San José State University
College of Science / Department of Computer Science
CS267 Topics in Database Systems, Spring 2020

Course and Contact Information

Instructor:	Dr. Mike Wu
Office Location:	MacQuarrie Hall 211
Email:	Ching-seh.Wu@sjsu.edu
Office Hours:	Tuesday 2:30~3:30pm and Thursday 3:00~4:00pm **(Please drop me an email with time info and subject.)**
Class Days/Time:	TuTh 4:30-5:45pm
Class Room:	MacQuarrie Hall 233
Prerequisites:	CS 157B Database Management Systems II (with a grade of "C-" or better)

Faculty Web Page and OneSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found at **Canvas** of SJSU One. **You are responsible for regularly checking with the email system and Canvas through One.SJSU at http://one.sjsu.edu to learn of any updates.**

Course Description

General: Advanced topics in the area of database and information systems. Content differs in each offering. Possible topics include though not restricted to: Data Mining, Distributed Databases and Transaction Processing. (This description is from course catalog of CS Department Website)

This semester, topics include the following (time permits):

- Introduction to Big Data
- Big Data Mining
- Large-scale data processing platforms.
- HDFS
- Apache Hadoop architecture
- MapReduce model
- Scalable algorithms used to extract knowledge from Big data.
- Advanced scalable data analytics platforms.
- Stream data processing
- Google Big Table Platform
- Big data: NoSQL data modeling.
- Big data analytics using machine learning

Course Learning Outcomes (CLO)

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

- Gain knowledge and key concepts, algorithms, techniques related to Big Data.
- Familiar with Mining data streams.
- Familiar with Apache Hadoop architecture, and Map-Reduce.
- Gain hands-on experience to develop and implement Big Data analytical project.
- Use scalable algorithms to extract knowledge from Big data
- Become familiar with the different data models used by NoSQL Big Data platforms.
- Become familiar with tradeoffs between SQL and NoSQL: Data model, Query language, guarantees provided.
- Gain experience and skill in big data analytics research project using machine learning models

Required Texts/Readings

No Required Textbooks

Optional Textbooks

Mining of Massive Datasets, Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman, Cambridge University Press, ISBN: 978-1-107-01535-7.
Free download copy: **http://www.mmds.org**

Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale (Addison-wesley Data & Analytics) 1st Edition: Ofer Mendelevitch, Casey Stella, and Douglas Eadline, ©2017 |Addison-Wesley Professional

Hadoop: The Definitive Guide, Tom White, O'Reilly, 4rd Edition, 2015, ISBN: 978-149-190-1687,
Free download copy http://grut-computing.com/HadoopBook.pdf

Hadoop MapReduce Cookbook. Recipes for analyzing large and complex datasets with Hadoop MapReduce. Srinath Perera. Thilina Gunarathne. BIRMINGHAM
Free download copy:
http://barbie.uta.edu/~jli/Resources/MapReduce&Hadoop/Hadoop%20MapReduce%20Cookbook.pdf

Cassandra: The Definitive Guide, Eben Hewitt, O'Reilly,
http://www.gocit.vn/files/Cassandra.The.Definitive.Guide-www.gocit.vn.pdf

Online Reading Materials and Tools:
Apache Hadoop: http://hadoop.apache.org/
Apache Spark: https://spark.apache.org/
Hadoop HDFS: http://wiki.apache.org/hadoop/HDFS
MapReduce Tutorial: http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
Mahout - Scalable Data Mining Algorithms Over Hadoop: http://mahout.apache.org/
Apache Hive Home Page: http://hive.apache.org/
Apache Pig Home Page: http://pig.apache.org/
Hbase Home Page: http://hbase.apache.org/
Cassandera Home Page: http://cassandra.apache.org/
CouchDB Home Page: http://couchdb.apache.org/
MongoDB Home Page: https://www.mongodb.com/

Course Requirements and Assignments

Assignments

You are expected to learn all of the material presented in the lectures. Assignments include written and programming. Assignments must be turned in on time; late submission will not be accepted with the exception of medical emergencies or similar exceptional circumstances that must be discussed in advance with the instructor. All assignments are due at the beginning of the class period on the announced due date.

Mid-Term and Final Exams

Exams will consist of questions and problems aimed at assessing student mastery of course topics. Conceptual questions may be in the form of essay or multiple-choice format and questions that require pseudo code and/or computations. All exams for this course are closed book.
If you are unable to attend any one of the exams, arrangements may be made only if you have a legitimate reason. You need to inform your instructor ahead of time and have written documentation available. If you are unable to attend the exam due to illness or emergency, you also need to inform your instructor before the exam and bring documentation afterwards to request a make-up exam, or the points for that exam will be allocated to other exams.

Team Project

- A topic of the project (development, implementation, analysis, or measurement) of your choice approved by the instructor. **(Description and examples of project will be posted on Canvas)**
- Projects will be carried out in groups of 2 (or 3 if the project is sophisticated enough, subject to instructor's approval). If you cannot find a team, contact instructor to assign you to a team. Every team will write a final report and present their work at the end of the semester.
- Stage:
    - Literature search
        - SJSU Library CD ROMs: Compendex, Books in Print, SJSU e-books, IEEE, ACM, WWW, etc.
    - Reading
    - Defining a project topic and writing up Proposal
    - Development and Implementation
    - Writing up final report in IEEE Journal or Conference paper format. (A sample of paper format will be provided)

Team recent research paper reading and oral presentation

The purpose of this assignment is to give you the opportunity of exploring what is being researched in the field of Big Data Analytics Using Machine Learning, methods, and results. This assignment will also allow you to research one topic or issue of your interest. (Specific instructions for this assignment will be posted on Canvas.)

Grading Information

Determination of Grades

The components of the final grade will be distributed as follows:

- Class Participation: 5% (pop quizzes, pop questions discussion, interaction with instructor, etc.)
- Homework Assignments: 25% (written and programming)
- Team Project: 25%
- Team Recent research paper reading and oral presentation: 5% (date will be assigned)
- Midterm exams: 20%

- Final exam: 20%

Digit number grades will be assigned according to the following policy:

97 ~ 100 ---- A+
93 ~ 96 ---- 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-
0 ~ 59 ---- F

- Each assignment and exam will be scored (given points) but not assigned a letter grade. Final individual class letter grades will be assigned based on the class curve. Your final class grade can be adjusted up or down depending on your level and quality of class performance.
- **Zero-Tolerance on plagiarism: any types of cheating will not be tolerable; a final course grade 'F' will be given and will be reported to the Department and the University. Sharing your homework solutions with any other students will be treated as cheating.**

Classroom Protocol and Other Notes

- **Absences in attending the first two lectures will be dropped out from the class.**
- Every student must attend class and participate actively.
- You will be called in most class sessions to discuss material contained in lectures.
- Pop questions will also be given by using Random Roster Checker.
- **Always start your email subject with "CS267" to get my attention.**
- To encourage participation from students, no recording is allowed.
- Students must be respectful of the instructor and other students. For example: turn off/silence **cell phones and other mobile devices**.
- Attendance is crucial to doing well on assignments and examinations.
- Students are responsible for all materials posted on Canvas and discussed in the class.

Attendance: 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.

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: 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. Course material cannot be shared publicly without his/her approval. **You are not allowed to publicly share or**

upload instructor generated material for this course such as exam questions, lecture notes, or homework solutions without instructor consent.

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/" Make sure to review these policies and resources.

Topics in Database Systems, CS267, Spring, 2020, Course Schedule

Tentative Course Schedule (This schedule is subject to change with fair notice.)

Week	Date	Topics, Readings, Assignments, Deadlines
1	1/23	Motivation, Orientation /Syllabus, Introduction: (Student Information Due)
2	1/28	Pre-course knowledge test (no credit/scoring) Introduction to Big Data (Big Data Systems Hadoop, Spark and Hive) Project Team formation
2	1/30	Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model
3	2/4	Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model Last day to Drop a Class without a "W" grade.
3	2/6	Introduction to Big Data Mining: Data Cleaning Outliers, Integration, Reduction and Transformation
4	2/11	Introduction to Big Data Mining: Data Cleaning Outliers, Integration, Reduction and Transformation
4	2/13	Online Analytical Processing (OLAP)
5	2/18	Online Analytical Processing (OLAP)
5	2/20	Scalable Data Mining Algorithms: Frequent Itemsets and Mahout Guest Speaker – Google Data Scientist
6	2/25	Scalable Data Mining Algorithms: Frequent Itemsets and Mahout Project Proposal Due: Title and Goal
6	2/27	NoSQL and Big Data Processing Hbase, Hive and Pig, etc.
7	3/3	NoSQL and Big Data Processing Hbase, Hive and Pig, etc. Project Outline of Approach Due
7	3/5	Finding Similar Items: Locality Sensitive Hashing and Theory of Locality Sensitive Hashing
8	3/10	Mining Social Network Graphs
8	3/12	Mining Social Network Graphs
9	3/17	Midterm Exam
9	3/19	Dimensionality Reduction
10	3/24	Mining Data Streams
10	3/26	Mining Data Streams Midterm Project Progress Report Due
	3/31	Spring Recess 3/31 ~ 4/3
	4/2	Spring Recess 3/31 ~ 4/3

Week	Date	Topics, Readings, Assignments, Deadlines
11	4/7	SPARK Architecture, and YARN vs. Mesos
11	4/9	SPARK Architecture, and YARN vs. Mesos
12	4/14	Big Data Document-based Data Model
12	4/16	Big Data K/V-based Data Model: Hive, Pig, HBase
13	4/21	Scalability Models (Strong vs. Eventual Consistent Models) and Big Data Issues
13	4/23	Big Data analytics using machine learning
14	4/28	Big Data analytics using machine learning
14	4/30	Tradeoffs between SQL and NoSQL
15	5/5	Project Presentation and Demo
15	5/7	Project Presentation and Demo Final project Report Due
Final Exam	5/13	Wednesday 2:45~5:00pm