CS 467/567 Introduction to Big Data

Course Information

Class time: T/R 11:00 - 12:15

Classroom: ME 300

Prerequisites: Fluent in at least Python, Java, or Scala

Preferred: Background in Machine Learning, Data Mining, or Statistics

Piazza link: piazza.com/unm/fall2019/cs467567

Instructor: Trilce Estrada

Office: FEC 2390

Office hours: Tuesday 1:00 to 3:00

Appointments: https://calendly.com/trilce-estrada

Teaching Assistant: TBD

Office: TBD

Office hours: TBD

Course Description

The field of computer science is experiencing a transition from computation-intensive to data-intensive problems, wherein data is produced in massive amounts by large sensor networks, new data acquisition techniques, simulations, and social networks. Efficiently extracting, interpreting, and learning from very large datasets requires a new generation of scalable algorithms as well as new data management technologies.

In this course we explore key data analysis and management techniques, which applied to massive datasets are the cornerstone that enables real-time decision making in distributed environments, business intelligence in the Web, and scientific discovery at large scale. In particular, we examine the map-reduce parallel computing paradigm and associated technologies such as distributed file systems, no-sql databases, and stream computing engines. Additionally we review machine learning methods that make possible the efficient analysis of large volumes of data in near real time.

This course is highly interactive and based on the problem-based learning philosophy; students are expected to make use of said technologies to design highly scalable systems that can process and analyze Big Data for a variety of scientific, social, and environmental challenges.

Core Topics

The course is divided into three main core topics: (1) Introduction to the Big Data problem. Current challenges, trends, and applications. (2) Algorithms for Big Data analysis. Mining and learning algorithms that have been developed specifically to deal with large datasets.(3) Technologies for Big Data management. Big Data technology and tools, special consideration made to the Map-Reduce paradigm and the Hadoop ecosystem.

Course Objectives

At the end of this course, the student will become familiar with the fundamental concepts of Big Data management and analytics; will become competent in recognizing challenges faced by applications dealing with very large volumes of data as well as in proposing scalable solutions for them; and will be able to understand how Big Data impacts business intelligence, scientific discovery, and our day-to-day life.

Text Books

- **Mining of Massive Datasets** by Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman
- **Spark: The Definitive Guide - Big Data Processing Made Simple** by Bill Chambers and Matei Zaharia
- **Learning Spark: Lightning-Fast Big Data Analysis** by Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia

Coursework

Participation

Participation is the barometer of the class. Based on it I can determine if the pace of the course is too fast or too slow, it helps me to spot pitfalls and misconceptions, and it helps you to reinforce the material you learned.

The student can expect to have simple exercises frequently. Some of these daily assignments will be done in groups specified by the instructor and they will account for the participation grade of the course. Make up assignments will be allowed only if the instructor or TA were informed of a documented absence before the quiz took place.

Participation accounts for 15% of your final grade and won't be given for granted. You are required to participate either in class or electronically (through Piazza).

Homework

There will be a series of coding homework during the semester. For every homework students will turn in a two-page report and well documented code through **UNM Learn** only, no emailed assignments will be graded and no late assignments will be accepted.

Exam

Exams are this course's formal evaluation tool. In the exams students will be tested with respect to the learning goals of this course. Exams will comprise a mix of practical exercises and concepts. There will be only one midterm exam at around 3/4 of the semester. **The exam is open notes but only handwritten notes are allowed.**

Undergraduate students will be graded on 80% of the exam, graduate studens will be graded on 100% of the exam

Project

The final project is entirely to the discretion of the student (upon instructor approval). Students are free to explore a problem of their interest and propose their own solution. The project has the following deliverables:

Proposal. Maximum 2 pages of project proposal, why the problem is important, what has been done so far in the field, and what are the expected outcomes.

Presentations. Expect 2 presentations during the semester, each one will detail different aspects of your project and preliminary results are expected.

Poster and report. Maximum 10 page report highlighting consisting on the traditional sections of introduction, motivation, method, results, and conclusion

During the course we will hold bi-weekly brainstorming sessions to discuss and strengthen every proposed project.

Projects will be done in teams of 3 grad students or 4 students if they include at least 1 undergraduate student.

Grading

Grades will be based on your earned points, following this grade scale. You need to get the specified number of points or more to obtain the grade from the same column. Scores will be rounded to the closest integer value.

Incomplete can be assigned only for a documented medical reason. Change of grade to CR/NC after the semester deadline will be granted ONLY under special, documented extenuating circumstances.

A (95), A- (90), B+ (87), B (83), B- (80) C+ (77), C (73), C- (70), D+ (67), D (63), D- (60), F (le 60)

Participation15%

Homework15%

Exam30%

Project30%

Final presentation10%

Policies

Academic Honesty

Unless otherwise specified, you must write/code your own homework assignments. You cannot use the web to find answers to any assignment. If you do not have time to complete an assignment, it is better to submit your partial solutions than to get answers from someone else. Cheating students will be prosecuted according to University guidelines. Students should get acquainted with their rights and responsibilities as explained in the Student Code of Conduct

http://dos.unm.edu/student-conduct/academic-integrityhonesty.html

Any and all acts of plagiarism will result in an immediate dismissal from the course and an official report to the dean of students.

Instances of plagiarism include, but are not limited to: downloading code and snippets from the Internet without explicit permission from the instructor and/or without proper acknowledgment, citation, or license use; using code from a classmate or any other past or present student; quoting text directly or slightly paraphrasing from a source without proper reference; any other act of copying material and trying to make it look like it is yours.

Note that dismissal from the class means that the student will be dropped with an F from the course.

The best way of avoiding plagiarism is to start your assignments early. Whenever you feel like you cannot keep up with the course material, your instructor is happy to find a way to help you. Make an appointment or come to office hours, but DO NOT plagiarize; it is not worth it!.

Attendance

Attendance to class is expected (read mandatory) and note taking encouraged. Important information (about exams, assignments, projects, policies) may be communicated only during lecture time. We may also cover additional material (not available in the book or in slides) during the lecture.

If you miss a lecture, you should find what material was covered and if any announcement was made. If you have unexcused absences, this may result in participation points being deducted. Excused absences include sickness, attending conferences, job interviews, and similar. Even if your absence is excused, it is your responsibility to find out what material you missed. The professor is happy to answer specific questions regarding the lecture, but cannot go through all of the missed material on a one-to-one basis.

Excused absences have to be notified to the TA and instructor (through a piazza private post) at least 24hrs in advance, sickness has to be justified with a doctor's note

Communication

In order to facilitate interaction between students and to promote a broader participation, I created a Piazza group. Use the Piazza public group to ask general questions about homework, exams, projects, and lectures. You can also paste small snippets of code to clarify an idea. Students are encouraged to answer each others questions. Recall that your thoughtful participation in this forum accounts through your final grade. Use Piazza private posts to ask for excused absences and other personal matters. Always cc the class TA in those cases. Piazza is a discussion forum for the class and members are expected to conduct themselves with respect by posting comments and replies only in the context of the course.

Feedback

I value student's opinions regarding the course and I will take them into consideration to make this course as exciting and engaging as possible. Thus, through the semester I will ask students formal and informal feedback. Formal feedback includes short surveys on my teaching effectiveness, preferred teaching methods, and the pace of the class. Informal feedback will be in the form of polls or in-class questions regarding learning preferences. You can also leave anonymous feedback in the form of a note in my departmental mailbox, under my office door, or using this form. Remember that it is in the best interest of the class if you bring up to my attention if something is not working properly (e.g the pace of the class is too slow, the projects are boring, my teaching style is not effective) so that I can make the corrective steps.

ADA

In accordance with University Policy 2310 and the Americans with Disabilities Act (ADA), academic accommodations may be made for any student who notifies the instructor of the need for an accommodation. If you have a disability, either permanent or temporary, contact Accessibility Resource Center at 277-3506 for additional information.

Schedule

Topic	Subtopics	Readings
Mining Big Data and Applications	• The evolution of Big Data • Technologies contributing to its rise • Statistical Limits on Data Mining • Applications of Big Data and its future • Things useful to know: recap	MMD:CH1
Systems foundations of Big Data	• Distributed systems • Distibuted file systems	Tanembaum:CH
MapReduce and the New Software Stack	• Map-Reduce • Algorithms and complexity • Extensions to MR	MMD:CH2
Introduction to Apache Spark	• Apache Spark's philosophy • Running Spark • Spark architecture • Language's API • Spark sessions, dataframes, transformations, and actions	SDG:CH1, SDG

	• **HW:** Simple spark example	
Spark's Toolset and How Spark runs on a CLuster	• Datasets: type-safe structured APIs • Structured streaming • Machine Learning and advanced analytics • Lower-Level APIs and Spark’s ecosystem and packages • The life cycle of a Spark application • Execution details	SDG:CH3, SD
Mining Data Streams	• Advantages and challenges of stream processing • Stream processing design points • The stream data model • Sampling data and filtering streams • Estimating moments	MMD:CH4, SDG:CH16
Mining Data Streams Practice	• Architecture and abstraction • Streaming transformations • Output operations and input sources • Event-time and stateful processing • **HW:** spark streaming (in ch 22)	LSPK:CH10, SDG:CH22
Finding similar items	• Applications of Set Similarity	MMD:CH3

	• Shingling of Documents • Similarity-Preserving Summaries of Sets • Locality-Sensitive Hashing for Documents • Distance Measures • Applications of Locality-Sensitive Hashing • Methods for High Degrees of Similarity	
Clustering	• Clustering techniques • Clustering in non-Euclidean spaces • Clustering for Streams and Parallelism • **HW:** clustering with Spark	MMD:CH7
Link Analysis	• Page Rank • Efficient computation of Page Rank • Topic sensitive Page Rank • Links and authorities • **HW:** Page Rank ch4 in LSPK	MMD:CH5, LS
Mining Social-Network Graphs	• Clustering of Social-Network Graphs • Direct Discovery of Communities and Simrank • Partitioning of Graphs	MMD:CH10

	• Counting Triangles • Neighborhood Properties of Graph • **HW:** GraphX	
Recomender Systems	• A Model for Recommendation Systems • Content-Based Recommendations • Collaborative Filtering • Spark practice: recommendation systems with ALS • **HW:** Recommender system with MovieLens	MMD:CH9, SDG:CH28
Large-Scale Machine Learning	• The Machine Learning model • Learning from Nearest Neighbors • Linear regression • Support-Vector Machines • Decision trees • Perceptrons • Comparison of Learning Methods • **HW:** Page Rank ch4 in LSPK	MMD:CH12
Machine Learning with MLib	• Data types • Working with vectors • Feature extraction • Classification and regression • Model evaluation • **HW:** Spam classifier	LSPK:CH11, SDG:CH26

Neural Networks	• Introduction to Neural Nets • Dense Feedforward Networks • Backpropagation and Gradient Descent • Regularization • **HW:** Intro to	MMD:CH13
More Neural Networks	• Recurrent Neural Networks • Long Short-Term Memory (LSTM) • **HW:** RNN	MMD:CH13

- **MMD:** Mining of Massive Datasets
- **SDG:** Spark: The Definitive Guide - Big Data Processing Made Simple
- **LSPK:** Learning Spark: Lightning-Fast Big Data Analysis