15-319 Introduction to Cloud Computing

Project 4
"a.k.a The Big One"

Assigned Date: **March 23rd, 2010**
Deadline: **Part I - April 1st** in class (+3 grace days)**,**
Part III - April 20th, 2010 at 11:59 p.m (+3 grace days).

Goals:

1. Implement an Algorithm/Application using MapReduce/Hadoop.
2. Real-world exposure to research and development in Cloud Computing.

Background:

Projects 1-3 were aimed at getting your feet wet with our cloud infrastructure and Hadoop Map/Reduce. In this project, you may work on a project of your choice, or consult CS faculty or others to identify a data-intensive scalable project to work on. Examples are NLP with Prof. Kemal Oflazer, cloud performance analysis with Suhail Rehman, image or video processing with Prof. Justin Carlson or large network flow log analysis with Chas DiFatta.

Part I: Project Proposal

You have developed a set of skills to tackle data-intensive, scalable applications. Your role is to identify a client (a professor/researcher) or choose a project that will allow you to solve a problem to completion on our existing cloud infrastructure in Hadoop.

Once you have chosen a project, create a project proposal that includes your problem statement, planned approach in Hadoop, any extra cloud infrastructure that may be required (see part II below), deliverables that are due at the end of the project and a detailed timeline. This proposal will be graded as part of the project.

Deliverable: A project proposal (at least 2-pages) in 2-coloumn ACM format.
Due: April 1st in class (+3 grace days)**.**

Part II: Working on the Cloud

You may work with your existing cloud infrastructure to complete the project. In case you believe that this may not be sufficient, you can contact your TA and request for additional resources or a separate cloud from our hardware. If you require additional resources, please plan ahead to ensure that we can provide them to you on schedule.

Part III: Paper and Presentation

We will hold student project presentations during the last two lectures of this course. We expect you to present your project in 45 minutes and hand-in a detailed project report on the same day.

Hand in a project write-up discussing the problem and a description of the techniques you used to solve it in Hadoop. Make sure to include a results section including a bar graph of all the runtimes of your application. Use the 2-column ACM format for this paper (abstract, problem definition, methods used, results and comparison, conclusions and references). Include a full set of references, including books, tutorials and blogs that you have used to complete this project.

Deliverables: A project report in 2-coloumn ACM format and your project presentation as a PPT file.

Submission:

Add all the deliverable reports and presentations from each part into a single zip file (project4.zip) and place it in:

```
/afs/qatar.cmu.edu/course/15/319/handins/username/
```

This file is to be submitted once and the final timestamp on the server will determine your submission time.

Grading:

As mentioned in the syllabus, this project is worth 25% of your final grade. You have about a month to finish the project. The following rubric will apply:

Projects	
Deliverables Proposal Paper Final Paper Final Presentation	**75**
+	
Student Written Code	**25**
consists of:	
Property	**Percentage**
Technique/Algorithm	70
Performance	15
Documentation/Cleanliness	15