Final project

Objectives:

1. Having a good grasp on the concepts taught in this course.
2. Revision and practical applications of using Big data technologies.

Basic instructions:

1. Students need to form teams of 3 members. There can only be two teams with 4 members. Please decide it among yourself and enter the details in this Google Sheet.
2. Based on the given details, your team need to work on the following dataset:

Team Name	Member 1	Member 2	Member 3	Member 4	Dataset
Dark Knight	ME16B147	Me16b174	Me16b173		1. NYC tickets
Lockdown	NA16B117	MM16B019	MM16B009	MM16B023	2. YELP
Excalibur	ME16B172	CE16B053	ME15b077*		1. NYC tickets
Illuminati	BE16B016	ME16B148	ME16B166		2. YELP
Tony_Spark	ME16B180	NA16B026	ME16B176		1. NYC tickets
Crème de la Crème	ME16B125	ME16B128	AE16B005		2. YELP
Quarantined Cops	CH16B024	CH16B033	CH16B045		1. NYC tickets
Le Cunctators	CH16B058	CH16B113	CH16B119		2. YELP
Mavericks	ME16B135	ME16B177	EE16B121	EE16B120	1. NYC tickets
DeepBlue	ME16B175	NA16B116	CE16B034		2. YELP

* Did not fill the google sheet.

3. Each team will need to submit one report along with the code in a zipped folder (name: **datatset_<dataset-number>_<team_name>.zip**) in the moodle and show a demo of their project. There is no deadline as of now.
4. Once the report is submitted, based on your and TAs availability, you need to show a live demo of your project. You will be given **max 10 minutes** for the demo/presentation.

General instructions:

Batch computation:

For this task, your team needs to perform the pre-processing and other necessary tasks in a DataProc Cluster using Spark. Your team needs to generate the model and then save the best model for evaluation. You will need to write code for evaluation that can access this model and make predictions. You need to run this evaluation code in the demo/presentation. **Note: We don't require you to perform the training during the demo/presentation**.

Overview:

1. Use a DataProc Cluster and submit a Spark job for data pre-processing and model training.
   Note: For pre-processing, you can find this link helpful to get a better understanding of data. It is about working with Jupyter Notebook on Google Cloud Platform.
2. Store the model in your GCS Bucket.
3. Submit a Spark job for evaluation on validation data stored on a GCS bucket.

Real-time computation:

This task will require your team to show us a working demo of how you can use the trained model to perform real-time predictions of test data streaming to Kafka. The basic idea is that one of the members will be feeding the data into Kafka cluster, another member needs to use the data in a Spark cluster to generate real time predictions.

Overview:

1. Stream the test data stored on the GCS bucket into Kafka.
2. Use Spark Streaming to read the data and make real-time predictions using your stored model.

Task-specific instructions:

1. **New York Tickets Dataset: link**
   This dataset contains information about various tickets issued in New York. The aim is to predict the part in the city where the ticket was issued. (Prediction column name:- Violation location)
2. **YELP-Dataset: link**
   This dataset contains the reviews posted by various users on the Yelp website. The aim is to use NLP to predict the ratings in terms of stars. (Prediction column name:- stars)
   [Don't use Deep Learning for the assignment]

Evaluation:

1. **Group based evaluation (GE):** Depending on the presentation, demo and report, your team will be evaluated and you will receive a score for the same. The parameters for the evaluation are:
    a. Batch computation:
        i. Relative* Accuracy (Validation)
        ii. Feature Engineering
    b. Real-time computation:
        i. Publish/Subscribe code
        ii. Message parsing
        iii. Relative* Real-time prediction/accuracy

    *Relative means with respect to other teams which have been assigned the same dataset

2. **Individual evaluation (IE):** You will receive an absolute score based on the viva voce. It will be based on your final project and lab assignments.

Score distribution:

- GE = 30 marks:
    - Batch Computation = 15 marks
    - Real Time Computation = 15 marks
- IE = 10 marks

Report:

The basic outline of the report should be as follows:

1. Objective
2. Pre-processing (with screenshots)
3. Performance of the best model (with visualizations) and Inferences
4. Real-time computation (with screenshots) and latency of processing each window
5. Conclusion