CS 4650 Team Project Assignment (100 points)

Goal: Use Hadoop/MapReduce to process a big data set on a cloud computing platform.

Team: (2-4 students)

Computing Platform Choice:

1. AWS EC2 (team's responsibility to create accounts and take care of costs if any.)
2. XSEDE bridges
3. Other (requires instructor's approval at the topic proposal stage.)

Big Data Set and Big Data Processing Problem:

1. Team's choice of any non-trivial exploratory big data analytics problem (e.g. project 1). Each team will propose a problem to be solved and get approved by the instructor.
2. Big data set requirement: >= 65 MB. (use only 1 set of data but okay to try multiple sets)
3. Execution time should be recorded.

Deleted: 1

Submission:

1. You need to push all your project artefacts to a GitHub Repo. If you haven't used GitHub before, please take an online tutorial/video to learn the basic usage.
2. Your GitHut project repo should include the following items:
   a. The well-commented program codes;
   b. The data set;
   c. The execution time specified in the README.md file (The default GitHub Repo readme file).
   d. The presentation slides (pptx preferred; with screenshots for key demo steps)
3. You need to email the GitHub Repo URL to lyang@cpp.edu

Demo and presentation required.

Milestones:

Stage 1: Problem identification and platform selection
Topic proposal (Team information, choice of computing platform, choice of big data processing problem, a link to the big data set source), post on blackboard discussion board by Tuesday, Nov. 13, 10am.

Stage 2: Project detailed design
Expectations:
(1) know how to develop a MapReduce program for the chosen problem
(2) able to run the Hadoop wordCount problem on the chosen computing platform).
Progress report during the class meeting: <= 5 min report/demo on Tuesday, Nov. 20.

Stage 3: Implementation
Complete the programming and testing by Thursday, Nov. 29.

Stage 4: Project demo and presentation.
Tuesday, December 4 and Thursday, December 6.

Project submission due: Tuesday, December 11.

Grading criteria:

(1) Topic proposal (10 points) – submit all required components on time.
(2) Progress check (10 points) – meet the expectation; all team members present.

(3) Implementation (50 points) – successful execution of program; required testing; execution time.
(4) Presentation and Demo (20 points) – provide a clear summary of the project. Show program execution.
(5) Team management (10 points) – how well a team works together
(6) Quality of the project (up to 10 bonus points) – how difficult/challenge is your task! set a high goal!