Lecture 1:
  • Course Overview

Course Overview

Goal 1: Cover underlying ideas of “modern” data and analytics systems
  • systems span entire data and processing “lifecycle”
  • focus on scalability as data, processing, users, and uses increase

Note: scalable if more resources give same/similar performance gain
  • eg, handle 2x more data with 2x machines, 10x apps with 10x CPUs
  • scale-out: increase number of machines (servers) used
  • scale-up: increase “power” of a machine (CPUs, memory, disk)

Goal 2: Hands-on experience using these systems
  • built for and/or leverage distributed/cloud computing (for scalability)
  • rapid change and growth in last several years
  • which will likely continue for several more! ... one reason for Goal 1
Mix of Assignments

1. **Homework Assignments (Goal 2)** ... will involve some programming
   - Labs, exercises, and examples using the **Google Cloud Platform** (GCP)
   - The tools are specific to GCP, but largely similar in AWS and Azure

2. **Reading assignments (Goal 1)**
   - To gain understanding of technology behind the tools (how they work)
   - You’ll learn technical ideas, trade-offs from original sources (papers)
   - More transferable (to other cloud platforms), distinguish marketing hype

3. **Project**
   - Explore additional tools
   - Develop end-to-end solution
   - Or dive deeper into technical approaches

Course Structure

**Basics:** ... see syllabus for additional info
- **Course Webpage:** schedule, assignments, lecture notes
- **Piazza:** for Q & A, announcements (see invite)
- **GitHub:** for homework and reading assignments
- **Canvas:** only for tracking grades!
- **Office Hours:** MW 3:00-4:30, F 12:00-1:00, or by appointment

**Grading:** 700 points total, individual work ... again, see syllabus
- **Homework:** 340 pts 6 homeworks (40 each), 5 readings (20 each)
- **Project:** 60 pts 2nd half of semester
- **Quizzes:** 60 pts 6 quizzes (10 each), 1 make-up quiz
- **Exams:** 200 pts 2 midsemester (50 each), 1 final (100)
- **Attendance:** 40 pts ≈ 40 lectures (1 each)
Course Structure (cont)

Note: Must score 60% on homework + project and 60% on exams + quizzes

Late Policy: 25% penalty up to 2 weeks after due date
- must notify me via email for late work to be graded
- will be using hard deadlines (see next note)

Note on Assignments: Due dates are evenly spread out ...
- for each, you’ll have two weeks (reading and homework overlap)
- plan your time wisely ... both will take many hours to complete
- expectation is that you will be working on each for full two weeks
- ask questions early to get help (my availability decreases)

Course Topics

What do we mean by “Big Data Analytics”? (*)
1. What do we mean by analytics?
2. What do we mean by big data? ... to be continued Friday
3. What will we be covering (high-level)? ... to be continued Friday

(*) Notes:
- A better course title might be “Data Engineering for Data Science”
- This is also same basic course as “Data Intensive Systems”
- Our emphasis is more on data systems than on new analytical techniques
Analytics

1. **Four basic types of analytics:**

   **Descriptive:** understand historical and/or current trends
   - total energy usage by month over last two years

   **Diagnostic:** understand *why* historical and/or current trends occurred
   - relationships between monthly energy use and utilization

   **Predictive:** guess future trends (by generating a model)
   - energy use for next year by month given expected utilization

   **Proscriptive:** use predictive models to take action or make decision
   - using the model, delay requests to reduce demand on local grid energy
   - something similar done in Google data centers (see this article)

Summary – Things to Know

- what is meant by scalability
- difference between scale-out and scale-up
- four basic types of data analytics