CS-GY 9223: Programming for Big Data Analytics
Fall 2018 - Thursdays 6:00-8:30 PM
JABS 474

Professor J. Rodriguez
Computer Science and Engineering
New York University Tandon School of Engineering

To contact professor: jcr365@nyu.edu

Office hours: Thursday 4PM – 5PM
2 MetroTech 10.054W

TAs: To be announced

TA's will each hold a weekly hour for questions, tutoring, etc.
Times/Locations to be arranged by the TAs.

Course Pre-requisites

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

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 and in addition to covering the specifics of different platforms, models, and languages, students will look at real applications that perform massive data analysis and how they can be implemented on Big Data platforms.

Topics discussed include: Map reduce/Hadoop, NoSQL stores, languages such as Pig Latin and JAQL, large-scale data mining and visualization. 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.

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, Pig Latin, 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 weekly lectures and complete weekly reading and programming assignments. Students demonstrate mastery of course topics by designing, developing, and demonstrating an analytics project of their choosing. Class time will be set aside for final project demonstrations during our Big Data Analytics Symposium.

Readings

Optional and recommended texts:

- **Hadoop: The Definitive Guide,** fourth edition, by Tom White
- **Programming Hive**, by Capriolo, Wampler, and Rutherglen
- **Hadoop Operations**, by Eric Sammer
- **Programming Pig**, by Alan Gates
- **HBase: The Definitive Guide**, by Lars George
- **HBase in Action**, by Nick Dimiduk and Amandeep Khurana
- **Mining of Massive Datasets**. Rajaraman and Ullman, Cambridge University Press, 2011. Available online at http://infolab.stanford.edu/~ullman/mmds/book.pdf
- **Data-Intensive Text Processing with MapReduce**. J. Lin and Chris Dyer, Morgan Claypool , 2010. Available online at http://lintool.github.io/MapReduceAlgorithms/

Course requirements

- Weekly lecture attendance
- Homework submitted to NYU Classes by assigned due date
- Quizzes (Based on several reading assignments handed out during class)

- Midterm Exam
- Final Project - self-chosen analytics project, with Professor's approval

Policy on Academic Dishonesty

Please see the NYU policy on academic dishonesty at our school's website:

http://engineering.nyu.edu/academics/code-of-conduct/ academic-dishonesty

Each student is expected to complete homework independently unless otherwise noted.

- Homework and project submissions are required to be your **original** work, or in the case of the project, your team's **original** work.
- Proper references and citations are required if you use, or leverage, statements or ideas that were developed by someone other than you. **You are required to credit the author.**
- Failure to credit the author will result in a grade of **F**.
- Anyone observed to be collaborating during an exam or quiz will receive a grade of **F.**
- Anyone whose exam or quiz solution(s) is/are indicative of collaboration, will receive a grade of **F for the course**.
- Cheating and/or failure to abide by the school's code of conduct will fail the course.

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 be found at:

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

Late Policy

A homework assignment submitted the day after the due date starts out with a B.

Grading (tentative)

Readings, quizzes, lab assignments, class participation, attendance	10%
Midterm	25%
Homeworks (4 to 6)	30%
Project	35%

Course Topics – *** May be subject to change

Class 1	Course Introduction Introduction to Hadoop Tools used in the course
Class 2	Introduction to Distributed and Parallel Computer Systems Distributed File Systems
	HDFS
	Introduction to MapReduce
Class 3	HDFS and MapReduce MapReduce Architectures – MR1 and YARN/MR2
Class 4	Introduction to Pig, Analytics Examples Programming in Pig Apache Spark
Class 5	Apache Spark Project Tee-up
Class 6	New Alternatives to Traditional Database Systems and Access Methods NoSQL Introduction to Flume
Class 7	*Midterm Exam*
Class 8	Programming with Hive Spark-SQL
Class 9	Streaming Systems Spark-Streaming Project Breakout
Class 10	Autonomic Systems Internet of Things BigData Machine Learning
Class 11	Distributed Coordination Zookeper
Class 12	Special Topics in Big Data Programming
Class 13	Coordination Sqoop OOzie
Class 14	Big Data Analytics Symposium - Project Demonstration Day

Class 15	Big Data Analytics Symposium - Project Demonstration Day