DS403.3 – Big Data Programming Group Assignment

Group Project

Team size: Up to 3 members
Grade: 30/100

Objective:
The project should showcase a novel and creative use of data, and recommended to follow the lambda architecture for Analytics

Proposal: A proposal write-up will be due on Mar 29, 2025.
Intermediate report: Apr 10, 2025.
Presentation: Apr 20/27, 2025.
Final report: The final report will be due on May 11, 2025.

The details and deliverables are as follows:

1. Create groups (up to 3 members) and assign a group number to each team, forum created in learning system and due on Mar 17th 11.59 pm.
2. A maximum two-page (single space) project proposal write-up will be due on Mar 29, 2025. It should be a single file with the filename groupXX_proposal.pdf, where XX is your group number. All groups are required to submit the proposal by Mar 29, 2025, 09.00 am to Learning platform. The failure to submit the proposal on time will lead to reduction in grade for the whole project.
   It should include the following:
   a. What is the problem or question?
   b. Who is the consumer of the data and analytics?
   c. Identify sufficiently "**large**" data sets, data from APIs or other sources.
   d. The method of analysis to be used
3. There will be a one to one discussion with each group to discuss about the proposal on Mar 29/Apr 6. Time slot 4-5 pm, 15 minutes for each group.
4. The intermediate submission is due on Apr 10, 2025. Upload the intermediate work as groupXX_Intermediate_submission.zip. where XX is your two digit group number. Include the intermediate report inside the zip folder.

5. Present your work in class on April 20th and 27th. Schedule will be posted in learning system.
6. The final submission is due on May 11 at 11.59 pm. Upload the final project as a single zip file with the filename bdp_groupXX_Final_submission.zip, where XX is your two digit group number. Include the final report inside the zip folder.
7. Intermediate and Final Report details
    a. Students can aim for 5-10 pages for intermediate and 10-12 pages for final single spaced (Font 12 Times New Roman).
    b. The report format should be similar to the proposal, but include more details about the method used, conclusions from the analysis. Source code (data modelling) should be provided in the appendix, and does not count towards your page limits.
    c. Your data sets
    d. It should include all source codes such that the results are reproducible from the raw dataset.

Expected Deliverables

1. **Architecture Diagram**
    1. High-level design of Lambda Architecture/Any other suitable architecture (Detailed analysis for choosing a tool. i.e.: Spark over Map Reduce) for the usecase.
    2. Usecase diagram, component diagram, depoyment diagram
2. **Benchmark of Processing Engines and databases**
    1. Performance of MapReduce vs Spark
3. **Codebase**
    1. Source code in GitHub/GitLab repository
    2. Archived folder of working code
4. **Data Pipeline Demo**
    1. Working ingestion, processing, and visualization
5. **Documentation/Report**
    1. Includes data schemas, pipeline details, visuals, APIs etc.
6. **Final Presentation** – Summary of design, implementation, challenges, and results.

Evaluation Criteria

Criteria	**Description**	**Weight**
Architecture & Design	Effective use of Lambda Architecture	20%
Implementation & Code Quality	Efficient, scalable, and fault-tolerant system	30%
Data Processing Accuracy	Correct batch & real-time aggregations	15%
Visualization & Insights	Interactive dashboard & data usability	15%
Team Collaboration & Documentation	Clear documentation & teamwork	20%

****Note: There would be a reduction in grade for late submissions and online copying**

Recommended Technologies: Spark, Hive, MongoDB/Redis, Spark Streaming/Kafka, Graffana