USN NUMBER: 1RVU23CSE264
NAME: Meghana G

Ex No: 2 Date: 14/8/25	Building a Basic ETL Data Pipeline Using Python

Objective:

This lab experiment provides practical experience in building a basic ETL (Extract, Transform, Load) data pipeline using Python. It guides participants through the core stages of the pipeline from extracting raw data, transforming it into a usable format, and loading it into a target system. The activity also simulates the roles and responsibilities of key stakeholders, such as data engineers, data scientists, and business analysts, to demonstrate their collaborative contributions in the data pipeline.

Outcomes:

1. Identify and describe the stages of the data engineering lifecycle.
2. Explain the roles and responsibilities of different stakeholders at each stage.
3. Perform basic data engineering tasks within a simulated environment.
4. Collaborate across simulated stakeholder roles to design and implement a data-driven solution.

Materials

A pre-packaged CSV file representing raw, messy data (e.g., "sales_data_raw.csv"). This file should contain missing values, incorrect data types, and inconsistent formatting.

A pre-packaged JSON file representing a different data source (e.g., "customer_feedback.json").

A simple, mock "data warehouse" or "data lake" environment (can be a designated folder structure).

Lab Procedure

Stage 1: Problem Definition and Requirements Gathering (Business Analyst)

Business Analyst Task:

1. Review the provided `sales_data_raw.csv` and `customer_feedback.json`.

2. Formulate a business question based on this data. For example: "What are the top 5 products by revenue in the last quarter, and how does customer sentiment vary for these products?"
3. Write down the specific data points needed to answer this question (e.g., `product_id`, `sale_price`, `sale_date`, `customer_id`, `sentiment_score`).
4. Create a simple requirements document outlining the business problem and the desired outcome (e.g., a final report with the top products and their average sentiment).

Stage 2: Data Ingestion and Cleansing (Data Engineer)

Data Engineer Task:

1. **Ingestion:** Read the `sales_data_raw.csv` and `customer_feedback.json` files using `pandas`.
2. **Cleansing:**
   Handle missing values in the `sales_data_raw.csv` (e.g., impute with a default value or drop rows).
   Correct incorrect data types (e.g., ensure `sale_price` is a numerical type).
   Address inconsistent formatting (e.g., standardize date formats).
3. **Transformation:**
   Calculate a new column, `total_revenue`, by multiplying `quantity` and `sale_price`.
   Join the cleaned `sales_data` with the `customer_feedback` data on a common key (`customer_id` or `product_id`).
4. **Loading:** Store the final, cleaned, and transformed data in a new file (e.g., `processed_sales_data.csv`) within the mock "data warehouse" folder.

Stage 3: Data Analysis Task:

1. **Access Data:** Retrieve the `processed_sales_data.csv` file from the "data warehouse" folder.
2. **Analyse:**
   Group the data by `product_id` and sum the `total_revenue` to find the top-performing products.
   Calculate the average `sentiment_score` for each of the top products.
3. **Communicate:** Create a simple table or a bar chart (using a library like `matplotlib` or just text-based) that visualizes the top 5 products and their corresponding average sentiment scores.
4. **Provide Feedback:** Based on the analysis, identify any data quality issues or new data points that would be helpful for future analysis and communicate these back to the Data Engineer.

Stage 4: Reporting and Business Insights (Business Analyst)

USN NUMBER: 1RVU23CSE264
NAME: Meghana G

Business Analyst Task:

1. **Review:** Examine the analysis and visualization provided by the Data Analyst.
2. **Interpret:** Translate the technical findings into business-relevant insights. For example: "Product X is our highest revenue generator, but it has a lower-than-average customer sentiment score. This indicates a potential quality issue or a need for improved customer support."
3. **Report:** Write a concise final report summarizing the initial business question, the key findings from the data analysis, and actionable recommendations for the business.

ML engineer Task

In this lab exercise, the objective is to classify VIP customers based on their purchasing behavior and update this information back into the raw data using a reverse ETL process. Begin by analyzing the raw dataset to identify relevant features such as customer ID, total purchase amount, purchase frequency, and average transaction value. Preprocess the data to handle missing values and normalize features as needed. Use a machine learning model such as logistic regression or clustering (e.g., K-Means) to label customers as VIP or non-VIP. Once the classification is complete, enrich the original dataset by adding a new column indicating the VIP status. Finally, use a reverse ETL approach to export the enriched dataset into a destination such as a CSV file, database, or customer relationship management (CRM) tool, ensuring the updated VIP tags can be used by business teams for personalized offers and targeted marketing.

GitHub Link: https://github.com/meghana1653/Data-Engineering