

## **PROJECT #1 - MapReduce/Hadoop - DUE DATE: March 31<sup>st</sup>, 2019**

In this project you will implement the 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  $(x_1, y_1)$ ,  $(x_2, y_2)$  and  $(x_3, y_3)$  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. Implement KNN algorithm as described in the class and apply it in your data set