Advanced Data Systems (COSI 167A)

Class timings: **Tue/Fri 12:45 PM – 2:05 PM @ Gerstenzang: 121**
Class Website | Gradescope | Moodle

Course Description

Data is everywhere. As scientists, users, and citizens we are both generating and exploiting large, ever-growing, diverse sets of data. For several applications – ranging from scientific discovery to business analysis, governance, and everyday activities – we are directly using and indirectly affecting hundreds of data systems! The big challenge is to turn data into useful knowledge, and to do so quickly, in order to increase the impact of the new insights. Achieving these goals comes with a number of technical challenges. How to exploit the continuously evolving hardware (storage, computation, network)? How to collect all incoming data efficiently? How to query dynamic collections of data that keep accumulating incoming data? How to parallelize query processing from one core to a few (scale-up), and then to thousands (scale out)? What are the needs of evolving workloads (hybrid transactional/analytical processing, graph analytics, Internet-of-Things, micro-payments, monitoring)? In this course, we will discuss how to design data systems that can address these challenges. We will see in detail the two driving forces behind innovation in data systems: hardware and workloads, and we will discuss recent and future trends of both. We will use examples from several data management areas including relational systems, distributed database systems, key-value stores, newSQL and NoSQL systems, data systems for machine learning (and machine learning for data systems), interactive analytics, and data management as a service. In a quickly moving industry and research landscape, such skills are essential.

Prerequisites

COSI 127B – Database Management Systems. A working knowledge of **C/C++** and **Java or Python** programming and a fundamental understanding of data structures and algorithms is required. Please see the instructor if you are not sure about the level of your preparation.

Instructor Information

	Contact Information	**Student Hours**
Subhadeep Sarkar	Office: Volen 259 Email: subhadeep@brandeis.edu	Tu/Th: 2:15 PM - 3:15 PM Fr: 4:00 PM - 5:00 PM (*by appointment*)*

Recitation for this class is ***optional***. The recitation slot on Friday from 4 PM to 5 PM will be used as **extra student hours** (by appointment) to discuss project progress, and the student hours will be **held in Volen 259** (instead of Gerstenzang: 121).

Class Resources & Contact Policy

The class website is updated in real-time with the class. **This is your go-to** for any general resources about the class. This is also where all **project details will be posted**, along with other relevant resources for the class.

Students are highly encouraged to regularly interact with the teaching staff to clarify doubts, ask questions, seek help for presentation review, and discuss project progress. Frequent interaction with the teaching staff is crucial to success in the course.

Emails sent during weekdays should expect a response within 48 hours. Be aware that not receiving a response is not a valid reason not to hand in homework/assignments on time, so start your work early to make sure you have no questions.

We will be using Gradescope to grade assignments. Please register to Gradescope, if you are not enrolled already (access code: **PYG88X**).

Credit Hours Statement

Success in this four-credit course is based on the expectation that students will spend a **minimum of nine hours of study time per week** in preparation for class (reading papers, preparing and presenting research papers, working on projects, etc.).

Required Textbook

There is no textbook that covers cutting-edge research; however, the data management community has produced top-quality textbooks that can serve as references to provide background material. The class is based on recent research papers which will be available to you through the Brandeis network.

An excellent paper in the database field is the following:

> **Readings in Database Systems.** P. Bailis, J. Hellerstein, M. Stonebraker, editors.

Other good background material includes:

1. *Architecture of a Database System*. J. Hellerstein, M. Stonebraker, and J. Hamilton, Foundations and Trends in Databases, 2007.
2. *The Design and Implementation of Modern Column-store Database Systems*, D. Abadi, P. Boncz, S. Harizopoulos, S. Idreos, and S. Madden. Foundations and Trends in Databases, 2007.
3. *Data Structures for Data-Intensive Applications*, Manos Athanassoulis, Stratos Idreos, and Dennis Shasha, Foundations and Trends in Databases, 2023.

Learning Goals

Students who successfully complete all components of this course will be able to demonstrate the following by the end of the semester.

a) Familiarization with the history and evolution of **NoSQL data systems** design over the past decades.
b) Understanding of the **challenges and tradeoffs** associated with large-scale data management and analysis.
c) Knowledge about the key **design principles** of state-of-the-art NoSQL systems.
d) Ability to **interact and tune** with modern key-value stores.
e) Understanding how **large-scale commercial data systems work** and how to **optimize such systems** for a given workload and performance target.
f) Ability to **program, experiment, evaluate complex, scalable data systems,** and **reason about their performance**.
g) **Read and interpret state-of-the-art research papers**, with the ability to **criticize** them constructively.

Topics
In this class, we will cover data systems design principles from the following different angles.
1. What affects new data systems designs (data and applications, emerging hardware, and new workloads)
2. Traditional data systems for modern hardware
3. Distributed database systems
4. NoSQL, newSQL, and key-value stores
5. Hardware-conscious systems design

Components of Course Work

Project 1: The first requirement for the class is a small implementation project at the beginning of the semester. Project 1 will be carried out by **each student independently** during the first three weeks of the semester. Its goal is to prepare you for the semester project by sharpening your development skills.

Paper Presentation: After the initial 15 classes all students will take turns presenting papers. In each class, we will discuss one (or two) main paper(s) (and there will be a few background papers), and each student will present once in the semester, either alone or as a group of two students). The student(s) presenting will be responsible for outlining the strong and the weak points of the paper and proposing at least one idea for improving the approach presented in the paper. **All students** will read the presented paper.

Paper Reviews & Technical Questions: All students should read all papers. Reading the paper and writing a review is very important to help the students prepare for the class presentation and discussion. Every student is expected to deliver a **review of 3 papers and answer 8 technical questions from (a subset of) all other papers**. Each paper will be clearly marked as a paper for review or a paper for a technical question (which will be provided well before the class). Every review or answer to a technical question for a given paper has to be submitted **before the class**, having the class starting time as a hard deadline.
A ***review*** consists of a few paragraphs answering the following questions: (i) what is the problem, (ii) why it is important, (iii) why is it hard, (iv) why older approaches are not enough, (v) what is

the key idea and why it works (a list of at least three key points), (vi) what might be missing and how can we improve this idea (a list of at least three key points), (vi) an evaluation as to whether the paper supports its claims, and (vii) possible next steps of the work presented in the paper. The ideal size of the review is about *1 page, single column, 10pt font, 1-inch margin* (and it may only exceed 1 page if the student wants to elaborate on how to improve the ideas on the paper).

Class Project: Finally, this class requires a semester-long project and a final report in the style of a conference paper. The project will be either implementation-heavy or research-oriented. Students will work **individually or in groups of 2** (for the implementation project or the research project) and after the first two weeks, each team will have been associated with a specific project. Students can propose their own research project upon approval by the instructor. For each class project the student(s) must produce (i) a **proposal**, (ii) a **mid-term progress report**, and finally, (iii) a **final presentation** associated with (iv) a **code review**.

Evaluation and Grading

The course grade will break down as follows (minor alterations may occur).

Class Element	**Grade Percentage**
In-class participation	*5%*
Project 1	*15%*
Paper reviews	*9%*
Technical questions	*16%*
Paper presentation	*15%*
Project proposal	*5%*
Mid-semester project progress report	*5%*
Class Project	*30%*

Tentative Schedule

Week	**Topic**	Readings*
1	Introduction to COSI 167A	Architecture of a Database System, *Foundations and Trends in Databases*, 2007
2	Data Systems Fundamentals	Massively Parallel Databases and MapReduce Systems, *Foundations and Trends in Databases*, 2013 Column-Stores vs. Row-Stores: How Different Are They Really?, *SIGMOD* 2008
3	Data Layouts	Bridging the Archipelago between Row-Stores and Column-Stores for Hybrid Workloads, *SIGMOD* 2016 Data Structures for Data-Intensive Applications, *Foundations and Trends in Databases*, 2023
4	Modern Data Structures - I	LSM-based Storage Techniques: A Survey, *VLDB Journal*, 2019 Anatomy of the LSM Memory Buffer: Insights & Implications, *DBTest*, 2024

5	Modern Data Structures - II	Monkey: Optimal Navigable Key-Value Store, *SIGMOD*, 2017 Constructing and Analyzing the LSM Compaction Design Space, *VLDB*, 2021
6	Tuning Data Systems	LSM-Tree Under (Memory) Pressure, *ADMS*, 2022 ENDURE: A Robust Tuning Paradigm for LSM Trees, *VLDB*, 2022
7	Deletion in Modern Systems	Lethe: A Tunable Delete-Aware LSM Engine, SIGMOD 2020 Enabling Timely and Persistent Deletion in LSM-Engines, *TODS*, 2023
8	Indexing - I	The Adaptive Radix Tree: ARTful Indexing for Main-Memory Databases, *ICDE*, 2013 Adaptive Adaptive Indexing, *ICDE*, 2018
9	Indexing - II	Indexing for Near-Sorted Data, *ICDE*, 2023
10	Modern Hardware	ACEing the Bufferpool Management Paradigm for Modern Storage Devices, *ICDE*, 2023 A Parametric I/O Model for Modern Storage Devices, *DaMon*, 2021
11	Hardware-Conscious System Design - I	FASTER: A Concurrent Key-Value Store with In-Place Updates, *SIGMOD*, 2018 Cosine: A Cloud-Cost Optimized Self-Designing Key-Value Storage Engine, *VLDB*, 2022
12	Hardware-Conscious System Design - II	Relational Memory: Native In-Memory Accesses on Rows and Columns, *EDBT*, 2023
13	Project Presentations + Code Review	

*This reading list will be updated during the class.

Late Policy

The late policy stated below is only applicable for paper reviews and technical questions, and **NOT** for paper presentations, project presentations, project evaluations (mid-term and final), and code review. Every student is granted a cumulative allowance of 4 late days, serving as extensions for their individual assignments, and no penalties are incurred during this period. These late days can be utilized across various assignments. However, once the allotment of 4 late days is exhausted, a penalty of 20% per day is imposed on overdue paper reviews and technical questions.

Important Course Policies

Academic honesty

You are expected to be familiar with, and to follow, the University's policies on academic integrity. You are expected to be honest in all of your academic work. Please consult Brandeis University Rights and Responsibilities for all policies and procedures related to academic integrity. Allegations of alleged academic dishonesty will be forwarded to Student Rights and Community Standards. Sanctions for academic dishonesty can include failing grades and/or suspension from the university. Citation and research assistance can be found on the university library website.

Accommodations

Brandeis seeks to create a learning environment that is welcoming and inclusive of all students, and I want to support you in your learning. If you think you may require disability accommodations, you will need to work with Student Accessibility Support (SAS). You can contact them at 781-736-3470, email them at access@brandeis.edu, or visit the Student Accessibility Support home page. You can find helpful student FAQs and other resources on the SAS website, including guidance on how to know whether you might be eligible for support from SAS.

If you already have an accommodation letter from SAS, please provide me with a copy as soon as you can so that I can ensure effective implementation of accommodations for this class. In order to coordinate exam accommodations, ideally you should provide the accommodation letter at least 48 hours before an exam.

Respectful environment

Brandeis University is committed to providing its students, faculty, and staff with an environment conducive to learning and working, where all people are treated with respect and dignity. Please refrain from any behavior toward members of our Brandeis community, including students, faculty, staff, and guests, that intimidates, threatens, harasses, or bullies.

Laptop computer and cell phone use

To create a focused and distraction-free learning environment, we have implemented a no laptop/mobile policy during the class. This policy is aimed at maximizing student engagement, fostering active participation, and promoting a conducive atmosphere for effective learning. By minimizing the use of laptops and mobile devices, we aim to enhance the overall quality of the learning experience and encourage direct interaction with course materials and fellow students.

Financial aid for purchasing course materials

If you have difficulty purchasing course materials, please make an appointment with your Student Financial Services or Academic Services advisor to discuss possible funding options and alternative solutions.

Student Support

Success in this course depends heavily on your personal health and well-being. Recognize that stress is an expected part of the college experience, and it often can be compounded by unexpected setbacks or life changes outside the classroom. Your other professors and I strongly encourage you to reframe challenges as unavoidable pathways to success. Reflect on your role in taking care of yourself throughout the academic year, before the demands of exams and projects reach their peak. Please feel free to reach out to me about difficulties you may be having that may impact your performance in this course as soon as it occurs and before it becomes too overwhelming.