Data Engineer Assignment

Objective:

The goal of this assignment is to design and implement a scalable ETL pipeline and work with databases.

Prerequisites

1. PostgreSQL database set up on your local machine or a cloud instance.
2. Python environment with necessary libraries installed (Pandas, SQLAlchemy, Metaflow, etc.).
3. GitHub account for code repository.

Assignment Overview

You will design and implement a scalable ETL pipeline using a publicly available dataset. The pipeline will involve data ingestion, transformation, and loading into a PostgreSQL database. Additionally, you will use Python for data processing and Metaflow to manage the ETL workflow.

Tasks

Step 1: Data Ingestion and Storage

1. **Dataset Selection:** Use the Airbnb New York City dataset from Kaggle.
2. **Database Setup:** Set up a PostgreSQL database.
3. **Data Loading:**
    a. Write a script to load the dataset into a PostgreSQL table.
    b. Ensure the table schema is designed to handle the dataset efficiently.

Step 2: ETL Process

1. **Data Extraction:** Extract the data from the PostgreSQL database using SQLAlchemy or a similar library.
2. **Data Transformation:**
    a. Normalize the data (e.g., separate the date and time into different columns).
    b. Calculate additional metrics (e.g., average price per neighborhood).
    c. Handle missing values appropriately (e.g., fill, remove, or flag them).
3. **Data Loading:** Load the transformed data into a new table in the PostgreSQL database.

Step 3: Workflow Management with Metaflow

Workflow Implementation:

1. **Use Metaflow to manage the ETL workflow.**
2. **Implement steps in Metaflow** to handle the ETL process from data ingestion to loading the transformed data into the PostgreSQL database.
3. **Ensure the workflow is reproducible and can handle failures gracefully.**

Deliverables

1. **GitHub Repository:**
    a. All code related to the assignment.
    b. Well-organized repository with a clear directory structure.
    c. README file with instructions on how to set up and run the project.
2. **Documentation:**
    a. Detailed explanation of the ETL process and data transformations.
    b. Instructions for running the Metaflow workflow.
3. **Demonstration:**
    a. Short video or series of screenshots demonstrating the working ETL pipeline.

Submission Guidelines

1. **GitHub Repository Link:** Submit the link to your GitHub repository.
2. **Additional Documentation:** Include any additional documentation or demonstration videos/screenshots in the GitHub repository or provide separate links if needed.

Evaluation Criteria

1. **Code Quality and Organization:**
    a. Clean, readable, and well-documented code.
    b. Proper use of version control with meaningful commit messages.
2. **Functionality:**
    a. Correct implementation of the ETL pipeline and data processing tasks.
3. **Scalability and Performance:**
    a. Design and implementation of scalable solutions.
    b. Performance optimization for database queries and data processing tasks.
4. **Documentation and Demonstration:**
    a. Clear and comprehensive documentation.
    b. Effective demonstration of the implemented tasks.

References/Resources

- Metaflow - Metaflow Docs
- Dataset - New York City Airbnb Open Data | Kaggle
- Metaflow - GitHub - ashishtele/MetaFlow_MLOps: End to end example of Metaflow and Prefect pipelines (Python)

We look forward to seeing your innovative solutions and technical prowess. Good luck!