COSC2633/2637 Assignment Presentation

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

Name: **Vikas Virani** **Student ID:** **S3715555** **Lab Time:** Fr (AM)

1. **What is the problem that your assignment aims to solve?**
   It aims to do k-means clustering over NYC Taxi pickup & drop off points. To compare the performance & scalability of dataset over the number of nodes & number of clusters.

2. **What is the input data?**
   My input data is <Centroid, Datapoint> as <Key, Value> pair, Where, each Datapoint is a pickup & drop off location for each observation. Centroid is initialised using a random Datapoint from the range between minimum & maximum values of each feature.

3. **Are there iterations where each iteration has its own Map and Reduce?**
   Yes, as I'm implementing k-means clustering, Execution of job terminates when all the clusters are converged, to calculate new Centroids after finding all the nearest Data points to each cluster, it need to have iterations which has its own Map & Reduce. Number of iterations depends on how fast all the clusters are converged.

4. **What is the key and value in the <key, value> pair of Map output?**
   I'm emitting **<Centroid, DatapointWrapper>** as Map output. Where, DatapointWrapper is a wrapper class which holds **list of Data points** nearest to the **Centroid in Key**.

5. **Did you implement location aggregation?**
   Yes, as mentioned above, I've maintained an Associative array of **<Centroid, DatapointWrapper>** which holds Centroid as key & list of Data points nearest to that centroid as value in Wrapper class for the whole dataset. And then it emits those values after all map of particular iteration are executed, instead of emitting values in each map execution.

6. **What is the output of Reduce tasks?**
   The output of Reduce tasks is <Centroid, Datapoint> as <Key, Value> pair, which will be fed as input to next iteration if all Clusters are not converged.

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?**
   Initially, I took a sample of 50,000 & 100,000 rows and tested on it. It increases the time to process the records if number of nodes are same in both case. First one took 6-7 min each iteration (15 iterations to converge), second one took 10 min for each iteration (24 iterations to converge), Almost linear increase in terms of timing.

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?**
   I've tested the dataset on 2, 4 & 6 nodes in Hadoop cluster on a sample data of 50,000 rows to compare the performance. It doesn't affect the time to process the records that much for same number of Clusters. First one took 6-7 min each iteration (15 iterations to converge), second one took 6 min for each iteration (16 iterations to converge), third one took ~5-6 min for each iteration (14 iterations to converge), Almost same results in terms of timing.