NYU CS-GY-6513 Big Data Syllabus

Course Description

This course introduces the architectures and technologies at the foundation of the Big Data movement. These technologies facilitate scalable management and processing of vast quantities of data collected through near real-time sensing and bulk data ingest.

Big Data requires the storage, organization, and processing of data at a **scale** and **efficiency** that go well beyond the capabilities of conventional information technologies. The course reviews the state of the art in Big Data analytics frameworks and in addition to covering the specifics of different platforms, models, and languages, students will look at real applications and patterns that perform massive data analysis and how they can be implemented on Big Data platforms.

Topics discussed include: Distributed Data processing platforms: Map reduce/Hadoop, Apache Spark and Dask; Distributed data stores: NoSQL and MongoDB; and large-scale data mining patterns.

The curriculum will primarily consist of technical readings and discussions and will also include programming projects where participants will prototype data-intensive applications using existing Big Data tools and platforms.

This course is a Project Course!

You are expected to do and complete a course project. Details on the project will be discussed during the semester.

Course Pre-requisites

- Graduate or undergraduate courses in the following areas: Operating Systems, Data structures, Programming Languages
- **Programming experience is required** in one of the following programming languages for assignments and final project: Java, Python, Scala.
- Familiarity with databases, Linux scripting, and distributed systems will be useful.

NYU CS-GY-6513 Big Data Syllabus

Course Objectives

1. To learn about basic concepts, technical challenges, and opportunities in big data management and big data analysis technologies.
2. To learn about common algorithmic and statistical techniques used to perform big data analysis.
3. To learn and get hands-on experience in using some data analysis and management tools such as Hadoop MapReduce, and others.
4. To learn about different types of scenarios and applications in big data analysis, including for structured, semi structured, and unstructured data.

We will use a suitable combination of technologies, including virtual machines, containers and cloud-based technologies (VM, NYU HPC Hadoop, AWS, Azure) for completing homework and projects. Students may also opt to create their own cloud-based Hadoop clusters. Since some students may not have experience with cloud technologies, we will cover the practical details of using the VMs and the NYU HPC.

Course Structure

Students are required to attend the ***weekly lectures*** and complete reading and/or programming assignments. Students will demonstrate mastery of course topics by designing, developing, and demonstrating a final analytics project of their choosing. Class time will be set aside for live final project demonstrations.

Optional and recommended texts:

- **Mining of Massive Datasets**. Rajaraman and Ullman, Cambridge University Press, 2011. Available online at http://infolab.stanford.edu/~ullman/mmds/book.pdf
- **Hadoop: The Definitive Guide,** fourth edition, by Tom White
- **Data Mining: Concepts and Techniques**, 3rd edition (2011) by Han, Kamber, and Pei

NYU CS-GY-6513 Big Data

Syllabus

Course requirements

- Weekly lectures and office hours participation
- Homework submitted to NYU website by assigned due date
- Midterm Exam
- Final Project - self-chosen analytics project, with Professor's approval

Participation

Contributing in a significant way to both discussions in class and online during the whole semesteris of great importance to your success in the course. Over the semester, you should make at least

3 substantive, interesting posts to the discussion board either by initiating a new topic or responding to statements made by others. These posts should be directly related to the course topics. Some examples could include a review of a paper, tool, tutorial, or online data science course.

Re-grading Policy

After the grade of an assignment or quiz is available, you have **ONE** week to request a review fora new grade. Once the 1-week re-grading period has passed, the opportunity for a new grade will no longer be available for the assignment or exam. On reviewing your assignment, *you may be assigned a grade that is higher or lower than the one given.*

Policy on Academic Dishonesty

See the NYU policy on academic dishonesty at our school's website:
http://engineering.nyu.edu/academics/code-of-conduct/ academic-dishonesty

As an NYU community member, it is your responsibility to know your rights and responsibilities around academic misconduct. Please click here to review the NYU Tandon Student Code of Conduct. Any questions about how this policy relates to this class, please ask your professor.

A. Introduction: The School of Engineering encourages academic excellence in an environment that promotes honesty, integrity, and fairness, and students at the School of Engineering are expected to exhibit those qualities in their academic work. It is through the process of submitting their own work and receiving honest feedback on that work that students may progress academically. Any act of academic dishonesty is seen as an attack upon the School and will not be tolerated. Furthermore, those who breach the School's rules on academic integrity will be sanctioned under this Policy. Students are responsible for familiarizing themselves with the School's Policy on Academic Misconduct.

B. Definition: Academic dishonesty may include misrepresentation, deception, dishonesty, or any act of falsification committed by a student to influence a grade or other academic evaluation. Academic dishonesty also includes intentionally damaging the academic work of others or assisting other students in acts of dishonesty. Common examples of academically dishonest behavior include, but are not limited to, the following:

1. Cheating: intentionally using or attempting to use unauthorized notes, books, electronic media, or electronic communications in an exam; talking with fellow students or looking at another person's work during an exam; submitting work prepared in advance for an in-class examination; having someone take an exam for you or taking an exam for someone else; violating other rules governing the administration of examinations.

2. Fabrication: including but not limited to, falsifying experimental data and/or citations.

3. Plagiarism: intentionally or knowingly representing the words or ideas of another as one's own in any academic exercise; failure to attribute direct quotations, paraphrases, or borrowed facts or information.

4. Unauthorized collaboration: working together on work meant to be done individually (this includes sharing your work with other students for any purpose).

5. Duplicating work: presenting for grading the same work for more than one project or in more than one class, unless express and prior permission has been received from the course instructor(s) or research adviser involved.

6. Forgery: altering any academic document, including, but not

limited to, academic records, admissions materials, or medical excuses.

NYU School of Engineering Policies and Procedures on Excused Absences

Students who miss class for any reason should complete this form, through the Office of Student Advocacy. Any questions about what qualifies as an excused absence should email advocacy.tandonstudentlife@nyu.edu. Please note: attendance policies vary from class to class, so please check in with your professor if you have any questions.

A. Introduction: An absence can be excused if you have missed no more than 10 days of school. If an illness or special circumstance has caused you to miss more than two weeks of school, please refer to the section labeled Medical Leave of Absence.

B. Students may request special accommodations for an absence to be excused in the following cases:

1. Medical reasons
2. Death in immediate family
3. Personal qualified emergencies (documentation must be provided)
4. Religious Expression or Practice

Deanna Rayment, advocacy.tandonstudentlife@nyu.edu, is the Director of Student Advocacy & Compliance and her office handles excused absences and general connection to resources. The Office of Student Advocacy is located in 5 MTC, LC240C and they can assist you should it become necessary or if you have any questions.

Moses Center Statement of Disability

If you are student with a disability who is requesting accommodations, please contact New York University's Moses Center for Students with Disabilities (CSD) at 212-998-4980 or mosescsd@nyu.edu.

The Moses Center is located at 726 Broadway on the 2nd and 3rd floors. You must be registered with CSD to receive accommodations. Information about the Moses Center can

NYU CS-GY-6513 Big Data Syllabus

be found at:
http://www.nyu.edu/students/communities-and-groups/students-with-disabilities.html

ChatGPT

It is important that the written work required by the course is yours. You should not use ChatGPT or other AI tools for any purpose other than idea generation. Use of these tools is considered academic misconduct.

Grading (tentative)

Class participation, attendance/participation in forum/slack.	5%
Midterm	30%
Homeworks (4 to 6)	30%
Project	35%

Late Policy

A homework assignment submitted the day after the due date starts out with a maximum grade of B.
****** Assignments submitted more than one day late will get a 0 (F) grade. *****

Course Topics – *** subject to change

Session 1	Course Introduction Tools used in the course Introduction to Hadoop
Session 2	Introduction to Distributed and Parallel Computer Systems Distributed File Systems HDSF Intro to Map/Reduce
Session 3	HDFS and Map/Reduce Map/Reduce Architectures Apache Spark
Sessions 4,5	Apache Spark
Session 6	The Data Tier New Alternatives to Traditional Database Systems and Access Methods NoSQL

NYU CS-GY-6513 Big Data

Syllabus

Session 7	*Midterm Exam*
Session 8	Big Data - Next Generation Frameworks Ray/Dask/Prefect
Session 9	BigData Machine Learning
Session 10	Streaming Systems Apache Kafka
Session 11	Distributed Coordination
Session 12	Special Topics in Big Data Programming -Web Search
Session 13	Special Topics in Big Data Programming
Session 14	Project Demonstration Day
Session 15	Project Demonstration Day