Assignment No. 2 (SPRING - 2025)

CT – 528 – Advanced Database Techniques
Program: MS (CS)
Instructor: Dr. Syed Saood Zia
Name: Shujaatmallick
Name: DS W-23

Question No. 1

a. Difference between Data Warehouse and Data Mart

A **data warehouse** is a centralized repository that aggregates data from multiple sources across an entire organization. It supports complex queries, analytics, and historical data analysis.

A **data mart**, on the other hand, is a subset of a data warehouse focused on a specific business line or department such as sales or inventory.

Benefits to StoreMart:

- A **data warehouse** will provide StoreMart with a centralized, consistent, and integrated view of enterprise-wide data.
- **Data marts** will allow departments like sales and inventory to access relevant, customized, and faster data for analysis and decision-making.

b. High-Level Architecture for StoreMart's Data Warehouse

1. **Data Sources**:
   - POS systems (Sales)
   - Inventory Management Systems
   - CRM (Customer Data)
   - SCM systems
   - External market data (e.g., competitor prices)

2. **ETL Process**:
    - **Extract**: Pull data from heterogeneous systems
    - **Transform**: Cleanse, validate, and conform to business rules
    - **Load**: Insert into the central data warehouse
3. **Data Warehouse Structure**:
    - **Staging Area**: Temporary storage for raw extracted data
    - **ODS (Operational Data Store)**: For real-time or near-time data
    - **Data Warehouse**: Centralized, historical, subject-oriented
    - **Data Marts**: Sales and Inventory marts derived from the warehouse
    - **BI Tools**: Dashboards, reporting, OLAP tools

c. Sales Data Mart Design

Fact Table: `Fact_Sales`

- Facts: Total_Sales, Quantity_Sold, Discount, Profit

Dimension Tables:

- `Dim_Date`: Date_ID, Day, Month, Quarter, Year
- `Dim_Store`: Store_ID, Location, Manager
- `Dim_Product`: Product_ID, Category, Brand
- `Dim_Customer`: Customer_ID, Name, Segment

Interaction with Data Warehouse:

- The sales data mart extracts its data from the central warehouse using periodic ETL jobs, enabling focused and efficient reporting for the sales team.

d. Challenges in Data Integration & Ensuring Quality

Challenges:

- Data inconsistency across sources
- Missing or duplicate data
- Incompatible data formats
- Real-time integration complexities

Solutions:

- Implement robust **data quality checks** during ETL
- Use **metadata management** to track data lineage
- Apply **master data management (MDM)** for consistency
- Use data profiling and cleansing tools
- Implement **audit trails and logging**

Question No. 2

a. Explanation of Star Schema

A **star schema** is a type of data warehouse schema that consists of a central **fact table** connected to multiple **dimension tables** in a star-like formation.

Advantages:

- Simplified structure for end users
- Faster query performance due to fewer joins
- Ideal for OLAP and BI tools

b. Star Schema Design for ConnectTel

Fact Table: `Fact_Calls`

- Primary Key: Call_ID
- Foreign Keys: Time_ID, Customer_ID, Location_ID
- Measures: Call_Duration, Call_Cost, Satisfaction_Score

Dimension Tables:

1. `Dim_Time`
   - ○ Time_ID (PK)
   - ○ Date, Day, Week, Month, Year
2. `Dim_Customer`
   - ○ Customer_ID (PK)
   - ○ Name, Age, Gender, Income_Level, Subscription_Type
3. `Dim_Location`
   - ○ Location_ID (PK)
   - ○ City, State, Country

Relationships:

- `Fact_Calls` references `Dim_Time`, `Dim_Customer`, and `Dim_Location` via foreign keys.

c. Query Performance Optimization

Indexing Strategies:

- Use **bitmap indexes** on foreign keys for low-cardinality dimensions.
- Create **clustered indexes** on date columns for time-series queries.

- Apply **materialized views** for common aggregations.
- Partition fact table by time or location.

Other optimizations:

- Use **columnar storage** if supported
- Enable **parallel processing** in ETL and queries

d. Limitations of Star Schema

- Not ideal for complex many-to-many relationships
- Redundancy in denormalized dimension tables
- Lacks support for slowly changing dimensions (SCDs)

When to use other schemas:

- Use **snowflake schema** for normalized dimension tables and reduced data redundancy.
- Use **galaxy schema** (fact constellation) when multiple fact tables share dimension tables.