Northeastern
College of Engineering

DAMG 7245: Big-Data Systems and Intelligence Analytics
FALL 2024

Course Information
Course Title: Big Data Systems and Intelligence Analytics
Course Number: DAMG 7245
Term and Year: Fall 2024
Credit Hour: 4
CRN: 22907 (Seattle, WA)
Course Format: Traditional

Instructor Information
Full Name: Salman Yousaf
Email Address: salmanyz6@gmail.com
Office Hours: Thursday, 4-5pm

Instructor Biography
Salman currently works at Google on cutting-edge data processing platforms, and has expertise in developing and optimizing high-throughput, low-latency, and scalable pipeline infrastructure that powers modern A.I and analytics. Previously, he was involved with Argonne National Laboratory where he built and deployed Machine Learning algorithms. As an instructor, Salman is passionate about empowering students to build real-world applications, leveraging his knowledge to demystify the process of developing end-to-end data pipelines and implementing intelligent algorithms. Salman holds a Master's in Applied Data Science from the University of Chicago, and a Master's in Computer Science from the University of Illinois at Urbana-Champaign.

Teaching Assistant Information
Full Name: TBD
Email Address: TBD
Office Hours: TBD

Course Prerequisites
DAMG 6105 with a minimum grade of B or INFO 6105 with a minimum grade of B.

Course Description
Offers students an opportunity to learn a hands-on approach to understanding how large-scale data sets are processed and how data science algorithms are adopted in the industry through case studies and labs. This project-based course focuses on enabling students with tools and frameworks primarily to build end-to-end applications. The course is divided into three parts: building the data pipeline for

data science, implementing data science algorithms, and scaling and deploying data science algorithms.

Course Learning Outcomes

- **Understand the fundamentals of Big Data:** Define big data, its characteristics, and its importance in today's data-driven world.
- **Grasp the concepts and applications of data pipelines:** Explain the role of data pipelines in data processing, analysis, and decision-making.
- **Work with various big data technologies:** Demonstrate proficiency in using tools like Hadoop, Apache Beam, and Spark for processing and analyzing large datasets.
- **Master data cleaning, preparation, and transformation:** Apply techniques to clean, prepare, and transform data for analysis.
- **Implement data pipelines using industry-standard frameworks:** Design and build data pipelines using Apache Beam and other relevant frameworks.
- **Apply machine learning techniques to big data:** Integrate machine learning models into data pipelines for predictive analytics and decision support.
- **Build and deploy real-time data processing pipelines:** Develop pipelines using streaming platforms like Apache Kafka or Flink to process data in real time.
- **Design and implement scalable and efficient data pipelines:** Optimize pipelines for performance, scalability, and cost-effectiveness.
- **Understand data pipeline infrastructure considerations:** Dimensions of scalability and performance, latency, reliability and availability, cost and efficiency, complexity, and security
- **Ensure data quality and governance:** Implement data governance practices to maintain data quality and integrity.
- **Understand the ethical implications of big data:** Be aware of privacy concerns, bias, and fairness in big data applications.
- **Apply data anonymization techniques:** Protect sensitive data using anonymization methods.
- **Stay updated on emerging trends in big data:** Be knowledgeable about the latest developments in big data technologies and applications.
- **Collaborate effectively in a team environment:** Work effectively with others to design, implement, and maintain data pipelines.
- **Communicate technical concepts effectively:** Clearly explain complex technical concepts to both technical and non-technical audiences.
- **Demonstrate critical thinking and problem-solving skills:** Apply analytical and problem-solving skills to address challenges in big data projects and engineering problems.

Required Tools and Course Textbooks.

Textbooks: Materials and tutorials will be provided to students.

The following books are optional, but recommended for theoretical concepts:

- Mining of Massive Datasets, (http://www.mmds.org/)
- Data Pipelines Pocket Reference: Moving and Processing Data for Analytics, James Densmore
- Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, Martin Kleppmann

Tools: The following access and tools will be required to perform your assignments:

- Access to a coding environment / IDE with Python
- Access to AWS, Azure, or GCP. Guidelines on signing up for free student accounts will be provided.

Course Schedule/Topics Covered.

Week	Date	In Class Topic	Assignment Due
1	**09/09**	- Course Overview - What is Big Data? - An overview of Data Pipelines, and how they are crucial for intelligent systems - Foundational concepts and terminology - Case Study: Introducing the Need for Big Data	● Class Participation Exercise
2	**09/16**	- Types, sources and formats of Big Data - MapReduce fundamentals - processing large datasets with a parallel, distributed algorithm - Cluster computing for Big Data - Introducing Hadoop - Concepts of Data Ingestion and Data Stores	● Assignment 1 released ● Class Participation Exercise
3	**09/23**	- Introduction to Apache Beam, Spark - Loading datasets - Hands-on examples - Case Study + Technology Landscape - Data cleaning and preparation - Data transformations and feature engineering: concepts and methodologies - Case Study + Technology Landscape	● Quiz 1 ● Class Participation Exercise
4	**09/30**	- Introduction to Data Warehouses and Data Lakes: Differences, Use Cases, and Architectures - ETL Processes (Extract, Transform, Load): Key components and their role in data integration for Data Warehouses and Data Lakes - Big Data Analytics Frameworks: Introduction to Apache Hive, Presto, and HBase for querying large datasets within Data Lakes and Data Warehouses - Case Study: Integrating Data Warehouses, Data Lakes, and ETL in a Data Pipeline - Engineering a data pipeline using Apache Beam	● Assignment 1 due ● Assignment 2 released ● Class Participation Exercise

5	**10/7**	- Introduction to Machine Learning (ML) pipelines - ML pipeline components: data ingestion, feature engineering, model training, evaluation, and deployment - Hands-on exercise: Building a simple ML pipeline using a popular ML framework	● Quiz 2 ● Class Participation Exercise
6	**10/14**	- Building ML pipelines (continued) - Hyperparameter tuning and optimization - Case study: Building a predictive analytics pipeline - Discussion on challenges and best practices in ML pipeline development	● Final Project Check-In ● Assignment 2 due ● Assignment 3 released ● Class Participation Exercise
7	**10/21**	- Case Study with Hands-On Exercises: Building a Data Pipeline for a Recommender System	● Quiz 3 ● Class Participation Exercise
8	**10/28**	- Real time intelligent systems: concepts and foundations - Streaming data pipelines - Understand the concepts and challenges of real-time data processing and analysis. - Explore popular data streaming platforms like Apache Kafka and Flink. - Build a simple real-time data processing pipeline to analyze streaming data.	● Assignment 3 due ● Assignment 4 released ● Class Participation Exercise
9	**11/4**	- Advanced Data Streaming Concepts and Case Studies - Explore state management, windowing, machine learning, and complex event processing in streaming pipelines.	● Quiz 4 ● Class Participation Exercise

10	**11/11**	- Data processing pipeline infrastructure
		  - Scalability and Performance; Reducing Latency
		  - Reliability and Availability
		  - Cost and Efficiency
		  - Complexity and Maintainability
		  - Security and Privacy
		- Evaluate the performance and security of data pipelines.
		- Discuss strategies for improving the scalability and reliability of big data systems.
		- Analyze the cost-benefit trade-offs of different data pipeline architectures.
			● Assignment 4 due
			● Assignment 5 released
			● Class Participation Exercise

11	**11/18**	- Productionizing, Monitoring and	● Quiz 5

		Maintaining Data pipelines - Pipeline Deployment Strategies and Considerations - Infrastructure Considerations: Cloud vs. on-premises deployment, resource allocation, scalability, availability - Monitoring and Alerting: Key metrics to track: data ingestion rates, processing time, error rates, resource utilization. - Alerting mechanisms - Visualization tools: Use dashboards to monitor pipeline performance and identify trends	● Class Participation Exercise
12	**11/25**	- Testing and Validation: Unit and integration testing for data pipelines - MLOps: Introduction to MLOps for productionizing machine learning models within data pipelines; automating deployment, monitoring, and management of models in production environments - Version Control and CI/CD: Implement continuous integration and continuous delivery (CI/CD) pipelines for automated testing and deployment of data pipelines and machine learning models - Infrastructure Automation: Automating infrastructure deployment, scaling, and management with tools like Docker and Kubernetes	● Assignment 5 due. ● Assignment 6 released. ● Class Participation Exercise
13	**12/2**	- Data Governance and Quality - Understand the importance of data governance - Developing a data governance plan - quality checks and validation procedures - Data Anonymisation and PII (Personally Identifiable Information) - Ethical Considerations in Big Data - Synthetic data generation, masking - Case Study: Modern Data Laws and Acts	● Quiz 6. ● Class Participation Exercise
14	**12/9**	- Frontier topics: Future of Big Data - Case Studies on Advanced Topics	● Assignment 6 due. ● Class Participation Exercise
15	**12/16**	- Final Projects - Demo Day	● Final Project Due

Assignment Grading

- Assignments: 35%

- Quizzes: 20%
- Class Participation: 10%
- Final Project: 35%

Note: The lowest scoring assignment and quiz may be dropped and not counted towards the final grade calculation.

Grading Scale

	87-89.9% B+	77-79.9% C+	
95-100% A	84-86.9% B	74-76.9% C	
90-94.9% A-	80-83.9%B-	70-73.9% C-	69.9% or below F

Attendance/Late Work Policy

Attendance Policy

Students registered in MGEN courses (INFO, CSYE, and DAMG) are allowed **a maximum of 2 absences per course, with 3 or more absences resulting in an automatic 'F' for that course.** Students are expected to inform their instructors of any absences in advance of the class; if a student is sick long-term or experiences a medical issue that prevents class attendance, it is strongly encouraged that they speak with their Academic Advisor (coe-mgen-gradadvising@northeastern.edu) to learn more about the Medical Leave of Absence. Should a student anticipate being unable to attend 3 or more classes, they should discuss their situation with their Academic Advisor to explore other types of leave in accordance with the University's academic and global entry expectations. International students should review the Office of Global Services webpage to understand their visa compliance requirements.

Teaching Assistants (TAs) or Instructional Assistants (IAs) will be present at each class to collect student attendance.

Late Work Policy

Students must submit assignments by the deadline in the time zone noted in the syllabus. Students must communicate with the faculty prior to the deadline if they anticipate work will be submitted late. Work submitted late without prior communication with faculty will not be graded.

End-of-Course Evaluation Surveys

Your feedback regarding your educational experience in this class is particularly important to the College of Engineering. Your comments will make a difference in the future planning and presentation of our curriculum.

At the end of this course, please take the time to complete the evaluation survey at **https://neu.evaluationkit.com**. Your survey responses are **completely anonymous and confidential**. For courses 6 weeks in length or shorter, surveys will be open one week prior to the end of the courses; for courses greater than 6 weeks in length, surveys will be open for two weeks. An

email will be sent to your Northeastern University Mail account notifying you when surveys are available.

Academic Integrity
A commitment to the principles of academic integrity is essential to the mission of Northeastern University. The promotion of independent and original scholarship ensures that students derive the most from their educational experience and their pursuit of knowledge. Academic dishonesty violates the most fundamental values of an intellectual community and undermines the achievements of the entire University.

As members of the academic community, students must become familiar with their rights and responsibilities. In each course, they are responsible for knowing the requirements and restrictions regarding research and writing, examinations of whatever kind, collaborative work, the use of study aids, the appropriateness of assistance, and other issues. Students are responsible for learning the conventions of documentation and acknowledgment of sources in their fields. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards, as set forth either explicitly or implicitly in this Code or by the direction of instructors.

Go to http://www.northeastern.edu/osccr/academic-integrity-policy/ to access the full academic integrity policy.

MGEN Student Feedback
Students who would like to provide the MGEN unit with anonymous feedback on this particular course, Teaching Assistants, Instructional Assistants, professors, or to provide general feedback regarding their program, may do so using this survey: https://neu.co1.qualtrics.com/jfe/form/SV_cTIAbH7ZRaaw0Ki

University Health and Counseling Services
As a student enrolled in this course, you are fully responsible for assignments, work, and course materials as outlined in this syllabus and in the classroom. Over the course of the semester if you experience any health issues, please contact UHCS.

For more information, visit https://www.northeastern.edu/uhcs.

Student Accommodations
Northeastern University and the Disability Resource Center (DRC) are committed to providing disability services that enable students who qualify under Section 504 of the Rehabilitation Act and the Americans with Disabilities Act Amendments Act (ADAAA) to participate fully in the activities of the university. To receive accommodations through the DRC, students must provide appropriate documentation that demonstrates a current substantially limiting disability.

For more information, visit https://drc.sites.northeastern.edu.

Library Services
The Northeastern University Library is at the hub of campus intellectual life. Resources include over 900,000 print volumes, 206,500 e-books, and 70,225 electronic journals.

For more information and for education specific resources, visit https://library.northeastern.edu
Network Campus Library Services: Northeastern University Library Global Campus Portals

24/7 Canvas Technical Help
For immediate technical support for Canvas, call 617-373-4357 or email help@northeastern.edu

Canvas Student Resources: https://canvas.northeastern.edu/student-resources/

For assistance with my Northeastern e-mail, and basic technical support:
Visit ITS at https://its.northeastern.edu
Email: help@northeastern.edu
ITS Customer Service Desk: 617-373-4357

Diversity and Inclusion
Northeastern University is committed to equal opportunity, affirmative action, diversity, and social justice while building a climate of inclusion on and beyond campus. In the classroom, members of the University community work to cultivate an inclusive environment that denounces discrimination through innovation, collaboration, and an awareness of global perspectives on social justice.

Please visit http://www.northeastern.edu/oidi/ for complete information on Diversity and Inclusion

Title IX
Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty, and staff.

In case of an emergency, please call 911.

*Please visit **https://www.northeastern.edu/ouec** for a complete list of reporting options and resources both on- and off-campus.*