San José State University
CS274, Topics in XML and Web Intelligence, Section 2, Spring, 2017

Course and Contact Information

Instructor: H. Chris Tseng

Office Location: MH213

Telephone: (408) 924-7255

Email: chris.tseng@sjsu.edu

Office Hours: Tue/Thur.:8:45 AM -9:45 AM and by appointment or email

Class Days/Time: Tuesdays/Thursday 7:30 – 8:45 AM

Classroom: MH 225

Prerequisites: **Introduction to Database (CS 157A)**

Course Format

Technology Intensive, Hybrid, and Online Courses

This course combines theories with hands-on assignments. Students are expected to work in teams to accomplish big data query and analytics project(s) using open source technologies, including Cassandra, Spark, and Solr.

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on your Canvas account (https://sjsu.instructure.com). You are responsible for regularly checking with the messaging system through Canvas account to learn any updates.

Course Description

The rise of cloud and big data applications demands a new approach to data query and analytic. These applications require a highly scalable and distributed database with powerful query and analytic capabilities. This course focuses on the main theories and technologies of query and analytics in distributed databases where big data resides. We will study various aspects of underlying theories and applications in this topic. These include distributed architecture for distributed databases, query based data modeling techniques, internal architecture of read path, write path, and data compaction and how it affects data query and analytics. We will also learn how Spark and Solr can couple with distributed databases to provide additional query and analytic features.

Learning Outcomes and Course Goals

Course Learning Outcomes (CLO)

This course aims to introduce students to the underlying theories and solutions for big data query and analytic of big data in distributed databases.

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

• XML and JSON

• Understand the basic theories of ACID, CAP, and BASE for distributed databases

• Know the internal architecture of a distributed database like Cassandra

• Learn the concept of read path, write path, and data compaction in distributed database

• Understand the relation of replication and eventual consistency for data query

• Understand anti-entropy operations for read repair

• Design query based data model for query and analyzing big data

• Learn Spark RDD and related data analytic operations

• Learn how Solr can perform complex and rich text data query and analytic

Required Texts/Readings

Textbook (you may click on the books to buy online)

Cassandra: The Definitive Guide, 2nd Ed., by Carpenter and Hewitt, O'Reilly Publishing

Optional books for the class are

1. XML for the World Wide Web: Visual QuickStart Guide, (2nd Edition), Goldberg, Peachpit Press; 2008, ISBN: 0321559673
2. Learning Spark, by Karau, Konwinsky, Wendell, and Zahari, O'Reilly Publishing
3. Solr in Action, by Grainger and Potter, Manning Publications

Other Readings

A list of online 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 and approximately 30GB disk space free. You will be required to have a wireless-network ready laptop computer to take online quizzes in the class. You will also need to use your own laptop with wireless access to submit your software assignment inside SJSU campus. Your laptop needs to have wireless capability and you need to register a free wireless account at http://www.sjsu.edu/sjsuone/. For information on typical laptop requirement see Wireless Requirements update announcement under http://www.cs.sjsu.edu/. The instructor is not responsible for providing either laptops or alternatives.

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.

a. Projects:
A final team project will be provided for you to practice big data query and analytic principles in distributed database. Self-selected teams of 3 people will work together to solve some selected problems discussed in the course.

This team project will be a collaborated group project. You are free to choose your own partners but you cannot change your partners in the middle of the project. Progressive design and implementation of the term project will be done through assignments as part of the learning objectives.

b. Exams:
There will be one midterm

c. Quizzes:
There will be 2-4 quizzes and each will be counted as a HW. Some of the quizzes are part of the online lesson activities.

d. Homework:
There will be 3-4 HWs. Intermediate milestones of your team project will also be counted as HW grades.

e. Tentative course exam and HW due dates:
(Please note that this is "subject to change with fair notice")
HW/Quiz/Practice problems: One of these will be assigned every 2-3 class meetings.
Midterm: Thursday, March 23, 2017
Final: (Per SJSU final schedule)

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

Grades:

HW assignments and quizzes	30 %
Midterm	30 %
Final Team Project	40 %

Grading information:

Grades will be assigned as described below. These intervals, however, may change (i.e., either way!) according to the performance of the class as a whole. C- is a passing grade.

- A: [93, 100]
- A-: [90, 93)
- B+: [87, 90)
- B: [83, 87)
- B-: [80, 83)
- C+: [75, 80)
- C: [70, 75)
- C-: [65, 70)
- D+: [60, 65)
- D: [55, 60)
- D-: [50, 55)
- F: [0, 50)

Policies

Penalty (if any) for late or missed work:
No credit will be given for assignments turned in late. No makeup exams or quizzes will be given.
Incomplete grade will only be assigned to students with sudden events such as medical or personal emergency. Written proof is needed in all cases.

Classroom Protocol

You are expected to attend classes. If you cannot attend, it is your responsibility to get a copy of the lecture notes and class announcements from a reliable classmate. The instructor reserves the right to ignore frivolous or inappropriate e-mail inquiries.

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.

CS274 Topics in XML and Web Intelligence, Spring 2017, Course Schedule

(Please note that the course calendar is "subject to change with fair notice")

Course Schedule

Week	Topics, Readings, Assignments, Deadlines
1	Introduction to XML
2	Introduction to JSON
3	Understand big data and distributed databases (HW/Quiz)
4	ACID, CAP, and BASE theories
5	Distributed database architecture
6	Read Path and Write Path (HW/Quiz)
7	Querying and analyzing big data with Cassandra CQL
8	Review/Midterm
9	Query based data modeling and read repair
10	Replication factor and consistency level consideration in data query and analytic (HW/Quiz)
11	Complex queries and rich text queries in Solr
12	Designing Solr schema for complex query and analytic
13	keyword search, wildcard, fuzzy, range, inequality data queries with Solr (HW/Quiz)
14	Spark RDDs and basic analytic operations
15	Pair RDDs and related operations (HW/Quiz)
Final Exam (Project Presentation)	Per SJSU final schedule