CS 535 - Large Scale Data Analysis

Department of Computer Science - Boise State University

Syllabus

Catalog Description - CS 535

CS 535 LARGE-SCALE DATA ANALYSIS (3-0-3)(F)(Odd years). Covers algorithms and infrastructures for managing large-scale data, applying efficient algorithms based on MapReduce and other paradigms using current software packages for distributed data analysis. Storage of large-scale data using distributed file systems and distributed databases. Identifying and handling common pitfalls in large-scale data analysis. COREQ: CS 533

Instructor contact information

- Section 1: Amit Jain
  - **Office:** CCP 355 **Office hours:** posted on blackboard
  - **Email:** ajain@boisestate.edu **Phone:** 208-426-3821

Final Exam

Take home final exam will be assigned before or in the last week of classes.

Textbooks

No required textbook

Optional reference books

- *Learning Spark: Lightning-fast data Analysis* by Holden Karau, Andy Kowinski, Patrick Wendell, and Matei Zaharia
- *Advanced Analytics with Spark: Patterns for Learning from Data at Scale* by Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills
- *Data Intensive Text Processing with Mapreduce* by Chris Dyer and Jimmy Lin
- *Hadoop: The Definitive Guide*: *Storage and Analysis at Internet Scale* by Tom White

Software and Hardware

Having a Linux Virtual Machine installed on your personal machine will be handy though not required.

- Hardware - Minimum of 8GB of RAM and an i5 or equivalent processor (if you will run a virtual machine)

Learning Objectives

At the end of this course, the student will be able to:

- explain how large scale data analysis differs from normal data analysis
- understand different algorithms, distributed platforms, and tools and techniques used to analyze big data
- understand how to setup and deploy a Hadoop cluster for MapReduce and Hive
- understand how to setup and use Spark on a cluster
- utilize Spark and Hadoop to analyze large scale data for actionable insights

Major Topics Covered in the Course

- Introduction (0.5 week)
- Overview of various frameworks for distributed storage and analysis (0.5 week)
- Hadoop ecosystem: Hadoop base system, Hadoop Distributed File System (HDFS), MapReduce (2 weeks)
- More on MapReduce paradigm (1 week)
- Spark ecosystem (5 weeks)
- Using Spark with Hadoop HDFS and other distributed storage (2 weeks)
- Scaling and performance tips (1 week)
- Privacy and ethics (0.5 weeks)
- More on NoSQL Databases (1.5 weeks)
- Exams and review (1 week)

Online discussion forum

We will be using Piazza for class discussion. Rather than emailing questions to the teaching staff, please post your questions on Piazza. Find our class discussion forum page on the Piazza link in Blackboard. Please note that all important announcements will be made on Piazza so it is your responsibility to keep up. You should have already received an invitation to join Piazza. Please contact the instructor if you cannot locate the invite. The direct Piazza link for the class is: https://piazza.com/configure-classes/fall2019/cs535

Homework and Programming Projects

There will be several types of assignments throughout the semester. Documentation is an integral part of projects. Writing is expected to be professional, which includes adhering to grammar, spelling, capitalization, formatting and punctuation standards.

Programming assignments require the implementation of working programs using the language constructs and techniques described in class. Late homework will not be accepted. Make up exams will not be granted other than for exceptional reasons. All work is to be done individually unless explicitly allowed by the instructor. There may be group assignments during this class and those will be clearly marked.

Unless otherwise stated, hiring a tutor that has not been approved by the instructor or by the University could be considered academic dishonesty and could result in immediate failure of the course. If you wish to hire your own personal tutor then you must first contact your instructor.

Grading

You have **1000 points** available to earn over the semester.

- Projects: 600 points
- Homework: 100 points
- Midterm: 100 points
- Final exam (take home): 200 points
- Participation: -30 to +30 points

Final grade breakdown

Grades for this class are calculated by dividing the number of points earned over the semester by the total number of points available. There are no weighted categories. Final letter grades will be assigned according to the table below.

93% - 100%	A+
90% and less than 93%	A
87% and less than 90%	A-
84% and less than 87%	B+
81% and less than 84%	B
78% and less than 81%	B-
75% and less than 78%	C+
72% and less than 75%	C
69% and less than 72%	C-
66% and less than 69%	D+
63% and less than 66%	D

60% and less than 63%	D-
0 and less than 60%	F

Institutional Policies and Accessibility

- As a student you are responsible for knowing the policies that have been set at Boise State. Please review them at: http://registrar.boisestate.edu/general-information-and-policies/
- If you need help with accessibility you can visit the educational access center at: https://eac.boisestate.edu/
- Review the safety document located at: http://coen.boisestate.edu/cs/safetydocument
- Review the University attendance policy at: https://registrar.boisestate.edu/registration/attendance-policy/

Academic Honesty

Students are expected to work on their own unless explicitly instructed otherwise. Students who copy from each other or from any other source on assignments will be considered to be cheating as will students who allow their work to be copied. **Cheating is grounds for immediate failure of the course**. This includes trying to find answers to problems, programs, and exams from the Internet or other sources (and uploading your completed assignments to Internet sites that are publicly accessible). For more information, please visit the University's web page regarding academic integrity: http://registrar.boisestate.edu/general-information-and-policies/academic-integrity/