PROJECT #2 - MapReduce/Hadoop - DUE DATE: April 23th, 2023

In this project you will implement the first iteration of K-Means clustering algorithm in MapReduce and apply it on synthetic data that you will create.

1. Install Hadoop (e.g. Cloudera distribution)

2. Create using whatever language you want a very large text file, containing at least 1M data points in the form of (x, y), where x and y are real numbers. The generation of should be biased toward the creation of three clusters. In other words, choose a-priori three centers (x1,y1), (x2,y2) and (x3,y3) and generate the rest of the data points around these, using some random distance following a skewed distribution (towards 0)

3. Move your file to HDFS

4. Write a MapReduce job that distributes the centers you have chosen at Step 2, (x1,y1), (x2,y2) and (x3,y3) to all mappers and reads in the file that you have created in Step 2 and maps each pair to the closest center. Reduce function should compute the new center for each constructed list. All distances are Euclidean. This is the iteration step in K-means algorithm.

5. Now implement the full K-means algorithm (stop when centers are close enough to previous step).