DS5102: Big Data Lab

Lab-2: MapReduce

In this lab session, you will learn how to perform basic MapReduce jobs on data. You need to design your own MapReduce programs using the `mrjob` Python package. Please get your results verified by the TAs.

Task-1: Estimate Pi

You will use the file `estimate_pi.txt`, which contains $100,000$ lines. Each line consists of two random numbers, independently generated and uniformly distributed between 0 and 1. Please remember to take the output screenshot for the post-lab report.

Task-2: Movie Rating Analysis

You will use the `ml-100k` dataset for this task. Please make sure that the outputs are stored in separate HDFS directories (you will need the output screenshots for the post-lab report)

(i) Using the `ratings.py` Python code, count the number of times each star rating was given.

(ii) Write a program to count the number of ratings given by each user.

(iii) Write a program to obtain the average rating for each movie.

Task-3: Matrix Multiplication

You will use the `mm.json` file for this task, which contains two 5×5 matrices. You need to multiply these two matrices using MapReduce. Please make sure that the outputs are stored in separate HDFS directories (you will need the output screenshots for the post-lab report)

(i) Multiply the matrices using two Map tasks and a Reduce task (do not use dummy/identity map tasks ☺)

(ii) Multiply the matrices using one Map and Reduce task each.

Post-lab report

Prepare a post-lab report (submission link will be made available on Moodle) addressing the following points;

- Include Python codes for all the tasks.
- Include screenshots of all the output files (if the file is large, it is enough to show the initial parts).