CS F320 – Foundations of Data Science – Assignment #1
Comparative Performance Study of MapReduce/Hadoop & Spark
Total Marks – 20 (weightage 10%) Submission Deadline: 11th Nov. 2019

Aim:

The aim of this assignment is to give you an exposure to programming in MapReduce and Spark. It also aims to give you a glimpse of the performance difference you get while executing programs on these platforms.

To do:

1. Install Hadoop and Spark on a single machine and emulate 1 master + 4 slaves' configuration.
    - You can follow any online tutorial for this.
    - Use the most recent versions of Hadoop and Spark.
2. Implement **k-NN Classifier** algorithm and **Bisecting K-means** Clustering algorithm to work over MapReduce and Spark.
    - You can use the dataset uploaded at this link to execute your programs. The dataset consists of 10M million data points of 13 numerical dimensions. The first two lines in the dataset indicate the number of data points and the number of dimensions respectively.
    - For k-NN, you can take 80% of the dataset for training and 20% for testing.
    - You can use any number of Map and Reduce tasks as you want for the above configuration in case of Hadoop.
    - Don't use Mahout or MLlib. Your code must be implemented by you and has to be original.
3. Compare the execution time of each algorithm over both the platforms.
    - You can use in-built functions or graphical interface available with Hadoop and Spark to measure the execution time.

Expected Outcomes:

1. **Your implementation along with a readme in a zip file**
2. **Written Report**
   Your written report must clearly explain your design of the algorithm for each of the above platforms. The design should also answer the following questions:
    - What are the **Map** and **Reduce** tasks in case of Hadoop?
    - What **RDD transformations** and **actions** have you applied to get the desired output in case of Spark?
    - Must contain a **flow chart** of the algorithm for both the platforms.
    - Must clearly state what are the metrics or system functions you have used to measure the execution time for each case.
    - What is the time difference observed? What inferences do you make from these observations?
    - Clearly identify which step of the algorithm is actually leading to the difference of execution time observed.
3. **A presentation explaining your design and implementation**

PS: There will be a viva, where you will have to present your design and show the execution of your code. You can use any of the machines in teaching labs to do the assignment. You can also do it on your laptop, if you wish to.

Plagiarism of any kind shall lead to zero marks.