Data Engineering W24 - Milestone 4 Description

Deadline Monday 30th December at 11:59 pm. Presentation time to be decided

IN THIS MILESTONE YOU WILL BE WORKING IN TEAMS OF 4

1 Objectives:

For this milestone you are required to

1. Orchestrate the tasks performed in milestone 1 and 2 using Airflow in Docker. The tasks were as follows:

   Read csv file $\gg$ clean and transform $\gg$ load to csv(both the cleaned dataset and the lookup table) $\gg$ load both csv files to postgres database as 2 separate tables.

2. Create a dashboard in web to present the data using (dash package in Python). Your dashboard should preview atleast 3 graphs that are properly labeled and represented (You are free to reuse the graphs created in milestone 1).

3. Make a presentation to present your work.

2 Airflow

1. Start off by following the instructions in lab 10 on how to work with airflow in docker.
2. Create directory for milestone called DE_Milestone4
3. Under this dir create the directories needed following same hierarchy as lab 10.
4. Inside the same directory, set airflow uid to be same as localhost id and save this variable in .env file (this file is read when airflow-init runs) : `"echo -e "AIRFLOW\_UID=\$(id -u)" > .env"`
5. To create your dags, you should:
   - Create an airflow python script, where you initialize the dag.
   - Use the python scripts you created in milestone 2 and call the appropriate functions for each task in your airflow file
   - You will also create a new function for creating a dashboard.
6. You should have 2 docker-compose files for this milestone (placed in different folders).
   - 1 for the PostgreSQL image (that holds the tables in a PostgreSQL database)
   - 1 for airflow application (where you will execute your dag).
7. You will need to connect both yaml files through a network as shown in lab 10.

3 Presentation

Here are some key points to guide you in the presentation:

1. Identify a use case or a real-life problem that you can solve using the fin-tech dataset and our project in general (e.g.enhancing decision-making for loan approvals, improving customer segmentation).
2. Dataset
   - Briefly describe the dataset
   - Key findings from the analysis and how they relate to the problem.
   - Use graphs and diagrams from the dashboard to help you illustrate.
3. Your proposed solution
   - Discuss your solution
   - Discuss the technical implementation
   - Include diagrams for the pipeline
4. Business Strategy
   - Explain how your project could be turned into a business
   - Include potential users (ex: Banks), value propositions, and scalability.

4 Deliverables and submission guidelines

4.1 Deliverables

For this milestone each team will need to submit

1. Zipped folder named ”Milestone4” containing the following:
   (a) Folder named ”Airflow” containing all files required to run docker-compose up.
      i. Docker-compose to run airflow
      ii. Additional dockerfile created,
      iii. requirements.txt (the dependencies for your dag),
      iv. All folders mounted from your host (ex: dag, data, ... etc).
      v. Similar to milestone 2, DO NOT include your cleaned dataset.
   (b) Folder named ”Database” containing the docker-compose to run the database for your data
   (c) Folder named ”Dashboard” containing screenshots from your dashboard
2. A 2-3 minute video of your pipeline being run, starting from running ’docker compose up’ for both yaml files upto running the dag and showing the dashboard.
3. The presentation slides

4.2 Submission

Upload your deliverables to a google drive and submit the link to the drive through this Form .

MAKE SURE YOUR DRIVE IS ACCESSIBLE!

5 Weight Distribution

1. Airflow working properly with scripts, mounting and file hierarchy (30%)
2. Dashboards: Insightful representation with correct labels and meaningful outputs and relations. (20%)
3. Presentation (50%)
    - Problem Identification (20%):
      Creativity and significance (scope and impact) of the chosen real-world problem.
    - Data Utilization and Insights (30%):
      Depth of analysis and meaningful insights derived from the dataset.
      Relevance of findings in addressing the problem.
    - Proposed Solution (10%):
      Clear, practical and innovative solution that matches the insights from data, to tackle the problem.
    - Technical Explanation (15%):
      Clear discussion of the technical aspects of the solution.
      Explanation of tools, methodologies, and pipeline implementation.
    - Business Aspect and Feasibility (20%):
      Viability of transforming the project into a business solution considering target users, scalability, value, ... etc.
    - Visual Representation (5%):
      Using dashboards, graphs, and diagrams in presentation
4. ***BONUS* (5%): Be creative**

Best of luck!