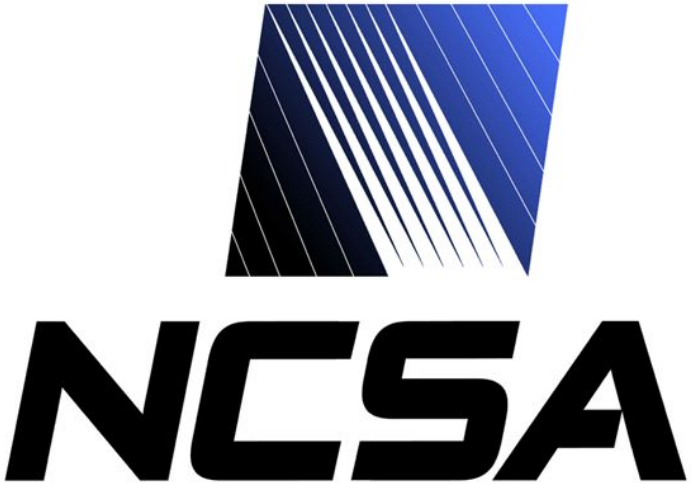
CS 199
Applied Cloud Computing

Prof. Robert J. Brunner

Quinn Jarrell
Tyler Kim



Illinois Data Science Initiative

Course Info

- Led by Professor Brunner
- Instructors
  - Quinn Jarrell - qjarrel2@illinois.edu
  - Tyler Kim - tkim139@illinois.edu
- Course Staff
  - Saisaket Potluri
  - Sameet Sapra
  - Bhuvan Venkatesh
  - Benjamin Congdon
  - Drake Eidukas

Course Info

- Moodle - learn.illinois.edu
  - Grades
  - Syllabus
  - Lectures
- Slack
  - Questions
  - Group chats
  - Majority of communications

Times

This is the only meeting at 5:30.

All other lectures will be on Thursday from 7:00 PM to 7:50 PM in **Siebel 1131**

Office hours are every Thursday from 8 - 8:50 pm right after lecture

Course Content

- What are we going to cover?
  - Hadoop
  - Spark
  - HDFS (Hadoop Distributed File System)
  - NoSQL Databases
  - Graph processing
- This is **applied** cloud computing, not cloud computing
  - Very little theory

Grading

Attendance	10%
Labs	30%
Technical Report	60%

Tentative Schedule

- First few weeks will be learning how Hadoop works and how to write programs to run on Hadoop
- Next few weeks will be introducing how to handle large datasets and what we can do with them
- The middle few weeks will depend on how far we get in the prior weeks
- The last few weeks will be dedicated to running a project using Hadoop/Spark/NoSQL and writing a technical report. This will be the majority of your grade.

Labs

Lab 1: Intro to MapReduce

Lab 2: MapReduce

Lab 3: Data Cleaning using Hadoop

Lab 4: Data Mining

Lab 5: Intro to Spark

- Part 2: Advanced Spark Lab

Lab 6: Spark Query Language

Technical Reports + Course Goal

The goal of this course beyond teaching you cloud computing technologies is to write technical reports.

- You will work in groups writing one or more technical reports with the assistance of one of the RAs
- These can vary widely but they should present something non trivial which others can build upon.

Registration

http://bit.ly/cs199register

&&

http://bit.ly/cs199ncsa

Do both

What is cloud computing?

What is cloud computing?

- It is NOT an actual cloud

- Computing using many computers
- Abstracts away the physical hardware
- Typically made of up of a bunch of virtual machines
- Instead of one super powerful computer, use many weak computers
- Most of the time you will not know where your code is running physically

Virtual machines (VMs)

What’s a virtual machine?

Virtual machines (VMs)

Virtual machines are simulated computers

- Allows us to simulate multiple computers on the same physical computer
- Each virtual computer shares the resources of the physical computer
- Lets you run an operating system on top of another operating system

Virtual machines (VMs)

Advantages?

Disadvantages?

Virtual machines (VMs)

Advantages

- Simulated computers are cheap to restart and destroy
- Each simulated computer is identical!
  - If it runs on one VM, you can duplicate the VM and run hundreds of identical VMs
- They provide security for running untrusted code
- They can be used to scale horizontally by running duplicates of your code on many weak machines.

Disadvantages

- Primary problem is computing is at lowest common denominator (of the hardware available, thus no special purpose GPUs, etc.).
- Slower than physical computers
  - One more layer of software between your code and the transistors

NCSA Nebula VMs

- You are going to run programs on part of the NCSA Nebula cluster
- NCSA is providing a compute cluster for us to run hadoop on.
- It will work similarly to EWS but most of the time your code will run on more than one computer in the cluster

VirtualBox

- For the first few weeks you will be running your labs on your own computer.
- VirtualBox is a piece of software which runs virtual machines
- The virtual machine we will give you has Hadoop installed already

Goal for this week

- Download VirtualBox and install it
- Download the virtual machine image you will use to run lab 1
  - http://bit.ly/CS199Lab1
- Register for NCSA access
- Register for this class!