**B534-Distributed Systems Project 2B: Harp & Kmeans**

**Fall 2016**

**Team Members:**

Abhijit Karanjkar(aykaranj)

Rohit Nair(ronair)

**Main steps:**

1. Parse program arguments and read parameters.
2. Create directory structure for data, centroid and output
3. Generate data points and centroids in local file system and copy it to hdfs.
4. Configure Job using parameters
5. Run the job.
   1. Get data points & Centroids in mapper.
   2. Calculate Euclidean distance of data points from centroids.
   3. Assign data points to nearest centroids.
   4. Update centroid value to average of all data pointes assigned to it
6. Repeat step 4 - 5 till number of iterations & Update final result to hdfs.

**Data Flow:**

1. Data points & centroids are generated into local file system.
2. This data is then copied to hdfs file system.
3. Each iteration is responsible for loading the data from hdfs, performing business logic on that data and then putting that data back to hdfs

**Input:**

hadoop jar harp3-app-hadoop-2.6.0.jar edu.iu.km.KmeansMapCollective 1000 10 10 2 10 /kmeans /tmp/kmeans

Note: Generated data files, centroid files and output files are included in zip