CS 444: Big Data Systems

General Information

Instructor: Jaini Bhavsar
Department office: GITC 2119
Department Phone: 973 642-7030
Email: jkb37@njit.edu
Class Timings: Thursday, 8:30 AM to 11:20 AM, FMH 321
Office Hours: Thursday, 12:00 PM to 1:00 PM, GITC 2119 or by appointment

Course Description

This course provides a broad coverage of topics on big data generation, transfer, storage, management, computing, and analytics with a 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 with the permission of the 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:

Discussion	10%
Homework	20%
Midterm	25%
Project	20%
Final	25%

Grading scale*:

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

*Final grades will not be curved unless necessary.

Online Discussions

You are expected to actively engage in weekly discussion forums on Canvas. These forums are designed to foster a collaborative learning environment where all students contribute to discussions. Participation will not only help deepen your understanding of the course material but also enhance your overall success in the class. Your contributions should be thoughtful, demonstrate critical thinking, and reflect on the content covered in lectures and readings.

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: NJIT Academic Integrity Code.

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".

Policy on AI Usage

Student use of artificial intelligence (AI) is permitted in this course for certain assignments and activities. It is not permitted to be used in the assignment of writing a survey paper. Additionally, if and when students use AI in this course, the AI must be cited as is shown within the NJIT Library AI citation page for AI. If you have any questions or concerns about AI technology use in this class, please reach out to your instructor prior to submitting any assignments.

Office of Accessibility

The Office of Accessibility Resources and Services works in partnership with administrators, faculty and staff to provide reasonable accommodations and support services for students with disabilities who have provided our office with medical documentation to receive services.

Course Syllabus

Topics

• Introduction
• Trends of Computing for Big Data
  ○ High-performance Computing (Supercomputers and Clusters)
  ○ Grid Computing
  ○ Continuum Computing: from Edge to Cloud
  ○ Mobile Computing

• Big Data Overview   ○ Drivers of Big Data

- Big Data Attributes and Data Structures
- Big Data Ecosystem
- Big Data Use Cases

- Big Data Tools, Techniques, and Systems
  - HDFS, HBase, and NoSQL (Document Store, Graph DB, etc.)
  - MapReduce, Spark, Oozie, Tez, Hive, Pig, etc.
  - Hadoop 1 and Hadoop 2 (YARN)

- Analytical Theories and Methods for Big Data
  - Hadoop/Mahout
  - Machine Learning with Big Data
    - Recommendation
    - Clustering
    - Classification
    - Regression
- Advanced Topics
  - Big Data Volume and Information Visualization
  - Big Data Movement in High-performance Networks
  - Big Data Scientific Workflow Management and Optimization