CS6240: Large-Scale Parallel Data Processing

For all general course information such as credit hours, format, meeting times and location, please refer to the registrar system for the latest information.

Instructor Information:	Dr. Mirek Riedewald
Office Hours:	this information will be posted on Canvas
Email:	m.riedewald@northeastern.edu
T.A.:	this information will be posted on Canvas

Special policies and requirements due to the ongoing COVID crisis:

This course will be delivered using the Hybrid NUflex learning modality and I will be teaching remotely. I will join you virtually in the class at the scheduled class time using Zoom with some students in the classroom and others joining remotely. You will be able to ask questions, discuss, and interact with me and other students in real time. Remember that on your scheduled days in the classroom, you will need to practice healthy distancing and wear a face mask or face covering. I will also be available for virtual office hours.

This course, or parts of this course, might at some point be recorded for educational purposes. These recordings will be made available only to students enrolled in the course, instructor of record, and any teaching assistants assigned to the course.

Only students who have arranged an accommodation with the Disability Resource Center may use mechanical or electronic transcribing, recording, or communication devices in the classroom. Students with disabilities who believe they may need such an accommodation may contact the Disabilities Resource Center.

- **This is not an online course!** If you believe you will have difficulty attending the lectures "live," please contact the instructor immediately.
- Synchronous lectures: You are expected to attend all lectures during the regular class times. Real-time interaction with the professor and other students in class are an essential aspect of the learning experience in this course.
    - For students interested in attending lectures in the classroom, a scheduling system managed by the university will control room occupancy. Pay close attention to university announcements about this.
- Video recording: We wanted to record lectures for offline viewing, but following advice by university legal experts, we will unfortunately not be able to do so for the time being. In short, whenever students appear in a recording, even if it is via chat or by drawing on a shared whiteboard, there are privacy implications. Things become even more challenging when a person is located in a different state or country that have their own privacy laws and requirements. Unless the university can guarantee the instructor absolute immunity from any possible recording-related legal issues, no lectures will be recorded.
    - For the same reasons, we also ask that no student or TA record any lectures or other course-related interactions, e.g., office hours. This is very important

and serious. We understand that you may want to have those videos for study purposes, but you are risking serious legal consequences.

- Office hours will be held remotely by default. We will use Zoom, but all interactions are still happening in real-time. We will use screen sharing and virtual whiteboards.
- Exam format: It is your responsibility to ensure that you will have a stable and reliable Internet connection to Gradescope during the exam time window.
    - We will provide a practice exam for you to explore and test your setup.
    - We strongly advise to have a backup option ready. For example, if your default Internet access is via a cable provider, set up Internet access via your cell phone service as well. Maybe you can plug your phone's SIM card into your laptop or tablet, or you can use tethering to let your cell phone serve as an Internet access point for your computer. Or maybe there is a Café in your neighborhood that provides good Internet access. Taking the exam requires very little bandwidth and even a slow connection should be fine.
    - Important: We understand that these are difficult circumstances and we want to help you get through them as much as possible. However, we also must ensure a fair exam environment that discourages any cheating attempts. Hence, we unfortunately cannot consider accommodations or exceptions due to technical difficulties on your side. The reason is that we have no way of verifying if somebody really suffered from an Internet outage or just made it up because they were not sufficiently prepared and wanted to force a time extension.
    - If you believe that you may likely suffer from Internet-connection issues, contact the instructor immediately. One option would be for you to take the exam in the classroom. However, seating capacity is limited due to COVID restrictions and this option may or may not be available, e.g., if the governor or mayor prohibit on-campus instruction.

End of special COVID part.

Please be aware of the following policies:

- There are **no** deadline extensions or make-up assignments/exams, except if you have a major emergency. You have to provide evidence in order to claim such an emergency and you have to inform the instructor *as soon as possible*. The following are examples for situations that do *not* qualify as emergencies:
    - I have a job/co-op/internship interview scheduled.
    - My other course has an exam.
    - My other course has a major homework or project deadline.
- We understand that some weeks are busier than others, but that's how things will be in your future job as well. By announcing deadlines well in advance, we give you the opportunity to plan and schedule your work accordingly. Make sure you start early so that you have the flexibility for dealing with unexpected issues.
- **Honor Code**: All students must adhere to the Northeastern University honor code available on the Northeastern web site and the graduate student handbook.

    - Please note that you are *not* allowed to share homework solutions with others, or copy anybody else's homework entirely or in parts. We will check for originality during the grading process.
    - Violations will be reported to OSCCR.

Course Prerequisites and Description: See the official information in the course catalog.

Course Format & Methodology: This course runs for a total of 15 weeks and contains online content accessible through http://khoury.northeastern.edu/~mirek/teaching.htm and https://canvas.northeastern.edu/ Each week (or module) contains one or more lessons, which need to be completed by Sunday of the week *before* the module is discussed. **Please note that all due dates and times are specified according to the local Boston time (Eastern US time zone)**.

Recommended Textbook & Materials: To gain a deeper understanding of the material covered in this course, we recommend the following books, most of which are available online (and for free) for Northeastern University students from Safari Books Online:

- *MapReduce Design Patterns* by Donald Miner and Adam Shook
- *Hadoop: The Definitive Guide* by Tom White
- *High Performance Spark* by Holden Karau and Rachel Warren
- *Spark in Action* by Petar Zecevic and Marko Bonaci
- *Programming Elastic MapReduce* by Kevin Schmidt and Christopher Phillips

For some topics we will work with research papers or other online resources, e.g., the Hadoop and Spark API doc.

Course Outcomes: This course has the following main objectives and content:

- Get an overview of the big-data-processing landscape.
    - We will discuss some trends and challenges and briefly survey alternative approaches.
- Learn how to design distributed algorithms for processing big data, and how to implement them in Hadoop MapReduce and in Spark. While MapReduce or Spark might be replaced at some point by other systems, the algorithm design patterns taught in this course will remain relevant, because they are concerned with partitioning of a problem, assigning data to many machines, and then performing local computation in parallel on these machines.
    - We will cover a variety of fundamental problems and design patterns, including join computation, graph algorithms, information retrieval and data mining techniques, and analyze how they can be implemented in a scalable manner.
- Get hands-on practice writing code and running it on many processors.
    - We will work with Hadoop MapReduce and Spark.
    - We will use the Amazon Cloud to run the code. **You need to have your own Amazon Web Services (AWS) account** to do this, for which you might need

to register with your personal credit card. Amazon typically offers $100 in free credit, so be sure to explore this option. However, there is no guarantee that Amazon will give out this credit.

- Understand the system architecture and functionality below MapReduce and Spark.
    - We will discuss features and limitations of MapReduce and Spark.

Notice that we cannot cover all possible parallel-computation approaches. You are encouraged to explore other courses on related topics. Also note that new approaches for big-data processing keep appearing, often trying to address some weakness of existing ones. We will not be able to cover them at this point, but a solid understanding of parallel-data-processing principles will help you evaluate their tradeoffs—something the marketing people probably will not tell you about…

Participation and Engagement: Your presence in peer-to-peer activities serves as an indicator of your level of engagement and effort throughout the course. Frequent and varied (e.g., synchronous/asynchronous/face-to-face) opportunities to receive feedback, help, and clarification on course material from the instructor are provided throughout the term. The following activities count towards class participation:

1. Asking or answering questions in class.
2. Submitting solutions for in-class exercises when requested by the instructor.
3. Answering questions or posting relevant information in the discussion boards. Participation points are awarded based on quality and quantity of contributions.

Communication/Submission of Work: Make sure you receive course-related announcements the day they are made. Guidelines for completing and submitting each assignment are posted along with the assignment. Late and early homework submission policies will be announced with the individual assignments.

Course Activities and Assignments:

- **Weekly reading/viewing** Weekly readings provide the background knowledge, terminology, and examples you need to understand and apply fundamental course concepts. You must complete/view all assigned readings, presentations, and demonstrations included in the lessons. All materials should be completed by the due dates specified.

- **Self-checks** When available, complete self-checks about the online lecture material designed to enhance your current understanding and ability to correctly apply concepts covered in weekly readings and presentations. The grading is based on how many self-check questions you have answered correctly in the *first* self-check you submit for the module. Getting a few questions wrong does not result in any deduction, unless it looks like you are guessing. Notice that you must complete the self-check for a module *by midnight on Sunday, **before** the module is discussed*. As a rule of thumb, if you have carefully studied the material and made a serious attempt to answer all the questions,

then you will earn full marks.

- **Exam** You will complete an exam designed to test your understanding of the course concepts. The exam is **closed-book**, i.e., you cannot bring any material other than a writing instrument. ~~Students in hybrid sections of the course have to be present in the lecture room for the exam. Online students on other campuses have to attend the proctored exam there in person at the announced date and time.~~ Due to the COVID crisis, you will take the exam online on Gradescope during the time window that will be announced as we approach the exam date.

- **Homework/project** You will complete multiple homework assignments that give you the opportunity to practice the concepts you learn. More information about these assignments and the course project is available in Canvas.

Course Grading Criteria:

- Self-checks: 5%
- Participation: 15%
- Exam: 60%
- Homework/project: 20%

Class Schedule / Topical Outline:
This schedule is subject to updates.

Module	Topics	Assignments
1	Trends, Cloud Computing, Parallel Processing Basics	
2	Distributed Services: Distributed File System, Resource and Application Management	Begin Homework 1
3	MapReduce and Spark Overview	**Homework 1 due**
4	Fundamental Techniques	Begin Homework 2
5	Joins	**Homework 2 due**
6	Common Algorithm Building Blocks	Begin Homework 3
7	Graph Algorithms	**Homework 3 due**
8	Data Mining 1 (K-Means, Decision Trees)	Begin Homework 4
9	Data mining 2 (Ensembles)	**Homework 4 due**

10	Intelligent Partitioning	Begin Project
11	More About Spark	**Project Progress Report due**
12	**Exam**	
13	CAP, HBase, and Hive; Flexible Topics	
14	Flexible Topics	**Project reports due**
15	**Project Presentations**	

How to Succeed in this Course

This is an advanced graduate course about an evolving topic. It is therefore essential that you go through the online material carefully and methodically, attend the lectures and participate in online discussions. Homework is designed to help you understand the material and prepare for the exam. The following often works well:

1. When going through the online material, make notes about questions you have or about material you find difficult to understand. Then share these questions through the online forum or in class.
2. When you get a question in a check-your-knowledge quiz wrong or were not sure about the answer, go back to the corresponding online material and try to find the answer.
3. After going through an online lecture, try to explain the material to yourself or to a friend. This way you can better judge if you understand it. Once you identified things that need clarification, try to find the answer yourself by consulting one or more of the recommended books. If you cannot find the answer with reasonable effort, ask others for help (online discussion forum, office hours, and in-class discussions).
4. Start working on homework assignments as soon as they come out. This way you have time to ask questions and get help.

Is This the Right Course for You?

This really is an *algorithms* course at heart. You will write plenty of code, but the main emphasis is on learning how to approach big-data analysis problems. You will need solid Java programming skills to succeed, but we are not teaching any Java basics in this course. You do not need advanced Scala skills, and should be able to pick up what you need on-the-fly with reasonable effort.

- If you believe that programming in Java or Scala presents an insurmountable barrier for you, contact the instructor during the first week of classes to find a solution. It is possible to program in other languages, but we generally cannot provide any support for them—so you may be on your own if you get stuck. Students in the past completed their homework successfully using Python for both MapReduce and Spark. Python is well supported in Spark and the programs often look similar to those written in Scala.

We are learning about novel techniques that are only partially understood and explored by the research community. Hence in many cases there are no "certain truths." At times we might find better solutions that could be publishable in a research paper.

We are working with complex cutting-edge software from the open-source community. This means that there will be bugs, lack of documentation, and simply inexplicable behavior at times. Hadoop and Spark also keep changing and updating their API, therefore some code you find in books or on the Web might be outdated or use deprecated features.

When dealing with big data in a complex environment such as MapReduce/Spark and AWS, developing and debugging code is different compared to traditional settings. Sometimes a task might appear easy but turns out to be much harder and more time-consuming (or the other way round).

You should only take this course if you are prepared to deal with such issues and are willing to put in extra time when necessary. Do not take this course if you want a well-polished and well-tested course without any uncertainty. If you are genuinely interested in the topic and are ready to work around the inevitable frustrations, then this will be a rewarding experience.

Special Accommodations: If you have specific physical, psychiatric or learning disabilities that may require accommodations for this course, please contact Northeastern's Disabilities Resource Center (DRC) at (617) 373-2675. The DRC can provide you with information and assistance to help manage any challenges that could affect your performance in the course. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.

If the Disability Resource Center has formally approved you for an academic accommodation in this class, please present the instructor with your "Professor Notification Letter" *during the first week of the semester*, so that we can address your specific needs.

Northeastern University Copyright Statement