Cloud Computing for Data Analysis
ITCS 6190/8190 – Fall 2024

Welcome to *ITCS 6190/8190 – Cloud Computing for Data Analysis*! This is a challenging course, encompassing a substantial amount of technical content and programming. Consistent study, timely completion of assignments and projects, class attendance, and seeking assistance when necessary are the key that will lead you to great success in this course.

This syllabus contains the policies and expectations established for the course. Please read the entire syllabus carefully before continuing in this course. These policies and expectations are intended to create a productive learning atmosphere for all students. Unless you are prepared to abide by these policies and expectations, you risk losing the opportunity to participate further in the course. Any modifications will be communicated through in-class announcements and/or updates on *Canvas*.

Course Description

This is a foundational course on cloud computing technology for data intensive applications. The course provides students with essential knowledge and practical skills on scalable and efficient data analysis. The topics were thoughtfully selected to guide students from fundamental concepts to more advanced aspects, providing a solid understanding of the domain. The course is grounded on the Apache Software Foundation's ecosystem, which hosts a vast and diverse ecosystem of open-source software projects covering a wide range of domains, in particular the Hadoop and Spark open-source frameworks. This ecosystem includes projects related to web servers, big data processing, data storage, machine learning, development frameworks, and more. However, besides focusing on the Hadoop and Spark open-source frameworks, the course also includes hands-on experience with Amazon Web Services (AWS) cloud infrastructure, a leading cloud service provider that offers a wide array of cloud computing services, including storage, computing power, machine learning, and big data processing tools.

Hadoop is a widely used open-source implementation of Google's MapReduce technology designed for distributed storage and processing of large sets of data using a cluster of commodity hardware. On the other hand, Apache Spark is an open-source distributed computing system designed for fast and large-scale data processing that provides an alternative to the traditional MapReduce model used in Hadoop, offering improved performance and ease of use. Both Hadoop and Spark have played significant roles in the development of commercial solutions available in the cloud, particularly for big data processing and analytics. In fact, many cloud service providers, including AWS, offer managed services and platforms that incorporate or support Hadoop and Spark to simplify the deployment and management of large-scale data processing tasks.

This course employs a balanced approach, combining concepts with hands-on exercises. Students will apply the learned principles to design and implement data analysis jobs on top of cloud computing technology. Emphasis is on collaborative learning, with opportunities for group work, discussions, and presentations.

Learning Outcomes

Upon completion, students will have a comprehensive understanding of principles on cloud technology for data analytics and the practical skills necessary to design and implement large-scale data processing tasks. This course sets the foundation for further specialization in the dynamic field of cloud computing and data analysis.

Location and Time

Tuesday, 11:30am-2:15pm, Dubois Center (Uptown) 801

Instructor

Marco Vieira
Office: Woodward 205C
Office Hours:
Tuesday, 10:30am to 11:30am, Dubois Center (Uptown) 713
Wednesday, 1pm to 2pm, Woodward 205C

Teaching Assistants *(see canvas for details)*

Aravinda Reddy Gangalakunta
Lakshmi Prayuktha Mudumba

Textbook(s)

Thomas Erl, Eric Monroy, "Cloud Computing: Concepts, Technology, Security, and Architecture", 2nd Edition, ISBN-13: 978-0138052256, Pearson, 2023.

Sridhar Alla, "Big Data Analytics with Hadoop 3", 1st Edition, ISBN-13: 978-1788628846, Packt Publishing, 2018.

Jules Damji, Brooke Wenig, Tathagata Das, Denny Lee, "Learning Spark: Lightning-Fast Data Analytics", 2nd Edition, ISBN-13: 978-1492050049, O'Reilly Media, 2020.

(optional) Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, "Mining of Massive Datasets", 3rd Edition, ISBN: 978-1108476348, Cambridge University Press, 2020.

Course Topics

Getting Started: Virtualization, containerization, and cloud concepts and models.

Hadoop: Big data analytics and Hadoop, HDFS and MapReduce, MapReduce programming model, and big data with Python and Hadoop.

Spark: Big data analytics and Spark, Spark structured APIs, Spark SQL and DataFrames, streaming analytics with Spark, and Machine Learning with MLlib.

AWS: AWS concepts, core AWS services, and Advanced AWS Services

Grading

Students will be evaluated through a combination of hands-on activities, practical assignments, presentations, course projects, and exams.

10-minute In-class Quizzes (10 points) - in most classes, students will complete 10-minute quizzes, which collectively account for 10 points out of the total 100. These quizzes are designed to assess students' understanding of the topics discussed in class and to encourage consistent engagement with the course material. Makeup quizzes will not be permitted unless prior arrangements have been made, or exceptional circumstances are documented and approved.

In-class Hands-on (10 points) - throughout the semester, students will participate in hands-on activities directly related to the topics discussed in class. These activities, which account for 10 points out of the total 100, will be conducted as group exercises, with ad hoc groups of 2 or 3 students. The outcome of each activity must be submitted on *Canvas* before the end of the day to receive credit. Points will be awarded not only based on the quality of the submission but also on attendance and punctuality, as attendance will be collected at the beginning of each class. This ensures that students who arrive on time and actively participate in the hands-on activities are recognized for their efforts.

In-class Presentation (15 points) - students will be organized into 10 groups of approximately 6 members each. Each group will be responsible for delivering one presentation on a designated course topic, with topics and presentation dates assigned randomly. The presentations, which should last between 40 to 50 minutes, will cover topics such as HDFS and MapReduce, Big Data with Python and Hadoop, and AWS Core and Advanced Services, among others. It is essential that all group members contribute meaningfully to the presentation, as individual performance will dictate the final grade. The presenting group is also expected to engage the class by posing questions to those in attendance. Guidelines for each topic will be provided to help structure the presentations. This activity accounts for 15 points out of the total 100. Grades may vary among group members based on their individual contributions. Groups will be defined during the first class, so everyone should be prepared to begin collaborating soon.

Assignments (10 points) - students will be required to complete 5 individual assignments, which together will account for 10 points out of the total 100. These assignments will be made available in advance, allowing students ample time to work on them. Each assignment is designed to reinforce the concepts covered in class and will require students to apply their knowledge independently. All assignments must be submitted on *Canvas* before the specified deadline to receive credit. Timely submission is crucial, as late work will not be accepted without prior approval.

Course Project #1 (10 points) + Course Project #2 (15 points) - students will complete two major group projects, each worth 15 points. The first project will focus on Hadoop, while the second will center on Spark. These projects will be completed in the same groups assigned for the presentations, allowing students to build on their collaborative efforts. Each group must submit their completed projects on Canvas before the specified deadline. Additionally, each group will present and discuss their projects with the TAs on a date and time to be determined and agreed

upon by both the group and the TAs. This review session will provide an opportunity for groups to explain their approach, answer questions, and receive feedback on their work.

Midterm (10 points) + Final Exam (20 points) - the course includes two major exams: a midterm and a final. The midterm, worth 10 points, will take place during class on October 8. The final exam, which will account for 20 points, is scheduled for a date yet to be determined. Both exams will be closed book and conducted on paper, requiring students to rely solely on their knowledge and understanding of the course material. These exams are designed to assess students' comprehension of key concepts and their ability to apply what they have learned throughout the semester.

Standard grading -

100%-90%: A
<90%-80%: B
<80%-70%: C
<70%: F

Prerequisites

Familiarity with Java, Python, SQL, Linux, Data Structures, and ML; good programming skills and a solid computer science background.

Required: ITCS 6114 or permission from department.

In-class Hands-on, Assignments & Project Submissions

Canvas will be used for assignment and project submissions. Regularly check canvas for important dates, materials, and class announcements.

Late submissions of assignments and projects will lead to a reduction of the grade, unless authorized by the course instructor. Grade reduction will be as follows: 20% after one day, 50% after two days, and 100% after three days. In other words, submitting three days after the deadline results in a zero grade.

Students may request to be regraded. Regrading of assignments can be requested by posting a message on *Canvas*. Regrading of exams or of the course project must be requested by email to the instructor. Grading of group work will consider the output of the entire group and each individual contribution. A final presentation of the work may be part of the assessment.

Policies

I. Course Materials

All lectures and course material will be available in *Canvas*. Lectures and course materials, including presentations, assignments, exams, outlines, and similar materials, are protected by copyright. You are encouraged to take notes and make copies of course materials for your own educational use. However, you may not, nor may you knowingly allow others to reproduce or distribute lecture notes and course materials publicly without my express written consent. This includes providing materials to commercial course material suppliers such as CourseHero, Chegg, and other similar services. Students who publicly distribute or display or help others publicly distribute or display copies or modified copies of an instructor's course materials may be in violation of University Policy 406, The Code of Student Responsibility, or University Policy 407, Code of Student Academic Integrity. Similarly, you own copyright in your original papers and exam essays.

II. Classroom Conduct

We will conduct this class in an atmosphere of mutual respect. Active participation in class discussions is greatly encouraged. Each of us may have strongly differing opinions on the various topics of class discussions. The conflict of ideas is encouraged and welcome. The orderly questioning of the ideas of others, including mine, is similarly welcome. However, I will exercise my responsibility to manage the discussions so that ideas and argument can proceed in an orderly fashion. You should expect that if your conduct during class discussions seriously disrupts the atmosphere of mutual respect I expect in this class, you will not be permitted to participate further.

III. Attendance and Absences

Students are expected to attend every class and remain in class for the duration of the session. Failure to attend class or arriving late may impact your ability to achieve course objectives, which could affect your course grade. An absence, excused or unexcused, does not relieve a student of any course requirement. Regular class attendance is a student's obligation, as is a responsibility for all the work of class meetings, including tests and written tasks.

The instructor has the authority to excuse a student's class absence(s) and to grant a student an academic accommodation (turn in a late assignment, provide extra time on an assignment, reschedule an exam, etc.). However, under Academic Affairs Policy on Course Attendance and Participation, University-sanctioned events or activities are considered excused absences. A University-sanctioned event or activity is one in which a student formally represents the University to external constituencies in athletic or academic activities. This policy does not supersede individual program attendance and/or participation requirements that are aligned with accreditation or licensure. For more information and student responsibilities to account for such an absence, see provost.charlotte.edu/policies-procedures/academic-policies-and-procedures/course-attendance-and-participation.

IV. Instructor's Absence or Tardiness

If I am late in arriving to class, you must wait a full 20 minutes after the start of class before you may leave without being counted absent, or you must follow any written instructions I may give you about my anticipated tardiness.

V. Non-Discrimination

All students and the instructor are expected to engage with each other respectfully. Unwelcome conduct directed toward another person based upon that person's actual or perceived race; color; religion (including belief and non-belief); sex; sexual orientation; gender identity; age; national origin; physical or mental disability; veteran status; genetic information; or for any other reason, may constitute a violation of University Policy 501, Nondiscrimination. Any student suspected of engaging in such conduct will be referred to the Office of Civil Rights & Title IX.

VI. University Policy on Withdrawals

Students are expected to complete all courses for which they are registered at the close of the add/drop period. If you are concerned about your ability to succeed in this course, it is important to make an appointment to speak with me as soon as possible. The University policy on withdrawal allows students only a limited number of opportunities available to withdraw from courses. It is important for you to understand the financial and academic consequences that may result from course withdrawal. See: provost.charlotte.edu/policies-procedures/academic-policies-and-procedures/withdrawal-and-cancellation-enrollment-policy

VII. Cell Phones or other Mobile Devices in the Classroom

The use of cell phones, smart phones, or other mobile communication devices is disruptive, and is therefore prohibited during class. Except in emergencies, those using such devices must leave the classroom for the remainder of the class period.

VIII. Computer use in the Classroom

Students are permitted to use computers during class for note-taking and other class-related work only. Those using computers during class for work not related to that class must leave the classroom for the remainder of the class period.

IX. Syllabus Policies, Academic Integrity, Plagiarism

All students are required to read and abide by the Code of Student Academic Integrity. Violations of the Code of Student Academic Integrity, including plagiarism, will result in disciplinary action as provided in the Code. Definitions and examples of plagiarism are set forth in the Code and on the Student Accountability & Conflict Resolution website. The Code is available from the Dean of Students Office or online at legal.charlotte.edu/policies/up-407. Additional resources are available on the Student Accountability & Conflict Resolution website.

Violation of these syllabus policies may result in appropriate academic penalties, including reduction of grade in the relevant assignment, project, or exam. If violation of these syllabus policies also implicates the Code of Student Academic Integrity because of alleged academic misconduct, I will follow the process outlined in the Code to address such cases.

X. Reporting Expectations

UNC Charlotte is committed to maintaining an environment conducive to learning for all students and a professional workplace for all employees. The University takes active measures to create or restore a respectful, safe, and inclusive environment for community members that is free from discrimination, discriminatory harassment, and interpersonal violence. If you (or someone you know) has experienced any of these incidents, know that you are not alone. UNC Charlotte has staff members trained to support you in navigating campus life, accessing health and counseling services, providing academic and housing accommodations, helping with civil protective orders, and more.

Please be aware that all UNC Charlotte employees, including faculty members, are expected to relay any information or reports of discrimination, discriminatory harassment, or sexual and interpersonal misconduct they receive to the Office of Civil Rights and Title IX. This means that if you tell me about a situation involving these matters, I am expected to report the information. Although I am expected to report the situation, you will still have options about how your case will be handled, including whether or not you wish to pursue a formal complaint. Our goal is to make sure you are aware of the range of options available to you and have access to the resources you need.

If you wish to speak to someone confidentially, you can contact the following on-campus resources, who are not required to report the incident to the Office of Civil Rights and Title IX: (1) Center for Counseling and Psychological Services (CAPS) (caps.charlotte.edu, 7-0311); or (2) Student Health Center (studenthealth.charlotte.edu, 7-7400). Additional information about your options is also available at civilrights.charlotte.edu under the “Students” tab.