# San José State University
# College of Science / Department of Computer Science
# CS267 Topics in Database Systems, Spring 2024

## Course and Contact Information

| | |
|---|---|
| **Instructor:** | Dr. Ching-seh (Mike) Wu |
| **Office Location:** | MacQuarrie Hall 211 |
| **Email:** | Ching-seh.Wu@sjsu.edu |
| **Office Hours:** | Tuesday & Friday 2:30pm-3:30pm<br>**(Please drop me an email with time info and subject.)** |
| **Class Days/Time:** | Tuesday and Thursday 6:00pm ~ 7:15pm |
| **Classroom:** | Duncan Hall 450 |
| **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)

The topics for this course will be focusing on **Big Data/Data Science Using Machine Learning.** Both theoretical and practical aspects including tools and models of Big Data Using ML Algorithms will be introduced. A significant semester-long project reinforces lectures and is designed by applying Google's Team Project Based Learning (PBL) derived from Google's software engineering best practices. In this team project, you will apply concepts presented in the lectures and obtain practical hands-on experience by using the tools with ML algorithms. Students, in randomly selected, 2 member teams, will complete a practical real-world application or a research-oriented project. Team may choose any Big Data with ML applications to solve a problem that are appropriate in size and complexity. Appropriateness of the project will be determined by the instructor.

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

- Introduction to Big Data
- Big Data Mining
- Large-scale data processing platforms.
- MapReduce model
- HDFS vs Spark

- Google File System (GFS)
- Apache Hadoop architecture
- 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, and tools related to Big Data.
- Familiar with mining data streams.
- Familiar with Apache Hadoop architecture, Google File System (GFS), 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

**Data-Intensive Text Processing with MapReduce,** Jimmy Lin and Chris Dyer, University of Maryland, College Park, ISBN-13: 978-1608453429

**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

**Online Reading Materials, Tools and Datasets:**
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/
UCI Machine Learning Repository (Datasets): https://archive.ics.uci.edu/ml/index.php
The list of datasets will be posted on the Course Canvas.

## Course Requirements and Assignments

### Assignments

You are expected to learn all 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.
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 explained and posted on Canvas)
- Projects will be carried out in groups of randomly-selected 2 members (or 3 if the project is sophisticated enough, subject to instructor's approval). Every team will write a final report and present their work at the end of the semester.
- Project Stages:
    - Literature search
        - SJSU Library: Compendex, Books in Print, SJSU e-books, IEEE, ACM, WWW, etc.
    - Reading
    - Defining a project topic and writing up a proposal
    - Development and implementation
    - Writing up final report in IEEE Journal or Conference paper format. (A sample of paper format will be provided)
    - Project demo and oral presentation

## 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:

- Active Class Participation: **15%** (Class participation, pop quizzes, pop questions, discussion, and hands-on exercises)
- Homework Assignments: **20%** (5 written and programming assignments)
- Team Project (including one assigned seminar research paper reading and oral presentation): **25%**
- Midterm Exam: **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.** A "Honesty Pledge" form must be signed and submitted by each student before the second day of the class**.**

## Classroom Protocol and Other Notes

- This course is an in-person class. Please be noticed that students who have absented in attending the first two class lectures will be automatically instructor-dropped out of the class. If you are unable to attend the first two lectures, I suggest that you should drop this course by yourself immediately so that people who are in the waiting list can add to this course.
- **There will be no lecture recordings for later review/study**. **Recording a lecture is prohibited**. Students are prohibited from recording class activities (including class lectures, office hours, advising sessions, etc.), distributing class recordings, or posting class recordings. Materials created by the instructor for the course (syllabi, lectures and lecture notes, presentations, etc.) **are copyrighted** by the instructor. This university policy (S12-7) is in place to protect the privacy of students in the course, as

well as to maintain academic integrity through reducing the instances of cheating. Students who record, distribute, or post these materials will be referred to the Student Conduct and Ethical Development office. **Unauthorized recording may violate university and state law**. It is the responsibility of students that require special accommodations or assistive technology due to a disability to notify the instructor.

- You will be called in most class sessions for pop questions and to discuss material contained in lectures by using Random Roster Checker.
- **When emailing me, please always start your email subject line with "CS267: XXXXX" to get my attention. (XXXXX: Subject, for example: CS267: HW1 Question)**
- **Plagiarism/Cheating will not be tolerable: 'F' will be given to your FINAL COURSE GRADE and will be reported to the Department and the University. (Please be noted: obtaining HW solutions from someone or giving/showing your HW solutions to someone is also treated as plagiarism/cheating.)**
- **Attendance/participation is crucial to doing well on pop quizzes/questions, assignments and examinations.**
- **Students are responsible for all materials distributed/posted on Canvas and discussed in the class. I also reserve the right to make announcements in class that may not appear on Canvas.**

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.

**Tentative Course Schedule** (This schedule is subject to change with fair notice.) (Assigned Seminar research paper presentations will be added.)

| Week | Date | Topics, Readings, Assignments, Deadlines |
|---|---|---|
| 1 | 1/25 | Course Introduction, Motivation, Orientation/Syllabus<br>(Upload unofficial transcript with highlighted CS157B prerequisite course to Canvas by 6pm next class, Tuesday, 1/30/2024)<br>(Sign the Honesty Pledge from and upload it to Canvas by next class, Tuesday, 1/30/2024) |
| 2 | 1/30 | Course Activities<br>Pre-course knowledge test (no credit/no scoring)<br>(Student Information Due) |
| 2 | 2/1 | Introduction to Big Data (Big Data Systems Hadoop)<br>Team Projects Introduction |
| 3 | 2/6 | Introduction to Big Data (Big Data Systems Hadoop)<br>(Project Team formation)<br>(Research papers reading and presentation assignments) |
| 3 | 2/8 | Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model + Google File System |
| 4 | 2/13 | Hadoop Anatomy: HDFS + MapReduce Parallel Computing Model + Google File System<br>HW1 |
| 4 | 2/15 | Introduction to Big Data Mining: Data Cleaning Outliers, Integration, Dimensionality Reduction and Transformation<br>(Project Proposal Due) |
| 5 | 2/20 | Introduction to Big Data Mining: Data Cleaning Outliers, Integration, Dimensionality Reduction and Transformation |
| 5 | 2/22 | Scalable Data Mining Algorithms: Frequent Itemsets and Mahout/Spark |
| 6 | 2/27 | Scalable Data Mining Algorithms: Frequent Itemsets and Mahout/Spark<br>Guest Speaker – Google Data Scientist (What Data Scientist s do at Google?)<br>HW2 |
| 6 | 2/29 | NoSQL and Big Data Processing Hbase, Hive and Pig, etc. |
| 7 | 3/5 | NoSQL and Big Data Processing Hbase, Hive and Pig, etc. |
| 7 | 3/7 | Big Data with Online Analytical Processing (OLAP) |
| 8 | 3/12 | Big Data with Online Analytical Processing (OLAP) |
| 8 | 3/14 | Midterm Exam |
| 9 | 3/19 | Midterm Solutions<br>Mining Social Network Graphs |
| 9 | 3/21 | Mining Social Network Graphs<br>HW3 |
| 10 | 3/26 | Mining Data Streams |
| 10 | 3/28 | Mining Data Streams |
| | 4/2 | Spring Recess 4/1 ~ 4/5 (No Class) |
| | 4/4 | Spring Recess 4/1 ~ 4/5 (No Class) |
| 11 | 4/9 | SPARK Architecture, and YARN vs. Mesos<br>(Midterm Project Progress Report Due) |
| 11 | 4/11 | SPARK Architecture, and YARN vs. Mesos<br>HW4 |

| Week | Date | Topics, Readings, Assignments, Deadlines |
|---|---|---|
| 12 | 4/16 | Big Data Document-based Data Model |
| 12 | 4/18 | Big Data K/V-based Data Models |
| 13 | 4/23 | Scalability Models (Strong vs. Eventual Consistent Models) and Big Data Issues |
| 13 | 4/25 | Big Data analytics using machine learning |
| 14 | 4/30 | Big Data analytics using machine learning<br>HW5 |
| 14 | 5/2 | Project Presentation and Demo |
| 15 | 5/7 | Project Presentation and Demo |
| 15 | 5/9 | Project Presentation and Demo |
| Final Exam | 5/16 | Thursday 5:15~7:30pm<br>(Final project Report Due) |