ADTA 5240/IPAC 4240 Harvesting, Storing, and Retrieving Data
Fall 2022
Dr. Boyce

COURSE INFORMATION
Welcome to ADTA 5240/IPAC 4240. I am so excited to have you in class this semester and I look forward to working with all of you throughout the course.

- Harvesting, Storing, and Retrieving Data
- Credit Hours: 3
- Class Times: 100% Online

Instructor Contact Information
- Dr. LeAnn K. Boyce
- Office Hours: By appointment
- Email: leann.boyce@unt.edu

Communication With Instructor
Communication is essential for a successful semester!

Interaction with Instructor: I look forward to getting to know all of you and working with you. Contact me anytime using my UNT email (Leann.Boyce@unt.edu). I will check the email daily and will make every effort to respond as quickly as possible. If you have not heard back from me in 24 hours, please feel free to send another email. Here is a great website provided by CLEAR to give you some communication tips for communicating online: CLEAR has a webpage for students that provides Online Communication Tips.

Teaching Philosophy
I believe students must be given clear direction in order for them to succeed. It is important that they understand the course learning objectives as documented in the syllabus, and the deadlines for submission of all course work as documented on the course calendar. I have found that if a student understands the learning objectives and the contribution to their professional development of each assignment, the quality of their work improves.

I also believe that students need to be given timely feedback so that they know how to improve and respond to communications from their professor in a timely manner. Timely feedback help students gain self-confidence and inspire them to fully engage in the course.
material and to believe that they can do well. It is important for students to know that as a professor, I understand that “life happens” and that I will be flexible when a legitimate problem arises. I also think that it is important to establish performance guidelines in the syllabus and to establish a safe and welcoming learning community in each course. It is my responsibility as a teacher to get my students ready for the workforce and to ensure that they are prepared to meet the professional challenges ahead.

Lastly, I believe we can all learn from one another, and I have assignments built around this concept. Research shows that engaging students in the learning process is important to increase their attention and focus and is a tool to practice critical thinking skills which promotes more meaningful learning experiences. I look forward to a wonderful semester and learning along with you!

**Course Pre-requisites, Co-requisites, and/or Other Restrictions**

Required prerequisite courses: NONE

**Course Description**

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, as well as data that ranges from clean and structured to dirty and unstructured.

The undergraduate class IPAC 4340, will follow along with this same syllabus but assignments and exams will be modified.

**Course Objectives**

- **Outcome 1**: Develop an understanding of the fundamental concepts of modern data management, including data analytics life cycle, data scaling, structuring data, and data lakes.

- **Outcome 2**: Develop knowledge and skills in harvesting, storing, retrieving, and processing data using the cloud technology.

- **Outcome 3**: Develop knowledge and skills in working with the Apache Hadoop framework including Hadoop Distributed File System (HDFS), MapReduce, Hive, and Spark, as well as Colossus.
Outcome 4
Develop knowledge and skills in working with HDFS, Spark, Linux, SQL, BigQuery, MySQL, and Cloud Spanner

Outcome 5
Develop knowledge and skills in cleansing/wrangling data with Google/Open Refine

Outcome 6
Introduce students to querying data in BigQuery, Hive, and Spark.

Course Topics

- Big Data
- Data analytics life cycle
- Data preprocessing
- Linux
- Distributed File Systems
- Hadoop Ecosystem Foundation
- Data Ingestion/Migration
- Introduction to SQL
- Data Preparation using OpenRefine
- Introduction to Google Cloud Platform
- Google Cloud Platform: Cloud Storage
- Google Cloud Platform: Creating a Bucket
- Google Cloud Platform: Cloud SQL
- Google Cloud Platform: Cloud Spanner
- Google Cloud Platform: BigQuery
- Queries with Hive, Spark, and BigQuery

Materials – Text, Readings, Supplementary Readings

No textbook is required for this course, but we will have articles to read throughout the semester (these are listed in the weekly modules under content).

These books are NOT required but you might find them beneficial for extra reinforcement of the material. If you are thinking about purchasing some of the Packt books, check out their website for a subscription as it is well worth the money.


Course Requirements

- The student will be responsible for checking the announcements in the UNT email and other types of class communication daily.
- The student will access and follow all course instructions found in the syllabus, announcements, assignments, and all other class-related documents.
- 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. Please note only 1 extension will be granted without penalty so be sure to only use the one extension when truly needed. After one extension, there will be a 25% deduction for each day the work is submitted late. This will be strictly adhered to throughout the semester unless prearranged with the professor. This is a very fast-paced class, and all assignments must be submitted on time to receive full credit for each submission.
- The student must complete all the assessment tests and exams in the time frame specified in the class documents, including the course calendar. There are NO extensions for mid-terms or finals. Please note receiving a zero for a midterm or final will have a major impact on your final grade.

Assessment & Grading

- There will be ten GCP-centered exercises throughout the course.
- There will be two discussions.
- There will be one quiz on plagiarism in the course.
- There will be one midterm take-home exam.
- There will be one take-home project/exam for the final.

Make-Up Policy

No make-up assignment or exams will be offered except for being approved in advance. Students will be required to provide the necessary documentation.
Late-work Policy

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. You are allowed one late assignment, without penalty. After the one late submission, the subsequent assignments will incur a 25% deduction for each day thereafter.

NOTES: Late work is subject to the penalty described above unless previously approved by the instructor. No Exceptions!

GRADING POLICY

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

- **Homework Assignments**: 30%
- **Discussions**: 10%
- **Midterm Exam**: 30%
- **Final Project/exam**: 30%

The final letter grade will be determined as follows:

- **A**: 90 – 100
- **B**: 80 – 89
- **C**: 70 – 79
- **D**: 60 – 69
- **F**: < 60

Class Schedule

The following is a tentative schedule. Should any change become necessary, it will be announced via the UNT email. It is the student’s responsibility to check for changes in the schedule.
<table>
<thead>
<tr>
<th>Module</th>
<th>Date: Due at 11:59 pm (Central Time) on the last day of the module.</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aug. 29 - Sept. 2</td>
<td></td>
</tr>
</tbody>
</table>

**Lectures:**
- Welcome & Overview
- Plagiarism
- Introduction to Big Data
- Google Cloud Platform Overview

**Readings:**
- *6 Consequence of Plagiarism*. iThenticate.

**Homework:**
- Plagiarism Quiz
- Personal Introduction
- Class Survey
- Setting up a GCP Account and Creating a Project in GCP
<table>
<thead>
<tr>
<th>2</th>
<th>Sept. 3 - Sept. 9</th>
</tr>
</thead>
</table>

### Lectures
- Introduction to Cloud Storage
- Introduction to Linux

### Readings:
- Pletcher, S. (2022). Storage services compared: [AWS vs Azure vs GCP](#).

### Homework:
- How to Create a New Storage Bucket, 3 Folders, and Load Data into a folder and Create a Cluster in GCP.
- Please get started on Exploring Hadoop and Spark Ecosystem. This is not officially due until next week but notice next week there are 3 assignments due. You will need to watch the video on Distributed File Systems to fully understand what is happening in this homework, but it is an easy assignment to complete, and it might help you understand the content of the video on distributed file systems in next week's module.
<table>
<thead>
<tr>
<th>Week</th>
<th>Dates</th>
<th>Lectures</th>
<th>Readings:</th>
<th>Homework:</th>
</tr>
</thead>
</table>
| 3    | Sept. 10 - Sept. 16 | • Distributed File System  
• Setting up a Virtual Machine via GCP with Linux Commands  
• MySQL and Cloud Spanner in GCP |
<p>| 4    | Sept. 17 - Sept. 23 | Midterm                                       |                                                                           |                                                                                                      |</p>
<table>
<thead>
<tr>
<th></th>
<th>Sept. 23 - Sept. 30</th>
<th>Lectures</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• Analytics Life Cycle</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data Ingestion vs. Data Migration</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Readings:</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Law, J. (2022). <a href="#">Take control of your data ingestion in your BigQuery project</a>.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Pierce, S. &amp; Tekiner, F. (2021). <a href="#">What type of data processing organization are you?</a>.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Homework:</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Joining Data from Different Publicly available datasets.</td>
</tr>
<tr>
<td></td>
<td>Oct. 1 - Oct. 7</td>
<td>Lectures</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data Preparation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Introduction to SQL</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Readings:</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Finances Online. (2022). <a href="#">OpenRefine Review</a>.</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Homework:</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data Prep with OpenRefine</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Discussion: To Clean or Not to Clean</td>
</tr>
<tr>
<td></td>
<td>Oct. 8 - Oct. 14</td>
<td>Lectures:</td>
</tr>
<tr>
<td>----</td>
<td>----------------</td>
<td>-----------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>• Data Queries with BigQuery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Creating Table and Querying in Hive and Spark</td>
</tr>
</tbody>
</table>

|    |                | Readings:                                                                                      |
|    |                | • Geeks for Geeks. 2022. [Difference between Apache Hive and Apache Spark SQL](https://databricks.com/apache-hive) | |
|    |                | • Databricks. 2022. [Apache Hive](https://databricks.com/apache-hive)                          | |
|    |                | • Sigmoid. 2022. [Apache Spark on DataProc vs Google BigQuery](https://sigmoid.com/apache-spark-on-dataproc-vs-google-bigquery) | |

<table>
<thead>
<tr>
<th></th>
<th>Oct. 14 at midnight - Oct. 21 at 5 pm</th>
<th>Homework:</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td></td>
<td>• Querying Data in BigQuery</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Querying Data in Hive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Querying Data in Spark</td>
</tr>
</tbody>
</table>

**Final Project/exam**

**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.

**Course Evaluation**

Student feedback is important and an essential part of participation in this course. The student evaluation of instruction is a requirement for all organized classes at UNT. The survey will be made available from October 10th - October 20th to provide students with an opportunity to evaluate how this course is taught. Students will receive an email from "UNT SPOT Course Evaluations via IASystem Notification" (noreply@iasystem.org) with the survey link. Students should look for the email in their UNT email inbox. Simply click...
on the link and complete the survey. Once students complete the survey, they will receive a confirmation email that the survey has been submitted. For additional information, please visit the SPOT website or email spot@unt.edu.

**Technical Requirements/Assistance**

**UNT Help Desk:**

The University of North Texas provides student technical support in the use of Canvas and support resources. The student help desk may be reached at:

Email: helpdesk@unt.edu

Phone: 940.565-2324

In-Person: Sage Hall, Room 130

**Hours** are:

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

**Canvas technical requirements:**

Other related hardware or software is necessary for the course: such as a headset/microphone for synchronous chats, a word processor, etc.

**Student Academic Support Services**

- **Code of Student Conduct:** provides Code of Student Conduct along with other useful links
- **Office of Disability Access:** exists to prevent discrimination based on disability and to help students reach a higher level of independence
- **Counseling and Testing Services:** provides counseling services to the UNT community, as well as testing services, such as admissions testing, computer-based testing, career testing, and other tests
- **UNT Libraries**
- **UNT Learning Center:** provides a variety of services, including tutoring, to enhance the student academic experience
- **UNT Writing Center:** offers free writing tutoring to all UNT students, undergraduates, and graduates, including online tutoring
- **Succeed at UNT:** information regarding how to be a successful student at UNT
Course Policies

Online Assignments and Examinations Policy

The University is committed to providing a reliable online course system to all users. However, in the event of an 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 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.

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 or TA will grade submitted assignments and will post grades for students within 10 days of the assignment due date.

UNT Policies

Academic Integrity Policy

Academic Integrity Standards and Consequences. According to UNT Policy 06.003, Student Academic Integrity, academic dishonesty occurs when students engage in behaviors including, but not limited to cheating, fabrication, facilitating academic dishonesty, forgery, plagiarism, and sabotage. A finding of academic dishonesty may result in a range of academic penalties or sanctions ranging from admonition to expulsion from the University.

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.

PLEASE NOTE: “Plagiarism is a very serious offense at UNT. Even if you don't mean to plagiarize, you can still commit academic dishonesty. When you are accepted to UNT, you are expected to adhere to our policies and standards of Student Academic Integrity in the Code of Student Conduct (as defined in University 07.012). Major violations can
result in expulsion from the university. More information on plagiarism can be found through the university library.

**ADA Policy**

UNT makes reasonable academic accommodations 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 a student with an accommodation letter to be delivered to the faculty to begin a private discussion regarding one’s specific course needs. Students may request accommodations at any time; however, ODA notices of accommodation should be provided as early as possible in the semester to avoid any delay in implementation. 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. For additional information see the ODA website at [disability.unt.edu](http://disability.unt.edu).

**Emergency Notification & Procedures**

UNT uses a system called Eagle Alert to quickly notify students with critical information in the event of an emergency (i.e., severe weather, campus closing, and health and public safety emergencies like chemical spills, fires, or violence). In the event of a university closure, please refer to Canvas for contingency plans for covering course materials.

**Retention of Student Records**

Student records pertaining to this course are maintained in a secure location by the instructor of record. All records such as exams, answer sheets (with keys), and written papers submitted during the duration of the course are kept for at least one calendar year after course completion. Course work completed via the Canvas online system, including grading information and comments, is also stored in a safe electronic environment for one year. Students have the right to view their individual records; however, information about students’ records will not be divulged to other individuals without proper written consent. Students are encouraged to review the Public Information Policy and the Family Educational Rights and Privacy Act (FERPA) laws and the University’s policy. See UNT Policy 10.10, Records Management and Retention for additional information.

**Acceptable Student Behavior**

Student behavior that interferes with an instructor’s ability to conduct a class or other students’ opportunity to learn is unacceptable and disruptive and will not be tolerated in any instructional forum at UNT. Students engaging in unacceptable behavior will be directed to leave the classroom and the instructor may refer the student to the Dean of Students to consider whether the student’s conduct violated the Code of Student Conduct. The University's expectations for student conduct apply to all instructional
forums, including University and electronic classrooms, labs, discussion groups, field trips, etc. The Code of Student Conduct can be found at University Policy Office.

Access to Information - Eagle Connect

Students’ access point for business and academic services at UNT is located at: my.unt.edu. All official communication from the University will be delivered to a student’s Eagle Connect account. For more information, please visit the website that explains Eagle Connect and how to forward e-mail: eagleconnect.unt.edu/.

Sexual Assault Prevention

UNT is committed to providing a safe learning environment free of all forms of sexual misconduct, including sexual harassment sexual assault, domestic violence, dating violence, and stalking. Federal laws (Title IX and the Violence Against Women Act) and UNT policies prohibit discrimination on the basis of sex and therefore prohibit sexual misconduct. If you or someone you know is experiencing sexual harassment, relationship violence, stalking, and/or sexual assault, there are campus resources available to provide support and assistance. UNT’s Survivor Advocates can assist a student who has been impacted by violence by filing protective orders, completing crime victim’s compensation applications, contacting professors for absences related to an assault, working with housing to facilitate a room change where appropriate, and connecting students to other resources available both on and off-campus. The Survivor Advocates can be reached at SurvivorAdvocate@unt.edu or by calling the Dean of Students Office at 940-565-2648. Additionally, alleged sexual misconduct can be non-confidentially reported to the Title IX Coordinator at oeo@unt.edu or at (940) 565 2759.

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 online or through distance education and does not require the student’s physical attendance for classes, examination or other purposes integral to the completion of the class. An online or distance education course is a course that is offered principally through the use of television, audio, or computer transmission including open broadcast, closed circuit, cable,
microwave, satellite, audio conferencing, or computer conferencing. If the F-1 student's course of study is in a language study program, no online or distance education classes may be considered to count toward a student's full course of study requirement.

**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 experiences 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. ISIS 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.

**Student Verification**

UNT takes measures to protect the integrity of educational credentials awarded to students enrolled in distance education courses by verifying student identity, protecting student privacy, and notifying students of any special meeting times/locations or additional charges associated with student identity verification in distance education courses.

See [UNT Policy 07-002 Student Identity Verification, Privacy, and Notification and Distance Education Courses](#).

**Use of Student Work**

A student owns the copyright for all work (e.g., software, photographs, reports, presentations, and email postings) he or she creates within a class, and the University is not entitled to use any student work without the student's permission unless all of the following criteria are met:
• The work is used only once.
• The work is not used in its entirety.
• The use of the work does not affect any potential profits from the work.
• The student is not identified.
• The work is identified as student work.

If the use of the work does not meet all of the above criteria, then the University office or department using the work must obtain the student’s written permission.

Download the UNT System Permission, Waiver, and Release Form

Transmission and Recording of Student Images in Electronically Delivered Courses

• No permission is needed from a student for his or her image or voice to be transmitted live via videoconference or streaming media, but all students should be informed when courses are to be conducted using either method of delivery.
• In the event an instructor records a student presentation, he or she must obtain permission from the student using a signed release in order to use the recording for future classes in accordance with the Use of Student-Created Work guidelines above.
• Instructors who video-record their class lectures with the intention of re-using some or all of the recordings for future class offerings must notify students on the course syllabus if students’ images may appear on the video. Instructors are also advised to provide accommodation for students who do not wish to appear in class recordings.

Example: This course employs lecture capture technology to record class sessions. Students may occasionally appear on the video. The lecture recordings will be available to you for study purposes and may also be reused in future course offerings.

No notification is needed if only audio and slide capture is used or if the video only records the instructor's image. However, the instructor is encouraged to let students know the recordings will be available to them for study purposes.