**Feedback Data Warehouse & Sentiment Analysis Pipeline**

**Team Members:**

* **Aly Hesham** ETL processes and Prefect integration
* **Mohamed Walid** - Database Schema Design
* **Mohamed Osama** - Data Preprocessing
* **Zeina Ahmed** - Sentiment Analysis Model
* **Ziad Tarek** - Visualization

**1. Introduction**

This project involves developing an ETL pipeline to process customer feedback data, transform it, and load it into a data warehouse. The data warehouse uses a star schema design to facilitate complex querying for data analysis and business insights. The pipeline also incorporates a machine learning model to analyze customer sentiments.

**2. Data Pipeline Overview**

The data pipeline consists of four main steps:

* **Data Extraction and Loading:** Raw feedback data is extracted from CSV files and loaded into a raw database using Python and PyODBC.
* **Data Transformation and Warehouse Loading:** The extracted data is cleaned, transformed, and loaded into a data warehouse using a star schema.
* **Sentiment Analysis:** A logistic regression model performs sentiment classification on the review texts to predict sentiments.
* **Data Export and Visualization:** The processed data and model predictions are exported to CSV files for further analysis and visualization.

**3. ETL Process Details**

* **Extraction:** Data from CSV is imported into the raw database, including tables like Customer, Feedback, and Category.
* **Transformation:** Data is cleaned to address missing values, outliers, and incorrect formats. Dimensional tables (CustomerDim, ProductDim, etc.) and the fact table (FeedbackFact) are populated.
* **Loading:** The transformed data is loaded into the star schema-based data warehouse to support data analysis and reporting.
* **Star Schema Choice:** We used a star schema to simplify complex queries and optimize the performance of data aggregation operations.

**4. Database Schema Design**

* **Fact Table:** FeedbackFact stores transactional data with references to dimensions.
* **Dimension Tables:** Include CustomerDim, CategoryDim, DateDim, OrderDim, ProductDim, and ReviewDim, providing descriptive data.

**5. Data Preprocessing**

* **Cleaning Steps:** Removed duplicates, handled missing values, and validated email and phone formats.
* **Feature Engineering:** Added columns like Year, Month, Quarter for temporal analysis.

**6. Sentiment Analysis Model**

* **Model Choice:** Logistic Regression, trained on TF-IDF features of review texts.
* **Performance Evaluation:** Classification report and confusion matrix were used for model evaluation.
* **Result Mapping:** Sentiments are mapped to numeric values: positive (1), neutral (0), negative (-1).

**7. Visualization and Analysis**

* **Data Export:** Data is exported as CSV files for external visualization tools (e.g., Power BI).