THE GEORGE WASHINGTON UNIVERSITY
School of Engineering and Applied Science
Online Programs Office

Syllabus for
SEAS 8515, DA2
Data Engineering for AI
Spring 2 - 2024

Instructor:	Dr. Adewale Akinfaderin
eMail:	waleakinfaderin@gwu.edu
Credit Hours:	3 credit hours
Course Website:	On Blackboard

Class Time and Dates:

- Day and Time: 1 - 4 pm, Saturday (Eastern)
- All Class Meeting Dates: March 23, 30; Apr 6, 13, 20, 27; May 4, 11, 18; June 1
- Attendance is normally expected at all sessions. If an absence from a class meeting is needed (due to family/medical or work-related emergency) students must contact the instructor in advance.
- Online classes are conducted via Zoom; Links are provided in Blackboard.
- Zoom link for Office Hours: http://gwu-edu.zoom.us/my/akinfaderin

Office Hours: For 3 hours every week I will be available for drop-in office hours, as follows:

- Every Monday, 18:00 to 20:00 ET
- Every Friday, 18:00 to 19:00 ET

Bulletin Description of the Course:

Designing, developing, and implementing data engineering solution for sourcing data into AI applications. Topics that will be discussed in this course will be about how to leverage SQL and Python to define and schedule pipelines that incrementally process new data from a variety of data sources to power AI applications and dashboards. This course offers hands-on instruction in Data Science & Engineering Workspace, Delta Lake, Distributed Computing, Data Ingestion using Python and SQL, Data Processing, Data Warehousing, and Scalable Machine Learning for AI Applications.

Course Learning Objectives:

Upon completing the course, students will know how to:

1. Data Ingestion using Python and SQL (Structured and Batched Data Steaming)
2. Use SQL and Python to write production data pipelines to extract, transform, and load data into tables and views
3. Data Preprocessing with Python
4. Scalable Machine Learning for AI Applications
5. Orchestrate production pipelines to deliver fresh results for ad-hoc AI analytics and dashboarding

Required Textbook and Other Materials:

- Textbook: Learning Spark, 2nd Edition by Jules S. Damji, Brooke Wenig, Tathagata Das, Denny Lee, O'Reilly Media, Inc. ISBN: 9781492050049

Average Amount of Out-of-Class or Independent Learning Expected per Week:

Over 10 weeks, students will spend 3 hours per week in lecture, 1 hour per week in Blackboard discussion, and 6 hours in two exams given outside class hours (about 46 hours of guided instruction for the semester). Homework and other out-of-class work is estimated at around twice the classroom time (92 hours) for a total of about 138 hours of work.

Class Schedule and Assignments

Class	Topic/Activity	Assignment Due
1	Understanding the business of data - Why data engineering? Fundamentals of Distributed Computing, Data Storage Systems, The Data Science Workflow	None
2	Defining computation resources for different data workloads	Discussion1 (March 30, 9am ET)
3	SQL for querying data and sharing business insights Data Ingestion (Structured, Batched and Data Steaming)	Discussion 2 (April 6, 9am ET)
4	Data Processing with Apache Spark and Delta Lake	Individual Project 1 Discussion 3 (April 13, 9am ET)
5	Python Foundation for Data Engineering and AI	Discussion 4 (April 20, 9am ET)
6	Data Warehousing, Create and Manage Schemas for Databases	Individual Project 2 Discussion 5 (April 27, 9am ET)
7	Building reliable, maintainable, and testable data processing pipelines	Discussion 6 (May 4, 9am ET)
8	Machine Learning Overview	Individual Project 3 Discussion 7 (May 11, 9am ET)
9	Machine Learning Advanced Modeling XGBoost, Random Forest and NLP	Discussion 8 (May 18, 9am ET)
10	Scalable Machine Learning and AutoML for AI Applications	Individual Project 4 (June 1, 9am ET)

Course recordings: Downloadable recordings of each class session will be available within about 2 hours of the conclusion of class meetings and will be available for the duration of the course. These recordings are to be used exclusively by registered students in that class for their own private use. *Releasing these recordings is strictly prohibited.*

Weekly Discussion on Blackboard:
At the beginning of the course, I will post an assignment prompt on the discussion board, and you will be randomly assigned to a discussion group. Throughout the course, there will be milestones that need to be met by each discussion group. You are responsible for spending at least one hour each week collaborating within your group and individually posting a one-paragraph response on the Blackboard discussion board for your discussion group to see. During the final week of class, you will submit a 1-2 page summary of the collaborative group project. Mandatory. Calculated as part of grade (includes your weekly posting as well as your end-of-semester report). Responses are due by 9:00 am ET every Saturday.

Contact Mark Griffith at seasonline@gwu.edu (202-422-2806) and copy the instructor's email regarding issues related to the Blackboard.

Online Engineering Programs Labs: Students can remotely access most computer labs of the School of Engineering and Applied Science and work with a variety of engineering design and analysis software packages. See https://www.seas.gwu.edu/remote-access-labs

Grading:

GW's grading system for graduate students is: ***A,*** Excellent; ***B,*** Good; ***C,*** Satisfactory; ***F,*** Fail; other grades that may be assigned are ***A–, B+, B–, C+*****, C-***.* In this course, grades are determined by weighted average values and based on a standard curve relative to the class average:

- **Individual Projects (80%):**
  - o Individual Project 1 (20%)
  - o Individual Project 2 (20%)
  - o Individual Project 3 (20%)
  - o Individual Project 4 (20%)
- **Discussion Grade (20%)**

Written work must comply with the Academic Integrity Policy of the George Washington University policy. Any plagiarized material will receive a grade of 0. No late submission of homework or discussion board will be accepted.

Withdrawals:

- Students may drop from courses through the day after the second class meeting without any academic or financial penalty. After that time, students may withdraw through the day after the eighth class meeting and will receive a designation of "W" and are responsible for full tuition.

Incomplete

- Students who cannot complete a course due to deployment overseas/called to active military duty/death in the immediate family/debilitating illness may seek an incomplete with proper documentation.

University Policies

University Policy on Observance of Religious Holidays: Students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance. See https://registrar.gwu.edu/university-policies#holidays
Student Disability Support Services (DSS) 202-994-8250: Students needing an accommodation based on the potential impact of a disability should contact Disability Support Services. See https://disabilitysupport.gwu.edu/.
Student Mental Health Services 202-994-5300: GW offers 24/7 assistance and referral for students needing crisis and emergency mental consultations, confidential assessment, and counseling services. See https://counselingcenter.gwu.edu/.
Online Engineering Programs Office Policies: https://seasonline.gwu.edu/about-us/policies-procedures-masters/
Emergencies: In case of emergency, students will be notified on Blackboard.
Academic Integrity Code: Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and fabricating information. All academic work is subject to GW University and SEAS Online Programs policy and may be scrutinized electronically. For more information, see https://studentconduct.gwu.edu/.

Policy Violation Consequences

Homework/labs and other written material: Written work must comply with the Academic Integrity Policy of the George Washington University policy. Any plagiarized material will receive a grade of 0.