San José State University
Department of Computer Science
CS185C, Solving Big Data Problems, Section 03, Spring 2017

Course and Contact Information

Instructor:	James Casaletto
Office Location:	MH 422
Telephone:	(408) (394-5748)
Email:	james.casaletto@sjsu.edu
Office Hours:	Tuesday Thursday 7:00-7:30
Class Days/Time:	Tuesday Thursday 7:30-8:45
Classroom:	MH 422
Prerequisites:	CS146 or equivalent Java programming experience

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on the Canvas learning management system course website. You are responsible for regularly checking with the messaging system through MySJSU (or other communication system as indicated by the instructor) to learn of any updates.

Course Description

This course is a comprehensive overview of solving big data problems using Apache Hadoop and is comprised of three main parts. The first part of this course explores the core of Apache Hadoop. The second part of the course explores the Apache Hadoop ecosystem. The third part of the course explores topics in machine learning using Apache Spark. All programming assignments and coding examples are in Java.

Course Goals

The goal of this course is to gain a working knowledge of solving big data problems from end to end using Apache Hadoop and its ecosystem.

Course Learning Outcomes (CLO)

Upon successful completion of this course, students will be able to:

1. Use HDFS to store and retrieve data at scale
2. Write MapReduce programs in Java to transform, filter, and enrich data at scale.
3. Use Hadoop ecosystem components to ingest, transform, store, and analyze big data.
4. Use machine learning algorithms to derive insights from big data

Required Texts/Readings

Textbook

No textbook is required for this class. Optionally, students may use Hadoop, The Definitive Guide (4th edition) from O'Reilly Publishing as a supplement in learning Hadoop.

Other Readings

A list of other readings will be provided on the CANVAS page associated with this class.

Other equipment / material

Students are required to have a 64-bit laptop running either Windows, MacOS, or Linux with at least 8GB memory installed, 2 CPU cores, and approximately 30GB disk space free.

Course Requirements and Assignments

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on. More details about student workload can be found in University Policy S12-3 at http://www.sjsu.edu/senate/docs/S12-3.pdf.

- 1 x midterm exam
- 2 x individual labs
- 1 x team project
- 1 x final exam

Final Examination or Evaluation

The final exam will cover all the material discussed in class after the midterm exam. The format for the final exam will be all multiple-choice.

Grading Information

Determination of Grades

The final grade will be calculated as follows:
0.3 * (midterm exam score) + 0.3 * (average score of 2 labs) + 0.3 * (final exam score) 0.1 * (project score)

98-100 => A+, 93-97 => A, 90-92 => A-
87-89 => B+, 83-86 => B, 80-82 => B-
77-79 => C+, 73-76 => C, 70-72 => C-
67-69 => D+, 63-66 => D, 60-62 => D-
below 60 => F

- There will not be any extra credit given in this class
- 50% will be deducted for any assignment turned after the day after it's due – no credit is given after the second day.

Grading Information

"Passage of the Writing Skills Test (WST) or ENGL/LLD 100A with a C or better (C- not accepted), and completion of Core General Education are prerequisite to all SJSU Studies courses. Completion of, or co-registration in, 100W is strongly recommended. A minimum aggregate GPA of 2.0 in GE Areas R, S, & V shall be required of all students."

Classroom Protocol

Class begins promptly at 7:30 and end abruptly at 8:45. Please silence all cell phones during class. Your active participation in the lecture discussions is greatly encouraged.

University Policies

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of Graduate and Undergraduate Programs' Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/"

CS185C / Solving Big Data Problems, Sp17, Course Schedule

The following schedule is subject to change. All changes will be verbally communicated in class as well as messaged using the CANVAS system.

Course Schedule

Week	Date	Topics, Readings, Assignments, Deadlines
1	26.jan	Introduction to this course
2	31.jan	Introduction to big data
2	2.feb	Introduction to distributed file systems I HDFS
3	7.feb	Introduction to distributed file systems II: NFS, MapR-FS
3	9.feb	Building and managing a 1-node virtual machine Hadoop cluster
4	14.feb	MapReduce programming I: conceptual overview (lab 1 assigned)
4	16.feb	MapReduce programming II: writing a MapReduce program
5	21.feb	MapReduce programming III: using the API
5	23.feb	Introduction to the Hadoop ecosystem I: overview
6	28.feb	Introduction to the Hadoop ecosystem II: use cases
6	2.mar	Apache Spark I: RDD, transformations, actions
7	7.mar	Apache Spark II: streaming (lab 1 due)
7	9.mar	Apache Spark III: sql
8	14.mar	Apache Kafka I: consumers, producers, topics, keys, values, partitioning
8	16.mar	Apache Kafka II: cursor management, using offsets
9	21.mar	Apache Kafka III: using Apache Spark with Apache Kafka
9	23.mar	Midterm exam
10	28.mar	No class (spring break)
10	30.mar	No class (spring break)
11	4.apr	Introduction to data science I: machine learning (lab 2 assigned)
11	6.apr	Introduction to data science II: use cases

Week	Date	Topics, Readings, Assignments, Deadlines
12	11.apr	Linear regression I: overview, use cases, demo
12	13.apr	Linear regression II: algorithm
13	18.apr	Naïve Bayes I: overview, use cases, demo (group project proposal due)
13	20.apr	Naïve Bayes II: algorithm
14	25.apr	Decision trees / random forests I: overview, use cases, demo (lab 2 due)
14	27.apr	Decision trees / random forests II: algorithm
15	2.may	K-means I: overview, use cases, demo
15	4.may	K-means II: algorithm
16	9.may	Group project presentations I
16	11.may	Group project presentations II
17	16.may	Group project presentations III
18	19.may	Final exam (7:15 – 9:30) in MH 422