DAMG 7374-03 DATA ENGINEERING: Impact of Generative AI with LLM's

Course Information

Course Title: Data Engineering: Impact of Generative AI with LLM's
Course Number: 7374-03
Term and Year: Spring 2024
Credit Hour: 4
CRN:
Course Format: Onsite with Virtual components

Instructor Information

Full Name: Kishore Aradhya
Email Address: k.aradhya@northeastern.edu

Instructor Biography

Kishore Aradhya is a senior technology leader with over twenty years of experience developing and scaling technology organizations across diverse sectors, from startups to Fortune 500 companies. His career spans multiple organizations such as Stanley Black & Decker, Bose, Adobe, edX/MIT, Staples, Monster Search, and Fidelity Investments.

His expertise is in creating production-ready Enterprise Data Platforms for Analytics, Machine Learning, and BI applications, covering the entire data lifecycle. This includes the design, development, and architecture of highly scalable, high-volume enterprise SaaS cloud services, encompassing Customer and Analytics Data Platforms, Data Engineering, Search, and e-commerce solutions. At Stanley Black & Decker, Kishore led the AI and Data Platform Engineering teams, bringing GenAI and LLM technologies to the forefront by working with executive leadership. He directed Bose's global data engineering team, laying the foundation for a Data Center of Excellence (CoE). In addition, he led a nascent NLP and Computer Vision-driven document extraction research product initiative at Adobe, delivering AI-driven PDF Document product features.

Kishore holds an MBA, an MS (Computer Science), and a BS (Electrical Engineering), complemented by various executive and academic certifications. He is also a CDO Magazine Editorial Board Member, an Industry Advisory Board Member, a Product Advisory Council Member at various organizations, and an invited speaker at various data conferences and panels.

Teaching Assistant Information

Full Name:
Email Address:
Office Hours:

Course Prerequisites

- Essential Python, Streamlit, and SQL programming skills.
- Essential understanding of data concepts like Data Quality, Data Transformation, and Data Integration.
- Basic understanding of AI concepts - Machine Learning, Deep Learning & NLP

Course Description

This seminar-style course explores the evolving relationship between Data and Artificial Intelligence (AI), explicitly emphasizing Generative AI with Large Language Models(LLM) in the evolving data platform architecture landscape. You will hear from data leaders and working practitioners in the industry share their insights, experiences, and challenges as they navigate this highly dynamic world. Students will engage in dynamic discussions and critical analyses of data challenges in LLM Application development with a specific impact on the Enterprise Data Platform and the ever-evolving role of Data Engineers.

Course Learning Outcomes

1. Master basic and advanced data engineering principles, focusing on AI with LLM applications.
2. Data lifecycle implementation from Business problem formulation to solution delivery.
3. Learn about current trends and challenges in modern data platform architectures through invited speakers.
4. Study real-world case studies to grasp practical tradeoffs, challenges, and limitations.
5. Build cutting-edge SOTA model (Llama 2, GPT4) based LLM applications to address data challenges inherent in a Data Engineer role.
6. Learn to read, analyze and present NLP/LLM research papers in practical implementation lens.
7. Build professional connections with field experts for future collaboration and growth.

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.*

Required Tools and Course Textbooks.

We will be using the following services and tools as needed:

1. VS Code IDE (free app)
2. AWS Cloud Services
3. Snowflake
4. Data Orchestration tool: Airflow (both free and managed service)
5. Data Integration Tool: Airbyte (free; we will use this as needed)
6. Data Transformation Tool: dbt (free, but will probably use a cloud service, dbtCloud)
7. API Services from OpenAI and/or Hugging Face.
8. Time permitting, governance, and Data Observability tools like BigID, Monte Carlo, or Kensu.

Reference Books:

1. **Fundamentals of Data Engineering**
   Authors: Joe Reis, Matt Housley
   Published by O'Reilly Media, Inc.
2. **Data Quality Fundamentals**
   Barr Moses, Lior Gavish, Molly Vorwerck
   Published by O'Reilly Media, Inc.

Topics Covered. Please include a bulleted list of what topics will be covered in the course.

- Data Engineering
- Data Integration

- Data Orchestration
- Data Visualization
- Data Transformation
- Data Monitoring
- Data Observability
- Data Ingestion
- Data Platform Architecture
- Large Language Models (LLM)
- Generative Artificial Intelligence (AI)
- Data Governance
- Data Quality

Course Activities

1. **Activity #1 (Data Problem Formulation)**
   Data Driven Decision Making: What Business Question are we trying to solve with Data?
2. **Activity #2 (Data Platform Design & Architecture)**
   Deliver an overall data platform design and the data model and it's associated implementation architecture.
3. **Activity #3 (Phased Implementation Deliverable #1)**
   Develop a 1st iteration of this deliverable and present this for feedback from the class.
4. **Activity #4 (Phased Implementation Deliverable #2)**
   Develop a 2nd iteration of this deliverable and present this for feedback from the class.
5. **Activity #5 (Final Project Implementation Deliverable)**
   Present the final project deliverable to the class for feedback and presentation grades
6. **Group Research Paper (discussed throughout the class):** Selection, Analysis and Presentation

Course Schedule (will change due to the dynamic nature of the class & speaker availability)

Week 1	- Class Introductions - Course overview, objectives, and expectations - Introduction to Data Engineering – Data Platform, GenAI, and LLM
Week 2	*Modern Data Engineering: Evolving approaches to support modern AI-driven Product Development* - Data Engineering principles - The role of a Data Engineer in the world of GenAI - The evolving role of Data Engineers -- *Break* – *Modern Data Platform Architectures* - Understanding data architectures - Trends in data platform architectures - How AI and LLM are shaping data platform architectures
Week 3	*AI and Data: A strong interdependency between them* - The role of data in machine learning and AI - Understanding data needs for AI - Challenges in data preparation for AI -- *Break* – *Large Language Models and Data Engineering* - Understanding LLM - The importance of data engineering in LLM - Data needs and challenges for LLM

	-- *Break* – Invited Speaker #1: TBD (Topic & Speaker) - Q&A session
Week 4	Prep and Discussion for: **Activity #1** (Data Problem Formulation) Data Driven Decision Making: What Business Question are we trying to solve with Data? Miscellaneous Topics: Deep Dive into different Data Engineering areas as needed. Invited Speaker #2: TBD (Topic & Speaker) - Q&A session
Week 5	Class Presentation & Feedback: **Activity #1** (Data Problem Formulation) Miscellaneous Topics: Deep Dive into different Data Engineering areas as needed. -- *Break* – Invited Speaker #3: TBD (Topic & Speaker) - Q&A session
Week 6	Prep and Discussion for: **Activity #2** (Data Platform Design & Architecture) Deliver an overall data platform design and the data model and its associated implementation architecture. Miscellaneous Topics: Deep Dive into different Data Engineering areas as needed. -- *Break* – Data Engineering in different Industries: - Manufacturing - Life Sciences - Finance - Retail (B2B & B2C)
Week 7	Class Presentation and Feedback: **Activity #2** (Data Platform Design & Architecture) Miscellaneous Topics: Deep Dive into different Data Engineering areas as needed. -- *Break* – Invited Speaker #4: TBD (Topic & Speaker) - Q&A session
Week 8	Prep and Discussion for: **Activity #3** (Phased Implementation Deliverable #1) Develop a 1st iteration of this deliverable and present this for feedback from the class. Miscellaneous Topics: Deep Dive into different Data Engineering areas as needed. -- *Break* – Invited Speaker #5: TBD (Topic & Speaker) - Q&A session
Week 9	Class Presentation and Feedback: **Activity #3** (Phased Implementation Deliverable #1) Miscellaneous Topics: Deep Dive into different Data Engineering areas as needed. -- *Break* – Invited Speaker #6: TBD (Topic & Speaker)

	- Q&A session
Week 10	Ethics, Privacy, and Security in Data Engineering - Data ethics in AI and LLM - Data privacy concerns - Security practices in data engineering -- *Break* – Prep and Discussion for: **Activity #4** (Phased Implementation Deliverable #2) Develop a 2nd iteration of this deliverable and present this for feedback from the class.
Week 11	Class Presentation & Feedback: **Activity #4** (Phased Implementation Deliverable #2) Develop a 2nd iteration of this deliverable and present this for feedback from the class. Miscellaneous Topics: Deep Dive into different Data Engineering areas as needed. -- *Break* – Invited Speaker #7: TBD (Topic & Speaker) Managing ethics, privacy, and security in Data Platforms
Week 12	What's with all the *Ops and how is that related to Data Engineering: - DataOps - MLOps - DevOps What is Data Observability: - Why is this so critical and the glue that holds everything together. -- *Break* – Invited Speaker #8: TBD (Topic & Speaker) - Q&A session
Week 13	Future of Data Engineering: Emerging Trends and Challenges - Impact of AI advancements on Data Engineering - The future of data platforms with AI and LLM - Open problems and challenges in Data Engineering -- *Break* – Invited Speaker #9: TBD (Topic & Speaker) - Q&A session
Week 14	Final Class Presentation & Feedback: **Activity #5** (Final Project Implementation Deliverable) Present the final project deliverable to the class for feedback and final refinement.
Week 15	Final Class Presentation & Feedback: **Activity #5** - Final Class Presentations (contd., from Week 14) Course Review and Wrap-up - Recap of major course themes - Where to go from here and discuss the challenges and do a class retrospective.

Grade Breakdown:
Class Participation: 15%
Project Completion: 40%
Project Presentation: 20%
Group Research Paper Analysis & Presentation: 25%

Grading Scale: 95-100% A	87-89.9% B+	77-79.9% C+	
	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. Please insert what is applicable for your class. See sample provided below.

Attendance Policy

Students are expected to complete course readings, participate in class discussions or other learning activities during the unit, and complete written assignments for each unit during the time of that unit. It is understood that there might be one week when active participation in ongoing class conversations and learning activities might be delayed. Beyond one week's time, if there is an absence or lateness in participation (1) faculty must be notified in advance; (2) grades will be adjusted accordingly.

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 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 Husky 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.

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.

24/7 Canvas Technical Help

For immediate technical support for Canvas, call 617-373-4357 or email help@northeastern.edu
Canvas Faculty Resources: https://canvas.northeastern.edu/faculty-resources/
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.*