**Redshift Data Warehouse Transformation Project**

**Objective:**

Students will develop skills in data cleaning, transformation, and schema design by organizing raw, fragmented data into a cohesive and valid data warehouse schema suitable for Redshift. This project focuses on a fictional large clothing company that offers both discount and luxury brands of men's and women's clothing and accessories.

**Deliverables:**

Data Preparation: Students must clean, deduplicate, and standardize fragmented and unformatted data from multiple CSV files.

Data Warehouse Schema Design: Students will design a schema appropriate for a data warehouse (e.g., star or snowflake schema).

Redshift Implementation: Students will load the cleaned and transformed data into Redshift, creating tables that adhere to their schema design.

Analysis: Students will write sample SQL queries to demonstrate the usability and insights of their schema.

**Materials Provided:**

Raw input CSV files (detailed below)

**Assignment Description:**

The fictional clothing company operates globally, with distinct product lines, organizational structures, and customer demographics. Students are tasked with cleaning, merging, and organizing fragmented data into a valid schema for analysis in Redshift. The provided raw files contain issues such as missing values, duplicates, and unclear relationships between data points.

**Input Data Description:**

The raw data is divided into multiple CSV files with unclear names and fragmented contents, simulating a real-world scenario where data is unorganized and requires significant preprocessing. The files include:

**File A** (prod\_info\_part1.csv, discount\_data.csv, etc.)

Fragmented product data across several files.

Columns may include: ProductID, ProductName, Brand, Category, Gender, Price, DiscountedPrice, WarehouseLocation, StockQuantity, DateAdded.

**Issues:** Duplicates, incomplete rows, inconsistent category naming, split across multiple files.

**File B** (emps\_partial.csv, salaries\_data.csv, etc.)

Employee data spread across several files.

Columns may include: EmployeeID, Name, Position, Department, Salary, HireDate, Email, StoreID.

Issues: Inconsistent naming conventions, duplicates, and mismatched data between files.

**File C** (loyalty\_program.csv, cust\_data.csv, etc.)

Customer data distributed across files.

Columns may include: CustomerID, Name, Email, Phone, Gender, LoyaltyPoints, FavoriteCategory, FavoriteStore, DOB.

Issues: Missing values, overlapping records, ambiguous relationships.

**File D** (stores.csv, locations.csv, etc.)

Store location data in multiple files.

Columns may include: StoreID, StoreName, Address, City, State, Country, Region, OpeningDate, ManagerID.

Issues: Inconsistent state abbreviations, duplicates, incomplete rows.

**File E** (sales\_data\_part1.csv, transactions.csv, etc.)

Sales and transaction data split across files.

Columns may include: TransactionID, CustomerID, StoreID, ProductID, Date, Quantity, TotalAmount.

Issues: Missing relationships, mismatched CustomerID for anonymous purchases.

**File F** (web\_activity.csv, page\_views.csv, etc.)

Website activity data spread across files.

Columns may include: SessionID, CustomerID, PageViewed, Duration, Date, Device, Referrer.

Issues: Missing CustomerID for non-logged-in users, overlapping session data.

Instructions for Students:

**Data Cleaning and Preparation:**

Explore the fragmented raw data files.

Identify and resolve duplicates, inconsistencies, and missing values.

Merge related data into coherent datasets.

**Schema Design:**

Analyze the cleaned data and design an optimized data warehouse schema (e.g., star or snowflake schema).

Provide an ERD (Entity-Relationship Diagram) illustrating the design.

**Data Loading:**

Load the prepared data into Redshift using appropriate tools (e.g., COPY command).

Validate that the data adheres to the designed schema.

**SQL Analysis:**

Write and execute queries to extract meaningful insights (e.g., sales trends, customer demographics, product performance).

**Evaluation Criteria:**

Completeness and quality of cleaned data.

Logical and efficient schema design.

Successful implementation in Redshift.

Insightful SQL queries and analysis.

Clarity and detail in the final presentation.

Tables that we want as output would include things such as Fact\_sales, dim\_orders, dim\_customers, dim\_stores, dim\_products, and so on, for all of the different types of data.

You can organize the data into either a star schema or a snowflake schema