San José State University
Engineering/Computer Science
CS286, Solving Big Data Problems, Section A, Spring, 2016

Course and Contact Information

Instructor: James Casaletto

Office Location: Duncan Hall Room 282

Telephone: (408) 394-5748

Email: james.casaletto@sjsu.edu

Office Hours: Tuesdays/Thursdays 18:30 – 19:15

Class Days/Time: Tuesdays/Thursday 19:30 – 20:45

Classroom: SCI 311

Prerequisites: **Java Programming (CS 146)**

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 machine learning topics using Apache Spark. All programming assignments and coding examples are in Java.

Learning Outcomes and Course Goals

Course Learning Outcomes (CLO)

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

1. Install a 1-node Hadoop cluster in a virtual machine running on your laptop
2. Write MapReduce programs in Java to transform big data
3. Use Hadoop ecosystem components to ingest, transform, and store big data.
4. Use machine learning algorithms to analyze big data

Required Texts/Readings

Textbook

None required. Optional book for the class is Hadoop, The Definitive Guide (4th edition) by O'Reilly Publishing

Other Readings

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

Other equipment / material requirements

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 in-class exams 30%

The exam is comprised of multiple choice and short answer and covers HDFS, MapReduce, and the ecosystem

· 2 x individual labs 30%

The labs include a Java MapReduce programming assignment and a machine learning programming assignment

· 1 x team project 10%

Teams of 3 people will create an end-to-end solution using the Hadoop tools discussed in the course.

· 1 x final exam 30%

The exam is comprised of multiple choice and short answer.

NOTE that University policy F69-24 at http://www.sjsu.edu/senate/docs/F69-24.pdf states that "Students should attend all meetings of their classes, not only because they are responsible for material discussed therein, but because active participation is frequently essential to insure maximum benefit for all members of the class. Attendance per se shall not be used as a criterion for grading."

Grading Policy

A = 91-100 / B = 81-90 / C = 71-80 / D = 61-70 / < 61 = F

Your grade is calculated as the weighted average of the in-class exam (30%), individual labs (30%), a team project (10%), and the final exam (30%).

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

General Expectations, Rights and Responsibilities of the Student

As members of the academic community, students accept both the rights and responsibilities incumbent upon all members of the institution. Students are encouraged to familiarize themselves with SJSU's policies and practices pertaining to the procedures to follow if and when questions or concerns about a class arises. See University Policy S90–5 at http://www.sjsu.edu/senate/docs/S90-5.pdf. More detailed information on a variety of related topics is available in the SJSU catalog, at http://info.sjsu.edu/web-dbgen/narr/catalog/rec-12234.12506.html. In general, it is recommended that students begin by seeking clarification or discussing concerns with their instructor. If such conversation is not possible, or if it does not serve to address the issue, it is recommended that the student contact the Department Chair as a next step.

Dropping and Adding

Students are responsible for understanding the policies and procedures about add/drop, grade forgiveness, etc. Refer to the current semester's Catalog Policies section at http://info.sjsu.edu/static/catalog/policies.html. Add/drop deadlines can be found on the current academic year calendars document on the Academic Calendars webpage at http://www.sjsu.edu/provost/services/academic_calendars/. The Late Drop Policy is available at http://www.sjsu.edu/aars/policies/latedrops/policy/**.** Students should be aware of the current deadlines and penalties for dropping classes.

Information about the latest changes and news is available at the Advising Hub at http://www.sjsu.edu/advising/.

Consent for Recording of Class and Public Sharing of Instructor Material

University Policy S12-7, http://www.sjsu.edu/senate/docs/S12-7.pdf, requires students to obtain instructor's permission to record the course and the following items to be included in the syllabus:

- "Common courtesy and professional behavior dictate that you notify someone when you are recording him/her. You must obtain the instructor's permission to make audio or video recordings in this class. Such permission allows the recordings to be used for your private, study purposes only. The recordings are the intellectual property of the instructor; you have not been given any rights to reproduce or distribute the material."
    - It is suggested that the greensheet include the instructor's process for granting permission, whether in writing or orally and whether for the whole semester or on a class by class basis.
    - In classes where active participation of students or guests may be on the recording, permission of those students or guests should be obtained as well.
- "Course material developed by the instructor is the intellectual property of the instructor and cannot be shared publicly without his/her approval. You may not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, or homework solutions without instructor consent."

Academic integrity

Your commitment, as a student, to learning is evidenced by your enrollment at San Jose State University. The University Academic Integrity Policy S07-2 at http://www.sjsu.edu/senate/docs/S07-2.pdf requires you to be honest in all your academic course work. Faculty members are required to report all infractions to the office of Student Conduct and Ethical Development. The Student Conduct and Ethical Development website is available at http://www.sjsu.edu/studentconduct/.

Campus Policy in Compliance with the American Disabilities Act

If you need course adaptations or accommodations because of a disability, or if you need to make special arrangements in case the building must be evacuated, please make an appointment with me as soon as possible, or see me during office hours. Presidential Directive 97-03 at http://www.sjsu.edu/president/docs/directives/PD_1997-03.pdf requires that students with disabilities requesting accommodations must register with the Accessible Education Center (AEC) at http://www.sjsu.edu/aec to establish a record of their disability.

Accommodation to Students' Religious Holidays (Optional)

San José State University shall provide accommodation on any graded class work or activities for students wishing to observe religious holidays when such observances require students to be absent from class. It is the responsibility of the student to inform the instructor, in writing, about such holidays before the add deadline at the start of each semester. If such holidays occur before the add deadline, the student must notify the instructor, in writing, at least three days before the date that he/she will be absent. It is the responsibility of the instructor to make every reasonable effort to honor the student request without penalty, and of the student to make up the work missed. See University Policy S14-7 at http://www.sjsu.edu/senate/docs/S14-7.pdf.

CS286 / Solving Big Data Problems, Spring 2016, Course Schedule

The schedule is subject to change. It will be posted on the CANVAS web site.

Course Schedule

Week	Date	Topics, Readings, Assignments, Deadlines
1		
1	1/28	Introduction to big data
2	2/2	Installation and configuration of MapR distribution of Hadoop
2	2/4	Introduction to the Hadoop core
3	2/9	Using HDFS, MapR-FS, and NFS
3	2/11	MapReduce programming in Java I
4	2/16	MapReduce programming in Java II
4	2/18	MapReduce programming in Java III
5	2/23	Introduction to the Hadoop ecosystem
5	2/25	Using Sqoop, Flume, and Kafka

Week	Date	Topics, Readings, Assignments, Deadlines
6	3/1	Using Pig, Hive, and Drill
6	3/3	Using Spark I (RDD); lab 1 due
7	3/8	Using Spark II (Streaming)
7	3/10	Using Spark III (SQL and GraphX)
8	3/15	Building batch-based solutions with Hadoop
8	3/17	Building streaming-based solutions with Hadoop
9	3/22	In-class exam covering HDFS/MapR-FS, MapReduce, and Hadoop ecosystem
9	3/24	Introduction to data science
10	3/29	No class (spring break)
10	3/31	No class (spring break)
11	4/5	Introduction to machine learning
11	4/7	Introduction to recommendation engines
12	4/12	Using Naïve Bayes for classification
12	4/14	Using decision trees for classification
13	4/19	Using K-nearest neighbors for classification
13	4/21	Using linear and logistic regression
14	4/26	Using K-means for clustering
14	4/28	Using Principal Components Analysis for dimensionality reduction
15	5/3	Understanding neural networks
15	5/5	Understanding PageRank
16	5/10	Project presentations I
16	5/12	Project presentations II; lab 2 due
Final Exam	5/18-5/24	MH422 from 19:45 to 22:00