CUSP-GX-5008.001: Big Data Management & Analysis

SPRING 2016

Lectures: Thursday, 9:30am – 12:20pm
Location: 2 MTC, Room 820

Instructor:
Huy T. Vo, NYU-CUSP & CUNY-CCNY

Teaching Assistant:
Himanshu Kumawat
Kunal Barde

Office Hours
Office hours: Thursdays, 2pm – 4pm
Office location: NYU CUSP, 1 Metrotech Center, 19th Floor , 1940

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. These enthusiasts could be policy makers, government employees or managers, who would like to draw insights and (business) value from big data. Thus, 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 build big data programming frameworks that can deal with big data in as close a paradigm as the way it deals with "small data." Also such a framework should be as simple as possible, even if not as efficient as custom-designed parallel solutions. Users should expect that if their code works within these frameworks for small data, it will also work for big data.

The 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, Pig, and NoSQL solutions. Hands-on labs and exercises will be offered throughout to bolster the knowledge learned in each module.

Prerequisites
Graduate standing in CUSP. Non-CUSP students must request permission from the CUSP Program Director.

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

Required Text

None, but supplemental and copyrighted materials may be posted on NYU Classes or distributed in class.

Recommended/Suggested Readings

- *Data Science and Big Data Analytics* (John Wiley & Sons, Indianapolis IN, 2015)
  by EMC Education Services
- *Hadoop: The Definitive Guide* (O'Reilly, Sebastopol CA, 2015)
  by T. White
- *Learning Spark: Lightning-Fast Big Data Analysis* (O'Reilly, Sebastopol CA, 2015)
  by H. Karau, A. Konwinski, P. Wendell, and M. Zaharia
- *Advanced Analytics with Sparks: Patterns for Learning from Data at Scale* (O'Reilly, Sebastopol CA, 2015)
  by S. Ryza, U. Laserson, S. Owen, and J. Wills

Course Requirements

Each class session is divided into a 60-70 minute lecture and a hands-on lab for the rest of the time. Please bring your laptops to all class lectures. Class participation are recorded through lab submission.

This course will use Python as the main programming language; however, other languages may also be accepted where applicable, e.g. using Java for Hadoop. Please make sure to check with the instructor **before** planning to submit your homework with a non-Python language.

All assignments should be submitted via NYU Classes (unless otherwise noted). Please refrain from posting your work (assignments and projects) onto public spaces such as github. If you must do so, please only do it after the deadline or with access control.

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 – 45%
- Project Proposal – 15%
- Final Project – 30%
- Class participation and attendance – 10%

NYU Classes

You must have access to the NYU Classes site (http://classes.nyu.edu/). All announcements and class-related documents (supplemental and suggested readings, discussion questions, etc.) will be posted there.

Some class announcements will be distributed via NYU e-mail. Thus, it is important that you actively use your NYU e-mail account, or have appropriate forwarding set up on NYU Home (https://home.nyu.edu/).

Statement of Academic Integrity

NYU CUSP values both open inquiry and academic integrity. Students graduate programs are expected to follow standards of excellence set forth by New York University. 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 University's policy on academic integrity and CUSP's policies on plagiarism as they will be expected to adhere to such policies at all times – as a student and an alumni of New York University.

The University'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.

Course Outline (subject to change)

Week/Date		Topic	Assignment	
			Out	Due
W. 1	1/28	Introduction to Big Data		
W. 2	2/04	Big data life cycle and data warehouse architecture	HW 1	
W. 3	2/11	Dealing with data volume: streaming computation	HW 2	HW 1
W. 4	2/18	High order functions and parallel programming paradigm	HW 3	HW 2
W. 5	2/25	MapReduce	HW 4	HW 3
W. 6	3/03	Apache Hadoop – **time to form teams for final projects**	HW 5	HW 4
W. 7	3/10	Apache Hadoop Query: Pig, Hive & HBase	HW 6	HW 5
W. 8	3/17	*NO CLASS: Spring Recess*		
W. 9	3/24	**PRESENTATION: Project Proposal**		Project Proposal
W. 10	3/31	Apache Spark	HW 7	HW 6
W. 11	4/07	Apache Spark	HW 8	HW 7
W. 12	4/14	Dealing with big spatial data OR Machine Learning on Spark	HW 9	HW 8
W. 13	4/21	Dealing with data variety: NoSQL	HW 10	HW 9
W. 14	4/28	Big data ecosystem		HW 10
W. 15	5/05	**PRESENTATION: Final Project**		Project Report