**CS4650 – Project #2 (50 points)**

**Goal:** Practice MapReduce programming on various software/hardware platforms.

**Suggested Platforms:**

1. Hadoop on XSEDE
2. Spark on XSEDE
3. Hadoop or Spark on EMR

**Suggested Problems to be solved:**

1. WordCount

Test files:

* 1. Complete\_Shakespeare.txt
  2. Bibles data files – download from <https://www.sacred-texts.com/bib/osrc/index.htm>)

Output Requirement:

Output must be sorted in descending order of frequency, i.e. most frequently occurred words first.

1. Adult Census Income data analysis -- run the attached Spark program with data from <https://www.kaggle.com/uciml/adult-census-income>
   1. Attached Programs:
   2. Spark program: RDDCensusData.py
   3. Help program for organizing data: CensusData.py
   4. Feel free to modify the programs to find more information from the dataset.

**Specific Tasks and Grading Criteria:**

1. WordCount using Hadoop on XSEDE -- output sorted in descending order of frequency. Test with Bibles data files. Hint: modify the programs given in XSEDE activity. (25 points)
2. WordCount using Spark on XSEDE – output sorted in descending order of frequency. Test with both Complete\_Shakespeare.txt and Bibles data files. See lecture notes for WordCount Spark program. (25 points)
3. WordCount on AWS EMR using either Spark or Hadoop (check AWS tutorials for using EMR.) (Bonus problem – may substitute for either problem 1 or problem 2.)
4. Adult Census Income data analysis. (Bonus problem – a sample for Capstone project.)

**Note:** While the sample programs are written in Python, you are welcome to use Java version map-reduce programs.

**Submission Requirements: A zip file containing the following to blackboard project 2 link.**

1. A pdf file with step by step instructions (from login to the system till completion of execution) for solving each problem.
2. Mapper and Reducer programs for each problem.
3. Output files.