San José State University
Department of Computer Science
CS185C, Solving Big Data Problems, Section 02, Spring 2018

Course and Contact Information

Instructor:	James Casaletto
Office Location:	DH 282
Telephone:	(408) 394-5748
Email:	james.casaletto@sjsu.edu
Office Hours:	Tuesday Thursday 09:00-09:45
Class Days/Time:	Tuesday Thursday 7:30-8:45
Classroom:	SCI 311
Prerequisites:	CS146 or equivalent Java programming experience Basic Linux administration 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 four 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. The fourth part of the course is the group project presentations. All programming assignments and coding examples are in Java.

Course Goals

The goal of this course is to develop 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. Build, maintain, and use a Hadoop virtual machine sandbox environment
2. Use HDFS to store and retrieve data at scale
3. Write MapReduce programs in Java to transform, filter, and enrich data at scale.

4. Use Hadoop ecosystem components to ingest, transform, store, and analyze big data.
5. Use machine learning algorithms to derive insights from big data

Required Texts/Readings

Textbook

No textbook is required for this class.

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. All labs are to be developed using a virtual machine installed on the laptop with these and other requirements.

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.

- 4 x individual labs
- 1 x midterm exam
- 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.

Grading Information

Note that this course may be used as an elective for graduate students in the computer science graduate program at SJSU.

Determination of Grades

The final grade will be calculated as follows:
0.2 * (midterm exam score) + 0.4 * (average score of labs) + 0.2 * (final exam score) 0.2 * (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 provided
- 10% will be deducted for each day that any lab is turned in after it's due.

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 ends 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, Sp18, 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.jan	Introduction to this course
2	30.jan	Introduction to big data
2	1.feb	Building and managing a 1-node virtual machine running Linux
3	6.feb	Building and managing a 1-node Hadoop cluster on virtual machine
3	8.feb	Developing big data applications using Eclipse and Maven (lab 1 assigned)
4	13.feb	Introduction to distributed file systems I: HDFS architecture
4	15.feb	Introduction to distributed file systems II: HDFS CLI and API
5	20.feb	MapReduce programming I: conceptual overview
5	22.feb	MapReduce programming II: writing a MapReduce program (lab 2 assigned)
6	27.feb	Introduction to the Hadoop ecosystem: overview of Spark and Kafka
6	1.mar	Apache Spark I: RDD, data sets, data frames
7	6.mar	Apache Spark II: transformations (map and reduce variants) and actions
7	8.mar	Apache Spark III: developing Spark applications (lab 3 assigned)
8	13.mar	Apache Kafka I: consumers, producers, topics, keys, values
8	15.mar	Apache Kafka II: partitioning/keys, cursor management/offsets
9	20.mar	Putting it all together: using Spark with Kafka on Hadoop
9	22.mar	Midterm exam

Week	Date	Topics, Readings, Assignments, Deadlines
10	27.mar	No class – spring break
10	29.mar	No class – spring break
11	3.apr	Introduction to machine learning and data science (lab 4 assigned)
11	5.apr	Naïve Bayes I: algorithm
12	10.apr	Naïve Bayes II: use cases
12	12.apr	Decision tree / random forest I: algorithm (group project proposal due)
13	17.apr	Decision trees / random forest II: use cases
13	19.apr	K-means and GMM clustering I: algorithm
14	24.apr	K-means and GMM clustering II: use cases
14	26.apr	Group project presentations I
15	1.may	Group project presentations II
15	3.may	Group project presentations III
16	8.may	Group project presentations IV
16	10.may	Final review
17		Final exam