Data Engineer Test

Overview

You're aiming to create a data pipeline that involves generating data using the sampledata.py script and loading it into a PostgreSQL database. The pipeline will be orchestrated using technologies such as Airflow, Mage.ai, Kubeflow, and MLflow. The entire project will be containerized using Docker. You have 5 days to complete the test, which counts from the day HR sends you the test.

Deliverables

You've outlined several deliverables for this project:

1. Docker-Compose:
   - The project should include Docker Compose configurations for running orchestrating pipeline tools such as Airflow/Mage.ai/Kubeflow/MLflow
   - Database must be PostgreSQL run as a container.
2. Usable Data Pipeline:
   - The data pipeline should be functional and capable of completing its tasks within 30 minutes. This implies efficient data processing and loading.
3. Video Record:
   - A video recording demonstrates the testing of the data pipeline. This could involve running the pipeline and showcasing its functionality.
4. Database Design Diagram:
   - A diagram illustrating the design of the PostgreSQL database schema. This helps others understand the structure of the data being loaded.
5. Readme.md:
   - A comprehensive Readme file provides explanations on how to set up and run the entire project. It should include instructions for running Docker containers, setting up the data pipeline, and any other relevant information.

Preparing Data

You've mentioned that the initial step involves generating data using the Sampledata.py script, which is located in the data_sample folder. This script likely generates the data that you'll later load into the PostgreSQL database.

To generate the data, you would run the following command:

```
pip install -r requirements.txt
python sampledata.py
```

This step is crucial for the testing process and to ensure that the pipeline functions as intended.

Test Sending

Please upload the test result and video in the shared drive, and then send the link back to HR. Kindly make sure to set the sharing settings to "public" so that our team can review.