RMIT Vietnam University
School of Science and Technology

EEET2574 | Big Data for Engineering

Assignment 3: Big Data Project

Report Due: 23:59, Monday in Week 12
Presentation: Friday in Week 11 (during scheduled tutorial time)
This assignment is worth 40% of your overall mark.

Introduction

In this assignment, you will build a full data pipeline on AWS and any other platform (such as Databricks, MongoDB, etc). 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.

Data

You have to choose three or more different data sources for this group project. At least one of them must be coming in as streaming data. The other datasets can be a file saved and loaded from a S3 bucket. Ideally, the three data sources must make some sense when being combined together. Feel free to consult me on your idea about this.

Data Pipelines

From the three raw data sources, you have to build ETL processes to clean and transform data for analysis. You can use any of the AWS services available to you for this purpose: Athena, Glue, Lambda or any other AWS services. You have to process them separately first and save the cleaned data back (e.g. S3, DynamoDB or MongoDB or any other database).

After having a cleaned data, you can choose 1 or 2 of the following layers in the big data stack:

- Load data into Sagemaker, Databricks or other platform for training a prediction model with Spark MLlib.

- Load data into a visualisation dashboard (Mongo Charts, Power BI, or any front-end web that you can deploy online).

Important: At this step, you must combine and utilize both datasets.

You must utilize some AWS cloud service(s) or LLM programming techniques (or both) in your project.

Task 1: Report (60 pts)

Write a report on your project (max 20 pages) That consists of 3 parts: Overview, Solution Design and Conclusion.

Overview
What is the problem you are trying to solve?
Why is the problem important?
How do you use big data to solve the problem?
What are the different roles and contributions of members in the group?

Solution Design
What is the proposed solution?
What are the datasets?
What is the data pipeline?
What is the chosen technology/infrastructure/models? And what are the reasons?
What is the estimated running cost per day/week?
What are your preliminary results?
Note: The proposed solution must be data-oriented.

Conclusion
Summarise the project as well as note for future work. Note any limitation of your solution, if any.

Important: You must include a diagram clearly illustrates the steps and components in your pipelines. Any references must be included with appropriate citation.

Task 2: Presentation and Demonstration (40 pts)

Present your project during tutorial on Week 11. Each member must speak in the presentation. Please make sure you explain the code/component of every single step clearly, not just quickly showing the results.

What to Submit, When, and How

The assignment is due at 23:59, Monday in Week 12.

Assignments submitted after this time will be subject to standard late submission penalties.

Please submit your report as pdf file on Canvas.
Please also submit all source code and/or documentation associated with your project.
Zip it into a zip file and submit on Canvas.