

**DSE 6220: Big Data: Spark & Hadoop**

**Instructor: Jeremiah Lowhorn**

**Course Term: Summer 1 2024**

**On-line live meeting: Wednesdays 7:00pm ET**

**Office Hours: lowhornj@merrimack.edu**

# Course Description

This course will provide students with a comprehensive understanding of the big data processing foundation and techniques. Students will understand basic concepts of parallel computing, big data, Hadoop, MapReduce, and Spark. Students will develop skills to solve big data processing problems.

**Credit Justification Statement**

In this course, students will be expected to do the following amount of work:

* Live session: 1 hour per week
* Readings: 8 hours per week (on average)
* Labs: 10 hours per week (on average)
* Course Project 4 hours per week (on average)
* Total: 23 hours per week (on average)

# Course Learning Objectives

Upon completion of this course, students should:

* Understand Hadoop fundamentals including the core components and architecture.
* Understand Spark basics including RDDs, syntax, and partitioning.
* Understand common bugs and how to tune large jobs.
* Understand Spark streaming.
* Utilize MLLib for distributed machine learning.
* Demonstrate a thorough understanding of how to solve big data problems.

# Important Deadlines

|  |  |
| --- | --- |
| 5/6/2024 | First day of classes |
| 5/10/2024 | Last day to add/drop a course |
| 6/17/2024 | Last day to withdraw from a course with a “W” |
| 6/28/2024 | Last day of classes |

## Required Course Materials:

1. *Practical Big Data Analytics, Naturaj Dasgupta. ISBN: 978-1-78355-439-3*
2. *Learning Spark 2nd Edition, Jules S Damji, Brooke Wenig, Tathagata Das, Denny Lee. ISBN: 978-1-492-05004-9*

## Software:

The instructor will utilize Spark to illustrate all techniques during the instructional process.

# Grading & Assignments

|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **B** | **C** | **F** |
| **Excellent** | **Good** | **Fair/Poor** | **Unacceptable** |
| 93+ A | 87-89 B+ | 77-79 C+ | Below 70 |
| 90-92 A- | 83-86 B | 73-76 C |  |
|  | 80-82 B- | 70-72 C- |  |

Although the graduate grading policy is similar to the undergraduate policy, it should be noted that the expectations for graduate students are much higher and therefore the grading is more rigorous. Candidates for any graduate-level degree or certificate must attain a final cumulative grade point average of 3.0 before the degree or certificate will be conferred.

The following letter and special grades are used across all graduate programs. Please note that this document sets the minimum standards for the College. Individual degree programs may have stricter GPA and course grade requirements.

A: “A” indicates outstanding work

B: “B” means that the work is satisfactory

C: “C” (2.0) is deemed unsatisfactory at the graduate level. No more than two courses at the C level (2.0 or higher) will be counted as acceptable toward a graduate degree. Students may be permitted to repeat only two courses and may repeat each course only one time. Those who receive more than two C level grades will be automatically dismissed from their program of study.

Any grade lower than a C - will not be acceptable for graduate-level work and cannot be counted as credit towards the degree. However, the grade will be counted toward the graduate student’s GPA. If a student receives a grade lower than a C in a required course, the student will be required to retake the course. A course may only be retaken one time. The most recent grade will replace the first grade and will be factored into the student’s GPA.

Please see “Academic Requirements and Policies” in the Graduate Catalog at <https://catalog.merrimack.edu/index.php> for more information.

Student requirements and their corresponding weights are summarized as follows:

|  |  |
| --- | --- |
| **Assignment** | **Weight** |
| Course Project | 30% |
| Labs | 70% |

# Course Project

# The goal of this project is to showcase the student’s understanding of Big Data and Spark concepts using real-world data.

# Labs

There will be 7 Labs/Chapter Exercises throughout the course, each providing hands-on experience implementing and reinforcing the techniques and methods discussed during the Live Session and Readings.

# Grading-Related Policies for this Course

**Assignments** must be **submitted on Canvas by 11:59 pm of the due date listed in the Weekly Topics & Assignments section and in Canvas**. Please note that there will be **NO late work accepted**. Final grades will be based on a percentage of possible points earned, but the instructor reserves the right to adjust the course grade.

# Class Policies

**1. Academic Integrity**

See Merrimack College Student Handbook for college’s Academic Integrity Policy. The academic integrity code and policy is also posted in the Graduate Catalog: <https://catalog.merrimack.edu/content.php?catoid=9&navoid=202#academic-integrity>

Here is a brief excerpt from that: “…Academic integrity is fundamental to creating and maintaining an atmosphere of cooperation and trust. It is thus a concern for everyone in the college community. The academic integrity code below is designed to help students understand what is not permissible in their academic and intellectual lives at the college. It seeks to protect students from unintentional acts of dishonesty and to preserve the trust inherent in the student- teacher relationship, which is compromised if suspicion arises regarding the integrity of a student’s work. The code is also designed to inform students of the rules which will be used to judge academic integrity infractions…”

Specifically concerning the production of code it is important to recall that using code, text or other forms of media, such as homework solutions, from previous offerings of the course is forbidden, except as a “reference” or “guide” to work from, and even as such, **it is necessary to explicitly cite all consulted sources**. Tutors or other students who have taken the course in the past may help current students, but should not show them the solutions that they have developed for assignments. The lack of detailed citation of previous offerings of the course or any other similar sources, notably from the Internet or Generative AI systems, will be considered as submission of copied material, and a serious breach of academic integrity. Simply put, any work that is not your own should be clearly identified either through citations in written work or in comments in submitted files. This includes not only things that are directly copied from other’s work, but also for ideas and procedures gathered from outside sources. It is not necessary to cite the materials presented in class.

# Academic Accommodations from the Accessibility Services Office

Merrimack College provides reasonable accommodations for students with documented disabilities. Students who have, or think they may have, a disability are invited to contact the Accessibility Services Office via the online request form found on the Accessibility Services website: [www.merrimack.edu/aso](http://www.merrimack.edu/aso), email: [accessibilityservices@merrimack.edu](mailto:accessibilityservices@merrimack.edu), or by visiting us on the third floor of McQuade Library (in the event we are open).

Students are encouraged to contact the office as soon as possible to ensure adequate time to meet and create a plan. Students already registered with Accessibility Services are encouraged semesterly to request for their letters to be emailed and students are responsible to then email the PDF to their instructors personally. Accommodations cannot be made retroactively.

# Live On-line Meeting Attendance

Participation in weekly live on-line meetings is highly encouraged, but it is not mandatory. All sessions will be recorded, and recordings will be posted; however, when not attending students forego the ability to directly ask questions about topics being discussed, or other parts of the course.

# Communication

I will be communicating through official college email (merrimack.edu) and Canvas announcements as needed. Those will be the primary methods for dissemination of information between us, so please check your email at least once a day. I check my email several times a day, so it is the quickest way to reach me with questions. I will also be archiving important announcements through Canvas. When emailing me, use your Merrimack account, include a greeting with the class code and a closing signature with your name, or I will likely not respond.

If you are having trouble with an assignment, please do not suffer in silence. Let me know if you are struggling, either by email or by scheduling an appointment. I’m interested in seeing students succeed and am more than happy to provide extra help.

# 5. Requests for Extensions

The general policy is that, outside of properly verified serious medical emergencies\* (as defined below), extensions are not given, which applies to all 4 assignments. Missing an assignment without an acceptable reason (to be clear, that means a serious medical emergency, as defined below) will result in 0 points for the exam or a project. The intent here is not to penalize anyone – quite to the contrary, it is to create a level playing field so that no one has a unique and an unfair advantage. All assignment due dates are published (see the Live Meetings, Topics & Assignments section below) and will not change, barring a natural or other emergency – please consider those dates when planning any non-class related activities.

\**Serious medical emergency is defined as an injury or illness that is acute and poses an immediate risk to a person’s life or long term health. To be “properly verified”, the said serious medical condition must be attested to by hospitalization and related medical treatment documentation*.

# 6. Mental Health

It is important to make your mental health a priority! We will do activities in class to support your mental health and wellness, and I also encourage you to engage in your own self-care habits outside of class. If you want more information or resources, please come see me! If you are struggling with your mental health, or you believe a classmate is, please talk to me so I can put you in touch with qualified and caring support to get you back on track and feeling better.

# 7. Suicide Prevention Lifeline:

We can all help prevent suicide. The Lifeline provides 24/7, free and confidential support for people in distress, prevention and crisis resources for you or your loved ones, and best practices for professionals.

http://www.ulifeline.org/stay\_well

National Suicide Prevention Lifeline 1-800-273-8255(TALK) or Call or Text 988

# 8. Counseling Services:

Counseling Services are available to Graduate students at the Counseling Center at Merrimack via Uwill for telecounseling services only. Telecounseling includes video, chat, phone, message. Uwill is HIPPA and FERPA compliant which means they follow the strictest privacy guidelines. To access Uwill, register with your Merrimack email address at [app.uwill.com](http://app.uwill.com/). To receive support after hours call 978-837-5444 to be connected to the Uwill 24/7/365 crisis line which offers phone assistance only. The Counseling Center is located on the third floor of the Sakowich Center.

# 9. Student Success Resources:

It takes a village to progress in your knowledge. The numerous resources available at Merrimack College to assist you in meeting your goals are outlined in the **Data Science Hub in Canvas**.

For assistance in navigating the student success resources please reach out to the graduate advising team via [ecs-grad-advising@merrimack.edu](mailto:ecs-grad-advising@merrimack.edu)

# Weekly Topics & Assignments

|  |  |  |
| --- | --- | --- |
| Date | Topic | Readings & Discussions |
| Week 1:  Intro to Big Data  5/6 to 5/12 | -What is Big Data0  -Big Data Mining | -Practical Big Data Analytics: chapters 1, 2  *-*Lab 1 |
| Week 2:  Hadoop  5/13 to 5/19 | -Big Data with Hadoop | -Practical Big Data Analytics: chapter 4  *-* Lab 2 |
| Week 3:  Intro to Spark  5/20 to 5/26 | -Spark for Big Data | - Practical Big Data Analytics: chapter 6  - Learning Spark: chapter 1, 2  *-* Lab 3 |
| Week 4:  Working with Data  5/27 to 6/2 | - RDDs  - Working with DataFrames  - Loading and Saving Data | - Learning Spark: chapter 3  *-* Lab 4 |
| Week 5:  Advanced Spark Programming  6/2 to 6/9 | -Tuning and Debugging  -Spark SQL | - Learning Spark: chapters 4, 7  *-* Lab 5 |
| Week 6:  Spark Streaming  6/10 to 6/16 | -Spark Streaming | - Learning Spark: chapters 8  *-* Lab 6 |
| Week 7:  MLLib  6/17 to 6/23 | -Distributed Machine Learning | - Learning Spark: chapter 10  *-* Lab 7 |
| Week 8:  Deployment  Final Project  6/24 to 6/28 | Managing, Deploying, & Scaling ML Apps  -Course Project Due | - Learning Spark chapter 11  -Course Project Due |