**Kmeans Assignment 2: ReadMe File**

Input Parameters for Program:

* Input Directory. Eg: “/home/gaurang/Kmeans/input”
* Output Directory. Eg: “/home/gaurang/Kmeans/output”

Input Files for the Program:

* Data file
* Centroid File

The MapReduce Program consists of the following classes:

1. **File Generator Class**

This class generates the input file for the MapReduce program. The input file is already generated and is located in the Input folder. Sample data for the input file is as follows:

<data-point> <tab> <comma separated attribute list>

1. **Driver Class**

The driver class sets the configuration parameters (Input / Output file, Mapper class, Reducer class, Partitioner class etc.) for MapReduce jobs through the Job object. Hadoop uses these configurations to run the MapReduce program.

This Assignment has 2 jobs (1 – Centroid file generation, 2 – Kmeans algorithm).

* Job 1 uses 1 Mapper class (Mapper1.class)
* Job 2 uses 1 Mapper (Mapper2.class), 1 Reducer (Reducer2.class) and 1 Partitioner class (Partitioner2.class).

1. **Mappers and Reducers Classes**

Mapper1.class

This Mapper takes the randomly generated Centroid file as the input and generates a global HashMap which contains the centroids and corresponding attribute list.

Mapper2.class

This Mapper takes the randomly generated Data file as the input and processes it to generate key value pairs. In this Mapper, the key is the corresponding centroid cluster the data-point is assigned to and the value is the corresponding attribute list for that data-point in the Input file.

Reducer2.class

The Reducer2 class takes input from Partitioner2.class and further processes, i.e., generates a single output line for multiple values of the same key (Cluster Centroids) based on aggregating the results for each centroid. Additionally, it assigns the new centroid values to another global HashMap that is used for comparison against the old centroid values.

The Mapper2 and Reducer2 classes are further chained in the Driver class until all the the convergence criterion for the centroid recalculation is reached, i.e, difference between old and new centroid values < 0.01.

1. **Partitioner Class**

The Partitioner2.class creates 3 partitions based on the values of cluster centroids. The Partitioner takes input from the Mapper2.class and generates 3 partitions for each of the 3 Reducer tasks defined in the Driver class.