ADTA 5240: Harvesting, Storing, and Retrieving Data

Professor/Instructor Information

Name: Dr. Richard Appiah
Virtual Office Hours: Tuesdays and Thursdays, 11:00 AM – 1:00 PM CST by appointment on MS Teams or Zoom
Email: Richard.Appiah@unt.edu
Communication Expectations: The best way to communicate with me is by emailing me directly, copying the attending TA where applicable, or setting up an appointment via my virtual office hours. Please email me and copy the TA with any questions or concerns. I do check my emails regularly and will try to respond as quickly as possible. However, if you do not receive a response within two business days, please send a follow-up email – a gentle nudge is always appreciated. For appointments, please let me know 8 hours in advance.

Course Description, Structure, and Objectives

This course introduces the fundamentals of data engineering, including harvesting (and processing), storing, retrieving, exploring, and visualizing data. The goal of this course is to provide students with both theoretical knowledge and practical experience, leading to mastery of the fundamentals of data engineering using both small and large datasets. As these fundamentals are introduced, exemplary technologies will be employed to illustrate how storage and processing architectures can be constructed. The problems are being considered in the context of big data analytics. Exercises and examples will consider both simple and complex data structures, and data ranges from clean and structured to dirty and unstructured.

This course is an 8-week online class covering one module per week.

Important Notice for F-1 Visa Students taking this online course: Federal regulations state that students may apply only three fully online semester credit hours (SCH) to the hours required for full-time status for F-1 Visa holders. Full-time status for F-1 Visa students is 12 hours for undergraduates and 9 hours for graduate students.

By the end of this course, students will be able to:

1. Develop an understanding of the fundamental concepts of modern data management, including data science life cycle, data scaling, structuring data, and data lakes.
2. Develop knowledge and skills in data harvesting, storing, retrieving, and processing using cloud technology.
3. Develop knowledge and skills in working with the Apache Hadoop framework, including Hadoop Distributed File System (HDFS), MapReduce, and Hive.
4. Develop knowledge and skills in working with HDFS, Spark, Linux, SQL, and BigQuery
5. Develop knowledge and skills in cleansing/wrangling data with Google/Open Refine
6. Understand the basics of querying data in BigQuery, Hive, and Spark

Topics included in this course are as follows:

1. Introduction to Big Data
2. Structured and Unstructured Data
3. Data Lifecycle
4. Introduction Google Cloud Platform
5. Exploring Hadoop Ecosystem
6. Introduction to the Linux Operating System
7. Distributed File Systems
8. Data Preparation and Using OpenRefine
9. Data Queries with BigQuery
10. Creating Tables and Querying in Hive and Spark

Required/Recommended Materials

No textbook is required for this course, but we will have articles to read and videos to watch throughout the semester.

These books are NOT required, but you might find them beneficial for extra reinforcement of the material.

1. Marr, B. (2016). Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Wiley. ISBN: 978-1119231387
2. Lakshmanan, V. (2022) Data Science on the Google Cloud Platform 2nd Edition. O'Reilly. ISBN: 978-1098118952
3. Mucchetti, M. (2020). BigQuery for Data Warehousing. Apress. ISBN-13 : 978-1484261859

How to Succeed in this Course

Complete Your Work

Complete your work on time and contact me and the attending TA, where applicable, when you have questions. Contact us through email and/or by attending my virtual office hours. The virtual office hours offer you an opportunity to ask for clarification or find support with understanding class materials. Additional virtual office hours are available by contacting me to schedule an appointment. Remember, your success is our goal.

Learn and Grow

UNT strives to offer you a high-quality education and a supportive environment, so you learn and grow. As a faculty member, I am committed to helping you be successful as a student. To learn more about campus resources and how to succeed at UNT, visit unt.edu/success and explore unt.edu/wellness. To get all your enrollment and student financial-related questions answered, go to scrappysays.unt.edu.

Engage

Every student in this class should have the right to learn and engage in an environment of respect and courtesy from others. We will discuss our classroom's habits of engagement, and I also encourage you to review UNT's student code of conduct so that we can all start with the same baseline civility understanding (Code of Student Conduct) (https://policy.unt.edu/policy/07-012).

Seek Accommodation if Needed

The University of North Texas makes reasonable academic accommodations for students with disabilities. Please review the ADA Policy section for more information.

Teaching Philosophy

I am dedicated to guiding my students towards becoming not just skilled professionals but also responsible, innovative, and ethically minded contributors to society. To achieve this, I encourage my students to utilize cutting edge technologies in the dispensation of their course work in ethical and academically acceptable submissions. Other major bedrocks to my teaching delivery include the integration of real-world experiences, and hands-on approach in a continuous improvement feedback loop. By this approach, I strive to demonstrate a commitment to providing an education that is both academically rigorous and richly aligned with the demands of the contemporary data analytics landscape.

Assessing Your Work

Assessment & Grading

1. There will be **six assignments** throughout the course. (45%)
    - Students are required to submit their homework on time.
2. There will be **one discussion post** throughout the course. (5%)
    - Students are required to submit their discussion post on time.
3. There will be **one midterm**. (25%)
    - Students are required to submit their midterm on time.
4. Lastly, there will be **a final project**. (25%)
    - Students will submit the final project by the deadline provided on the Course Calendar.
5. The total percentage for the course is 100%.

The final letter grade will be determined as follows:

A = 90-100%
B = 80-89.9%
C = 70-79.9%
D = 60-69.9%
F = 0-59.9%

Course Requirements/Schedule

This schedule is subject to change by the professor. Any changes to this schedule will be communicated in class or via class email or Canvas announcement. Additional readings and activities may be added, which will be noted in the weekly module overview instructions.

Week	***Date***	***Topic***	***Activities***	***Assignments Due***
Week 1	*06/03/2024 6:00 – 8:00 PM CST*	*Welcome & Overview Introduction to Big Data & Data Engineering*	*Lecture 1 Readings/Video*	*Discussion Post #1 Assignment #1*
Week 2	*06/11/2024 6:00 – 7:30 PM CST*	*Data Lifecycle Introduction Google Cloud Platform*	*Lecture 2 Readings/Video*	*Assignment #2*
Week 3	*06/18/2024 6:00 – 7:30 PM CST*	*Exploring Hadoop Ecosystem MIDTERM Review*	*Lecture 3 Readings/Video*	*Assignment #3*
Week 4	*TBD*	*Midterm*	*Midterm*	*Midterm*
Week 5	*07/02/2024 6:00 – 7:30 PM CST*	*Introduction to the Linux Operating System Distributed File Systems*	*Lecture 4 Readings/Video*	*Assignment #4*
Week 6	*07/09/2024 6:00 – 7:30 PM CST*	*Data Processing & Data Queries*	*Lecture 5 Readings/Video*	*Assignment #5*
Week 7	*07/16/2024 6:00 – 7:30 PM CST*	*Creating Tables and Querying in Hive & Spark*	*Lecture 6 Readings/Video*	*Assignment #6*
Week 8	*TBD*	*Final Project*	*Final Project*	*Final Project*

Course Policies

Online Course Attendance

Students are expected to participate actively each week and meet all course assignment deadlines as detailed in the Course Calendar. Each lecture will be recorded if students cannot attend lectures at the weekly scheduled time. For more information about the University of North Texas

Attendance Policy, go to: http://policy.unt.edu/policy/15-2-5.

Recordings

All the video recordings in this course, including lecture videos and live class activity recordings, are for students enrolled in the class section to refer to throughout the semester. Class recordings are the intellectual property of the University or instructor and are reserved for use only by students in this class and for educational purposes. Students may not post or otherwise share the recordings in any form outside the class or the Canvas Learning Management System. Failing to follow this restriction violates the UNT Code of Student Conduct and could lead to disciplinary action.

Class Participation

Online participation is expected for this class. To learn more about campus resources and information on how you can achieve success, go to https://succeed.unt.edu.

Late Work

All assignments are to be submitted by the due date and time. The deadline for submitting an assignment is 11:59 pm on the due date. No late midterm or final project will be accepted.

Online Assignment and Examination Policy

The University is committed to providing all users with a reliable online course system. However, if any unexpected server outage or any unusual technical difficulty prevents students from completing a time-sensitive assessment activity, the instructor will extend the time windows and provide an appropriate accommodation based on the situation. Students should immediately report any problems to the instructor and contact the UNT Student Help Desk at helpdesk@unt.edu or 940.565.2324. The instructor and the UNT Student Help Desk will work with the student to resolve any issues immediately.

Scholarly Expectations

All works submitted for credit must be original works created by the scholar uniquely for the class. It is considered inappropriate and unethical to make duplicate submissions of a single work for credit in multiple classes unless specifically requested by the instructor. Students must submit their own work. It is unacceptable to copy work from another student or copy and paste from a website or Generative AI such as ChatGPT, Bard, Bing, or similar. If your work is copied from another student, you will receive a Zero for the assignment and be reported to the Academic Integrity Office. If your work is copied from the web or any other source without a citation, you will receive a Zero for the assignment and be reported to the Academic Integrity Office. If caught a second time, you will get a Zero for the course and will be reported to the Academic Integrity Office again. Two academic integrity offenses will result in a suspension from the ADTA program, and a 3rd offense will result in dismissal from the ADTA program. Please review the UNT Student Academic Integrity policy for more information.

ADTA students must read and adhere to the course, department, and UNT Academic Integrity expectations. The consequences of violating Academic Integrity expectations are outlined below.

Advanced Data Analytics Integrity Policy

	Penalty	**Other**

1st Academic Integrity Offense	The minimum penalty is a 0 for the assignment AND a deduction of one letter grade from the final grade for the course. Other penalties may be assessed by the course instructor up to course failure, depending on the severity of the offense.	All Academic Integrity offenses will be reported to the UNT Academic Integrity Office.
2nd Academic Integrity Offense	Suspension from the ADTA program.	A second offense is defined as a separately reported offense either in the same class as the 1st offense or in a different course. Students suspended for a second Academic Integrity violation will not be allowed to enroll in ADTA courses for one calendar year. For students who had a single Academic Integrity violation prior to Fall 2023, a second violation will result in suspension from the ADTA program.
3rd Academic Integrity Offense	Dismissal from the ADTA program.	Students committing a 3rd Academic Integrity offense will be dismissed from the program. For students who had multiple Academic Integrity violations prior to Fall 2023, any additional violation will result in dismissal from the ADTA program.

Instructor Responsibilities and Feedback

The instructor is responsible for responding to student questions about assignments and projects, the course material presented, and providing additional resources to enhance understanding of course material. Timely feedback is essential for student success, and the instructor is responsible for providing timely feedback to students throughout the course. The instructor or TA will grade submitted assignments and post grades for students within ten days of the assignment's due date.

Syllabus Change Policy

Should any changes to the syllabus be necessary, the professor will provide ample notification to students to adjust for completing assignments in a timely manner without penalty.