Lab 10
Apache Airflow

What is Airflow?

Airflow is a platform to programmatically author, schedule, and monitor workflows or data pipelines.

Why do we need Airflow?

- Data grows fast, gets more complex and harder to manage as your company scales.
- In order to provide insights, you need to have some kind of visualization to explain your findings and monitor them over time.
- For these data to be up to date, you need to extract, transform, load them into your preferred database from multiple data sources in a fixed time interval (hourly , daily, weekly, monthly)

Scheduling your DAGs

There are two parameters that you can define when instantiating your DAG

to specify when your DAG will be run:

1- start_date

2- schedule_interval: dictates how often to run the DAG

https://crontab.guru/

This is a way to get the interval of the scheduling.

Some notes on the files

amazon-dag1

In this file the function are not having parameters nor returning values.

But if we want to return some values we can do this using “Context”.

amazon-dag2

It is the same functionality as in amazon-dag1 but we are passing the parameters(The dataframe).

We can handle this by adding a context to the function and then pull the parameters.

It was a pleasure for me to be your TA

:D