAP (Online Analytical Processing).
- Offering **dashboards and reports** for easy visualization of key performance metrics.
- Supporting **predictive analytics** and **data mining** for better insights.

Question 5: Explain the architecture of a data warehouse.

Solution: Data Warehouse Layers

A typical data warehouse architecture consists of:

- **Data Source Layer:** Collects data from multiple heterogeneous sources.
- **ETL (Extract, Transform, Load) Layer:** Extracts, cleans, and loads data into the warehouse.
- **Data Storage Layer:** Centralized repository for processed data.
- **Metadata Layer:** Stores information about the data structure and definitions.
- **OLAP Engine:** Supports multi-dimensional data analysis.
- **Presentation Layer:** Provides reports, dashboards, and visualization tools for end-users.

Question 6: Differentiate OLTP and OLAP.

Solution: OLTP vs. OLAP

Feature	OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)
Purpose	Supports real-time transaction processing.	Used for analytical and business intelligence applications.
Data Type	Operational data (current and detailed).	Historical, aggregated, and multi-dimensional data.
Query Type	Short and simple transactions.	Complex queries for decision-making.

Feature	OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)
Performance	Optimized for speed and high availability.	Optimized for read-intensive operations.
Storage	Uses normalized databases to avoid redundancy.	Uses denormalized databases for faster querying.

Question 7: Write a short note on Data Marts.

Solution: Data Marts Overview

A **Data Mart** is a subset of a data warehouse designed for a specific **department or business function**, such as finance, marketing, or sales. It contains **filtered and relevant data** for a particular user group, improving performance and efficiency. Data marts can be categorized into:

- **Dependent Data Marts:** Derived from a central data warehouse.
- **Independent Data Marts:** Built separately without relying on a data warehouse.

Question 8: What are the challenges faced during data warehousing implementation?

Solution: Implementation Challenges

Implementing a data warehouse presents several challenges:

- **High Initial Cost:** Significant investment in hardware, software, and skilled personnel.
- **Data Integration Complexity:** Consolidating data from multiple sources with different formats.
- **Performance Issues:** Managing large datasets efficiently for fast query processing.
- **Data Security and Privacy:** Ensuring data protection against unauthorized access.
- **Scalability Concerns:** Handling growing volumes of data over time.
- **Change Management:** Adapting to evolving business needs and technology updates.

Unit 2

Question 1: Explain Data Modelling.

Solution: Data Modelling

Data Modelling is the process of defining the **structure, relationships, and constraints** of data within a data warehouse. It involves creating a conceptual, logical, and physical model of the data, ensuring

efficient storage, retrieval, and analysis. The three key types of data models are:

- **Conceptual Model:** High-level overview of data entities and relationships.
- **Logical Model:** Detailed structure using entities, attributes, and relationships.
- **Physical Model:** Implementation of the logical model in a database system.

Question 2: Compare Star Schema, Snowflake Schema, and Fact Constellation Schema with an Example.

Solution: Comparison of Schema Models

Feature	Star Schema	Snowflake Schema	Fact Constellation Schema
Structure	Central fact table linked to dimension tables.	Dimension tables are further normalized.	Multiple fact tables sharing dimension tables.
Normalization	Denormalized structure.	Partially normalized.	Highly normalized.
Performance	Fast query execution.	Slightly slower due to joins.	Complex queries but optimized storage.
Complexity	Simple and easy to use.	Moderate complexity.	High complexity.
Example	Sales fact table linked to Product, Time, and Customer dimensions.	Sales dimension further split into Product Category, Region, etc.	Sales and Inventory fact tables sharing Time and Product dimensions.

Question 3: How Does Data Normalization Affect Data Warehouse Design?

Solution: Data Normalization in Data Warehousing

Data normalization reduces data redundancy and improves data integrity by organizing data into **multiple related tables**. However, in a data warehouse:

- **Too much normalization** increases query complexity and slows performance.
- **A balance is required** to optimize storage while ensuring efficient queries.
- **Star Schema (denormalized) is preferred** for fast query performance, whereas **Snowflake Schema (partially normalized)** is used when space optimization is needed.

Question 4: How Do Fact and Dimension Tables Work Together in a Data Warehouse? Explain with an Example.

Solution: Fact and Dimension Tables

Fact Tables store measurable business data (e.g., sales amount, quantity sold).
Dimension Tables store descriptive attributes (e.g., product name, customer location).

Relationship: Fact tables contain **foreign keys referencing dimension tables**, allowing data to be analyzed from different perspectives.

Example:

- **Fact Table:** Sales (Sales_ID, Product_ID, Customer_ID, Sales_Amount).
- **Dimension Tables:**
  - Product (Product_ID, Product_Name, Category).
  - Customer (Customer_ID, Name, Location).

This setup allows queries like **"Total sales by product category"** using a **JOIN between the fact and dimension tables**.

Question 5: What is the Role of Metadata in a Data Warehouse?

Solution: Metadata in Data Warehousing

Metadata is **"data about data"**, describing the structure, source, and usage of data in a warehouse. It is categorized into:

- **Technical Metadata:** Data types, indexes, table structures.
- **Business Metadata:** Meaning, purpose, and relationships between data.
- **Operational Metadata:** Data refresh frequency, ETL process details.

Metadata ensures **data consistency, easier maintenance, and better governance** in a data warehouse.

Question 6: Discuss the Challenges and Strategies for Handling Slowly Changing Dimensions (SCDs).

Solution: Slowly Changing Dimensions (SCDs)

Slowly Changing Dimensions (SCDs) refer to **dimension data that changes over time**, such as customer addresses. There are three main types of SCDs:

- **Type 1 (Overwrite):** Updates old data with new data, losing history.
- **Type 2 (Versioning):** Adds a new row for each change, preserving history.
- **Type 3 (Column Update):** Adds a column to store old and new values, limited history tracking.

Question 7: Explain the Concept of Data Granularity in a Data Warehouse.

Solution: Data Granularity

Data Granularity refers to the **level of detail** stored in a data warehouse.

- **High Granularity:** Detailed data (e.g., individual transactions).
- **Low Granularity:** Aggregated data (e.g., monthly or yearly sales summaries).
- **Trade-off:** Storing highly granular data increases storage but provides flexibility in analysis, whereas low granularity improves performance but reduces detail.

Question 8: What are the Best Practices for Designing a Data Warehouse?

Solution: Best Practices

- ✅ **Define Clear Objectives:** Align the warehouse with business goals.
- ✅ **Choose the Right Schema:** Star or Snowflake schema based on query needs.
- ✅ **Optimize ETL Processes:** Ensure smooth data extraction, transformation, and loading.
- ✅ **Ensure Data Quality:** Implement validation checks and data cleansing.
- ✅ **Manage Indexing and Partitioning:** Improve query performance.
- ✅ **Implement Security Measures:** Control access and protect sensitive data.
- ✅ **Use Incremental Updates:** Avoid reloading the entire warehouse frequently.
- ✅ **Monitor Performance Regularly:** Optimize query execution and storage usage.

Unit 3

Question 1: Write and explain the stepwise ETL process.

Solution: Stepwise ETL Process

ETL (Extract, Transform, Load) is a process used in data warehousing to extract data from various sources, transform it into a suitable format, and load it into a data warehouse. The steps involved are:

- **Extraction:** Data is extracted from multiple sources like databases, APIs, flat files, and cloud storage.
- **Transformation:** The extracted data is cleaned, validated, standardized, aggregated, and formatted to meet business requirements.

- **Loading:** The transformed data is loaded into the target data warehouse for analysis and reporting.
- **Data Validation & Quality Check:** Ensures accuracy and completeness before final storage.
- **Monitoring & Maintenance:** Continuous monitoring for data consistency and performance optimization.

Question 2: List down the advantages and disadvantages of different data extraction techniques.

Solution: Advantages and Disadvantages of Data Extraction Techniques

Extraction Technique	Advantages	Disadvantages
Full Extraction	Simple to implement, ensures complete data retrieval	Time-consuming, high processing cost
Incremental Extraction	Faster processing, reduced load on source systems	Complexity in tracking changes, requires additional metadata
Log-based Extraction	Real-time data updates, minimizes impact on source systems	Requires access to database logs, complex implementation

Question 3: How does data transformation impact the quality of data in a data warehouse?

Solution: Impact of Data Transformation on Data Quality

- **Standardization:** Converts data into a uniform format, reducing inconsistencies.
- **Data Cleaning:** Eliminates duplicates, errors, and missing values, improving accuracy.
- **Aggregation & Integration:** Combines data from multiple sources for comprehensive analysis.
- **Normalization & Denormalization:** Optimizes storage and query performance.
- **Validation Rules:** Ensures data integrity and compliance with business rules.

Proper transformation improves data usability, accuracy, and reliability for decision-making.

Question 4: What are the common ETL tools used in data warehousing?

Solution: Common ETL Tools

- **Informatica PowerCenter** – Enterprise-grade ETL tool with extensive transformation capabilities.
- **Talend Open Studio** – Open-source ETL tool with strong data integration features.
- **Apache Nifi** – Real-time data processing and integration tool.

- **Microsoft SQL Server Integration Services (SSIS)** – ETL tool for Microsoft environments.
- **IBM DataStage** – High-performance ETL tool for large-scale data integration.
- **Pentaho Data Integration (PDI)** – Open-source ETL tool for data warehousing.

Question 5: Discuss the role of data staging in the ETL process.

Solution: Role of Data Staging in ETL

- **Data Cleansing:** Prepares raw data for transformation by removing inconsistencies.
- **Performance Optimization:** Reduces processing load on source systems.
- **Data Integration:** Allows merging of data from multiple sources before transformation.
- **Backup & Recovery:** Stores intermediate data to prevent data loss in case of failures.

Data staging enhances ETL efficiency and reliability in large-scale data processing.

Question 6: Explain different types of immediate data extraction techniques.

Solution: Immediate Data Extraction Techniques

- **Change Data Capture (CDC):** Identifies and extracts only modified data from source systems.
- **Log-Based Extraction:** Reads database transaction logs to capture real-time updates.
- **Trigger-Based Extraction:** Uses database triggers to capture data changes automatically.
- **Event-Driven Extraction:** Extracts data based on predefined business events.

These techniques enable real-time analytics and reduce the overhead of full data extraction.

Question 7: How can data validation be ensured during the ETL process?

Solution: Ensuring Data Validation in ETL

- **Schema Validation:** Ensures data matches the expected structure.
- **Data Type Checks:** Confirms that data types are correct for each field.
- **Uniqueness Constraints:** Identifies duplicate records.
- **Range Checks:** Validates numeric or date values within acceptable limits.
- **Referential Integrity Checks:** Ensures relationships between tables are maintained.
- **Null Value Handling:** Identifies and fills missing values appropriately.

Implementing validation rules enhances data quality and prevents errors in the data warehouse.