INFO7250 – Engineering of Big-Data Systems – Fall 2018

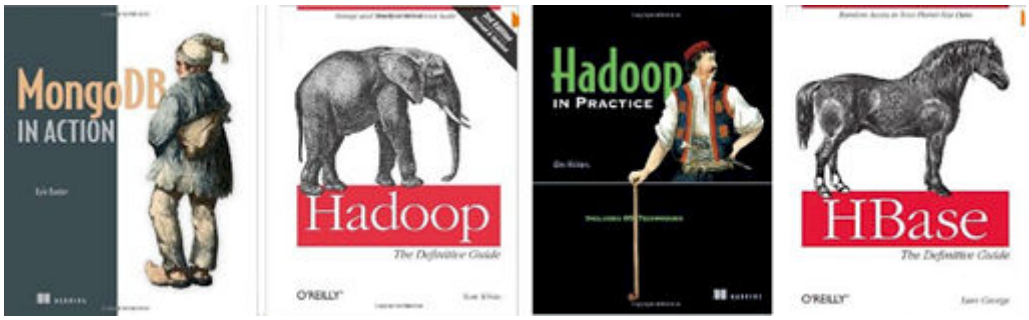
Instructor:
Yusuf Ozbek <yozbek@coe.neu.edu>

Description:
Services like social networks, web analytics, and intelligent e-commerce have promoted a rapid increase in the volume of data generated, analyzed and archived. These larger-volumes of data are too big for a traditional database, and are imposing challenges for storing, analyzing, and archiving. In order to solve the problems related to big data, a newer type of database products have emerged. These database products are collectively identified as NoSQL, and are quite varied with their unique features.

This class introduces a general framework for thinking about big data, and then shows how to apply technologies like Hadoop, HBase, MongoDB, and various NoSQL databases to build simple and efficient systems to manage and analyze big data, and describes an easy approach to big data systems that can be built and run by students. This course serves as an introductory course for Information Systems students who want to explore Big Data storage, processing, analysis, and application issues.

Prerequisites:
INFO5100 (or Java knowledge), and basic DBMS knowledge (or with permission).

Reference Books:
Note: Electronic versions are available on the Snell Library website, and Safari Books Online.



Core topics:
The course is divided into three main sections:

- Introduction to the Big Data, exploring challenges, trends, and applications in the Industry,
- Algorithms for Big Data analysis,
- Integrating Hadoop with other technologies.

Course Outline

- Introduction - a new paradigm for Big Data and NoSQL basics
- Getting Initial Hands-on Experience with MongoDB
- Replication in MongoDB
- Sharding in MongoDB
- Language Bindings - interacting and interfacing with NoSQL
- MapReduce
- Cassandra
- Hadoop
- Hadoop Integration
- Apache Pig
- Analyzing Big Data with Hive
- Hbase
- Mahout

Course objectives:
At the end of this course, students will

- understand the capabilities and pitfalls of Big Data systems, and how these issues are addressed.
- become familiar with the fundamental concepts of Big Systems and analytics.
- learn how to implement the MapReduce programming model.
- understand the challenges for Big Data Applications that deal with very large volumes of data.
- propose scalable solutions for problems related to Big Data.
- get hands-on experience on Big Data Analytics.

Assignments:
Project is the most important learning tool of this class. There will be weekly assignments during the semester, and one final project to apply all the technics learned in the class to make complex analyses on a given dataset.

Labs:
We will have multiple labs during the semester. These labs are based on Apache Hadoop, and we will use a virtual machine for most of the labs.

Final Project:
Students may use a given dataset, or explore a problem of their interest and propose their own solution.
The project has the following deliverables:

- 1 or 2 pages of project proposal, to be submitted after the mid semester, explaining why the problem is important, and what analyses are to be performed.
- Presentation of the project at the end of the semester
- A written report, to be submitted before the presentation, to highlight motivation, method, results, and conclusion.

Grading Policy:
Assignments: 10%
Midterm: 20-30%
Final Project: 20-30%
Final Exam: 20-30%

Attendance Policy:
Attendance is required. Students are responsible for any material covered in class. Lots of the materials covered in class will not be in the textbook. Announcements about homework, projects, programming assignments, etc. may be made in class or online or by emails.

Academic Dishonesty Policy:
Occurrences of academic dishonesty, such as copying and the submission of work that is not the student's own, will be dealt with according to the NEU's and COE's policies on academic dishonesty. In addition, students who allow their files or assignments to be copied are as guilty of academic dishonesty as those who copy and will be treated accordingly. Each student is responsible for taking reasonable precautions to ensure that his/her work is not available for unauthorized use. Students stealing or passing off class work as one's own will fail the class, and risk suspension from the MSIS program. Essential to the mission of Northeastern University is the commitment to the principles of intellectual honesty and integrity. Academic integrity is important for two reasons. First, independent and original school work ensures that students derive the most from their educational experience and the pursuit of knowledge. Second, academic dishonesty violates the most fundamental values of an intellectual community and depreciates the achievements of the entire University community. It is extremely important that the student understand that Northeastern University views academic dishonesty as one of the most serious offenses that a student can commit while in college. It is the student's responsibility to know and follow these standards/codes of ethics, which are part of the student's academic program. Please take the time to read what constitutes dishonesty and what the University is willing to do to respond to such incidents: http://www.northeastern.edu/osccr/academichonesty.html