EECS E6893 Big Data Analytics - Fall 2022

Homework Assignment 4: Data Analytics Pipeline

Due Friday, December 2nd, 2022, by 5:00pm

Task 1 Helloworld (35 pts)

Q1.1 Read through the tutorial slides and install Airflow either on your local laptop or on a VM of GCP. You can also use google cloud composer if you know how to use that. (20 pts)

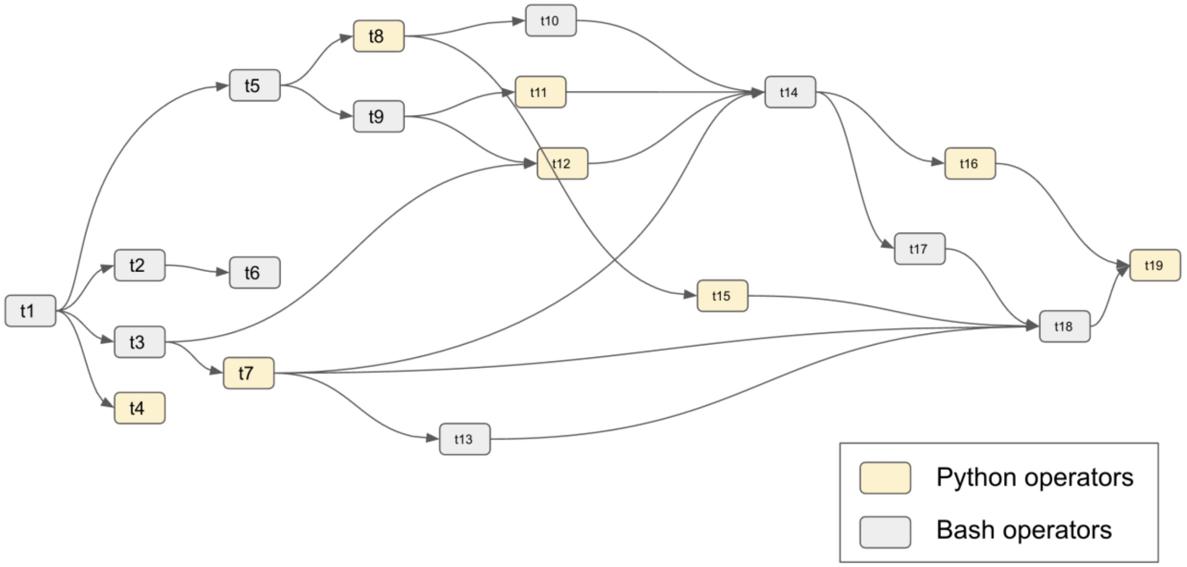
(1) Provide screenshots of terminals after you successfully start the webserver and scheduler.
(2) Provide screenshots of the web browser after you successfully login and see the DAGs.

Q1.2 Run helloworld with SequentialExecutor and LocalExecutor. (15 pts)

(1) Provide screenshots of Tree, Graph, and Gantt of each executor. (5 pts + 5 pts)
(2) Explore other features and visualizations you can find in the Airflow UI. Choose two features/visualizations (other than tree, graph, and Gantt), explain their functions and how they help monitor and troubleshoot the pipeline, use helloword as an example. (5 pts)

Task 2 Build workflows (50 pts)

Q2.1 Implement the DAG below (25 pts)



For each kind of operator, use at least 3 different commands. For example, you can choose sleep, print, count functions for Python operators, and echo, run bash script, run python file for Bash operators.

(1) Provide screenshots of Tree and Graph in airflow. (10 pts)
(2) Manually trigger the DAG, and provide screenshots of Gantt. (10 pts)
(3) Schedule the first run immediately and running the program every 30 minutes. Describe how you decide the start date and schedule interval. Provide screenshots of running history after two repeats (first run + 2 repeats). On your browser, you can find the running history. (5 pts)

Q2.2 Stock price fetching, prediction, and storage every day (25 pts)

(1) Schedule fetching the stock price of [AAPL, GOOGL, FB, MSFT, AMZN] at 7:00 AM every day. Use Yahoo! Finance data downloader https://pypi.org/project/yfinance/.
(2) Preprocess data if you think necessary.
(3) Train/update 5 linear regression models for stock price prediction for these 5 corporates. Each linear model takes the “open price”, “high price”, “low price”, “close price”, “volume” of the corporate in the past ten days as the features and predicts the “high price” for the next day.
(4) Every day if you get new data, calculate the relative errors, i.e., (prediction yesterday - actual price today) / actual price today, and update the date today and 5 errors into a table, e.g., a csv file.
(5) Provide screenshots of your code. Describe briefly how to build this workflow, e.g., what the DAG is, how you manage the cross tasks communication, how you setup scheduler…

Task 3 Written parts (15 pts)

Q3.1 Answer the question (5 pts)

(1) What are the pros and cons of SequentialExecutor, LocalExecutor, CeleryExecutor, KubernetesExecutor? (10%)

Q3.2 Draw the DAG of your group project (10 pts)

(1) Formulate it into at least 5 tasks
(2) Task names (functions) and their dependencies
(3) How do you schedule your tasks?

Homework Submissions

Put your screenshots and answers in one PDF. Submit the PDF and the code of Task 2.