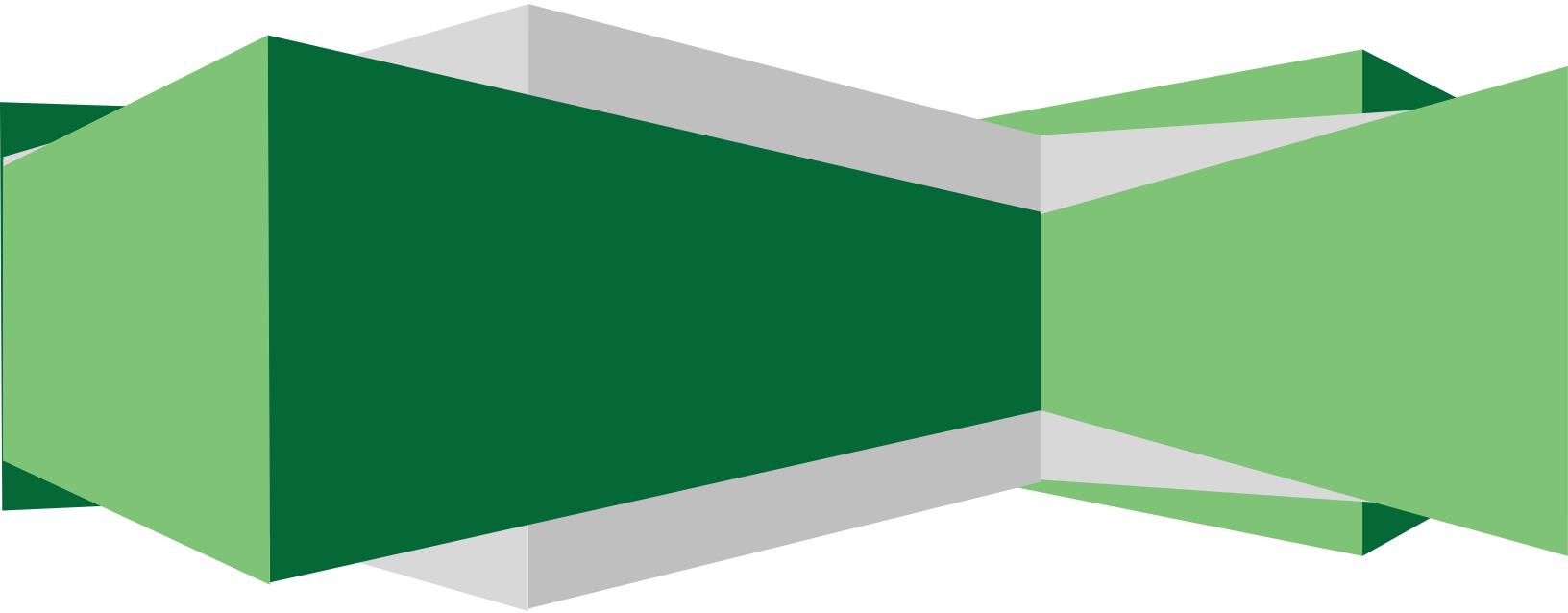
University of North Texas
Advanced Data Analytics – Toulouse Graduate School

Course Syllabus

ADTA 5240 - Harvesting, Storing, and Retrieving Data



ADTA 5240.701: Harvesting, Storing, and Retrieving Data
Spring 2020

COURSE INFORMATION

- ADTA 5240: Harvesting, Storing, and Retrieving Data – SPRING 2020 8W1
- ADTA 5240: 3 credit hours
- 100% online course. No scheduled meeting times are required.

Professor / Instructor Contact Information

- Professor: Thuan L Nguyen, PhD, Clinical Assistant Professor
- Office Location: UNT New College at Frisco #146
- Office Hours: Fridays: 10:00 AM – 12:00 PM; Online by appointment
- Email Address: Thuan.Nguyen@unt.edu

About the Professor / Instructor

Welcome to ADTA 5240.701 – Harvesting, Storing, and Retrieving Data. I would like to share a little information about my background. I finished my BS in computer science, MS in Management Information Systems, and PhD in Information Systems. I had nearly 20 years of work experience in software engineering and information systems (designing, developing, and administering software, hardware systems in telecommunication, networking, business information systems, and IT industries) and data analytics. After that, I started my research and teaching career, first at the University of Texas at Dallas (UTD) and then University of North Texas (UNT). Besides data analytics and machine learning, my current research interests also include the theories of knowledge management, intellectual capital, and their applications in firm operations.

Course Pre-requisites, Co-requisites, and/or Other Restrictions
Required prerequisite courses: NONE

Materials – Text, Readings, Supplementary Readings
No textbook is required for this course.

In addition to the articles listed in the document suggested_reading.docx (Getting Started - Canvas), the following books are for suggested reading:

- Guy Harrison (2016). Next Generation Databases: NoSQL, NewSQL, and Big Data. APress. ISBN: 978-1-4842-1330-8
- DJ Patil and Hilary Mason (2015). Data Driven: Creating a Data Culture. ISBN: 978-1-491-92119-7
- Tom White (2015). Hadoop: The Definitive Guide. ISBN: 978-1491901632
- Holden Karau and Rachel Warren (2017). High Performance Spark. ISBN: 978-1491943205

Course Description

This course introduces the fundamentals of data engineering, including collecting, wrangling, storing, retrieving, and processing 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, as well as data ranges from clean and structured to dirty and unstructured.

Course Objectives

- Develop an understanding of the fundamental concepts of the modern data management, including data science life cycle, data scaling, structuring data, and data lakes
- Develop knowledge and skills in storing, retrieving, and processing data with the Apache Hadoop framework using the cloud technology
- Develop knowledge and skills in working with the Apache Hadoop framework including Hadoop Distributed File System (HDFS), MapReduce, and Hive
- Develop knowledge and skills in working with HDFS and Spark/pySpark
- Develop knowledge and skills in cleansing/wrangling data with Google/Open Refine
- Develop knowledge and skills in collecting data using streaming technologies
- Introduce students to real-time big data using Spark Streaming

Course Topics

- Apache Hadoop framework and its major components
- Apache Hadoop framework and the cloud technology
- Storing & Retrieving data with Apache Hadoop HDFS, MapReduce, and Hive
- Storing & Retrieving data with Apache Hadoop HDFS and Spark
- Data lakes: A storage of choice for the modern data management
- Data cleansing and wrangling with Google/Open Refine
- Data collection with the streaming technologies
- Introduction to real-time big data with Spark Streaming

Teaching Philosophy

For teaching, it is my main goal to create a teaching/learning environment in which students feel respected and valued, and they believe that they are capable contributors. As an instructor, it is my responsibility to determine exactly what I expect students to understand after completing my course, then to facilitate student learning so that every student reaches this level. I believe the essence of effective teaching is to provide students with real-world examples, encourage them to discuss the material, and offer them opportunities to practice what they have learned. I also believe that creating an active learning environment is an essential part of teaching. Asking

questions, promoting discussion, and using real-world analogies are important in an interactive classroom that can enhance students' learning and sustain their enthusiasm. I expect all students to make the best efforts with their class work, respect others, and participate in the class activities so that their experiences can add to the overall learning experience.

COURSE REQUIREMENTS

1. The student will be responsible for checking the announcements that are sent to his/her UNT email.
2. The student will access and follow all course instructions found in the syllabus, announcements, assignments, and all other class-related documents.
3. The student will complete all the class assignments in the time frame specified in the class documents including the course calendar to participate effectively in class activities.
4. The student will complete all the assessment tests and exams – if required – in the time frame specified in the class documents including the course calendar.
5. The student will complete all the projects – if required – in the time frame specified in the class documents including the course calendar.

COMMUNICATIONS

Interaction with Instructor: I look forward to getting to know all of you and working with you. Contact me anytime using my UNT email (Thuan.Nguyen@unt.edu). I will check messages daily and will make every effort to respond as quickly as possible. If you would like to meet me personally, we can meet in the office at UNT New College in Frisco #126 during the office hours (10:00 AM – 12:00 PM Fridays). Please let me know in advance if you intend to have an online meeting.

My goal is for you to enjoy this course, to learn how to engage in meaningful and useful online course activities, to gain a greater understanding of the topics associated with the fundamentals of data engineering, and to help you in any way that I can to be successful.

ASSIGNMENTS, ASSESSMENTS, and PROJECTS

There will be **weekly discussions,** except for the midterm take-home week and the final week.

- The student will respond to posted online course discussion questions each week following the instructions for discussion forums. Each student should submit his/her initial posts to respond to the discussion questions by the deadline provided on the Course Calendar. Then he/she should continue by posting responses to his/her classmates' posts throughout the week to maximize points earned each week. Students are encouraged to enhance the conversation by providing complementary resource materials and properly referenced supplementary items.

There will be **five homework assignments** throughout the course.

- Students are required to submit their homework on time.

There will be **one midterm take-home assignment**.

- Midterm take-home assignment: Assigned and due in Week 4

There will be **a final project**.

- The student will complete a final project following the project instructions.
- Students will submit the final project by the deadline provided on the Course Calendar.

Make-Up Policy
No make-up assessment tests or exams will be offered except for being approved in advance. Students will be required to provide necessary documentation.

Late-work Policy
All assignments are to be submitted using the UNT email.
The deadline for submitting an assignment is 11:00 PM on the due date.
Late submissions can still be submitted up to 24 hours after the deadline. Assignments submitted within 24 hours after the due date/time will be subject to a 25% penalty. No submissions will be accepted later than 24 hours after the deadline.

NOTES: Late work is subject to penalty described above unless previously approved by the instructor.

Class Schedule

The following is a tentative schedule. Should any change become necessary, it will be announced in class as well as in the announcements sent via the UNT email. It is the student's responsibility to check for changes in the schedule.

Week	Date	Topics	Suggested Reading
1	01/13 - 01/19/2020	Course Overview Introduction to Apache Hadoop Framework Apache Hadoop Framework in the Cloud GOOGLE CLOUD PLATFORM: GCP for Hadoop & Spark Homework Assignment 1: Assigned	Readings: Week 1
2	01/20 - 01/26/2020	Fundamental Concepts: Data Analytics Life Cycle Fundamental Concepts: Structured & Unstructured Data Exploring Apache Hadoop System in the Cloud Discussion 1: Assigned (Monday) Homework Assignment 1: Due Homework Assignment 2: Assigned Discussion 1: Due (Sunday)	Readings: Week 2
3	01/27 - 02/02/2020	Introduction to Data Lakes Introduction to Apache Hive Storing & Retrieving Data with HDFS, Map Reduce, & Hive Discussion 2: Assigned (Monday) Homework Assignment 2: Due Homework Assignment 3: Assigned Discussion 2: Due (Sunday)	Readings: Week 3
4	02/03 - 02/09/2020	**Midterm Take-Home: Assigned** **Midterm Take-Home: Due**	NO Readings
5	02/10 - 02/16/2020	Introduction to Apache Spark Storing and Querying Data with HDFS & Spark API's Discussion 3: Assigned (Monday)	Readings: Week 5

		Homework Assignment 3: Due Homework Assignment 4: Assigned Discussion 3: Due (Sunday)	
6	**02/17 - 02/23/2020**	Cleansing & Wrangling Data with Google/Open Refine Discussion 4: Assigned (Monday) Homework Assignment 4: Due Homework Assignment 5: Assigned Discussion 4: Due (Sunday)	Readings: Week 6
7	**02/24 - 03/01/2020**	Introduction to Spark Streaming Collecting data with Streaming Technologies Discussion 5: Assigned (Monday) Homework Assignment 5: Due Discussion 4: Due (Sunday)	Readings: Week 7
8	**03/02 - 03/05/2020**	**Final Project Presentation**	

GRADING POLICY

The student's grade in the course consists of the following components:

Weekly Discussions:	15%
Homework Assignments:	30%
Midterm Take-Home:	25%
Final Project:	30%

The final letter grade will be determined as follows:

- **A: 90 – 100**
- **B: 80 – 89**
- **C: 65 – 79**
- **D: 50 – 64**
- **F: < 50**

STUDENT TECHNICAL SUPPORT

The University of North Texas UIT Student Helpdesk provides student technical support in the use of Canvas and supported resources (https://clear.unt.edu/services/lms-support). The student help desk may be reached at:

Email: helpdesk@unt.edu
Phone: 940.565-2324
In Person: Sage Hall, Room 130

Business hours are:

- Monday-Thursday 8am-midnight
- Friday 8am-8pm
- Saturday 9am-5p
- Sunday 8am-midnight

ACCESS & NAVIGATION

Access and Log in Information

This course was developed and will be facilitated utilizing the University of North Texas' Canvas. To get started with the course, please go to the website at https://canvas.unt.edu

You will need your EUID and password to log in to the course. If you do not know your EUID or have forgotten your password, please go to the website at http://ams.unt.edu.

TECHNICAL REQUIREMENTS / ASSISTANCE

The following information has been provided to assist you in preparation for the technological aspect of the course.

UIT Help Desk: http://www.unt.edu/helpdesk/index.htm
Web Browser
Word Processor
Creating and submitting files in Microsoft Office, the standard software for this course

Minimum Technical Skills Needed:

Using the learning management system
Using email with attachments
Copying and pasting
Downloading and installing software
Using spreadsheet programs

ACADEMIC POLICIES

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, particularly at the graduate level, to make duplicate submissions of a single work for credit in multiple classes, unless specifically requested by the instructor. Work submitted at the graduate level is expected to demonstrate higher-order thinking skills and be of significantly higher quality than work produced at the undergraduate level.

Instructor Responsibilities and Feedback

The instructor is responsible for responding to student questions about assignments and projects, about the course material presented, and for 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 will actively participate in each week's discussion forum and will provide feedback to students each week regarding their participation. The instructor will grade submitted assignments and will post grades for students within 10 days of assignment due date.

Class Participation

Students are required to log in regularly to the online class site. Students are also required to participate in all class activities such as discussion boards, chat or conference sessions, and group projects. To learn more about campus resources and information on how you can achieve success, go to succeed.unt.edu

Virtual Classroom Citizenship

The same guidelines that apply to traditional classes should be observed in the virtual classroom environment. Please use proper netiquette when interacting with class members and the professor.

Incompletes

This course will observe the UNT policy on incompletes, found here:
http://registrar.unt.edu/grades/incompletes

Policy on Server Unavailability or Other Technical Difficulties

The University is committed to providing a reliable online course system to all users. However, in the event of any unexpected server outage or any unusual technical difficulty which 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 also contact the UNT Student Help Desk: helpdesk@unt.edu or 940.565.2324. The instructor and the UNT Student Help Desk will work with the student to resolve any issues at the earliest possible time.

Copyright Notice
Some or all of the materials on this course Web site may be protected by copyright. Federal copyright law prohibits the reproduction, distribution, public performance, or public display of copyrighted materials without the express and written permission of the copyright owner unless fair use or another exemption under copyright law applies. Additional copyright information may be located at http://copyright.unt.edu.

Graduate Online Course Attendance Policy
Students are expected to participate actively each week and to meet all deadlines for course assignments as detailed in the Course Calendar. *Information about the University of Texas' Attendance Policy may be found at http://policy.unt.edu/policy/15-2-5*

Administrative Withdrawal
This course will observe the UNT policy on academic withdrawal found here:
https://deanofstudents.unt.edu/withdrawals

Syllabus Change Policy
Changes to the course syllabus or due dates are not anticipated but should they be necessary, the instructor will provide ample notification to students to allow them to complete assignments in a timely manner without penalty.

UNT GENERAL POLICIES

Student Conduct and Discipline: *Student Handbook*.
You are encouraged to become familiar with the University's Policy of Academic dishonesty found in the *Student Handbook*. The content of the Handbook applies to this course. If you are in doubt regarding the requirements, please consult with me before you complete any requirements of the course.

The UNT Code of Student Conduct can be found here:
https://deanofstudents.unt.edu/sites/default/files/code_of_student_conduct.pdf

ADA Policy
The University of North Texas makes reasonable academic accommodation for students with disabilities. Students seeking accommodation must first register with the Office of Disability Accommodation (ODA) to verify their eligibility. If a disability is verified, the ODA will provide you with an accommodation letter to be delivered to faculty to begin a private discussion regarding your specific needs in a course. … Note that students must obtain a new letter of accommodation for every semester and must meet with each faculty member prior to implementation in each class. Students are strongly encouraged to deliver letters of accommodation during faculty office hours or by appointment. Faculty members have the authority to ask students to discuss such letters during their designated office hours to protect the privacy of the student. For additional information see the Office of Disability Accommodation website at http://disability.unt.edu/. You may also contact them by phone at 940.565.4323.

Add/Drop Policy

The University of North Texas Add Drop Policy for Fall 2017 can be found at the following link: http://registrar.unt.edu/registration/fall-add-drop

Important Notice for F-1 Students taking Distance Education Courses:

Federal Regulation

To read detailed Immigration and Customs Enforcement regulations for F-1 students taking online courses, please go to the Electronic Code of Federal Regulations website at http://www.oea.gov/index.php/links/electronic-code-of-federal-regulations. The specific portion concerning distance education courses is located at "Title 8 CFR 214.2 Paragraph (f) (6) (i) (G)" and can be found buried within this document: http://www.gpo.gov/fdsys/pkg/CFR-2012-title8-vol1/xml/CFR-2012-title8-vol1-sec214-2.xml

The paragraph reads:

(G) For F–1 students enrolled in classes for credit or classroom hours, no more than the equivalent of one class or three credits per session, term, semester, trimester, or quarter may be counted toward the full course of study requirement if the class is taken on-line or through distance education and does not require the student's physical attendance for classes, examination or other purposes integral to completion of the class.

University of North Texas Compliance

To comply with immigration regulations, an F-1 visa holder within the United States may need to engage in an on-campus experiential component for this course. This component (which must be approved in advance by the instructor) can include activities such as taking an on-campus exam, participating in an on-campus lecture or lab activity, or other on-campus experience integral to the completion of this course. If such an on-campus activity is required, it is the student's responsibility to do the following:
(1) Submit a written request to the instructor for an on-campus experiential component within one week of the start of the course.
(2) Ensure that the activity on campus takes place and the instructor documents it in writing with a notice sent to the International Student and Scholar Services Office. ISSS has a form available that you may use for this purpose.
Because the decision may have serious immigration consequences, if an F-1 student is unsure about his or her need to participate in an on-campus experiential component for this course, s/he should contact the UNT International Student and Scholar Services Office (telephone 940-565-2195 or email internationaladvising@unt.edu) to get clarification before the one-week deadline.