DS4300 - Large-Scale Storage and Retrieval

About the Course

John Rachlin, PhD
Northeastern University

Description

This is a survey course in data engineering using polyglot persistence. We will go beyond the relational database model that has dominated industry since the 1970's and explore the world of **NoSQL** (document, key-value, column, and graph) databases.

Main Topics:

- Data Engineering Principles
- NoSQL Databases
- Distributed batch processing with Scala and Spark

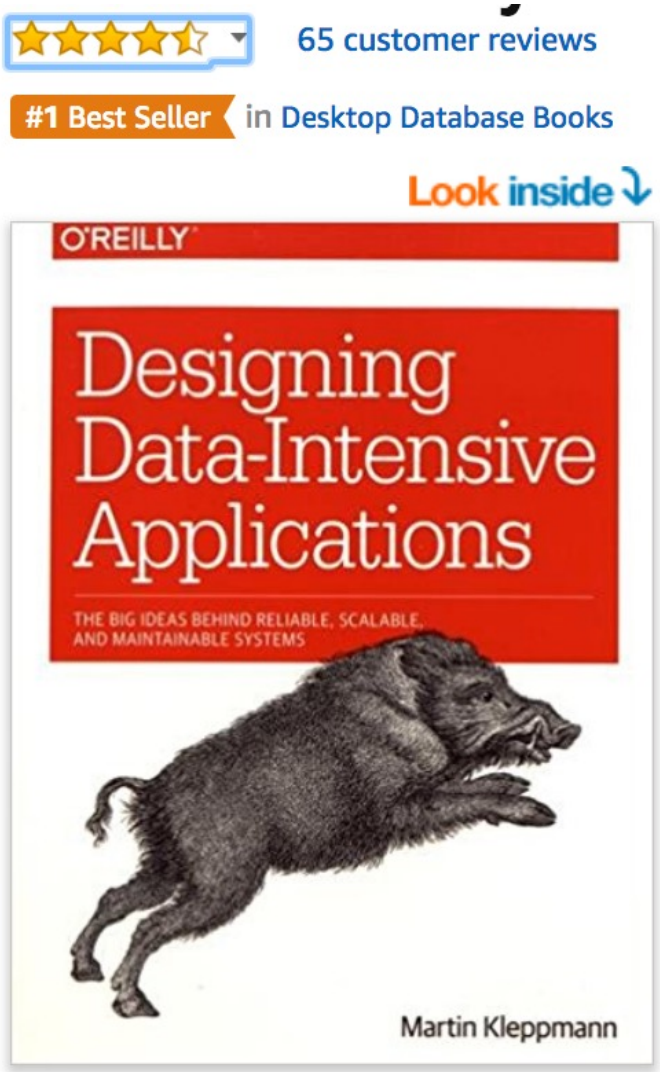
Bonus Topics (Time permitting):

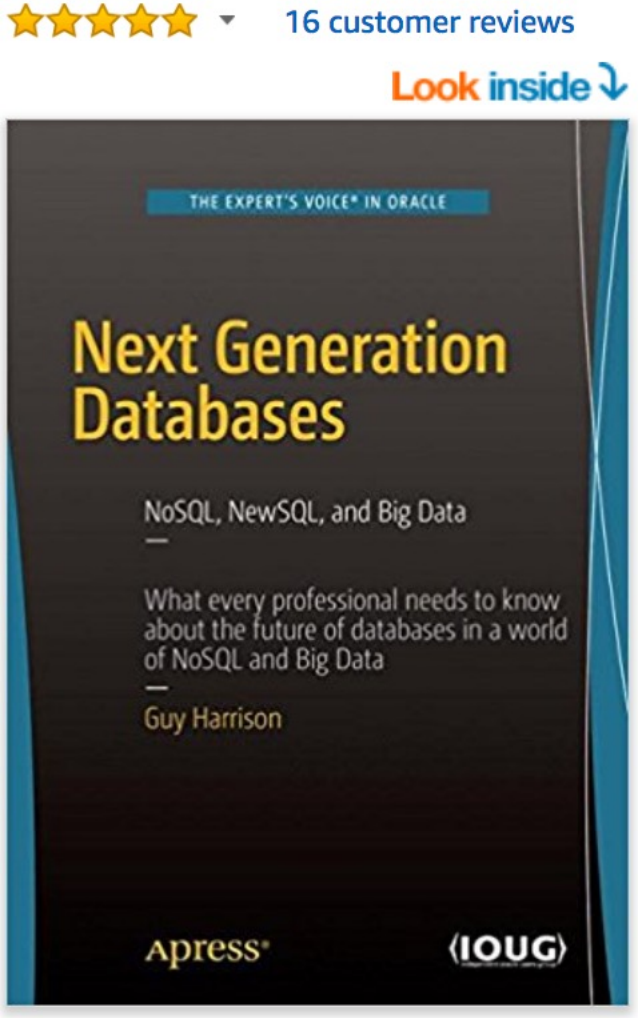
- AWS / Cloud computing
- Hadoop
- Elastic Search
- Event Processing using Kafka

Primary Textbooks:

http://proquest.safaribooksonline.com.ezproxy.neu.edu/

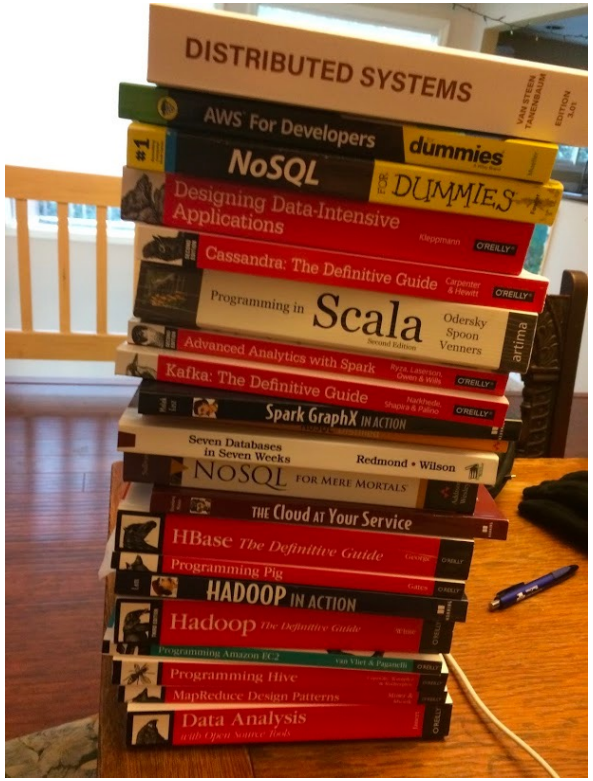
Both are freely available through your Safari / ProQuest account!





Other resources

Knowledge in this area is highly distributed and rapidly evolving!



Kleppmann, 2017. Designing Data-Intensive Applications, O'Reilly
Harrison, 2016. Next Generation Databases: NoSQL, NewSQL, and Big Data, Apress.
Sullivan, 2015. NoSQL for Mere Mortals, Addison Wesley
Fowler, 2015. NoSQL for Dummies, Wiley (Amazon $20)
Mueller, 2017. AWS for Developers for Dummies, Wiley
Steen & Tanenbaum, 2017. Distributed Systems, 3rd ed. (Optional - Amazon $35)
Redmond & Wilson, 2012. Seven Databases in Seven Weeks, Pragmatic Programmers
Odersky, Spoon, Venners, 2016. Programming in Scala, 3rd ed.
Ryza et al, 2017. Advanced Analytics with Spark, 2nd ed. O'Reilly
Carpenter & Hewitt, 2016. Cassandra: The Definitive Guide, 2nd ed. O'Reilly

… plus there are research papers, conference proceedings, on-line tutorials, video presentations, and a growing number of courses on Coursera, Udemy, EdX, and elsewhere!

Grading: Summer 2021

- 5-6 weekly homework assignments (70%)
  - Programming assignments
  - NoSQL database or Spark API evaluation

- Group Project (20%)
  - Application development using a NoSQL database
  - Integrated data modeling using NoSQL

- Quizzes (10% - take home / open book)
- No midterm / final

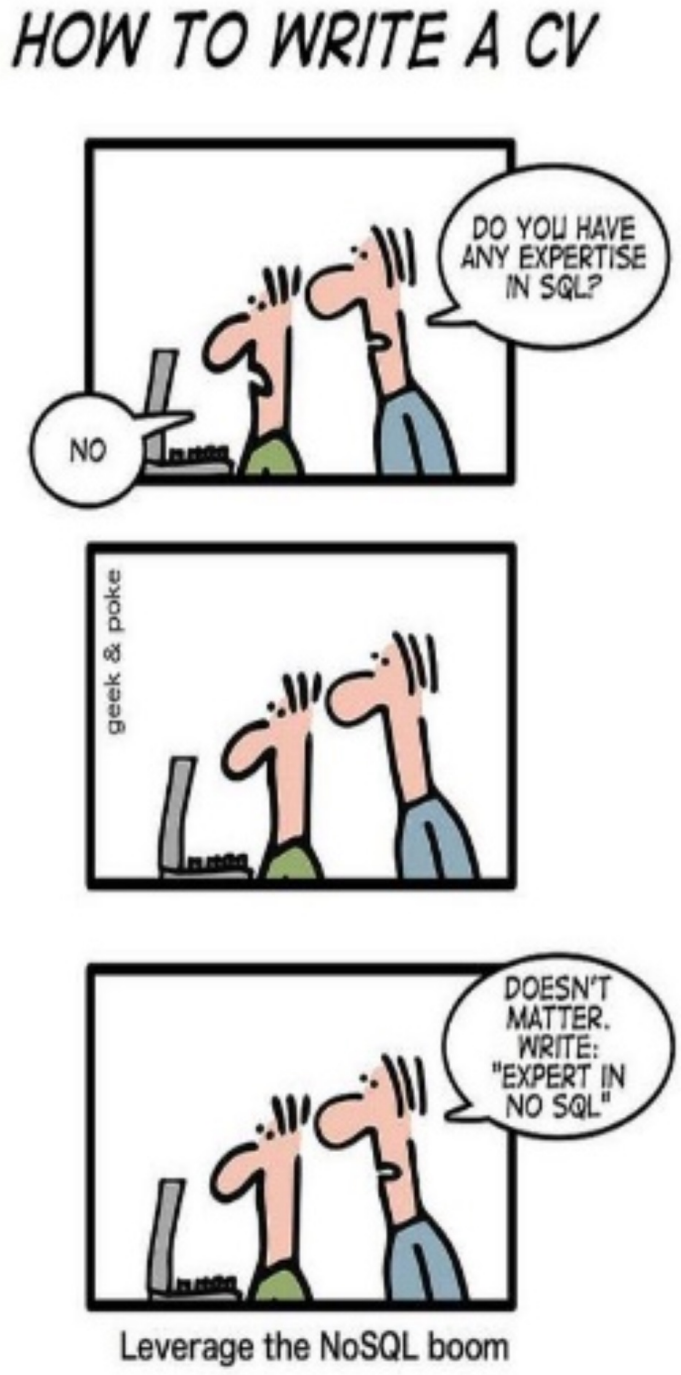
Research Assignment

From http://www.bigdata-startups.com/open-source-tools/

1. Pick a NoSQL database technology (SparkSQL, Redis, Neo4J, Mongo are excluded – covered in class.)
2. Learn about it
3. Prepare a two-page summary report or a 10-15 minute video presentation explaining salient features, learning resources, advantages and disadvantages
4. Prepare a re-usable demonstration script that a newcomer could use to understand the technology's unique features and capabilities.

What what will you gain?

- A deeper understanding of data processing and storage
- Appreciation for the rapidly changing world of big data
- NoSQL programming skills, data models, and use cases
- Speak intelligently about different data engineering architectures



You will enjoy this course if…

- You enjoy exploring new technologies and programming techniques
- You are interested in being able to architect and build data pipelines
- You want to take your data analytics skills to a deeper level
- You want to learn from other people's code (and are willing to share your own!)