CS 286: Topics - Solving Big Data Problems

The volume, velocity, and variety of data available today is astronomical. Organizations in almost every human endeavor can benefit from exploring and analyzing the abundance of data that is available from public and private sources. Analyzing data at scale using existing paradigms such as relational database management systems and high-performance compute clusters have proven either insufficient or too costly. This course explores the storage and analysis of large-scale data sets using clusters of commodity hardware.

This course is a comprehensive overview of Apache Hadoop for the Java programmer and is comprised of three main parts. The first part of this course explores the core of Apache Hadoop. The second part of the course explores the Apache Hadoop ecosystem. The third part of the course explores machine learning topics (collaborative filtering, clustering, and classification) using Apache Mahout.

instructor: James Casaletto

contact info: james.casaletto@gmail.com

meetings: Tu,Th from 7:30 to 8:45 pm (science bldg room 311)

office hours: Tu,Th from 6:45 to 7:30 pm (science bldg room 311)

pre-requisites: Java programming (CS 146)

grading:

91-100 / 81-90 / 71-80 / 61-70 / < 61

Your grade is based on a combination of in-class exams, individual labs, and team projects.

· 3 x in-class exams 60%

The exams are comprised of multiple choice and short answer.

· 3 x individual labs 30%

The labs include a Java MapReduce programming assignment (lab I) and a Hadoop ecosystem assignment (lab II).

· 1 x team project 10%

Randomly chosen teams of 2-3 people will work together to create an end-to-end solution using the Hadoop tools discussed in the course. Teams will choose from a list of solutions provided by the instructor. Grading is based on how completely and accurately the implementation provides the solution, as well as a presentation/demonstration of the solution in class. Teams are given 3 weeks to complete this project.

infrastructure:

Students will need a 64-bit laptop (Mac or Windows) with 8GB or more of RAM to run a VirtualBox virtual machine.

optional books:

Hadoop: The Definitive Guide (O'Reilly)

learning and performance objectives:

- install, configure, and use Apache Hadoop and its ecosystem
- write a MapReduce program in Java
- use machine learning libraries from Apache Mahout
- create an end-to-end big data solution using Apache Hadoop

tentative schedule:

Session	Description	Notes
Aug 20	Introduction to the course	Instructor and student intros; schedule; grading; policies; logistics
Aug 25	Introduction to big data	definition; use cases; paradigms
Aug 27	Introduction to the Apache Hadoop core (I)	HDFS and MapR-FS
Sep 1 (last day to drop)	Introduction to the Apache Hadoop core (II)	MapReduce historical; conceptual; data/execution flow
Sep 3	Installing and configuring MapR Hadoop	MapR deployment options; VirtualBox;
Sep 8	Using HDFS, NFS, and MapR-FS	hadoop fs/mfs; export/mount; resource management
	Managing MapReduce programs	MRv1 and MRv2
Sep 10	MapReduce	Write "hello world"

	programming in Java I	
Sep 15	MapReduce programming in Java II	Writable/comparable types; Input/output formats;
Sep 17	MapReduce programming in Java III	monitoring and testing MapReduce jobs
Sep 22	In-class exam I	MapReduce programming
Sep 24	Introduction to the Apache Hadoop ecosystem	Lab I due (MapReduce); high-level intro to components
Sep 29	Using Apache Sqoop, Flume, Kafka, and NFS	Data ingestion
Oct 1	Using Apache Pig, Hive, and Drill	Data transformation and analysis
Oct 6	Using Apache Spark I	RDD's; Spark Streaming
Oct 8	Using Apache Spark II	SparkSQL; MLLib
Oct 13	Guest speaker I	TBD
Oct 15	End-to-end solutions with Apache Hadoop (non-streaming)	data sources, ingestion, analysis
Oct 20	Introduction to data science	Project teams identified; project topic list distributed lab II ecosystem due
Oct 22	In-class exam II	Hadoop ecosystem
Oct 27	Introduction to machine learning	Historical; conceptual
Oct 29	Using Apache Mahout for recommendation	k-nearest neighbors
Nov 3	Using Apache Mahout for classification and clustering	Project selection due; Naïve bayes; K-means

Nov 5	Neural networking I	Perceptron
Nov 10	Neural networking II	Multi-layer perceptron
Nov 12	Using Apache Spark for classification and regression	Linear regression; decision trees
Nov 17	Guest speaker II	TBD
Nov 19	Project presentations II	Groups 1-6
Nov 24	project presentations II	Groups 7-12
Nov 26	no class meeting	Thanksgiving holiday
Dec 1	In-class exam III	Machine learning
Dec 3	Comprehensive review	lab III ML due
Dec 8	final exam	In-class