COMP 4334
Parallel and Distributed Computing for Data Science

Course Overview

This course will cover widely used parallel and distributed computing methods, focusing on datacenter-scale distributed software and methods such as Spark, MapReduce, distributed machine learning, and graph algorithms. We will study the types of algorithms that work well with these techniques and have the opportunity to implement some of these algorithms. We will also look at the types of hardware architectures that have been developed along with these computing methods.

Objectives

Students will be able to program in a distributed Spark environment. Students will be able to understand functional programming and the MapReduce framework. Students will be able to explain the fundamental hardware issues that led to current datacenter programming models. Students will have a general knowledge of issues in parallel programming and data movement within such systems. Students will gain knowledge of distributed machine learning and graph processing systems.

Textbooks and Materials

Although a textbook is not required for this class, the following book is recommended: Karau, Konwinski, Wendell, & Zaharia. (2015). *Leaning Spark* (1st ed.). O'Reilly.

Additional reading materials such as class slides and handouts will be provided.

Grading

9 labs: 36% total
9 quizzes: 36% total
2 assignments: 24% total
Participation/knowledge-check questions: 4% total

Assignment/Assessment	Points	Weight on Final Grade
Lab 1	10	4%
Lab 2	10	4%
Lab 3	10	4%
Lab 4	10	4%
Lab 5	10	4%
Lab 6	10	4%
Lab 7	10	4%
Lab 8	10	4%
Lab 9	10	4%

Quiz 1	10	4%
Quiz 2	10	4%
Quiz 3	10	4%
Quiz 4	10	4%
Quiz 5	10	4%
Quiz 6	10	4%
Quiz 7	10	4%
Quiz 8	10	4%
Quiz 9	10	4%
Assignment 1	20	12%
Assignment 2	20	12%
Participation/knowledge-check questions	100	4%

Grading Scale

A 93-100
A- 90-92.99
B+ 86-89.99
B 83-85.99
B- 80-82.99
C+ 76-79.99
C 73-75.99
C- 70-72.99
D+ 66-69.99
D 63-65.99
D- 60-62.99
F < 60

Assignment and Assessment Information

There are coding and testing components of the grade. For coding, there will be nine weekly programming labs and two larger programming assignments. The labs/assignments will all require you to write Python Spark code on your own machine and upload the .py file for grading. The details of each lab/assignment are provided in their own PDF files. For testing, there will be nine weekly quizzes and knowledge-check questions embedded into the asynchronous content. No late submissions of coding or testing are allowed without previous permission for special circumstances.

Weekly Schedule

Each week there will be asynchronous videos/PowerPoint slideshows and integrated knowledge-check questions to watch and complete. The knowledge-check questions are all multiple-choice and embedded at the end of each video section. You will have one chance to answer each question. They must be complete before the live session for the week starts. All the knowledge-check questions total up to 4% of the overall grade. Thus, to have the chance for full credit, one needs to complete the videos/knowledge-check questions each week before the live session starts.

Each week (except the first) there will be a lab due 12 hours before the live session. These labs are all programming based and focus on the content in the videos for the previous week. The labs will be discussed during the live session. For example, Lab 3 will be given out at Live Session 3 and be due right before Live Session 4. Each lab is worth 4% of your overall grade, for a total of 36% for all labs combined.

Each week (except the first) there will be a weekly quiz given after each of the nine labs are completed. The quiz will be released at the end of the live session in which the lab was due. You will have 48 hours to complete each quiz. Once you start a quiz, you will have 20 minutes to complete it. For example, Lab 3 is due 12 hours before Live Session 4. At the end of Live Session 4, Quiz 3 will be released. You will then have 48 hours to complete the 20-minute quiz. Each quiz is worth 4% of your overall grade, for a total of 36% for all quizzes combined.

There will be two larger programming assignments given at the middle and end of the quarter. They will be discussed when given out during a live session and be due approximately 2.5 weeks later. Each assignment is worth 12% of your overall grade, for a total of 24% for all assignments combined.

Week 1: Introduction and History

Videos/knowledge-check questions
No lab due this week

Week 2: Basic RDDs

Videos/knowledge-check questions
Lab 1 due 12 hours before the live session
Quiz 1 (take within 48 hours of finish of live session)

Week 3: Pair RDDs and MapReduce

Videos/knowledge-check questions
Lab 2 due 12 hours before the live session
Quiz 2 (take within 48 hours of finish of live session)

Week 4: Data Movement and Partitioning

Videos/knowledge-check questions
Lab 3 due 12 hours before the live session
Quiz 3 (take within 48 hours of finish of live session)

Week 5: Basic SparkSQL

Videos/knowledge-check questions

Lab 4 due 12 hours before the live session
Quiz 4 (take within 48 hours of finish of live session)
Assignment 1 assigned (due in Week 7)

Week 6: Advanced SparkSQL

Videos/knowledge-check questions
Lab 5 due 12 hours before the live session
Quiz 5 (take within 48 hours of finish of live session)

Week 7: SparkML

Videos/knowledge-check questions
Lab 6 due 12 hours before the live session
Quiz 6 (take within 48 hours of finish of live session)
Assignment 1 due 96 hours (4 days) after the start of the live session

Week 8: Spark Streaming

Videos/knowledge-check questions
Lab 7 due 12 hours before the live session
Quiz 7 (take within 48 hours of finish of live session)
Assignment 2 assigned (due in Week 10)

Week 9: GraphFrames

Videos/knowledge-check questions
Lab 8 due 12 hours before the live session
Quiz 8 (take within 48 hours of finish of live session)

Week 10: Distributed File Systems and Apache Spark

Videos/knowledge-check questions
Lab 9 due 12 hours before the live session
Quiz 9 (take within 48 hours of finish of live session)
Assignment 2 due 96 hours (4 days) after the start of the live session

Attendance Policy

Attendance at all live session meetings is mandatory.

Program Mission

Our MS in data science provides students with a broad course of study in programming, algorithms, statistics, and data management, as well as a depth of understanding in specific fields such as data mining, machine learning, and parallel systems. Graduates of the data science program go on to work in a wide variety of careers, including business, government, education, and the natural sciences.

Honor Code and Academic Integrity

All students are expected to abide by the University of Denver Honor Code. These expectations include the application of academic integrity and honesty in your class participation and assignments. Violations of these policies include but are not limited to

- Plagiarism, including any representation of another's work or ideas as one's own in academic and educational submissions
- Cheating, including any actual or attempted use of resources not authorized by the instructor(s) for academic submissions
- Fabrication, including any falsification or creation of data, research, or resources to support academic submissions

Violations of the Honor Code may have serious consequences including, but not limited to, a zero for an assignment or exam, a failing grade in the course, and reporting of violations to the Office of Student Conduct.

Diversity, Inclusiveness, Respect

DU has a core commitment to fostering a diverse learning community that is inclusive and respectful. Our diversity is reflected by differences in race, culture, age, religion, sexual orientation, socioeconomic background, and myriad other social identities and life experiences. The goal of inclusiveness, in a diverse community, encourages and appreciates expressions of different ideas, opinions, and beliefs, so that conversations and interactions that could potentially be divisive turn instead into opportunities for intellectual and personal enrichment.

A dedication to inclusiveness requires respecting what others say, their right to say it, and the thoughtful consideration of others' communication. Both speaking up *and* listening are valuable tools for furthering thoughtful, enlightening dialogue. Respecting one another's individual differences is critical in transforming a collection of diverse individuals into an inclusive, collaborative, and excellent learning community. Our core commitment shapes our core expectation for behavior inside and outside of the classroom.