Data Engineering Project
Milestone 3

Deadline on Tuesday,3rd of January.

Description -

For this milestone you are required to orchestrate the tasks performed in milestone 1 and 2 using Airflow in Docker. The tasks you have performed in milestone 1 and 2 were as follows.

Read csv file >> clean and transform >> load to csv(both the cleaned dataset and the lookup table) >> extract additional resources >> integrate with the cleaned dataset and load back to csv.

In addition to orchestrating these tasks in Airflow. You will perform 2 additional tasks. a)load both csv files(lookup and cleaned dataset) to postgres database as 2 separate tables. b) create a dashboard and present it in a web interface (using dash package in Python).The tables you create and populate in Postgres **MUST** be named as follows: UK_Accidents_YearNumber and lookup_table. Replace YearNumber with the year of the dataset,, i.e if your dataset is 2000. Your postgres database should have 2 tables. UK_Accidents_2000 and lookup_table.

You will be penalized if your table names are not named per the instructions.

Therefore your workflow(DAG) should ideally be as follows. Read, transform and load to csv (t1, milestone 1) >> extract additional resources(t2) >> integrate and load to both postgres and csv)t3) >> create dashboard (t4). Your dashboard should preview at least 5 graphs that are properly labeled and represented (You are free to reuse the graphs created in milestone 1).

Important notes

- You are more than free to create your DAG as you see fit as long as all the tasks aforementioned have been performed. However, your DAG should orchestrate the tasks and NOT your data flow. In other words, you should keep your data flow encapsulated within a single task and not orchestrate the data transformation steps. For instance, note that I have placed all of milestone 1 in 1 task and have not separated the tasks (cleaning and transformation) into separate tasks in my DAG as this is simply a flow of data and not tasks (such as extracting from 3rd party, loading to a database,loading to a cloud service,etc).

- Regarding milestone 2, those who used an API key special to them or have made a very large number of API calls could just store the extracted data as a csv file in the dag and simply just read the csv file (you should still convert the code you made in your notebook to a python script but not call it in your dag) . Otherwise, you should call your function that extracts the additional data in your dag.

How to get started -

To create your dags, you should convert your notebooks to python scripts and in your airflow python script,call the appropriate functions(from the notebooks you converted) for each task in your dag. You will also create two new functions for the 2 new tasks(loading to postgres and creating a dashboard).

You should have 2 yaml (docker-compose) files for this milestone(1 in each folder). 1 for the postgres image(that holds the tables in a postgres database) and 1 for airflow(where you will execute your dag). Note that you cannot place your postgres container, which will hold the tables,within the same yaml file of airflow, because airflow already uses a postgres image internally and it is only reserved for internal airflow services(to hold logs,metadata and other info regarding airflow in a database).For that you will need to connect both yaml files as shown in lab 10.

Deliverables and submission guidelines.

Send a zip folder by mail(badr.tarek@guc.edu.eg) by the deadline. Your zip folder must include 2 folders.

The zip folder AND subject of the mail **MUST** be named DE_M3_Faculty_GroupNumber_DatasetYear. Faculty is either IET or MET. Therefore if you are in MET and assigned to group number 20 working on the year 2018 dataset your folder would be named ‘DE_M3_MET_20_2018’.

Folder A **MUST** be named postgres_accidents_datasets must contain the following(but not limited to).

- yaml file for the postgres image where you load your database.
- Database folders mounted from your host host machine to container.
- Any additional folders mounted from your host machine to the container must be included in the folder you submit.

Folder B **MUST** be named airflow_milestone3.

- All files required to run docker-compose up. Such as: additional dockerfile created, requirements.txt (the dependencies for your dag), all folders mounted from your host

machine to the containers. (mainly /dags,/plugins,/logs).and most importantly the accidents dataset that you first process(the unprocessed dataset).

- **Important note**: only 1 csv file should be in folder B. **DO NOT** include the csv files created from the tasks. **I REPEAT,in the folder that you send me, DO NOT include the csv files created from the tasks. You will be penalized if you do so.** The only exception is for those who perform the ‘extraction of additional data task’ by reading from a csv file, check the second point in the important notes section.In such case you will have 2 csv files rather than 1.

You will be penalized if your folders are not named per the instructions.