CS4225: Massive Data Processing Techniques in Data Science
Course Project Guidelines

1. Overview

This is a team project. Each team comprises 2 or 3 members. Each team should identify a topic on Massive Data Processing Techniques in Data Science, and conduct a course project specified in one of the topics in the "Project Topics" section.

An important component of the course is the course project. Rather than having more homework, the course project allows students the flexibility to explore some aspect of reconfigurable computing on their own. The project involves the development of a proposal, experimental results, and a final report. In proposing a project please keep the following in mind: it is better to propose a reasonable project you can complete rather than a huge project that will be in an intermediate state at the end of the term. To receive a final grade for the project, there must be an experimental result that is shown (e.g. a graph, table, or chart) based on experiments that you perform.

You are encouraged to use a cloud platform for processing big data. Microsoft has kindly offered us the education grant on Windows Azure for this course. Therefore, you may be able to run your project and analytics on multiple virtual machines on Azure. If you are interested in using those resources, please contact our teaching assistant.

2. Project schedule

10 Feb 2018	Grouping on IVLE
10 March 2018	Submit a project proposal (See Section 6, upload to IVLE)
20 April 2018	Project presentation
28 Apr 2018	Project due – report must be submitted (upload to IVLE)

* all deadlines are 11:59pm of the date.

3. Project Topics

The topic of the course project can be ONE of the followings:

- **Data Analysis**

In big data processing, data analysis is a very critical step since it draws conclusions from new datasets that are important to a specific domain. Nowadays, trending application domain contains healthcare, insurance, transportation, social media, etc. As a part of this project, you need to identify: a) A trending topic that is interesting to you. b) Novel and influential analytic methods that are relevant to this topic. (You need to find some papers) c) New datasets that are not fully explored by data scientists. (*Better if the datasets are released within 1 year*).

Impact: you may create new data science insights that can impact the society and improve people's life. For example, by analyzing the taxi trajectory data carefully, you may help to cut down the waiting time for each user and make the trip more environment friendly.

Examples: a) A real-time sentiment analysis of twitter feeds with the NASDAQ index (i.e., analyze the correlation between tweeter feeds and hourly movements of the NASDAQ index), b) Use deep learning techniques to predict the stock price.

- **Scalable Data Science Tools and Platforms**

Scalable platforms and tools are very important for meeting the requirement of big data. Although there are many new algorithms that have been introduced, they may work on a single machine. In this project, we will develop scalable platforms and tools and make them open-sourced for public use. Here, scalability can be across multiple cores or

multiple machines. Examples include the new data mining and machine learning algorithms that was proposed recently.

Impact: you may create new platforms and tools that can address the big data challenge that have not been ever attacked. For example, by developing an easy-to-use and efficient graph processing platform (say finding the shortest path), you may help the graph analysts to improve their productivity.

Examples: a) Visualized graph analytics: the study implements different ways of graph visualization and offers an intuitive way of graph analytics. b) K-means optimization: the study optimizes K-means on multi-core machines, and later on a distributed environment. It studies different implementation strategies (e.g., early reduction) and hardware capabilities (e.g., Ethernet vs. RDMA).

- **Self-proposed Projects**

If you have some ideas that may not fit exactly what is listed above (e.g., you want to design a totally new algorithm, or new statistical measures of interestingness, etc), talk to the lecturer.

Impact: The only limit to your impact is your imagination and commitment. (by Tony Robbins).

Examples: a new machine learning platform.

4. Data Sets

You can use any datasets on the internet. We prefer to use the **new and big** data sets (*Better if the datasets are released within 1 year*). There are several possible URL links:

https://www.kaggle.com/datasets

https://cloud.google.com/public-datasets/

https://github.com/caesar0301/awesome-public-datasets

https://github.com/openimages/dataset

http://www.cs.cmu.edu/~enron/

https://webobservatory.soton.ac.uk/

……

5. More Sample Projects

Regarding sample big data projects, you can refer to the following links. However, we encourage you to think beyond that and explore significantly new ideas (for example, whether you should identify a trending topic that is interesting to you, and then apply novel and influential analytic methods that are relevant to this topic).

https://www.coursera.org/specializations/big-data

http://hadoopproject.com/big-data-projects/

https://blog.kaspersky.com/cool-big-data-projects/8186/

……

6. Project Proposal

The purpose of the project proposal is to provide background material for the work that you are to complete and to describe the actual experiments and expected results. The proposal should be sufficiently detailed so that I can understand specifically what you are going to do. An important part of the proposal is my understanding of your proposed work. If I think the project is too large I will ask you to trim it down. The following is a general outline for the proposal. *All page counts are for 11pt. font, single spaced, single column, not counting pictures, tables, or other graphics.*

1) Topic introduction (0.5 - 1 page)
2) Discussion of previous work and how it relates to your topic. (DO NOT CUT AND PASTE PICTURES OR TEXT FROM OTHER DOCUMENTS IN COMPLETING THE PREVIOUS WORK SECTION) (1-2 pages)
3) Discussion of your experimental approach (including specific tools, methodologies, experiments, etc.). Try to be as specific as possible. (1 to 2 pages)
4) Data sets to be used (Are they large and new? Justify it, 0.5-1 pages)
5) Expected results. What exactly will your result be? What will you show in a table or chart? Be specific. (1 page)
6) Project Summary (0.5 to 1 page)
7) References (should include more than just web pages)

You need to submit your project proposal to IVLE, and name of the file should be include all student IDs of your team. For example, if the group has three members, the file name should be:

[GroupID]-[Student1ID]-[Student2ID]-[Student3ID]-proposal.pdf

7. Submission Requirements

a) Final Report

You need to submit a report which is at most 20 single-columned page paper on the problem and solution(s) and what will be demonstrated. *All page counts are for 11pt. font, single spaced.*

You should compare the different solutions qualitatively as well as provide an experimental analysis. In general, the final report should contain the following contents:1. Project Introduction. This can be the same as for the proposal. 2. Methodology and experimentation. How did you perform the experimentation? This can be a revision of the proposal. 3. Discussion of results. 4. Problems encountered and lesson learnt. 5. Personal contribution (for each student written by the individual student). 6. Project summary. 7. References.

b) Code

Code can be in any programming languages (c, c++, java, matlab, R…).

c) File Name

You need to compress all the documents and code into a zip (or rar) file, and name of the file should be include all student IDs of your team. For example, if the group has three members, the file name should be:

[GroupID]-[Student1ID]- [Student2ID]-[Student3ID]-FinalReport.zip

Note: Please ensure that the code and report are complete before submission! You need to ensure that your work is recoverable by others using the code and report that you provided.

<u>For both project proposal and final report:</u>

No multiple submissions allowed: Each team should make sure that the team only submit exactly once. If a team submits two versions, we will **retain the earliest version and discard all later versions**. The team will then be grade on the earliest version. For such a reason, if you want to update your submission, you should delete your old submission first and then submit a new one.

Policy on late submission: For fairness, reports submitted after the deadline but no more than 48 hours after the deadline will still be graded, with a penalty of 20%. Namely, I will first grade the report normally, and then multiple the mark by 80% to get the final mark for that report. **Reports submitted more than 48 hours after the deadline will not be accepted and will get 0 mark.**

8. Presentation

The presentations will take place during lecture. Each group will get up to 20 minutes for the presentation. Your presentation should cover the same aspects of the project mentioned above. The final report will be due one week after oral presentation. Thus, you are encouraged to add more solid results and findings into the report, besides those presented in oral presentation.

9. Grading

Although students performing different works, to make the grading work fair and reasonable, we will evaluate your work from the following perspectives.

(a) Linguistic ability
(b) Complexity of the problem
(c) Tools and algorithms
(d) Comprehensiveness of the analysis and findings
(e) Impact of your project or findings
(f) Datasets used (Large? New?)
(g) Oral presentation and demo

10. Plagiarism

You are reminded that plagiarism is a very **SERIOUS** offense, and disciplinary action (including possibility of expulsion from the university) will be taken against any individual or team found plagiarizing. The individual or team that is being plagiarized will also be punished if it is found to have allowed the work to be plagiarized voluntarily.