Northeastern University
Khoury College of Computer Sciences

Large Scale Information Storage and Retrieval

DS 4300 - Spring 2025

Course Overview

The relational data model has dominated industry since the 1970s. We will explore aspects of the efficiency of relational database management systems, including how data organization can affect performance. We will then turn to NoSQL (not only SQL) databases, including document databases, graph databases, key-value stores, and vector databases (we'll explore others if there is time). The course will also explore tools for distributed data processing like Apache Spark, stream/event processing, and other Big Data technologies. We will finally explore deploying these various systems and technologies in a cloud environment like AWS. The course will require substantial programming in Python and SQL, learning query languages and processes for NoSQL databases, and exploring systems in a containerized platform. The world of data engineering and big data is one of the most exciting and fastest-changing areas in technology today! Time to get started!

Meeting Time

MWR 9:15 am - 10:20 am - Mugar 201

Instructional Team

Professor:

Mark Fontenot, PhD m.fontenot@northeastern.edu
Office: Meserve Hall 353
Office Hours: M&Th 1:30 - 3:00 pm
If the above times do not work for you, please reach out via Slack, and we can schedule a better time.

Teaching Assistants:

- Iker Acosta Venegas - acostavenegas.i@northeastern.edu
- Dallon Archibald - archibald.d@northeastern.edu
- Nathan Cheung - cheung.nat@northeastern.edu
- Aryan Jain - jain.aryan1@northeastern.edu
- Abhishek Kumar - kumar.abhishe@northeastern.edu
- Junxiang Lin - lin.junx@northeastern.edu
- Edward Liu - liu.ed@northeastern.edu
- Sevinch Noori - noori.s@northeastern.edu

Learning Outcomes

1. Understand the efficiency-related concepts (including limitations) of RDBMSs
2. Understand data replication and distribution effects on typical DB usage scenarios
3. Understand the use cases for and data models of various NoSQL database systems, including

storing and retrieving data. Data models include document-based, key-value stores, graph based among others.

4. Access and implement data engineering and big-data-related AWS services

Class Webpage: https://markfontenot.net/teaching/ds4300/25s-ds4300/

Communication and HW Submission Platforms

Join Slack: https://join.slack.com/t/fontenotsclasses/signup

- You're already in my Slack org if you've taken one of my classes before. No need to make a new account
- Use your Northeastern Email to sign up.
- Join the **#25s-ds4300** channel
- Complete your Slack profile including a clear, professional headshot for your profile picture

CampusWire for Q&A: https://campuswire.com/c/G21E59038/feed

- Everyone registered by Jan 3 should be added to the class. If you registered after, just let me know and I will add you.

GradeScope:

- https://www.gradescope.com/courses/939399
- I've added everyone registered as of Jan 5. If you add the class after that date, please let me know after class or DM me on Slack

Textbooks

Through the Northeastern Library, you have free access to the O'Reilly for Higher Education service, which hosts thousands of modern professional texts on countless technical topics. You can gain initial access to this resource by following > this < link. When creating your account on O'Reilly, you must use your Northeastern email address. I've created a playlist of books on O'Reilly that you can access from >here<.

Computing

- Personal Laptops
  - Update your laptops to the most current version of the operating system!
  - Make sure you have 10-15 GB of free space for data sets and software
- Amazon Web Services
  - You will need to make an AWS account - aws.amazon.com. You will need to enter a credit card number.
  - Most, if not all, of the AWS services we cover will fall under the free tier. However, there may be some modest charges if you go outside the free tier. **You're responsible for these charges to your account**. Nothing we do for the class should exceed $75, which is cheaper than most CS textbooks.

Evaluation

The relative weights of the various assessment types are given below:

- Homeworks: 30%
- Practicals: 20%
- Midterm Exam: 20%
- Semester Project: 30%

Final Grade Scale Mapping:

- A 93 - 100
- A- 90 - <93
- B+ 87 - <90
- B 83 - <87
- B- 80 - <83
- C+ 77 - <80
- C 73 - <77
- C- 70 - <73
- D+ 67 - <70
- D 63 - <67
- D- 60 - <63
- F <60

Homeworks and Practicals:

- All assignments and project materials will be posted on the class webpage.
- Submission details will be contained within the assignment itself. Assignments will be submitted via GradeScope and/or GitHub. No assignments will be accepted by means other than what is indicated in the assignment (nothing will be accepted via email, slack, etc.).
- When submitting your assignments via GradeScope, **it is your responsibility to properly complete the submission process by associating each assignment question with the specific part of the PDF that contains your solution** for assignments where you submit a PDF. Failure to do this will result in a grade of 0 on the assignment.

Submission Deadlines:

- Rather than penalize late submissions, I prefer to incentivize early submissions. You can earn an extra 3% on each assignment that is submitted 48 hours **BEFORE** the stated deadline. (This does not apply to project submissions)
- No late submission of assignments will be accepted, except…
    - In recognition of "life happens," everyone gets ***one free 48 hour extension***, ***no questions asked*** on **one** homework/implementation assignment. (This cannot be used on any course project deliverables or exams.)
    - It is your responsibility to let Dr Fontenot know that you want to use your free extension on a particular assignment **BEFORE** the original due date. A late submission option has to be entered in GradeScope for you to be able to submit.

Semester Project:

A team project will take place in the last third of the semester. It will allow you to build a cloud-based system to support some big data task. More information on the project will be released later in the semester.

Exams:

- There will be 1 midterm exam during the semester. The date is in the semester overview at the end of this syllabus.
- The midterm will include all material covered up to that point in the course
- If you need to miss the midterm for any reason, you must contact Dr. Fontenot **before the exam**. When a make-up exam is warranted, it may contain different questions and/or take a different form than the exam originally administered in class at the sole discretion of Dr. Fontenot.

There is NO FINAL EXAM during the finals week for this course.

Academic Conduct and Integrity

Submitting work that is not your own is **wrong**. Facilitating someone else in submitting work that is not their own is **wrong**. Unless expressly stated otherwise in an official course document or handout, I expect that all work you submit to be your own. You may not share any source code files, queries, other code, design documents, homework solutions, quiz or exam answers, etc. "Sharing" includes allowing (either actively or passively) someone access to your computer or to look at your screen where solutions might be displayed.

You must understand *everything* you submit for any assignment. For any submission, you should be prepared to explain it in detail to me in person.

I take academic integrity very seriously. **The penalty for any act of cheating or academic dishonesty will be a failing grade in the course and submission of the matter to OSCCR**. I reserve the right to impose a less severe penalty at my sole discretion. Any penalties that OSCCR imposes will be separate from the course penalties.

Classroom Environment

Northeastern University values the diversity of its students, staff, and faculty and recognizes the important contribution each makes to its unique community.

Respect is expected at all times throughout this course. In the classroom, it is expected that everyone is treated with dignity and respect. We realize everyone comes from a different background with different experiences and abilities. Our knowledge will always be used to better everyone in the class.

We strive to create a learning environment that welcomes students of all backgrounds. If you feel unwelcome for any reason, please let me or a TA know so we can work to make things better. If you feel uncomfortable talking to members of the teaching staff, please consider reaching out to your academic

advisor.

Accommodations

Northeastern is committed to providing equal access and support to all qualified students through the provision of reasonable accommodations so that each student may fully participate in the learning experience. If you have a disability that requires accommodations, please contact the Disability Access Services (DAS)

- https://disabilityaccessservices.sites.northeastern.edu/
- DASBoston@northeastern.edu
- 617-353-2675

Accommodations cannot be made retroactively and to receive an accommodation, a letter from DAS or LDP is required.

Northeastern University
Khoury College of Computer Sciences

Schedule of Topics (Tentative):

Unless otherwise stated on the handout, all homework assignments will be due on Tuesdays at 11:59 pm EST of the week listed.

Week:	**Topics:**	**HW/P:**
Week 1 (Jan 6)	Administrivia Processes for inserting and searching for data in RDBMSs	HW 1 Out
Week 2 (Jan 13)	Strategies/structures for efficiently searching large data sets	HW1 In Practical A Out
Week 3 (Jan 20)	*Monday - MLK Day - No Class* Other processes and structures that limit the efficiency of RDBMSs	HW2 Out
Week 4 (Jan 27)	Introduction to NoSQL Databases No SQL Data Models and Systems: K/V Stores	HW2 In
Week 5 (Feb 3)	No SQL Data Models and Systems: Document Databases	Practical A In
Week 6 (Feb 10)	No SQL Data Models and Systems: Graph Databases	Practical B out HW 3 Out
Week 7 (Feb 17)	*Monday - No Class - Presidents Day* Data Distribution and Replication	
Week 8 (Feb 24)	Introduction to Apache Spark & SparkSQL	HW3 In HW4 Out
Week 9 (Mar 3)	*Spring Break*	
Week 10 (Mar 10)	Cloud computing and AWS; AWS Service Categories Thursday - Exam Review	Practical B In
Week 11 (Mar 17)	**Midterm Exam on Monday March 17** AWS Compute; AWS Data Storage Options	HW 4 In
Week 12 (March 24)	Other important AWS services Intro to Data Engineering on AWS + Building Pipelines	HW 5 Out
Week 13 (Mar 31)	Data Engineering with AWS	
Week 14 (Apr 7)	Project Topics	HW 5 In
Week 15 (Apr 14)	Project Topics	Project Due April 15.

More information about the course project will be available later in the semester.