Class Project Guidelines for ETL & Data Warehousing

Objective:

The purpose of this class project is to provide students with hands-on experience in Extract, Transform, and Load (ETL) processes, data warehousing, and working with large datasets. Students will work with a dataset of their choice to design and implement a structured data warehouse. The project will focus on building a **star schema** with at least **one fact table** and a minimum of **four dimensional tables** to support analytical queries.

Project Scope:

Students are responsible for **finding their own dataset** that contains structured data suitable for an ETL and data warehousing project. The dataset should have multiple related tables and enough complexity to allow meaningful analysis. Recommended sources for datasets include:

- Kaggle
- Google Dataset Search
- Data.gov
- UCI Machine Learning Repository
- AWS Public Datasets

Students should select a dataset that allows for the analysis of trends, patterns, or performance indicators related to business, finance, healthcare, retail, or another domain of interest. The dataset should include transactional or event-based data that can be structured into a fact and dimension model.

Key Goals:

1. **Data Understanding and Exploration:**
    - Explore the dataset to understand its structure, relationships, and potential insights.
    - Identify key attributes to be used for fact and dimension tables.
2. **Data Extraction:**
    - Create an ETL pipeline using **SSIS** to extract relevant data from different sources.
    - Ensure the extracted data is in a suitable format for transformation and analysis.
3. **Data Transformation:**
    - Perform necessary transformations, such as data cleaning, handling missing values, and creating new attributes where appropriate.
    - Design transformations that prepare data for analytical processing.

4. **Data Loading & Data Warehouse Schema:**
    - Build a **star schema** with at least **one fact table** and **four dimensional tables** to support efficient querying and reporting.
    - Load the transformed data into the structured data warehouse.
5. **Data Analysis & Insights:**
    - Based on the prepared data, answer specific analytical questions, such as:
        - What factors influence key performance metrics?
        - How do different attributes (e.g., payment methods, customer location, product categories) impact business outcomes?
        - What trends or patterns can be identified from the data?

Project Deliverables:

- A complete **ETL solution**
- **SQL Scripts** to recreate all tables in the data warehouse
- A **Word document** detailing:
    - The **selected dataset** and rationale for its choice and the Data Dictionary
    - The **star schema design** (fact and dimension tables)
    - The **ETL process and key transformations**
    - **Insights and conclusions** drawn from the data