DSC 208R: DATA MANAGEMENT FOR ANALYTICS

Spring 2025

Course Syllabus

Instructional Team Information

Instructor: **Haojian Jin**
Email: **haojian@ucsd.edu**
Office hours: **Tuesdays at 4:00 – 5:00 p.m. PST or by appointment**

Instructor: **Arun Kumar**

Teaching Assistant: **Muchan Li**
Email: **mul005@ucsd.edu**
Office hours: **Wednesdays at 3:00 – 4:00 p.m. PST or by appointment**

Course Description

This course covers the principles, techniques, and tools of organizing, storing, querying, transforming, and using data for analytics and ML computations at scale. Students will learn the basics of data storage, acquisition, governance, organization, principles of the relational data model, relational algebra and its relationship to DataFrames, the Structured Query Language (SQL), relational database system features for faster querying and analytics, and basics of non-relational data systems. It will also cover major data quality issues and methodologies to clean data. An introduction to cluster and cloud computing, MapReduce and Spark, and the use of these tools and SQL to transform data at scale for ML feature engineering will be provided. Finally, methodologies to critically evaluate analytics results will be covered, including debugging ML results and reasoning about bias and fairness issues in the whole data science pipeline.

Course Objectives

At the end of the course, students should be able to:

- Explain the basic principles of managing large and complex datasets for analytics.
- Apply the relational model, relational algebra, SQL, and major relational DBMS features for data querying and analytics.
- Apply SQL and cluster programming techniques such as MapReduce and Spark to perform data transformations for ML at scale.
- Explain major data quality issues in data science pipelines and how to handle them.
- Analyze and evaluate tradeoffs of data science pipelines in terms of data quality, automation, scalability, accuracy, and fairness.

Course Format

This 10-week course, with one additional week for your final exam, is offered asynchronously via Canvas. Quizzes and exams will be offered on certain days within a time window. The lecture videos, quizzes, and exams will all be available on Canvas. We use Piazza for discussions.

Course Outline

Week	Topics & Learning Objectives	Video Lectures	Assignments Due
1	Module 1: **Introduction and Administrava** Module 2: **Data Collection and Governance**	▪ Course Information ▪ Data Collection and Governance Lectures	
2	Module 3: **Data Models and Querying**	▪ Data Models	1. Review Quiz 1 **Due – Sunday, April 13th**
3	Module 3: **Data Models and Querying**	▪ Introduction to SQL	
4	Module 3: **Data Models and Querying**	▪ Aggregation and Grouping ▪ Nested SQL Queries	1. Review Quiz 2 **Due – Sunday, April 27th** 2. Programming Assignment 1 **Due – Sunday, April 27th**
5	Module 3: **Data Models and Querying**	▪ Formal Query Languages and Constraints in SQL	1. Programming Assignment 2 Part 1 **Due – Sunday, May 4th**
6	Midterm Exam	▪ Query Evaluation, Indexing, and Optimization ▪ Review previous lecture videos	1. Midterm Exam **Due – Sunday, May 11th** 2. Review Quiz 3 **Due – Sunday, May 11th**
7	Module 3: **Data Models and Querying** Module 4: **Data Engineering for Machine Learning**	▪ Semi-structured Data Models ▪ Introduction to Parallelism ▪ Task Parallelism and Dask	1. Programming Assignment 2 Part 2 **Due – Sunday, May 18th**
8	Module 4: **Data Engineering for Machine Learning**	▪ Data Parallelism ▪ Data Parallel Data Science Operations ▪ Cloud Computing	1. Review Quiz 4 **Due – Sunday, May 25th**
9	Module 4: **Data Engineering for Machine Learning**	▪ DataFlow Systems: MapReduce ▪ DataFlow Systems: Spark ▪ Feature Engineering with MapReduce and Spark	
10	Module 5: **Data Quality and Cleaning**	▪ Basics of Data Cleaning ▪ Data Cleaning: Tasks, Methods, and Tools	1. Review Quiz 5 **Due – Sunday, June 8th** 2. Programming Assignment 3 **Due – Sunday, June 8th**
11	Final Exam	▪ Review previous lectures	1. Final Exam **Due – Friday, June 13th**

Learning Materials

Required: Lecture Videos

Each module consists of one or more topics linked by a technical theme. Each topic has one or more lecture videos. Each lecture video is about 20-30 min. long. Some of the topics have ungraded review questions, covered as Q&A in a video each.

Optional: Lecture Videos

The following textbooks are highly recommended (but not required) for this course. No single textbook covers all the topics of this course. And not all the content of the textbooks listed below are relevant for this course. So, use the textbooks as needed, aligned with the course topics.

***Database Management Systems**, by Raghu Ramakrishnan and Johannes Gehrke. McGraw-Hill; 3rd edition, 2002. ISBN-13: 978-0072465631*

***Spark: The Definitive Guide**, by Bill Chambers and Matei Zaharia. O'Reilly Media; 1st edition, 2018. ISBN-13: 978-1491912218*

***Data Management in Machine Learning Systems**, by Matthias Boehm, Arun Kumar, and Jun Yang. Morgan & Claypool Publishers, 2019. ISBN-13: 978-1681734989*

Assessments, Evaluation, and Grading

Required: Lecture Videos

This course will have 5 quizzes, all delivered via Canvas. Each quiz is worth 2% of the grade. Each quiz will be 20 minutes long and will have 5 multiple-choice questions. The quizzes are aligned with the modules of the course to help consolidate your learning as you go along.

The quizzes are open books/notes/Internet. The only requirement is that you must neither receive nor give help to anyone by any means. There will be no proctoring. If you find any questions ambiguous, make sound assumptions and proceed with them. If a question had genuine issues, it will be waived post-hoc for everyone who got that question.

To enforce Academic Integrity, we will use all the features of Canvas: display one question at a time, answer lock-in, shuffle answer options, and question groups from which Canvas picks a random subset for each student.

Quizzes will be available for a 3-day timeframe, from Friday at 12:00 a.m. to Sunday at 11:59 p.m.

Midterm and Final Exam

This course has a midterm exam and a cumulative final exam, both delivered via Canvas. The midterm exam will be held at the end of Week 6. It will cover Modules 1 and 2 and most of Module 3, up to the topic of "query evaluation, indexing, and query optimization." The final exam will have a higher weightage for the topics covered after the midterm exam.

The midterm exam is worth 15% of the grade. Its length is the equivalent of 2 hours of a written exam, but you will be given 2.5-hour time limit on Canvas. The final exam is worth 40% of the grade. Its length is the equivalent of 3 hours of a written exam, but you will be given 4-hour time limit on Canvas.

Both exams will feature a mixture of multiple-choice and yes-no/true-false questions, as well as longer essay questions, including writing SQL queries, relational algebra queries, and MapReduce jobs. Partial credits are possible for some of the essay questions.

The exams are also open books/notes/Internet. They will also use the same Academic Integrity features of Canvas as the quizzes.

The midterm will be available for a 3-day period, from Friday at 12:00 a.m. PST to Sunday at 11:59 p.m. PST, and the final exam will also be available for a 3-day period, from Wednesday at 12:00 a.m. PST to Friday at 11:59 p.m. PST.

Programming Assignments

This course features 3 hands-on programming assignments (PAs). In the first two, you will get to program with SQL using SQLite and PostgreSQL by installing them on your local machine. In the third, you will get to program with MapReduce/Spark and run them on a remote machine cluster at UC San Diego.

The first PA is worth 5% of the grade; the latter two 15% each. You have to work on the PAs in teams of 2 or teams of 1 (individual). You can use Piazza to find a teammate if you'd like. The Teaching Assistant (TA) will release a Google Form early on in the course to collect your team composition details. The TA will then confirm your team ID. You are not allowed to change teams between the PAs. But you are free to split up into individuals if you want to.

Late Policy

There are no late days in this course. The programming assignments will have set deadlines. Please plan your work well upfront and submit whatever you finish on time. Likewise, the quizzes and exams will have set time windows on Canvas during which you must take and submit them. All dates/time windows will be announced upfront. Make sure to add them on your calendar and adhere to them.

In case of medical or family emergencies, if you are able to provide university-approvable evidence, a one-day extension without penalty can be offered for the programming assignments; likewise, a makeup slot can be offered for a quiz/exam.

Assignments	% of Grade
Quizzes	10% (5 quizzes, 2% each)
Programming Assignments	35% (5 PA 10%, 2 PAs 15% each)
Midterm Exam	15%
Final Exam	40%
Total	**100%**

Grading Cutoffs

The grading scheme is a hybrid of absolute and relative grading. The absolute cutoffs are based on your absolute total score. The relative bins are based on your position in the total score distribution of the class. The better grade among the two (absolute-based and relative-based) will be your final grade.

Grade	Absolute Cutoff (>=)	Relative Bin (Use strictest)
A+	92	Highest 10%
A	85	Next 15% (10-25)
A-	80	Next 15% (25-40)
B+	75	Next 15% (40-55)
B	70	Next 15% (55-70)
B-	65	Next 5% (70-75)
C+	60	Next 5% (75-80)
C	55	Next 5% (80-85)

C-	50	Next 5% (85-90)
D	45	Next 5% (90-95)
F	<45	Lowest 5%

Example: Suppose the total score is 82 and the percentile is 43. The relative grade is B, while the absolute grade is A-. The final grade then is A-.

Course and UCSD Policies

Course Expectations

The quizzes and exams are all individual learning evaluation components. While they are open books/notes/Internet, you must not receive or give to anyone by any means. The course Piazza will also be closed during their time windows. We will cross-check answers across students to check for any potential collusion or other academic integrity violations.

It is okay to discuss the programming assignments (PAs) with your peers at a conceptual level. It is also okay to post conceptual or high-level questions, logistical questions, and useful references on Piazza. But do NOT share any code across teams, and do NOT post any of your own solution code for discussion.

A team's code submission must be entirely their own. Please review UCSD's honor code and policies and procedures on Academic Integrity here. Do **not** go searching for any code posted online by other students or prior editions. We will use advanced program analysis tools to compare your code submissions. These go well beyond basic string or syntactic comparisons to catch plagiarism.

If plagiarism is detected in your code or if any other form of academic integrity violation is identified on any graded course component, you will get zero for that component and also get downgraded substantially. We will also notify the University authorities for appropriate disciplinary action to be taken, up to and including expulsion from the degree program.

UCSD Code of Conduct

All participants in the course are bound by the **University of California Code of Conduct** (https://aisc.uci.edu/students/index.php).

Netiquette

Be respectful. Be sensitive. Be aware. Effective written communication and open academic dialogue are crucial for sustaining a learning community that is respectful, considerate, relevant, creative, and thought-provoking. In an online classroom, expressions, meaning, and tone can quickly be taken out of context, making it imperative that online learners adhere to the communication guidelines below:

- Treat your classmates with respect.
- Be thoughtful and open in a discussion.
- Be aware and sensitive to different perspectives.
- Build one another up and encourage one another to succeed.

The following behavior should be avoided:

- Using insulting, condescending, or abusive words.
- Using all capital letters, which comes across as SHOUTING.
- Contacting learners or posting advertisements and solicitations.
- Posting copyrighted material.

Academic Integrity

Academic Integrity is expected of everyone at UC San Diego. This means you must be honest, fair, responsible, respectful, and trustworthy in your actions.

Lying, cheating, or other forms of dishonesty will not be tolerated because they undermine learning and the University's ability to certify students' knowledge and abilities. Thus, any attempt to get, or help another get, a grade by cheating, lying, or dishonesty will be reported to the Academic Integrity Office and result in sanctions. Sanctions can include an F in the class and suspension or dismissal from the University.

So, think carefully before you act. Before you act, ask yourself the following questions: a: is my action honest, fair, respectful, responsible, and trustworthy, and b) is my action authorized by the instructor? If you are unsure, don't ask a friend; ask your instructor, instructional assistant, or the Academic Integrity Office. You can learn more about academic integrity at academicintegrity.ucsd.edu.

Accessibility

Students requesting accommodations for this course due to a disability must provide a current Authorization for Accommodation (AFA) letter issued by the UC San Diego Office for Students with Disabilities (OSD), which is located in University Center 202 behind Center Hall. Students are required to present their AFA letters to Faculty (please make arrangements to contact me privately) and to the OSD Liaison in the department in advance so that accommodations may be arranged. Contact the OSD for further information: https://disabilities.ucsd.edu/.

Religious Accommodation

See: EPC Policies on Religious Accommodation, Final Exams, Midterm Exams

It is the policy of the university to make reasonable efforts to accommodate students having bona fide religious conflicts with scheduled examinations by providing alternative times or methods to take such examinations. If a student anticipates that a scheduled examination will occur at a time at which his or her religious beliefs prohibit participation in the examination, the student must submit to the instructor a statement describing the nature of the religious conflict and specifying the days and times of conflict.

For final examinations, the statement must be submitted no later than the end of the second week of instruction of the quarter.

For all other examinations, the statement must be submitted to the instructor as soon as possible after a particular examination date is scheduled.

If a conflict with the student's religious beliefs does exist, the instructor will attempt to provide an alternative, equitable examination that does not create undue hardship for the instructor or for the other students in the class.

Accessibility

See: Nondiscrimination Policy Statement

CARE at the Sexual Assault Resource Center
858.534.5793 | sarc@ucsd.edu | https://care.ucsd.edu
Counseling and Psychological Services (CAPS)
858.534.3755 | https://caps.ucsd.edu

Subject to Change Policy

Note that the information contained in the course syllabus, other than the grade and absence policies, may be – under certain circumstances such as a modification to enhance student learning – subject to change with reasonable advanced notice, as deemed appropriate by the instructor.