(Large-Scale) Parallel Data Processing
CS4240 / CS6240
http://www.ccs.neu.edu/course/cs6240/

Spring, 2020
Prof. Gene Cooperman

Course Description:

The goal of the course is to gain an appreciation of the algorithms, hardware, databases, and big data languages being used for the analysis of "big data". The underlying hardware systems support for data and the algorithms to manipulate them together form the basis for a fundamental understanding of the subject. Among the concepts to be discussed are parallel file systems, disk partitioning, joins, parallel sorting algorithms, relational databases, graph algorithms (e.g., PageRank), and the CAP theorem for consistent distributed databases.

Finally, on the homeworks, I encourage students to share ideas orally, and even to share *small* excerpts of code. (Students often learn best from other students.) But the final work for the homework must be completely individual. Further, consulting the Internet for ideas is allowed only in the case of text-based articles (in English or another natural language), but *not* for code. Any violations will be considered as violations of academic integrity, and will be dealt with strictly.

Faculty Information:

Professor G. Cooperman
Office: 336 West Village H
e-mail: gene@ccs.neu.edu
Phone: (617) 373-8686
Office Hours: Tues. and Thurs.: 4:30 - 5:30; and by appointment.

Textbook: There is no single textbook. For readings, it is recommended to have done a first, cursory read of relevant material before class. The following resource materials are offered for texts:

1. "Hadoop: The Definitive Guide" by Tom White (Available from Safari Books Online.)
2. "MapReduce Design Patterns" by Donald Miner and Adam Shook (Available from Safari Books Online.)
2. "Hadoop in Practice" by Alex Holmes (Available from Safari Books Online.)
4. "Hadoop in Action" by Chuck Lam (Available from Safari Books Online.)
5. "Data-Intensive Text Processing with MapReduce" by Jimmy Lin and Chris Dyer. (Available online, see http://www.umiacs.umd.edu/∼jimmylin/book.html for info; or alternatively: http://lintool.github.io/MapReduceAlgorithms/)
6. "HBase: The Definitive Guide" by Lars George. (Available from Safari Books Online.)

Other online resources include documentation for the Hadoop and Spark APIs.

Exams and Grades:

The course grade will be based on 30% homework, 30% for the midterm exam, and 40% for the final exam. There are no deadlie extensions, except for a *major* emergency.

Topics (other side)

Topics (subject to updates):

Week	*Topics*
Jan. 6	Data and hardware trends, Cloud computing
Jan. 13	Amdahl's Law, Google File System, Hadoop's HDFS
Jan. 20	MapReduce, Hadoop, Spark
Jan. 27	Fundamental techniques in-mapper combining, sorting, secondary sorting (in-mapper combining, sorting, secondary sorting)
Feb. 10	Basic Algorithms (order inversion, per-record computation, group-by, global counters, random sampling and shuffling, quantiles, top-k)
Feb. 17	Joins: reduce-side join, replicated join, semi-join with Bloom filter
Feb. 23	Relational Databases
Mar. 10	CAP theorem, HBase, Hive
Mar. 17	Midterm exam
Mar. 23	Graphs: single source shortest path, PageRank
Mar. 30	Partitioning: Pairs and Stripes, theta-join
Apr. 7	Data Mining 1: clustering, classification
Apr. 14	Data Mining 2: ensemble methods, regression, matrix manipulation
Apr. 21	Week of Final Exams