CS6847: Cloud Computing

Assignment II

(Submission via Google Classroom)

Map Reduce is the popular programming paradigm used in Big data processing. The objective of the assignment is to understand and analyze the Map-Reduce programming model using the New York Taxi dataset **(Link)**.

Problem Description

- Set up a three node cluster for executing the Map Reduce program.
- Write a Map Reduce program to find out
  - Top five most popular route in the dataset (Any year, Example- 2013).
  - Top five most expensive route in the dataset. (Any year).
  - Top five most visited pickup and drop location. (Any year).
  - Top five most popular night life spots. (Time 8 P.M. to 2 A.M.)
- Experiment with different tuning parameter such as slow start, number of reducers, etc.

Evaluation

- Plot the graph of execution time of program by varying the number of reducers.
- Evaluate the behavior of program by varying the slow start parameter, using appropriate plots.
- For each of the program, write the output for every month in a separate file for any selected year.

Submission guidelines

- Submit the source code of Map Reduce program for the assignment. All other supporting files used for generating plots, logs, etc. should also be placed in the zip file (`Roll_number.zip`).
- Submit a `README` file containing the necessary details for running your program.
- Create a report explaining the plots and results in detail.
- Create separate folders for each of the program. The folder should contain `twelve files` representing the output for every month of the selected year. The folder should also contain a `README` file specifying the chosen year.

Academic Honesty

WARNING ABOUT ACADEMIC DISHONESTY: Do not share your work with anyone else. The work YOU submit SHOULD be the result of YOUR efforts.

Note: It is recommended that the assignment is done in a group of three students. In case you want to do it individually, you must be able to run the code on atleast two nodes.