Northeastern University
College of Engineering

DAMG 7245 Big-Data Systems and Intelligence Analytics

Course Information

Course Title: Big-Data Systems and Intelligence Analytics
Course Number: DAMG 7245
Term and Year: Spring 2022
Credit Hour: 4
Course Format: On-Ground

Instructor Information

Full Name: Srikanth Krishnamurthy
Email Address: s.krishnamurthy@northeastern.edu

Course Prerequisites

Databases, Python

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

Standard Learning Outcomes

Learning outcomes common to all College of Engineering Graduate programs:

1. *An ability to identify, formulate, and solve complex engineering problems.*
2. *An ability to explain and apply engineering design principles, as appropriate to the program's educational objectives.*
3. *An ability to produce solutions that meet specified end-user needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors.*

The Information Systems Program accepts students of different engineering backgrounds with minimum programming skills and produces first class Information Systems engineers that operate at the intersection of real-world complexity, software development, and IT management. Graduating students will be able to construct end-to-end advanced software applications that meet business needs.

Specific Learning Outcomes for the Information Systems program:

1. *Create a strong technical foundation through diverse, high-level courses*
2. *Built crucial interpersonal skills needed to succeed in any industry*
3. *Foster a deep level of applied learning through project based case studies*

Course Outcomes and Assesment Standards

Student learning outcomes are statements indicating the measurable outcomes of the course from the learner's perspective. They describe the intended purpose of learning: the end results of the learning experience at the course level which should be aligned with the program level outcomes recorded in the College AQA process. These statements answer the question "What should the students be able to do by the end of the course?" E.g., Based on satisfactory completion of this course, a student should be able to …

Technical/Course Materials Requirements:

- Materials and tutorials shall be provided to registered students
- You must be comfortable with Python as assignments expect you to work with Python.
- Optional reading
    - Data Pipelines Pocket Reference: Moving and Processing Data for Analytics, James Denmore
    - Mining of Massive Datasets: http://www.mmds.org/
    - Data Science in Production: Building Scalable Model Pipelines with Python –Ben Weber
    - Design of Web APIs https://www.manning.com/books/the-design-of-web-apis

About This Course

The field of Data engineering and data science is rapidly evolving. In the last decade, many industries have explored data science and machine learning for their use-cases. Despite interest, less than 20% of the big data and data science projects make it into production. While there are various reasons for this, many attribute to the non-availability of consistent data sets and robust, reproducible and performant engineering required to deploy and scale applications as primary reasons why machine learning hasn't seen increased adoption. In fact more than 80% of data science problems is spent on data engineering and pipeline activities when taking data-driven products to production. Recognizing this, as companies build more and more data-driven applications, significant emphasis is being placed on building pipelines for enabling data processing from ingestion to production.

This is a hand-on project-based course. We will work with real datasets. You will have to get your hands dirty. As you start looking at data sets, you will realize a couple of things.

- One, data is dirty and requires cleaning. We will look into various cleansing techniques and methods used in the industry.
- Second, information is hidden. We will look into ways in which you can explore data and try methods to extract information from data sets.
- Third, there is no correct answer.

In this course, we will focus on the key data engineering tools and techniques that are used in the industry to build, scale and serve robust data & big data applications. We will focus on components required to build reproducible and robust data engineering pipelines.

Student Learning/Course Outcomes (SLOs)

Upon completion of this course, a students should be able to:

- Understand how to design data-driven applications and pipelines.
- Design production-grade pipelines involving large-scale datasets
- Design and build REST APIs to facilititate machine-learning-as-a-service
- Understand considerations to build scalable applications including interpretability, auditability and reproducibility of Machine Learning and Data-driven applications

Attendance Policy

You are expected to attend all lectures and participate in class. If you plan to miss a class for a genuine reason, you must email the instructor of your absence. If you miss a class, to get class participation credit, you must submit a 2- page report on the class missed. If you don't submit the report, and you are absent more than 2 classes without reasonable excuse, you will automatically lose 5% of the class participation credit.

Late Work Policy

Students must submit assignments by the deadline in the time zone noted in Canvas 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 be subject to a late penalty of 10% per day.

Software:

We will use Python for exercises and to illustrate concepts, exercises and final projects. We will heavily use cloud services including GCP and AWS

Lectures:

Lectures would include discussion and illustration of methodologies. I will be posting required and optional reading. We will also have guest lectures and in-class exercises.

Grading:

5 Case study assignments : 50%

Final Project: 25%

Class Presentation topics: 15%;

Class participation & In-class quizzes/exercises: 10%

Case Studies:

You will work on five case studies to demonstrate your understanding of the topics covered in class. This will cover all aspects of a Data-science pipeline.

Final Project:

You will have the opportunity to choose the topic and work on an extended project. Additional information will be provided as the class progresses. Each Team is expected to build a fully functional Data-as-a-service/Model-as-a-service as a part of their final project

Class Presentation:

You will be presenting one topic and an associated example in class. You will have 20 minutes to present. A list of topics will be posted on Canvas

Class Participation:

You are expected to attendall lectures and participate in class. If you plan to miss a class for a genuine reason, you must email the instructor of your absence. If you miss a class, to get class participation credit, you must submit a 2- page report on the class missed. If you don't submit the report, and you are absent more than 2 classes without reasonable excuse, you will automatically lose 5% of the class participation credit.

Office hours:

Office hours: Friday 5.00-6.00pm over Zoom. Schedule via www.calendly.com & by appointment
TA will hold office hours weekly. Details announced on Canvas

Course Schedule

Week	Topic	Note
1	Introduction to Big data & pipelines	
2	Data driven applications and pipelines	Case study 1
3	Data pipelines: Pre-processing, feature engineering, feature selection, Missing value analysis, outlier detection and anomaly detection	

4	Working with Snowflake, Amazon S3, Salesforce	Case study 1 due Case study 2
5	Design of REST APIs for Data-as-a-service	
6	Server less functions, Lambda functions, Step functions	Case study 2 due Case study 3
7	Airflow, Dask, Orchestration	
8	Machine learning as a service for - Forecasting - Synthetic data generation - Synthetic data generation, masking, anonymization, deanonymization	Case study 3 due Case study 4
9	Pipeline versioning, maintenance	Case study 4 due Case study 5
10	Tuning parameters, performance and designing robust data pipelines	
11	Deploying Data and intelligent pipelines	Case study 5 due
12	Frontier topics	Final Project Proposal Due
13	Productionizing Data pipelines & monitoring	
14	Final Project presentations	

- Note: No class on Nov 26th

End-of-Course Evaluation Surveys

Your feedback regarding your educational experience in this class is very important to the College of Professional Studies. 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 HuskyMail 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.

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 http://www.northeastern.edu/drc/getting-started-with-the-drc/.

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 http://subjectguides.lib.neu.edu/edresearch.

24/7 Blackboard Technical Help

For immediate technical support for Blackboard, call 617-373-4357 or emailhelp@northeastern.edu

Within Blackboard, open a support case via the red support button on the right side of the screen, click Create Case

myNortheastern, e-mail, and basic technical support
Visit the Information Technology Services (ITS) Support Portal
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, member 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 www.northeastern.edu/titleix for a complete list of reporting options and resources both on- and off-campus.