CSC 369: Introduction to Distributed Computing
Spring 2020
Course Syllabus

April 5, 2020

Instructor: Alexander Dekhtyar
email: dekhtyar@csc.calpoly.edu
office: 14-210

Lecture	MWF	12:10 – 1:00pm	Zoom and Slack
Lab	MWF	2:10 – 3:00pm	Zoom and Slack

Office Hours

	When	Where
Monday	1:10pm - 2:00pm	Zoom and Slack
Tuesday	1:10pm - 3:00pm	Zoom and Slack
Wednesday	9:10am - 10:00am	Zoom and Slack

Additional appointments can be scheduled by emailing the instructor at *dekhtyar@calpoly.edu*.

Overview

In this course we study the design and implementation of a variety of data processing algorithms on distributed computing frameworks.

Mode of Teaching

All Spring 2020 courses are taught in a "distance" mode. CSC 369 will be taught in a **synchronous distance** mode. This means:

- We will have live lectures scheduled during the lecture part of the course and delivered via Zoom with Slack support.
- The lectures will be recorded for the benefit of the students who are unable to join live lectures, and will be made available to everyone.

- Lab periods will be used for structured activities or for work on weekly lab assignments and will take place over Zoom with Slack support.
- Office hours are all via Zoom with Slack support. Email also works.

Disclaimer

As everyone else, I had to figure out how to shift this course into a distance mode on a very short notice, and with limited resources. This may backfire. I reserve the right to change the mode of operation if the initial attempts at teaching the course prove ineffective.

Texbook

The course does not have a required textbook, primarily because no book known to the intructor has exactly the content covered in this course. However, there is an O'Reiley book for each component of our class. The books below are all recommended reading.

- Donald Miner, Adam Shook, *MapReduce Design Pattern: Building Effective Algorithms and Analytics for Hadoop and Other Systems*, O'Reiley Media, 1st Edition, 2012, ISBN: 978-1449327170.
- Mahmoud Parsian, *Data Algorithms: Recipes for Scaling Up With Hadoop and Spark*, O'Reiley Media, 2015, ISBN: 978-1491906187.
- Christina Chodorow, *MongoDB: The Definitive Guide*, O'Reiley Media, 2013, ISBN: 978-144924468
- Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, *Learning Spark: Lightining-Fast Big Data Analysis*, Packt, 2015, ISBN: 978-1449358624
- Tomasz Drabas, Denny Lee, *Learning PySpark*, O'Reiley Media, 2017, ISBN-13: 978-1786463708

Topics

The following will be covered in the course.

No.	Topic	Duration (weeks)
1.	Introduction: Distributed Systems and Computations	1
2.	Key-Value Relationships / MongoDB	2
3.	MapReduce/ Hadoop	3
4.	Resilient Distributed Datasets and Spark	3

Most of the topics will be covered in the order specified above, but some variations are possible during the course.

Grading

Homeworks	0–5%
Labs	50 – 60%
Exams/Written Assessments	35 – 50%

This quarter, the most difficult part is to figure out how to evaluate your work. The Lab assignments (see below) will remain roughly the same as before, rearranged to largely be for individual, rather than pair programming work.

The exams, though, will require a reassessment. I am lowering their impact on the class. The format of the exams will also change. More on this below.

Course Policies

Exams

The in-person version of the course traditionally had two paper-and-pencil exams: either two midterms, or one midterm and one final exam.

The exact exams that were administered in previous versions of this course are infeasible in a distance environment, the questions assumed closed-book, closed-notes format, which cannot be assumedd when running exams in a distance mode.

In this course we will have three "written assessments/exams" - one per each of the key topics we are studying: distributed DBMS, MapReduce, Resilient Distributed Datasets. At the moment, it looks like all three assessments will be done in a form of a lab exam - offered during the scheduled class time (lecture and/or lab period).

Additionally, we may have 3-4 short "paper-and-pencil" (more like Google Forms) quizzes, and a final exam that consits of both a paper-and-pencil and a programming part. (Alternatively, our third assessment can be moved to the finals week to give us more time in class).

Homeworks, Labs

The course will have 6-8 lab assignments, designed to let you test in practice what we have learned in class. Each lab assignment will span multiple lab sessions (typically 2 or 3). Due dates/times will be explicitly provided for each assignment. You are welcome to work on the lab assignments outside the lab hours, however, lab period attendance may be helpful as I will be there to answer questions.

Most labs will be individual. Some time in the middle of the quarter we may try pair programming assignments. If they work well, we may continue pair programming labs for the rest of the quarter - as most of my assignments from prior quarters are pair programming assignments.

In a typical course, I use paper-and-pencil homeworks as exam study guides. This may or may not be appropriate for this course, depending on the exact

nature of written assessments. Still, paper-and-pencil homeworks may be assigned - primarily to make sure you have seen the types of problems that might pop up on the quizzes and paper-and-pencil exam portions. I usually collect homeworks, but do not grade them - the grades are given for completion.

Late Submissions

Due to the nature of the course, I will try to be flexible with the due dates. However, please understand the following:

- When selecting the initial due dates, I **already** am trying to be as permissive as possible about the time to complete the assignment.
- I typically have a grace period of anywhere from two to twelve hours past the submission deadline during which any late submissions are not really considered late.
- If you need an extension, please talk to me **before** the deadline passes, not after.
- Deadlines are there ensure that you can switch from work on one assignment to working on the next. Most of the lab assignments in this class are independent of each other - failing to complete one in any way disadvantages your ability to complete the next one. Therefore, if you are having issues with a specific lab, or if you missed a portion of a class – my advice is often cut losses and move on to the next assignment.
- I **do award** partial credit for pretty much everything.

Communication

For this quarter, communication is **especially important**.

Mailing List

The class has an official mailing list:

csc-369-01-2204@calpoly.edu

All students enrolled in the class are automatically subscribed to the mailing list.

For the first two weeks, I am also maintaining a mailing list for the waitlist. All waitlisted students are receiving all communications and access to the course resources.

Mailing list is my primary means of getting information to you in between classes. It may be supplanted by Slack and other means, but I am old-school, so in times of need I revert to tired and true solutions.

Zoom

We will be using Zoom for lectures, labs, and office hours. Thtwo Zoom conferences related to this course:

- The lecture/lab conference call, scheduled for MWF 12pm — 3pm. This conference call will also be used for Monday's and Friday's office hour in between the lecture and the lab period. This conference call requires registration. Registration information has been sent to you.
- The Tuesday office hours conference call. You have received the Zoom link to it.

Slack

I have created the

https://calpolycsc369.slack.com

Slack organization to help run the course. Our Slack includes channels for communication with me during lectures, lab periods, and office hours. It also includes channels for discussions of each major technology covered in the course, as well as additional channels.

You are welcome to communicate with me and among yourselves via Slack. Please keep off-topic discussions to #random and #general. You are also welcome to use private channels to contact me one-on-one.

I will keep the Slack application open during lectures, office hours and labs.

Web Pages

We will have two web pages: a "regular" static web page for most of the traditional course content, and a Canvas page to distribute course materials (mostly recordings of lectures) that I want to restrict access to.

The statis class web page can be found at

http://www.csc.calpoly.edu/∼dekhtyar/369-Spring2020

Through this page you will be able to access all class handouts including homeworks, project information and lecture notes (should the latter be written).

The Canvas page will be made available to you via your MyCalPoly portal.

Academic Integrity

University Policies

Cal Poly's Academic Integrity policies are found at

http://www.academicprograms.calpoly.edu/academicpolicies/Cheating.htm

In particular, these policies define *cheating* as (684.1)

> "*. . . obtaining or attempting to obtain, or aiding another to obtain credit for work, or any improvement in evaluation of performance, by any dishonest or deceptive means. Cheating includes, but is not limited to: lying; copying from another's test or examination; discussion of answers or questions on an examination or test, unless such discussion is specifically authorized by the instructor; taking or receiving copies of an exam without the permission of the instructor; using or displaying notes, "cheat sheets," or other information devices inappropriate to the prescribed test conditions; allowing someone other than the officially enrolled student to represent same.*"

Plagiarism, per University policies is defined as (684.3)

> "... the act of using the ideas or work of another person or persons as if they were one's own without giving proper credit to the source. Such an act is not plagiarism if it is ascertained that the ideas were arrived through independent reasoning or logic or where the thought or idea is common knowledge. Acknowledgement of an original author or source must be made through appropriate references; i.e., quotation marks, footnotes, or commentary."

University policies state (684.2): "Cheating requires an "F" course grade and further attendance in the course is prohibited." (appeal process is also outlined, see the web site above for details.). Plagiarism, per university policies (684.4) can be treated as a form of cheating, although a level of discretion is given to the instructor, allowing the instuctor to determine the causes of plagiarism and effect other means of remedy. It is the obligation of the instructor to inform the student that a penalty is being assessed in such cases.

Course Policies

All homeworks are to be completed by each student **individually**. Lab assignments are to be completed by the appropriate units (individual, pair, group), and no code/solution-sharing between units is permitted. Students are encouraged to discuss class content among themselves but NOT in a manner that constitutes plagiarism and cheating as defined above (e.g., you can solve together a problem from the textbook that had not been assigned in the homework, but you should solve assigned problems individually).