SQL Data Warehouse Project Requirments

# 1. General Project Overview

* Implement an **ETL pipeline** using **SQL Server**.
* Data sources: **ERP and CRM systems** (each providing multiple CSV files).
* **Full load only** (no historical data management).
* Extensive **documentation** is required at every stage.

# 2. Data Sources

* **CRM System:** Provides sales related CSV files (sales, customer\_info, product\_info).
* **ERP System:** Provides additional information on customer and product in CSV files.
* **CSV File Characteristics:**
  + Delimited format : comma
  + Headers included.
  + No schema enforcement at the source

# 3. Data Model for Analysis & Reporting

* **Objective:** Provide a simple, **integrated schema** to support business intelligence and reporting.
* **Approach:**
  + Design a **star schema** or **denormalized tables** for easy querying.
  + Integrate **ERP and CRM data** based on business keys (e.g., customer ID, product ID).
  + Ensure **optimized performance** for analytics (indexing, pre-aggregated data if needed).

# 4. Infrastructure & Technology Stack

* **Database:** Microsoft SQL Server.
* **Storage:** CSV files stored on a file system or cloud storage (if applicable).
* **Automation & Scheduling (if needed):**
  + SQL Server Agent for scheduled execution.
  + PowerShell or Python for auxiliary automation.

# 5. Documentation

* **Data Dictionary:** Definition of all tables, columns, and data types.
* **ETL Pipeline Documentation:**
  + Step-by-step extraction process.
  + Transformation rules applied.
  + Load logic and schema design.
* **Error Handling & Logging Strategy:**
  + Log file storage location.
  + Handling of data failures (invalid records, missing data).
* **Performance Considerations:**
  + Bulk loading techniques for efficiency.

# 5. Deliverables

✅ Fully operational **ETL pipeline** in SQL Server.  
✅ Provide a simple, **integrated schema** to support business intelligence and reporting.

✅ **Documented** data sources, transformations, and database schema.  
✅ **Performance optimizations** for data processing efficiency.  
✅ **Error handling strategy** for data consistency and reliability.