Principles of Data-Intensive Systems

Winter 2021
Tue/Thu 2:30-3:50 PM Pacific

This course covers the architecture of modern data storage and processing systems, including relational databases, cluster computing systems, streaming and machine learning systems. Topics include database system architecture, storage, query optimization, transaction management, fault recovery, and parallel processing, with a focus on the key design ideas shared across many types of data-intensive systems.

Find the 2022 version of this course on Canvas

Course Staff

Instructor
Matei Zaharia (Office hours: by appointment, please email me)

Teaching Assistants
Cody Coleman (Office hours: Wednesdays 4-5:30 PM Pacific, Zoom)
Daniel Kang (Office hours: Sundays 4:30-6 PM Pacific, Zoom)
Gina Yuan (Office hours: Wednesdays 10-11:30 AM Pacific, Zoom)
Peter Kraft (Office hours: Mondays 2:30-4 PM Pacific, Zoom)
Xinyi Yu (Office hours: Fridays 8:30-10 AM Pacific, Zoom)

Schedule

T, Jan 12 — Introduction

Th, Jan 14 — Database System Architecture
Reading: A History and Evaluation of System R INSTRUCTIONS
Optional Reading: How to Read a Paper

T, Jan 19 — Database Architecture 2 & Storage
ASSIGNMENT 1 POSTED

Th, Jan 21 — Storage Formats and Indexing
Reading: Integrating Compression and Execution in Column-Oriented Database Systems INSTRUCTIONS
Optional Reading: C-Store: A Column-Oriented DBMS

T, Jan 26 — Storage Formats and Indexing 2

Th, Jan 28 — Query Execution

T, Feb 2 — Query Optimization

Th, Feb 4 — Query Optimization 2
Reading: Spark SQL: Relational Data Processing in Spark INSTRUCTIONS
ASSIGNMENT 1 DUE (AT MIDNIGHT) ASSIGNMENT 2 POSTED

T, Feb 9 — Guest Talk: Automatically Discovering Systems Optimizations for Deep Learning
Zhihao Jia, Carnegie Mellon University and Facebook
Optional Readings: Beyond Data and Model Parallelism for Deep Neural Networks, TASO

Th, Feb 11 — Transactions and Failure Recovery
TEST 1 POSTED (solutions)
Past Midterms for Reference: Winter 2020 (solutions), Spring 2019 (solutions), Winter 2017 (solutions)

T, Feb 16 — Failure Recovery 2
TEST 1 DUE (AT MIDNIGHT)

Th, Feb 18 — Concurrency

T, Feb 23 — Concurrency 2
Reading: Granularity of Locks and Degrees of Consistency in a Shared Data Base INSTRUCTIONS

Th, Feb 25 — Concurrency 3 & Distributed Databases
Optional Reading: Time, Clocks and the Ordering of Events in a Distributed System
ASSIGNMENT 2 DUE (AT MIDNIGHT) ASSIGNMENT 3 POSTED

T, Mar 2 — Distributed Databases 2
Optional Reading: Lessons from Internet Services: ACID vs. BASE

Th, Mar 4 — Cloud Database Systems
Optional Readings: Amazon Aurora, Dynamo, Delta Lake

T, Mar 9 — Streaming Systems
Optional Readings: The CQL Continous Query Language, The Dataflow Model

Th, Mar 11 — Security and Data Privacy
Reading: Privacy Integrated Queries INSTRUCTIONS
Optional Readings: Robust De-anonymization of Large Sparse Datasets, Opaque, Splinter
TEST 2 POSTED (pdf) (docx) (solutions)
Past Finals for Reference: Winter 2020 (solutions), Spring 2019 (solutions), Winter 2017 (solutions)

T, Mar 16 — Guest Talk: Lakehouse Technology as the Future of Data Warehousing
Reynold Xin, Databricks ABSTRACT

Th, Mar 18 — No Class (Work on Test 2)
ASSIGNMENT 3 DUE (AT MIDNIGHT) TEST 2 DUE (AT MIDNIGHT)

Logistics

Announcements

All announcements will be made on our Piazza page for the class. Make sure you sign up for Piazza!

Prerequisites

Students should ideally have taken CS 145 and CS 161, or their equivalent courses. In particular, we expect students to be familiar with SQL syntax. You can take a basic SQL tutorial for an overview of SQL if needed.

Lectures and Video Recordings

Lectures for the class will be given live on Zoom and recorded. You can find our Zoom link and the lecture recordings on Canvas. Please note that these recordings might be reused in other Stanford courses, viewed by other Stanford students, faculty, or staff, or used for other education and research purposes. If you have questions about video recording, please contact a member of the teaching team.

Assignments and Tests

We will have three programming assignments and two take-home tests. The programming assignments are designed to be runnable on your personal machine and should be submitted through Gradescope.

The tests are open-book, meaning that you can use your course notes, slides, books, or online resources, except that communication is not be allowed during them (e.g., you can't ask a question on Stack Overflow or contact another student). Tests will cover material in the lectures, readings and assignments.

Readings

We have occasional readings for the lectures. We expect students to complete these and think about the questions we list for each paper on their own (you do not need to turn in answers). Our tests will cover content in the readings.

Optional Textbook

Database Systems: The Complete Book (2nd Edition), by Garcia-Molina, Ullman and Widom, covers a lot of the technical material in the course and may be helpful as a study guide. We focus on chapters 13-20. We will also cover the material in lectures, but this book is a good source of additional information.

Grading

Assignments: 18% each (total: 54%)
Test 1: 20%
Test 2: 26%

Late Policy

Students each have up to 2 late days that they may use for assignments and tests. Assignments and tests submitted after these late days have been used up will incur a penalty of 10% per extra day late. In addition, we will not accept submissions after March 20th at midnight Pacific to give the staff enough time for grading.

Auditing

The course Zoom meetings and recordings are open for auditing to any Stanford student on Canvas.

Feedback

Please post public questions about the class on Piazza. For private questions to the staff, please open a private post on Piazza. You can also email professor Zaharia at matei@cs.stanford.edu.