RMIT University Vietnam
School of Science and Technology

EEET2574 | Big Data for Engineering

Assignment 1: Data Pipeline with Docker

Due: 23:59, November 29, 2024
This assignment is worth 20% of your overall mark.

Introduction

In this assignment, you will deploy an end-to-end data pipeline on your local computer using Docker containerized Kafka (data streaming), Cassandra (NoSQL database) and Jupyter Lab (data analysis Visualization). This assignment is intended to give you practical experience with a simple big data stack and data pipeline.

The "Big Data for Engineering" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis - it is your responsibility to stay informed with regards to any announcements or changes.

Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalised; there are no exceptions and no excuses. More information on Academic Integrity is available at https://www.rmit.edu.vn/students/my-studies/assessment-and-exams/academic-integrity

General Requirements

This section contains information about the general requirements that your assignment must meet. Please read all requirements carefully before you start.

Task 0: Video documentation

Include a demo video, showing

1. How you put up all the docker containers, set up the Kafka data pipeline, screen of the producers and consumers.
2. Check and print out all the data in Cassandra DB.
3. Check each data on Jupyter Lab (by printing out the df)
4. Briefly explain your visualization plots.

Please briefly describe orally what you're trying to demonstrate in the demo video. Please make sure you have already built the Docker images once before recording so the rebuild step is faster.

All the text must be readable (increase font size on your terminal/bash/cmd and on Jupyter Lab). You won't be graded on your presentation skill and you don't need to show your face in the video (don't cover some parts of the screen that I need to see). This is mainly for me to check the data pipeline.

Task 1: OpenWeatherMap (20pts)

Please repeat all the steps in tutorial 3 and create a OpenWeatherMap data pipeline for TWO NEW CITIES (i.e don't use Vancouver), using resources from the GitHub repo: https://github.com/vnyennhi/docker-kafka-cassandra

Task 2: Faker API (20pts)

Please repeat all the steps in tutorial 4, but you must include at least 10 different fields for the Faker API data. More about the fields:
https://faker.readthedocs.io/en/master/providers.html

Task 3: Another API (20pts)

Please repeat all the steps in tutorial 4, but for at least one more API that you can find online. Fake data API is fine. Try to find data that is interesting and meaningful.

Task 4: Visualisation and Analysis (10pts)

HD point: Please provide at least 2 useful visualization plots and analysis using any of these data in the Jupyter notebook. You can include your quick presentation on the data pipeline demo video, no need for a separate file submission.

Code Readability and Documentation (15pts)

Your code should be easy to follow and well-organized, adhering to clean coding principles in terms of comments, function/variable names, modularity, readme file, and jupyter notebook documentation.

Clarity in Video Presentation (5pts)

The demo video should clearly show how you meet the requirements.

Practical Implication of the Project (10pts)

Reflect on the practical value of your project, including description of the user case and scenario, and how your approach can help solve the problem described in the scenario.

What to Submit, When, and How

When:
The assignment is due at **23:59, 29 November 2024**. Assignments submitted after this time will be subject to standard late submission penalties. Network outage, computer crash, data lost with no backup, etc…, will not be considered as valid excuse of late submission.

What:
Please zip the whole **project folder** for submission, name it after your student id ("s1234567.zip"). The zip file should include:

- All the codes, files, etc.. required so that I can reconstruct your work in docker.
- A demo video, or a link to the demo video (e.g. link to youtube, google drive, etc).
- A README.md file with
    - Clear instruction on how to replicate the results on the video.
    - Information on what API you chose in Task 3.

- A brief description on your visualization plots in Task 4 (e.g. have you discover anything interesting in the plots?).

How:
Submit it through Canvas. Time stamp of submission will be used as your official submission time.

Please do NOT submit other unnecessary files.