San José State University
Department of Computer Science
CS185C, Solving Big Data Problems, Section 03, Fall 2016

Course and Contact Information

Instructor:	James Casaletto
Office Location:	DH 2222
Telephone:	(408) (394-5748)
Email:	james.casaletto@sjsu.edu

Office Hours:	Tuesday Thursday 18:30-19:15
Class Days/Time:	Tuesday Thursday 19:30-20:45
Classroom:	SCI 311
Prerequisites:	CS146 or equivalent 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. Install, configure, and maintain a 1-node Hadoop cluster in a virtual machine running on your laptop
2. Write MapReduce programs in Java to transform, filter, and enrich data at scale.
3. Use Hadoop ecosystem components to ingest, transform, store, visualize, 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 mostly multiple-choice with a few short-answer questions.

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
- 25% per day will be deducted from assignments submitted late

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 19:30 and end abruptly at 20:45. 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, Fa16, 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	25.aug	Introduction to this course; introduction to big data
1	30.aug	Introduction to Hadoop; creating a 1-node virtual cluster
2	1.sep	Introduction to distributed file systems, HDFS, MapR-FS (I)
2	6.sep	Introduction to distributed file systems, HDFS, MapR-FS (II)
3	8.sep	Introduction to MapReduce programming
3	13.sep	Writing a MapReduce program in Java
4	15.sep	Managing and testing MapReduce programs
4	20.sep	Writing a YARN application
5	22.sep	Introduction to the Hadoop ecosystem
5	27.sep	Introduction to Apache Hive
6	29.sep	Introduction to Apache Drill
6	4.oct	Introduction to Apache HBase (guest lecture); lab 1 due
7	6.oct	Introduction to Apache Kafka
7	11.oct	Introduction to Apache Spark
8	13.oct	Programming in Spark
8	18.oct	Spark streaming
9	20.oct	Midterm exam
9	25.oct	Introduction to Hunk (Splunk on Hadoop)
10	27.oct	Introduction to ELK (elasticsearch, logstash, and kibana)

Week	Date	Topics, Readings, Assignments, Deadlines
10	1.nov	Introduction to data science
11	3.nov	Introduction to recommendation engines (guest lecture); lab 2 due
11	8.nov	Introduction to decision trees
12	10.nov	Introduction to Naïve Bayes
12	15.nov	Introduction to K-nearest neighbors
13	17.nov	Introduction to logistic regression
13	22.nov	Introduction to linear regression
14	24.nov	No class (Thanksgiving holiday)
14	29.nov	Introduction to K-means
15	1.dec	Introduction to Gaussian Mixture Model
15	6.dec	Group project presentations I
16	8.dec	Group project presentations II
Final Exam	20.dec	SCI 311 (19:45-22:00)