**CS 4650 Capstone Project Assignment (100 points)**

**Goal: Big data processing with MapReduce programming paradigm on a cloud computing platform.**

**Team**: 2-3 students.

**Computing Platform Choice:**

1. AWS EC2 and/or EMR
2. XSEDE bridges
3. Other (requires instructor’s approval at the topic proposal stage.)

**Problem to be solved: Big Data Processing**

1. Team’s choice of any non-trivial exploratory big data analytics problem. Each team will propose a problem to be solved and get approved by the instructor.
2. Big data set requirement: >= 1 GB in one or multiple files.
3. Execution time should be recorded.

**Programming framework requirement:**

Use map-reduce programming on either Hadoop or Spark.

**Submission:**

1. Push all your project artefacts to a GitHub Repo or a shared drive, e.g. Google Drive or OneDrive.
2. The complete project submission package should include the following items:
   1. A README file on how to run your program (with detailed information including how to connect/interact with the system; compile and execution commands; etc.)
   2. The well-commented program codes;
   3. The data set;
   4. The presentation slides (in pptx format; with screenshots for key demo steps and display of execution time.)
3. Each team makes one submission on blackboard (on blackboard submission list all team member names and a link/URL to your complete project submission package.)

**Demo and presentation required.**

**Milestones:**

**Stage 1**: Problem identification and platform selection

Topic proposal (Team information, choice of computing platform, choice of big data problem, choice of Hadoop or Spark, a link to the data set source), post on blackboard discussion board.

**Stage 2:** Project detailed design and implementation with informal progress report during 12/3 class meeting.

**Stage 3:** Project demo and presentation. Finals week Tuesday, December 10th. Some teams may also present on Thursday, December 5th. Schedule to be set.

**Project submission due**: Thursday, December 12.

**Grading criteria:**

1. Stage 1 Topic proposal (10 points) – submit all required components on time.
2. Design and implementation (50 points) –quality MapReduce programs; level of difficulty of the big data task; successful execution of the program; impressive execution results (note: visualization not necessary); execution time.
3. Presentation and Demo (30 points) – Clear and logical presentation of project summary. Demo or show program execution and results (screen shots okay).
4. Team management (10 points) – how well a team works together in every stage; submission of complete project package. Late submission penalty applies.