San José State University
College of Science / Department of Computer Science
Topics in Database Systems, CS267, Spring, 2018

Course and Contact Information

Instructor:	Dr. Mike Wu
Office Location:	MacQuarrie Hall 214
Email:	Ching-seh.Wu@sjsu.edu
Office Hours:	Wednesday 1:15PM – 2:15PM Thursday 2:45PM– 3:45PM (Please drop me email in advance with time info and subject)
Class Days/Time:	TuTh 12:00-13:15
Class Room:	MacQuarrie Hall 225
Prerequisites:	CS 157B Database Management Systems II (with a grade of "C-" or better). Computer Science, Applied and Computational Math or Software Engineering Majors only; or Instructor Consent

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on MySJSU Canvas.
You are responsible for regularly checking with the email system through MySJSU at http://my.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.
- Big data: NoSQL data modeling.

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.

Required Texts/Readings

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**

Hadoop: The Definitive Guide, Tom White, O'Reilly, 4rd Edition, 2015, ISBN: 978-149-190-1687,
Free download copy: http://javaarm.com/file/apache/Hadoop/books/Hadoop-The.Definitive.Guide_4.edition_a_Tom.White_April-2015.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/
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: http://couchdb.apache.org/

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.

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 : 20% (written and programming)
- Recent research paper reading and oral presentation: 5% (date will be assigned)
- Project: 30%
- Midterm exams: 20%
- Final exam: 20%

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

92 -- 100 ---- A
90 -- 91 ---- A-
88 -- 89 ---- B+
82 -- 87 ---- B
80 -- 81 ---- B-
78 -- 79 ---- C+
72 -- 77 ---- C
70 -- 71 ---- C-
60 -- 69 ---- 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.

Project

- A topic of the project (development, implementation, analysis, or measurement) of your choice approved by the instructor. (Examples of project will be provided)
- A team of two members is allowed.
- Stage:
    - Literature search
        - CD ROMs:Compendex, Books in Print, SJSU e-books, WWW, etc.
- Reading
- Writing up Proposal
- Development and Implementation

- Writing up final report in Journal or Conference paper format. (A sample of paper format will be provided)

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.**
- **Cheating** will not be tolerable; a ZERO will be given to any cheated assignment/exam, and will be reported to the Department and the University.
- Your laptop must remain closed (preferably in your backpack and not on your desk)
- 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 distributed 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 may not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, or homework solutions without instructor consent.

Dropping and Adding

Students are responsible for understanding the policies and procedures about add/drop, grade forgiveness, etc. Refer to the current semester 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/.

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/

Topics in Database Systems, CS267, Spring, 2018, Course Schedule

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

Week	Date	Topics, Readings, Assignments, Deadlines
1	1/25	Motivation, Orientation /Syllabus, Introduction: (Student Information Due)
1	1/30	Big Data Introduction Hw0 Due
2	2/1	Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model
2	2/6	Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model Project Team formed
3	2/8	Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model
3	2/13	Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model
4	2/15	Introduction to Big Data Mining: Data Cleaning Outliers, Integration, Reduction and Transformation
4	2/20	Introduction to Big Data Mining: Data Cleaning Outliers, Integration, Reduction and Transformation
5	2/22	Introduction to Big Data Mining: Data Cleaning Outliers, Integration, Reduction and Transformation
5	2/27	Online Analytical Processing (OLAP)
6	3/1	Online Analytical Processing (OLAP) Project Proposal Due
6	3/6	Scalable Data Mining Algorithms: Frequent Itemsets and Mahout
7	3/8	Scalable Data Mining Algorithms: Frequent Itemsets and Mahout
7	3/13	Scalable Data Mining Algorithms: Frequent Itemsets and Mahout
8	3/15	Midterm Exam
8	3/20	Finding Similar Items: Locality Sensitive Hashing and Theory of Locality Sensitive Hashing
9	3/22	Finding Similar Items: Locality Sensitive Hashing and Theory of Locality Sensitive Hashing
9	3/27	**Spring Recess**
10	3/29	**Spring Recess**
10	4/3	Mining Social Network Graphs
11	4/5	Mining Social Network Graphs
11	4/10	Dimensionality Reduction
12	4/12	Mining Data Streams
12	4/17	Mining Data Streams
13	4/19	SPARK Architecture, and YARN vs. Mesos
13	4/24	Big Data Document-based Data Model
14	4/26	Big Data K/V-based Data Model: Hive, Pig, HBase
14	5/1	Scalability Models (Strong vs. Eventual Consistent Models) and Big Data Issues
15	5/3	Tradeoffs between SQL and NoSQL
15	5/8	Tradeoffs between SQL and NoSQL
16	5/10	Project Presentation and Demo

Week	Date	Topics, Readings, Assignments, Deadlines
16	5/15	Project Presentation and Demo Final project Report Due
Final Exam	5/17	Thursday, May 17 | 0945-1200