UNIVERSITAT DE
BARCELONA

Master in Fundamental Principles of Data Science

Dr Rohit Kumar

Final Assisgnment

Project Description

The Goal of this Project is to do a simple batch mode ML model in production.

Write a pipeline using airflow to train a ML model based on data in a s3 bucket and print the pridiction.

For ML model we will use a simple regression model using Iris data which is already done in attached ipynb file.

Data

- Training Data

There is data.txt file at https://ub-2021.s3-eu-west-1.amazonaws.com/data/data.txt

This file has one URL per line which has the data file link.

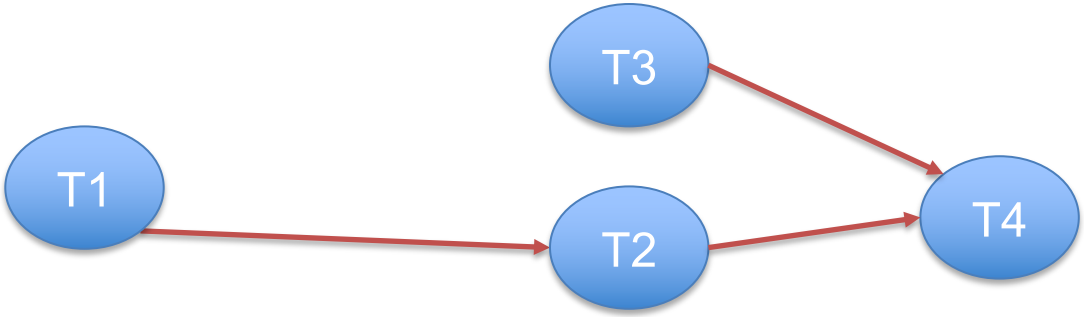
- Prediction File is available at https://ub-2021.s3-eu-west-1.amazonaws.com/data/predict.csv

Create the Pipeline

- Write a DAG code to do the following
  - T1. A task to download all csv from s3 bucket and store localy.
  - T2. A task to read all the downloaded csv and train the model and finally save the model locally.
  - T3. A task to download prediction.csv from S3 save it locally.
  - T4. Load the local model and read the downloaded prediction CSV and save a csv with prediction for each input.
- Dag Schedule: 8 PM every Monday
- Use the S3 path as Airflow Variable.
- Use the local path location for storing anything as Airflow Variable.

Run DAG

- Create a Dag like below using the Tasks
- Finally deploy your DAG test it and run it in airflow.



Deliverables to be uploaded

- Single Zip file
  - All Python Code for the airflow dag
  - Screenshot of Dag in Airflow.
  - Screenshot of one execution in airflow.

References

- https://airflow.apache.org/docs/stable/tutorial.html
- http://michal.karzynski.pl/blog/2017/03/19/developing-workflows-with-apache-airflow/
- https://www.polidea.com/blog/apache-airflow-tutorial-and-beginners-guide/
- https://towardsdatascience.com/getting-started-with-apache-airflow-df1aa77d7b1b