PROBLEM SET (MapReduce & SPARK)

1. For the given input file, calculate Wordcount using Hadoop MapReduce and Spark. Also, develop an equivalent conventional program (without spark RDDs) and compare the time taken by the two versions.

2. Consider two 10*10 matrix and perform Matrix Multiplication using Hadoop Mapreduce and Pyspark. (Note: matrix elements can be randomly populated)

3. Randomly populate 1000 numbers and calculate mean, variance, standard deviation for the generated data.

4. Compute correlation between the given two series using Pearson's and Spearman's Method.

(Use the Spark MLlib libraries and helper functions available)

Series A: 35, 23, 47, 17, 10, 43, 9, 6, 28
Series B: 30, 33, 45, 23, 8, 49, 12, 4, 31

5. Remove Stop words from the 'sentence' column given in the "StopWordRemoval_input" document.

6. In the given input file, binarize the column 'E' (fifth column). Set the threshold as 2.5

7a. Randomly generate 10k numbers and apply the following functions on the generated random numbers. Also, develop an equivalent conventional program (without spark RDDs) and compare the time taken by the two versions.

Exponential Function, $f1(x)=e^x$
Logarithmic Function, $f2(x)=\log(x)$
Square root Function, $f3(x)=\sqrt{x}$
Square function, $f(4)=x^2$

7b. Extension to the previous question: Instead of printing transformed values for all 10k words, print transformed values of the first 100 numbers only. Also, develop an equivalent conventional program (without spark RDDs) and compare the time taken by the two versions.