DS 4300
Large Scale Information Storage and Retrieval

Mark Fontenot, PhD
Northeastern University

Hi! 👋

- Mark Fontenot, PhD
  - Office: 353 Meserve Hall
  - Office Hours:
    - M & Th 1:30 - 3:00 pm

      (*If those times don't work, just DM me on Slack to set up an alternate time!)*
  - Usually, very available on Slack… so just DM me.
- m.fontenot@northeastern.edu

Teaching Assistants

Iker Acosta Venegas

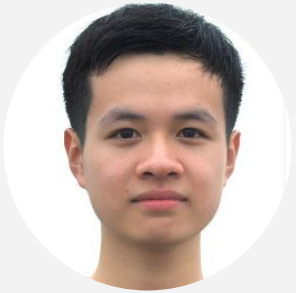
Dallon Archibald

Nathan Cheung

Aryan Jain

Abhishek Kumar

Eddy Liu



Junxiang Lin

Sevinch Noori

Where do I find … ?

- Course materials (Notes, Assignments, etc): https://markfontenot.net/teaching/ds4300/25s-ds4300/
- Assignment submissions (and grades): GradeScope
- Q & A Platform is CampusWire
- Quick DMs and Announcements will be on Slack

What's this class about?

- By the end of this class, you should
  - Understand the efficiency-related concepts (including limitations) of RDBMSs
  - Understand data replication and distribution effects on typical DB usage scenarios
  - 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.
  - Access and implement data engineering and big-data-related AWS services

Course Deliverables and Evaluation

Assignments

- Homeworks and Practicals
  - Usually due Tuesday Nights at 11:59 unless otherwise stated
  - 3% Bonus for submitting 48 hours early.
    (No… you can't get > 3% for submitting >48 hours early)
  - No Late Submissions accepted!
    - But… life happens…
      So everyone gets 1 free, no-questions-asked 48 hour extension.
    - DM Dr. Fontenot on Slack sometime ***before the original deadline*** requesting to use your extension.

Assignments

- Submissions will be via GradeScope and/or GitHub (unless directed otherwise)
  - Only submit PDFs unless otherwise instructed.
  - If only submitting a PDF, be sure to associate questions in gradescope with the correct page in your PDF.
    - Failure to do so may result in a grade of 0 on the assignment.
- All regrade requests must be submitted within 48 hours of grades being released on GradeScope. No Exceptions.

Monday, March 17

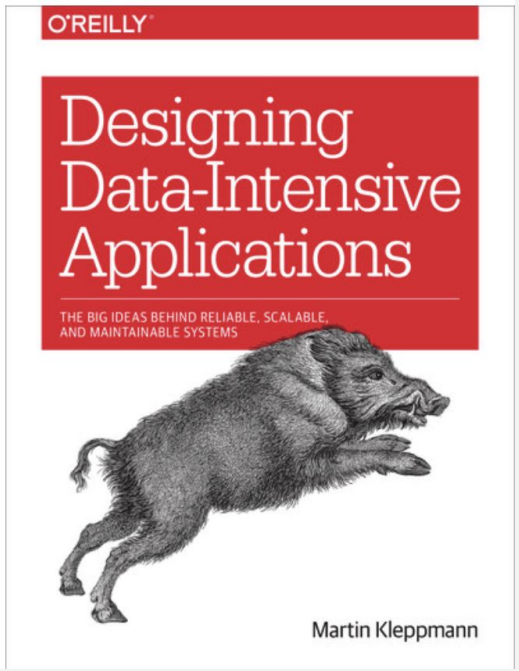
Mark it in your calendars now!

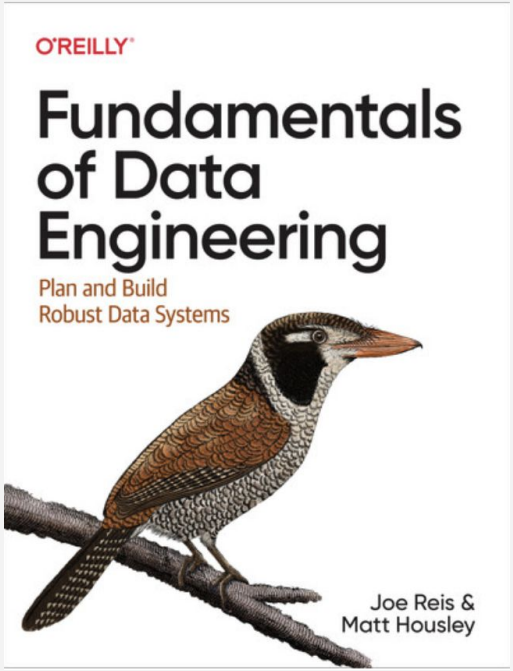
Final Grade Breakdown

- Homeworks (5) 30%
- Practicals (2) 20%
- Midterm 20%
- Semester Project 30%

Reference Materials

Primary Resources





Other books are in the playlist. I will add additional materials to the playlist or webpage as the semester progresses.

O'Reilly Playlist

Tentative List of Topics

- Thinking about data storage and retrieval at the data structures level
- How far can we get with the relational model?
- NoSQL Databases
  - Document Databases (Mongo)
  - Graph Databases (Neo4j)
  - Key/Value Databases
  - Maybe Vector Databases
- Data Distribution and Replication
- Distributed SQL DBs & Apache Spark/SparkSQL
- Big Data Tools and Services on AWS

Tools You Will Need to Install on your Laptop

- Docker Desktop
- Anaconda or Miniconda Python
  - *You're welcome to use another distro, but you're responsible for fixing it if something doesn't work (dependency conflicts, etc.)*
- A Database Access tool like Datagrip or DBeaver
- VS Code set up for Python Development
  - See > here < for more info about VSCode, Python, and Anaconda
- Ability to interact with git and GitHub through terminal or GUI app.

Topics to Review over the Next Few Days

- Shell/cmd Prompt/PowerShell CLI
  - Windows - if you want a Unix terminal: WSL2 or zsh on Windows
  - navigating the file system
  - running commands like pip, conda, python, etc
  - command line args
- Docker & Docker Compose
  - Basics of Dockerfiles and docker-compose.yaml files
  - port mapping
  - setting up volumes & mapping between host and guest OS

Is your Python Rusty *or* Haven't Done a ton with it?

- Python Crash Course by Net Ninja on YT
- On O'Reilly (See Python section of class playlist):
  - Python - Object-Oriented Programming Video Course by Simon Sez IT
  - E. Matthes - Python Crash Course, 3rd Edition - No Starch Press (not related to the YT video playlist listed above)

Expectations

- Conduct yourself respectfully
- Don’t distract your classmates from learning
- Don’t cheat!!
  - Do your own work unless group assignment
  - Discussing problems is encouraged, but you must formulate your own solutions
  - See Syllabus for details!

Let’s GOOO!