Northeastern University
Khoury College of Computer Sciences

ignLarge Scale Information Storage and Retrieval

DS 4300 - Fall 2024

Course Overview

The relational data model has dominated industry since the 1970s. We will explore aspects of 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 such as AWS. The course will require substantial programming in Python and SQL, learning query languages and processes for NoSQL databases, and exploration of 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 - Behrakis 310

Instructional Team

Professor:

Mark Fontenot, PhD   m.fontenot@northeastern.edu
Office: Meserve Hall 353
Office Hours: MW 3 - 4:30pm
If the above times do not work for you, please reach out via Slack and we can schedule a better time.

Teaching Assistants:

- Micah Pacis - pacis.m@northeastern.edu
- Amey Parab - parab.amey@northeastern.edu
- Tee Tesharojanasup - tesharojanasup.t@northeastern.edu

Learning Outcomes

1. Understand the limits of the relational model
2. Understand the data models of and how to use various NoSQL database systems
3. Understand data replication and distribution effects on typical DB usage scenarios
4. Ingest and query data with several database management systems
5. Access and implement big-data related AWS services

Northeastern University
Khoury College of Computer Sciences

Class Webpage: https://markfontenot.net/teaching/ds4300/24f-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 **#24f-ds4300** channel
- Complete your Slack profile including a clear, professional headshot for your profile picture

CampusWire for Q&A: https://campuswire.com/c/G2BF3D4D9/feed

- I've added everyone registered as of Sept 1, 2024. If you add the class after that date, please let me know after class or DM me on Slack.

GradeScope:

- https://www.gradescope.com/courses/851144
- I've added everyone registered as of Sept 1, 2024. 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<.

Evaluation

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

- Assignments: 55%
- Midterm Exam: 20%
- Semester Project: 25%

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

Assignments:

- 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 (not accepted via email, slack, etc.).
- When submitting your assignments via GradeScope, **it is your responsibility to properly complete the submission process by associating each question of the assignment 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 assignment. (This,too, 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:

There will be a team project in the 2nd half of the semester. It will be an opportunity for you to build something while exploring some of the course topics in more detail. More information on the project will be released later in the semester.

Exams:

- There will be 1 midterm exam during the semester. The dates are 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 our students, staff, and faculty, recognizing the important contribution each makes to our 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 is welcoming to 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.

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.

Schedule of Topics (Tentative):

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

Week:	Topics:	Assignments:
Week 1 (Sep 4 & 5)	Administrivia Relational Model - Indexing	Assignment 1 Out
Week 2 (Sep 9 - 13)	Relational Model OLTP vs OLAP Indexing and Transaction Processing	
Week 3 (Sep 16 - 20)	Relational Model - what are the limits?	
Week 4 (Sep 23 - 27)	NoSQL to the Rescue? The CAP Theorem Document Databases	Assignment 1 In; Assignment 2 Out
Week 5 (Sep 30 - Oct 4)	Key/Value Databases	
Week 6 (Oct 7 - 11)	Graph Databases	Assignment 2 In; Assignment 3 Out
Week 7 (Oct 14 - 18)	*Monday - No Class (Indigenous Peoples Day)* Vector Databases	
Week 8 (Oct 21 - 25)	Distributing Data Replication & Partitioning Introduction to Apache Spark & SparkSQL	Assignment 3 In; Assignment 4 Out
Week 9 (Oct 28 - Nov 1)	Spark & Spark SQL **Midterm Exam on Wednesday Oct 30** >> Project Kickoff <<	
Week 10 (Nov 4 - 8)	Spark Big Data in AWS	Assignment 4 In; Assignment 5 Out
Week 11 (Nov 11 - 15)	*Monday - No Class (Veteran's Day)* Big Data In AWS	
Week 12 (Nov 18 - 22)	Data Engineering in AWS	Assignment 5 In
Week 13 (Nov 25 - 29)	Project Work Time *No Class on Wed or Thursday for Fall Break*	
Week 14 (Dec 2 & 4)	Project Work and/or Project Presentations	Project Due

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