Fall 2019

CS 644-101: Introduction to Big Data

Chase Wu

CS 644: Introduction to Big Data

General Information

Instructor:	Chase Wu	Department office:	GITC 4100
Office/Lab:	GITC 4107	Department phone:	973-596-3366
E-mail:	chase.wu@njit.edu		
Phone:	973-642-4579		

Course Description

This course provides an in-depth coverage of various topics in big data from data generation, storage, management, transfer, to analytics, with focus on the state-of-the-art technologies, tools, architectures, and systems that constitute big-data computing solutions in high-performance networks. Real-life big-data applications and workflows in various domains (particularly in the sciences) are introduced as use cases to illustrate the development, deployment, and execution of a wide spectrum of emerging big-data solutions.

Required Background

Programming Skills

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

Prerequisite Courses

- CS 610: Data Structures and Algorithms
- Or permission of instructor

Textbook

- 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 ○ High-performance Computing (Supercomputers and Clusters) ○ Grid Computing ○ Cloud Computing ○ Mobile Computing
4, 5	• Big Data Overview ○ Drivers of Big Data ○ Big Data Attributes ○ Data Structures ○ Big Data Ecosystem ○ Examples of Data Analytics
6, 7	• Big Data Tools, Techniques, and Systems ○ Exascale Computing ○ HDFS, HBase, and NoSQL (Document Store, Graph DB, etc.) ○ MapReduce, Spark, Oozie, Tez, Hive, Pig, etc. ○ Hadoop 1 and Hadoop 2 (YARN)
8, 9	• Review and Midterm Exam
10, 11	• Advanced Analytical Theory and Methods ○ Hadoop/Mahout ○ Recommendation ○ Clustering ○ Classification ○ Regression
12, 13, 14	• Advanced Topics ○ Big Data Volume and Information Visualization ○ High-performance Networking for Big Data Movement ○ Big Data Scientific Workflow Management and Optimization
15	• Review