UNIVERSITAT DE BARCELONA

Master in Fundamental Principles of Data Science

Dr Rohit Kumar

UNIVERSITAT DE BARCELONA

Final Assisgnment

UNIVERSITAT DE BARCELONA

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 every hour based on data in an s3 bucket and print the pridiction.

Setup all required infra in AWS.

Create buckets and simple data (you can use sample data from github)

https://github.com/rohit-nlp/BigDataCourseUB/tree/master/assisgnment_3/data

UNIVERSITAT DE BARCELONA

Infrastructure setup

- Create an EC2 server and setup airflow
  - You can use AMI
  - Docker
  - User script
  - Or Manually
- Create two buckets in s3
  - One for storing training data (bucket name: YourFirstName.UB.Training)
  - One for storing prediction data (bucket name: YourFirstName.UB.prediction)
  - One for storing ML model (bucket name: YourFirstName.UB.ml)

Use your AWS Educate Login. If you do not have an account yet send me an email. You need to use your UB email to enrol for AWS resources

UNIVERSITAT DE BARCELONA

Prepare Data

- Upload train1.csv and train2.csv in S3 in bucket for training (YourFirstName.UB.Training)
- Upload prediction.csv in S3 in bucket for prediction. (YourFirstName.UB.prediction)

UNIVERSITAT DE BARCELONA

Prepare App

- Write a ML training function (you can use the one we used in class with Iris Data) Change the code to do the following
  - Take data as input (either as a string with path for csv or as a pandas dataframe)
  - Save the model and then upload it in S3. (YourFirstName.UB.ml)
- Write a ML prediction function (you can reuse the one we used for iris data)
  - Load the model in Python
  - Download prediction.csv from S3 (YourFirstName.UB.prediction)
  - Make prediction and upload a new csv back to YourFirstName.UB.prediction as *prediction_currentTimestamp.csv*

UNIVERSITAT DE BARCELONA

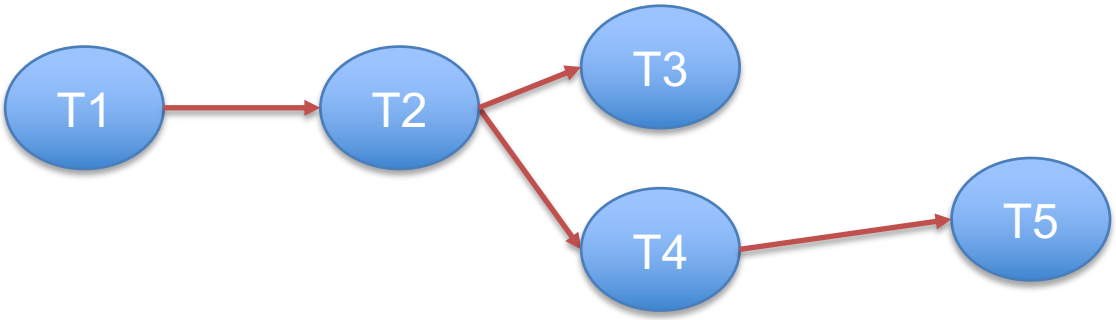
Create the Pipeline

- Write a DAG code to do the following
  - T1. A task to download all csv from s3 bucket and store locally (YourFirstName.UB.training)
  - T2. A task to read all the csv and to train the model and finally save the model locally.
  - T3. A task to Upload the model on s3.
  - T4. A task to download prediction.csv from S3 and run prediction using the model saved locally and créate a new csv with prediction.
  - T5. Upload the new csv with prediction to S3 as *prediction_currentTimestamp.csv*

UNIVERSITAT DE BARCELONA

Run DAG

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



UNIVERSITAT DE BARCELONA

Deliverables to be uploaded

- Single Zip file
- YourFirstName_Assignment3.zip
  - Python Code for the airflow dag
  - Python Code for all the task used in the dag
  - Screenshot of Dag in Airflow.
  - Screenshot of one execution in airflow.
  - Screenshot of S3 bucket with uploaded model
  - Screentshot of s3 bucket with uploaded prediction_currentimestamp.csv file.

UNIVERSITAT DE BARCELONA

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