Santa Clara University

Department of Computer Engineering

Big Data (COEN 242)

Group Project (15 pts)

Spring 2016

Group Project: Milestones as shown in the Syllabus

1. Introduction:

The purpose of this group project is to provide students the opportunity to develop their experience in programming Big Data solutions: implement a machine learning algorithm over Hadoop using MapReduce job. Each group will be assigned a Group number and can select any of the popular Machine learning algorithms (refer to WEKA and Mahout). Each group of ten students will jointly work on the above topic to be approved by the instructor upfront. Once a group selected a machine learning algorithm, you have one of two options: (1) select a dataset from the internet and run the machine learning algorithm from WEKA and from Mahout on the same dataset and do detailed analysis. (2) Implement the Machine learning algorithm as M/R job on Hadoop and compare performance between your implementation, WEKA and Mahout/Hadoop for the same dataset. The deliverables include a Word document describing your project, Power Point presentation and a demo showing the algorithm running. In addition, each group needs to submit hard & soft copy of their code. **P.S.** you need to get reasonable dataset from the internet.

2. Deliverables and due dates

Each group should submit on the due date both hard copy and soft copy of the PPT presentation. The PPT should describe the ML algorithm selected, dataset used, and the experiment results/analysis. The group number, project title, and member names should clearly appear on the PPT front page. If you encounter any difficulty, you should talk to the instructor as soon as possible. In addition, each group will give the instructor the presentation in the class for their project and will give a demo when appropriate.

• **Group formation** – on the date indicated in the syllabus, groups of ten will be formed and assigned Group #.
• **Project Proposal Outline** – On the due date indicated in the syllabus, each group will come back with the research topic chosen, if different from the default topic, with the headers' outline of the paper. In addition, each team should list who is doing what in your research paper. The length should be 3 to 5 pages.
• **Final group Project** – Each group will need to submit a Power Point presentation covering the problem being investigated, how it will be parallelized on Hadoop (description for the map() and reduce() functions, and preferably performance comparison between your own implementation, WEKA, and Mahout using the same dataset (due date is indicated in the syllabus).

3. *Remarks*

- You are encouraged before developing your project on Hadoop to run an existing MapReduce job such as word count or **teragen** and **tersort** to get familiar with the environment first.
- Select a public data set from the web:
  - Amazon 53 Datasets: http://aws.amazon.com/datasets/
  - Infochimps Datasets: http://www.infochimps.com/datasets
  - Wikipedia datasets: http://en.wikipedia.org/wiki/Wikipedia:Database_download
  - Weather large datasets: http://cdo.ncdc.noaa.gov/qclcd_ascii/
  - Hadoop Illuminated: http://hadoopilluminated.com/hadoop_book/Public_Bigdata_Sets.html
  - Stackoverflow.com: http://stackoverflow.com/questions/10843892/download-large-data-for-hadoop
  - Hadoop Lessons: http://www.hadooplessons.info/2013/06/data-sets-for-practicing-hadoop.html

- You are encouraged to discuss with the instructor about the project progress during the quarter; do not wait for the last week to raise issues. Given that enough time is given, therefore, you should take responsibility in finishing the project in a timely manner. Your group members are responsible for resolving conflicts (if any) within your group. If there are problems due to non-responsive group members or other factors that significantly impede the group's work, please report to the instructor immediately. Within the group, you can discuss the details of the project and related topics. You are also encouraged to discuss high-level thoughts with other groups. However, you cannot share your paper or research with students not belonging to your group. Nor you are allowed to work with people outside your group.
  Group Project Grading Sheet

Group #_____ Group Members:_____ Project Title:_____

	Presentation Quality	Rubric	Max	Score
1	Clarity	Writing easily understood, use of appropriate vocabulary, etc.		
2	Conciseness	Avoidance of verbosity or marketing slogans, appropriate use of words and expressions		
3	Flow and Structure	Use of transitions, organization of ideas, section headings and subheadings, paragraphs, etc.		
4	Logical coherence	Sanity of proposed solution, clear articulation of solutions, etc.		
5	Format	Cover page, executive summary, references, appendices (if any),		

		arrangement of figures, tables, and other illustrations, report layout.		
			3	
	Introduction			
6	Background	Explanation the algorithm the team is parallelizing on Hadoop demonstrating the benefits from parallelization to the selected algorithm.		
7	Current Approaches	Cite other efforts that addressed the same problem, if any.		
8	Potential contributions	Potential benefits and contributions made		
			2	
	Project Details			
9	Architecture Overview	Detailed description of the current leading approaches		
10	Design details to your approach	Survey of current research approaches to address the known issues in your topic		
11	Novelty	What do you think is novel about your design and any possible enhancements/extensions that were not addressed in your implementation.		
12	Benefits	Discuss the performance numbers if you have done that piece.		
			12	
	Conclusions			
13	Summary and contributions	Summarization of the paper, the work done, and contributions.		
14	Lessons learned	Lessons learned for students and suggestion for extension of this research work		
			3	
	Grand total		**20**	

4. Due date

All milestone dates are listed in the syllabus.