Project*

Project Proposal Due: Friday, August 30, 2013

1 Overview

Your project will consist of three elements.

- Project Proposal : Due Friday, August 30, 2013
- Intermediate Report : Due Wednesday, October 23, 2013
- Project Presentation : Either December 11 or 13, 2013.

You may work in teams of up to 3 (groups of 1 or 2 are also encouraged), but the amount of work you perform will need to scale accordingly. All students will need to have clearly defined roles as demonstrated in the final presentation.

1.1 Scale of Project

The specifics of the project will be very flexible. I expect each student to explore some topic in this class and to demonstrate or extend techniques from this class within that topic. Alternatively, student may explore a specific project using multiple large data paradigms. Projects may be experimental or analytical, as long as it is clear the student has demonstrated effort in exploring a topic.

2 Project Proposal

Due Friday, August 30, 2013

Prepare an at most **2 page** document detailing your plan. It can be less than 1 page.

If you plan to attack a specific problem in a large data paradigm, (1) describe the problem you plan to attack, and (2) why you think using that paradigm will be effective. I expect either this to have not to have been done before (in published form), or if it has been done, explain how you hope to extend this technique, or what unexplored aspect of the project you hope to explore.

If you plan to compare a certain problem using multiple paradigms, (1) describe the specific problems, (2) if it has been done within any of those paradigms before, (3) and which advantages you expect from each paradigm. (You may discover you expectations are wrong, but I want you to think about it ahead of time.)

If you plan for more analysis, (1) describe the specific problem, (2) what sorts of analysis bounds you expect to find. (Again, these can be wrong - thats research - but I want you to think through whether the answer is obvious).

The project can be related to or part of a larger ongoing research project. In this case, I also expect the proposal to describe (1) what has already been done in the larger project, (2) how this fits into the larger picture, and (3) what is new and different about the class project.

2.1 Timing

The drop deadline is September 4. Your project needs to be approved by that date. To ensure I have time to look at it, it should be submitted by Friday, August 30, in class. It is encouraged for students to discuss the planned project even earlier to increase the chance of approval.

If I do not approve the project, you will have to take the class for at most 1 credit.

2.2 Topics and Data

The best projects occur on topics that students have experience with and/or are passionate about. First look within your own experience.

If you still have trouble finding a topic, I suggest exploring these resources for interesting datasets to explore. You cannot just explore a data set, you must propose to find some structure in it.

- `http://snap.stanford.edu/data/`
- `http://www.census.gov/`
- `http://data.geocomm.com/catalog/`
- `http://meta.wikimedia.org/wiki/Data_dumps`
- `http://ngrams.googlelabs.com/datasets`
- `http://kdd.ics.uci.edu/`
- `http://www.cs.utah.edu/~lifeifei/datasets.html`
- `http://www.cise.ufl.edu/research/sparse/matrices/`
- `http://webscope.sandbox.yahoo.com/`
- `http://www.google.com/publicdata/directory`
- `http://www.infochimps.com/tags/twitter`
- `http://lib.stat.cmu.edu/datasets/`

If you have an advisor, they may also be good sources of problems and data.

Data. Unless your project is completely theoretical, you will need to find a data set at the time of your project proposal.

2.3 Computing Paradigm

A reason for and central focus of the project should be to explore a massive data computing platform or paradigm. You must specify which paradigm(s) you will explore. Unless you are doing a theory project, you will need to secure an actual computing platform to perform some experiments. The following list is a guide, but is not exclusive.

- I/O: To experiment with External memory algorithms, it is recommended that you use TPIE `http://www.madalgo.au.dk/tpie/`.
- Streaming: Streaming algorithms don't require much space and computing infrastructure by design. But a good large data set will be required to really see them work as designed.
- Hadoop: This is the standard open source version of MapReduce. A small number of students may be able to experiment with Hadoop on an internal cluster.
  In the past I have secured some credits on Amazon Web Services which has an installation of Hadoop. If the project is approved, I will attempt to help secure credits for student projects.

- Spark/Shark: An emerging (and perhaps much faster) version of MapReduce and related paradigms is called Spark and Shark: `https://github.com/amplab/shark/wiki`. A small number of students may be able to experiments with Spark/Shark on an internal cluster.
- GP-GPU: There are on campus GPU optimized machine with CUDA installed. A limited number of students may be able to experiment with these machines (however, there is an entire other class on this topic).
  Unless one can explain how this will be different than what is covered in CS 6235, the instructor may not approve.
- If one is interested in distributed streaming projects, they can often be simulated. If one is interested in a non-simulated system, speak with the instructor. Perhaps we can figure out something on AWS, Emulab `http://www.emulab.net`, or elsewhere.

2.4 Content

In summary, the project proposal needs to have:

1. Clear problem statement
2. Proposed data set
3. Proposed massive data paradigm you hope to explore
4. Proposed computing resources
5. Clearly defined goals

3 Intermediate Report

Due Wednesday, October 23, 2013

The report will be at most **1 page**, single columned at 11 point or larger font. You may also talk in person with the instructor and get approval (expect this to be at least 15 minutes with questions - and may lead to requesting a report anyways).

The report should:

1. Review the problem goal, computing paradigm being explored, and data sets.
2. Describe progress towards goal, and difficulties encountered.
3. Explain what left needs to be done.

4 Project Presentation

Given on Wednesday, December 11 or Friday, December 13.

The length and timing of each presentation will depend on the number of projects. They will scale in length linearly with the number of students. Each student will get equal time and will need to explain their contribution.

4.1 Content

I expect to hear three things in your presentation:

1. What is the problem, computing paradigm, and data you worked on?

2. What were the key ideas in your approach?
3. What did you learn?

This is a great opportunity for the class to learn about a large variety of topics. If you approach this presentation as a teaching experience, you will be more likely to succeed.