# CS 444: Big Data Systems

## General Information

| | | | |
|---|---|---|---|
| Instructor: | Chase Wu | Department office: | GITC 4100 |
| Office/Lab: | GITC 4418 | Department phone: | 973-596-3366 |
| E-mail: | chase.wu@njit.edu | | |
| Phone: | 973-642-4579 | | |

## Course Description

This course provides a broad coverage of topics on big data generation, transfer, storage, management, computing, and analytics with focus on state-of-the-art technologies and tools used in big data systems such as Hadoop. Real-life big-data applications and workflows in various domains are introduced as use cases to illustrate the development and execution of emerging big data-oriented solutions using HDFS, HBase, MapReduce/Spark, etc. deployed in cloud-based cluster environments.

## Required Background

Programming Skills

- Java, Python, or C/C++ in Linux

Prerequisite Courses

- CS 288 Intensive Programming in Linux AND CS 301 Introduction to Data Science
- Or permission of instructor

## Textbook (not required)

- Big Data Technologies for Business. By Arben Asllani, Prospect Press, 2020.
- Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph. By David Loshin, Elsevier, August 23, 2013.

## Resources

Additional reading materials including reference books and online resources will be assigned for some advanced topics as the course proceeds.

## Evaluation

Grading components:

| | |
|---|---|
| Attendance | 10% |
| Homework | 10% |
| Project | 20% |
| Midterm | 30% |
| Final | 30% |

Grading scale*:

| Grade | Score |
|---|---|
| A | 90 — 100 |
| B, B+ | 80 – 84, 85 – 89 |
| C, C+ | 70 – 74, 75 – 79 |
| F | Below 70 |

*Final grades will not be curved unless necessary.

## Late Policy

Students are expected to complete work on schedule. Late work is not accepted unless prior arrangements are made with the instructor.

## Academic Integrity and Student Conduct:

*“Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at:* http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf.

*Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office.* ***Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university.*** *If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at* dos@njit.edu*”*

## Course Syllabus

| Week | Topic |
|---|---|
| 1 | • Introduction |
| 2 | • In-class Presentation on 4 V’s of Big Data Applications |
| 3 | • Trends of Computing for Big Data<br>○ High-performance Computing (Supercomputers and Clusters)<br>○ Grid Computing<br>○ Continuum Computing: from Edge to Cloud<br>○ Mobile Computing |
| 4, 5 | • Big Data Overview<br>○ Drivers of Big Data<br>○ Big Data Attributes and Data Structures<br>○ Big Data Ecosystem<br>○ Big Data Use Cases |
| 6, 7, 8 | • Big Data Tools, Techniques, and Systems<br>○ HDFS, HBase, and NoSQL (Document Store, Graph DB, etc.)<br>○ MapReduce, Spark, Oozie, Tez, Hive, Pig, etc.<br>○ Hadoop 1 and Hadoop 2 (YARN) |
| 9 | • Review and Midterm Exam |
| 10, 11, 12 | • Analytical Theories and Methods for Big Data<br>○ Hadoop/Mahout<br>○ Machine Learning<br>▪ Recommendation<br>▪ Clustering<br>▪ Classification<br>▪ Regression |
| 13, 14 | • Advanced Topics<br>○ Big Data Volume and Information Visualization<br>○ High-performance Networking for Big Data Movement<br>○ Big Data Scientific Workflow Management and Optimization |
| 15 | • Review |