University of North Texas
Toulouse Graduate School

Course Syllabus
ADTA 5240 – Harvesting, Storing, and Retrieving Data
COURSE INFORMATION
- ADTA 5240: Harvesting, Storing, and Retrieving Data – Section 501 – Spring 2020
- Monday 6:30 PM – 7:50 PM, Frisco Campus Room #109 and using Canvas at (UNT.Instructure.com)
- ADTA 5240: 3 credit hours

Professor / Instructor Contact Information
- Instructor: Abdulrahman Habib, PhD
- Office Location: UNT New College at Frisco #109
- Office Hours: Monday 2:00 PM – 6:00 PM and online by appointment
- Email Address: habib@unt.edu

About the Professor / Instructor

Welcome to ADTA 5240 – Harvesting, Storing, and Retrieving Data. My name is Abdulrahman Habib, call me Habib. I have over 15 years of work experience in IT project management. I led projects in Smart City, e-government, e-commerce, and ERP implementation for small and medium businesses. I am passionate about how technology improves people lives and contribute to our society. I have a BS in computer science, MS in Engineering Systems, and a Ph.D. in Information Science. My current research interests include applied research on emerging technologies, innovation, and smart cities applications.
Course Pre-requisites, Co-requisites, and/or Other Restrictions

Required prerequisite courses: None

Minimum Technical Skills Needed:
- Using the learning management system
- Using email with attachments
- Copying and pasting
- Downloading and installing software
- Using spreadsheet programs

Materials – Text, Readings, Supplementary Readings

No textbook is required for this course.

The following books are for suggested reading:


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 using Spark Streaming
- Introduction to the Python API for Spark: pySpark

COURSE REQUIREMENTS

1. The student will be responsible for checking Canvas announcements. I recommend you using Canvas mobile application and setup your course notifications to your preferred email. Read more about Canvas: https://community.canvaslms.com/docs/DOC-10701-canvas-student-guide-table-of-contents

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 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 (habib@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 we can do an online conference call or we can meet in person at UNT New College in Frisco #109. In both cases please send me and email to schedule the 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 data visualization, and to help you in any way that I can to be successful.

ASSIGNMENTS, ASSESSMENTS, and PROJECTS

1. 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.

2. There will be five homework assignments throughout the course. Students are required to submit their homework on time.

3. There will be one midterm take-home assignment. Students have one week to complete it.

4. There will be a final project.
   a. The student will complete a final project following the project instructions.
   b. Students will submit the final project by the deadline.

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 11:59 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.

GRADING POLICY
The student’s grade in the course consists of the following components:

Weekly Discussions: 15%
Homework Assignments: 30%
Midterm Assignment: 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

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](http://www.unt.edu/helpdesk/index.htm)
- **Web Browser**
- **Word Processor**
- Creating and submitting files in Microsoft Office, the standard software for this course

**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

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 940-5652195 or email internationaladvising@unt.edu) to get clarification before the one-week deadline.

Class Schedule
The following is a tentative schedule. Should any change become necessary, it will be announced via the Canvas announcement. It is the student’s responsibility to check for changes in the schedule.
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topics / Required Reading</th>
<th>Assignment Due</th>
</tr>
</thead>
</table>
| 1    | Jan 13 – Jan 19 | Course Overview  
Fundamental Concepts: Structured & Unstructured data                             | Self-introduction      |
| 2    | Jan 21 – Jan 26 | * Monday Jan 21st is MLK day, No Class    
Fundamental Concepts: Data Analytics Life Cycle                          | Discussion #1          |
| 3    | Jan 27 – Feb 2  | Introduction to Apache Hadoop Framework  
Hadoop The Definitive Guide: Ch1                                         | Assignment #1          |
| 4    | Feb 3 – Feb 9   | Apache Hadoop Framework in the cloud  
Hadoop The Definitive Guide: Ch2                                          | Discussion #2          |
| 5    | Feb 10 – Feb 16 | Exploring Apache Hadoop System  
Hadoop The Definitive Guide: Ch3                                          | Assignment #2          |
| 6    | Feb 17 – Feb 23 | Introduction to Apache Hive  
Hadoop The Definitive Guide: Ch4                                           | Discussion #3          |
| 7    | Feb 24 – Mar 1  | Storing & Retrieving Data with HDFS, MapReduce, & Hive  
Hadoop The Definitive Guide: Ch 5, 7, & 17                                  | Assignment #3          |
| 8    | Mar 2 – Mar 8   | Midterm Exam                                                                            | Midterm Exam           |
| 9    | Mar 9 – Mar 15  | Spring Break                                                                            |                        |
| 10   | Mar 16 – Mar 22 | Introduction to Data Lakes                                                                |                        |
| 11   | Mar 23 – Mar 29 | Introduction to Apache Spark  
Apache Spark in the Apache Hadoop Ecosystem  
High Performance Spark: Ch 1                                              | Discussion #4          |
| 12   | Mar 30 – Apr 5  | Storing and Querying Data with HDFS & Spark API’s  
High Performance Spark: Ch 2                                              | Assignment #4          |
| 13   | Apr 6 – Apr 12  | Cleansing and Wrangling Data with Google/open Refine                                      | Assignment #5          |
| 14   | Apr 13 – Apr 19 | Introduction to Spark Streaming  
Collecting data with Streaming Technologies                                     | Discussion #5          |
| 15   | Apr 20 – Apr 26 | Final Project Submission  
Final Project Peer review                                                        | Final Project          |
| 16   | Apr 27 – May 3  | Final Project Peer review                                                                 | Final Project Peer review |