CS 33001 Data-Intensive Computing Systems Project
Chien -- Spring 2013
April 29, 2013

Project Mechanics

I. What is a data-intensive computing project (Due April 12)
- Objective: understand attributes and special challenges of DI systems

II. Assessing a Data-intensive computing infrastructure (Due 4/15)
- Objective: gain familiarity with candidate DI infrastructures for projects (what are their capabilities? Limitations?)

III. Sketch Project Plan (2-3 pages), Due May 3rd
- What is the research question your project will address?
- What software infrastructure will you use?
- What hardware infrastructure?
- What data will you run experiments on?
- Briefly describe the experimental design and argue that it will answer the research question.

IV. Full Project Plan (5-10 pages), Due May 10th
- Sketch plan with improvements + new items below
- List timeline and major development and experiment milestones (with dates)
- Committed choice of infrastructure, data set, workload/application. You should have all of these in hand, on disk, etc. and have experimented with them to ensure their adequacy
- Committed hardware resources for execution of experiments
- Justify soundness for significant choices

V. Run: Project Status Report (check-in, demos, learnings, in-class presentation+discuss), ~ May 31st
- Partial demonstrations
- Final adjustments to project plan
- 10-slide presentation, including 1-slide of learnings (4-5 bullets)

VI. Final Project Presentations and Demos, ~week of June 10th (schedule tbd)
- Full presentation and demo
- Final Project Report

Project Requirements

1. All projects should explore a data-intensive computing **systems** issue.
2. All projects will involve an **empirical evaluation** of systems question – and therefore a comparison of a set of alternatives.
    a. Evaluation metrics might include capability (what can be done), performance (scalability, latency, availability), flexibility/portability (what range of systems, heterogeneity, distribution), simplicity, cost-efficiency, etc.
3. All projects should include the use of **a realistic data-intensive computing data set (or sophisticated model thereof), and application computation** across it. These can be workloads designed and implemented by others.

Your project can be motivated by a particular data-intensive computing problem or domain, but should be focused around a data-intensive computing systems question. Your project write-ups should explicitly and clearly identify the question and how the study will provide insight into the answer to the question.