COSC2633/2637 Assignment Presentation

(No more than one page, font size 12, Time New Roman)

Name: Udeshika Dissanayake **Student ID:** s3400652 **Lab Time:** Tu, Th, Fr (AM), (Fr (PM))

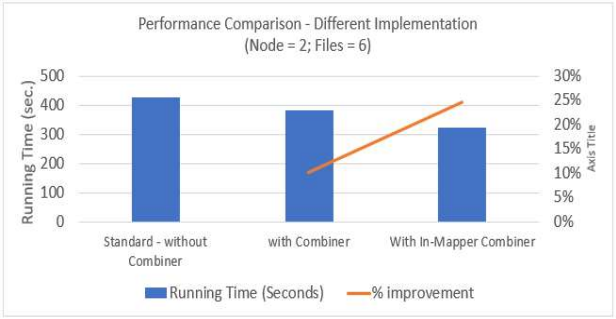
1. **What is the problem that your assignment aims to solve?**
   This BigData study intends to identify the most revenue generating Taxi zones in New York City for year 2019.
   Three MapReduce algorithms were developed and their performance were analyzed on different size of input datasets and on different size clusters in EMR.

2. **What is the input data?**
   NYC TLC Yellow Trip Data for year 2019.
   Six .csv files were used (one for each month). Each file is ~650MB.

3. **Are there iterations where each iteration has its own Map and Reduce?**
   Standard Map and Reduce classes were developed to find the total taxi fare for all taxi zones. No multiple iterations in the calculation was required.

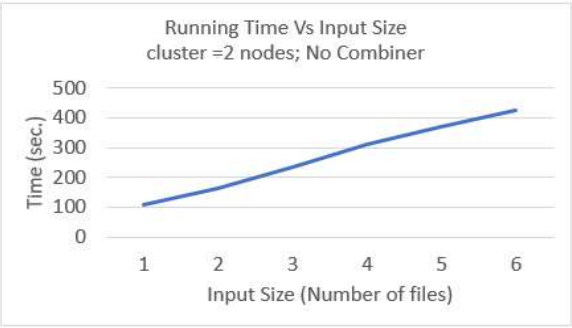
4. **What is the key and value in the <key, value> pair of Map output?**
   Map tasks emit the <key, value> pair of <'Pick-Up Location', 'Taxi Fare'> for each record

5. **Did you implement location aggregation?**
   Yes. In-Mapper Combiner (IMC) was implemented under the MapReduce Algorithm3. Also, standard Combiner was implemented under the Algorithm2. IMC showed 25% improvement in the performance.



6. **What is the output of Reduce tasks?**
   Reduce tasks emit <'Pick-Up Location', 'Total_Taxi Fare'> for each taxi zone

7. **What is the size(s) of input data you tested in the assignment? What is the impact of input data size to the processing efficiency?**
   Input data size was varied from 650MB to ~4GB.
   Running time linearly correlates with the size of the input data.



8. **What is the number of nodes in Hadoop Cluster you tested in the assignment? What is the impact of cluster node number to the processing efficiency?**
   Cluster node sizes were varied from 2 to 10.
   Running time speeds up by 36% when 10 node cluster was used compare to 2 node cluster.