The City College of New York | Department of Computer Science

CSC 445: Big Data Management & Analysis
FALL 2021

Lectures: Monday & Wednesday, 3:30pm – 4:45pm
Location: ONLINE (Zoom)

Instructor

Huy T. Vo (hvo@cs.ccny.cuny.edu)
Office Hours: Thursday, noon – 2pm (via Zoom), appointments required.

Course Description

Big data is sometimes defined as data that are too big to fit onto the analyst's computer. With storage and networking getting significant cheaper and faster, big data sets could easily reach the hands of data enthusiasts with just a few mouse clicks. It is crucial for big data to be made available to the non-expert users in such a way that they can process the data without the need of a supercomputing expert. One such approach is to use big data programming frameworks that can deal with big data in as close a paradigm as the way it deals with "small data." Users may expect that if their code works within these frameworks for small data, it will also work for big data. This course aims to provide a broad understanding of big data and current technologies in managing and processing them with a focus on the urban environment. General topics include big data ecosystems, parallel and streaming programming model, MapReduce, Hadoop, Spark, and NoSQL solutions. Hands-on labs and exercises will be offered throughout to bolster the knowledge learned in each module.

Prerequisites

- Basic knowledge of data analysis
- Proficiency in Python programming

Course Objectives

- Understand the big data ecosystem including its data life cycle
- Gain experience in identifying big urban data challenges and develop analytical solutions for them
- Understand the big data programming paradigm: streaming, parallel computing and MapReduce
- Gain knowledge in implementing analytical tools to analyze big data with Apache Spark & Hadoop

Textbook

There is no required textbook, but supplemental and copyrighted materials will be posted on NYU Classes and/or distributed in class.

Recommended/Suggested Readings

- *Hadoop : The Definitive Guide-Storage and Analysis at Internet Scale, 4th Edition* (O'Reilly Media, Incorporated, 2015)
  by T. White
- *PySpark Recipes A Problem-Solution Approach with PySpark2* (SpringerLink, Berkeley CA, 2018)
  by R. K. Mishra

- *Next generation databases : NoSQL, NewSQL, and Big Data* (Springer, Berkeley, CA, 2016) by G. Harrison
- *Data Science and Big Data Analytics* (John Wiley & Sons, Indianapolis IN, 2015) by EMC Education Services
- *reference* *Probabilistic Data Structures and Algorithms for Big Data Applications* (Books on Demand, 2019) by A. Gakhov

Grading

All requirements must be completed by the date specified and handed in at the beginning of class or they will not be counted toward the final grade. No late assignments will be accepted.

- Assignments – 60%
- Exam – 15%
- Final Challenge – 25%

CUNY Blackboard

You must have access to the CUNY Blackboard. All announcements and class-related documents (supplemental and suggested readings, discussion questions, etc.) will be posted there. Class announcements will be distributed via CUNY e-mail. It is important that you actively use your CUNY e-mail account, or have appropriate forwarding set up on your email.

Statement of Academic Integrity

The Department of Computer Science values both open inquiry and academic integrity. Students are expected to follow standards of excellence set forth by the City College. Such standards include respect, honesty, and responsibility. The program does not tolerate violations to academic integrity including:

- Plagiarism
- Cheating on an exam
- Submitting your own work toward requirements in more than one course without prior approval from the instructor
- Collaborating with other students for work expected to be completed individually
- Giving your work to another student to submit as his/her own
- Purchasing or using papers or work online or from a commercial firm and presenting it as your own work

Students are expected to familiarize themselves with the College's policy on academic integrity and the Department of Computer Science policies on plagiarism as they will be expected to adhere to such policies at all times – as a student and alumni of CUNY.

The College's policies concerning plagiarism, in particular, will be strictly followed. Please consult the *Chicago Manual of Style* for guidelines on citations. Do not hesitate to ask if you have any questions regarding writing style, citations, or any academic policies.

Tentative Course Outline (subject to change)

	Topic
Week 1	Introduction
Week 2	Streaming
Week 3	Distributed File System & Parallel Computing
Week 4	MapReduce & Apache Hadoop
Week 5	Apache Spark
Week 6	Hadoop/Spark on NYU-HPC
Week 7	Spatial Data
Week 8	Textual and Social Media Data
Week 9	Network and Graph Data
Week 10	Exam
Week 11	NoSQL and Cloud Computing
Week 12	Final Challenge & Wrap-Up