Course number, Course CRN, COURSE TITLE
Advanced Methods in Big Data Analytics – CIS 5354 – 110
Spring 2020
Texas A&M University-Central Texas

COURSE DATES, MODALITY, AND LOCATION
Course meeting dates: January 13, 2020 – May 08, 2020
Class meeting building and room number: Founder's Hall 207
Class meeting day and time: Monday 6:00 pm – 9:00 pm

INSTRUCTOR AND CONTACT INFORMATION
Instructor: Rahul Dwivedi
Office: Founders Hall 323N
Phone: 254 – 519 – 5784
Email: rahul.dwivedi@tamuct.edu (preferred) or Canvas inbox.

Office Hours
Monday 1:00 PM – 4:00 PM
Thursday 1:00 PM – 4:00 PM
At other times: By appointment through email

Mode of instruction and course access:

This a face-to-face course that makes extensive use of the TAMUCT Canvas Learning Management System (https://tamuct.instructure.com). The course syllabus, schedule, supplemental readings, class announcements, power point slides, learning modules, homework assignments, exams and other course related documents will be posted on Canvas. Each student is responsible for the posted material and should check Canvas several times a week for updates.

Student-instructor interaction
I typically respond to Canvas email within 24 hours except on weekends. Email is the best mode of communication (avoid phone calls or voice messages).

WARRIOR SHIELD
Emergency Warning System for Texas A&M University-Central Texas
Warrior Shield is an emergency notification service that gives Texas A&M University-Central Texas the ability to communicate health and safety emergency information quickly via email, text message, and social media. All students are automatically enrolled in Warrior Shield through their myCT email account.

Connect to Warrior Shield by 911Cellular [https://portal.publicsafetycloud.net/Account/Login] to change where you receive your alerts or to opt out. By staying enrolled in Warrior Shield, university officials can quickly pass on safety-related information, regardless of your location.

COURSE INFORMATION

Course Overview and description

Study advanced concepts and principles of Big Data Analytics and its role in supporting/enhancing organizational decision-making and predictions. Special emphasis on Hadoop Ecosystem, MapReduce, Pig, Hive, Spark, YARN, Sqoop, Flume and HBase.

Course Objectives and Student Learning Outcomes

Successful completion of this course should enable the student to:

- Describe data intensive computing
- Describe the Hadoop Ecosystem
- Implement Hadoop/MapReduce functionality
- Utilize tools such as Pig, Hive, Spark etc. for analyzing large data sets

Competency Goals Statements (certification or standards)
None

COURSE REQUIREMENTS

Required Reading and Textbook(s)

1. Data Analytics with Hadoop: An Introduction for Data Scientists, 1st edition, Benjamin Bengfort, Jenny Kim, ISBN: 9781491913703 (referred to as DAH from now onwards in the Syllabus).
2. Hadoop: The Definitive Guide, 4th Edition. Tom White. ISBN: 9781491901632 (referred to as HDG from now onwards in the Syllabus).

Optional (Reference) Textbook(s)

1. Hands-On Machine Learning with Scikit-Learn & TensorFlow, 1st edition, Aurelien Geron. ISBN: 9781491962299
2. Data Intensive Text Processing with MapReduce (Synthesis Lectures on Human Language technologies), Morgan and Claypool (pre-production manuscript freely available on GitHub at https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf)
3. The DataCenter as a Computer (Synthesis Lectures on Computer Architecture), Second edition, Morgan and Claypool (freely available from University of Michigan's website at http://web.eecs.umich.edu/~mosharaf/Readings/DC-Computer.pdf)

Note: A student of this institution is not under any obligation to purchase a textbook from a university-affiliated bookstore. The same textbook may also be available from an independent retailer, including an online retailer.

Reading Assignments: All assigned chapters along with assigned readings will be used for class and/or canvas discussions. Students are expected to study the assigned readings before each class session.

Supplementary Material: The course textbook will be supplemented with other materials that may include research papers readings, handouts, industry articles, videos and video lectures including TED talks, case studies, power point presentations etc. Students must know how to search for and download research papers / articles from the Web (more specifically from Google scholar (www.scholar.google.com)

COURSE REQUIREMENTS

Examinations: There will be two exams. First exam will be worth 15 points and second exam will be worth 25 points. The exams may have two parts: multiple choice questions (answered online via Canvas) and programming problems (executable code submitted via Canvas). For the programming part, you have to submit your executable Python programs along with screen shot of any output(s). You will earn partial credits for work demonstrating your efforts at attempting programming problems (i.e. programs with errors or incorrect output(s)).

The exams will all be open book / open notes and will be available via Canvas. You may also use your submitted home works but use of Internet not allowed. For the exams, you will be required to know not only the material from the textbook chapters, but also material from the class lectures such as power point slides and any supplementary/additional material provided as well. Both the exams will be in-class. You will get entire class duration of three hours to take the exams. Essentially, for the exams you must also know how to write and debug Python programming code independently. If you cannot take the exam(s) in class, inform me via email to plan to take during my office hours within 10 days of in-class exam(s).

Individual home works: There will be four individual home works. Students will get approximately two weeks to work on the individual homework assignments. Each homework will be worth 10 points and may cover one or more of the following topics (in no particular order):

- Hadoop Fundamentals (Ch. 1, 2 and 3 of DAH; Ch. 1, 2 and 3 of HDG)
- YARN (Ch. 2 of DAH and Ch. 4 of HDG)
- MapReduce (Ch. 6, 7 and if time permits Ch. 8 and 9)
- Data Processing with Spark, Hive and Pig (Spark covered in Ch. 4 of DAH and Ch. 19 of HDG, Hive in Ch. 6 of DAH and Ch. 17 of HDG and Pig covered in Ch. 8 of DAH and Ch. 16 of HDG)
- Data ingestion with Sqoop and Flume (Ch. 7 of HDG and Ch. 14, 15 of HDG)

- Data Storage with HBase (Ch. 6 of HDG and Ch. 20 of HDG)
- If time permits, following machine-learning and artificial intelligence topics will be covered: K-nearest neighbors, Support Vector Machines, Neural Networks and Deep Learning. Other topics of interest are Social Network Analysis and Data Center Architecture.

Home works turned in after due date are considered late. **2 points deducted for each day the homework is late.** Special circumstances need to be discussed with the instructor ahead of time when possible.

Semester wide team research project: There will be one semester long data analytics project using big data worth 20 points (15 points for the project report / research paper and 5 points for project proposal). Students can self-select into teams of 2 to 3 members using the Canvas team tool. Students must choose their teammates before the end of the third week. In case of students not able to form groups, instructor will assign students to random groups during the fourth week. The team must chose a freely available secondary data source from the Web (more on this during the second or third week of the class). There are many sources of secondary data available on the Web, you are free to explore and are not restricted to a specific type of data or application domain. Students are required to form a team and submit project proposal by the end of fourth week.

The aim of the project is to use one or more of the big data analytics techniques learned during the class. Students are not required to turn in the data or the programming code but must submit a written report in the form of a research paper (guidelines for the team project write up will be provided soon). **1 point will be deducted for each day the team project proposal is late.**

Projects turned in after due date are considered late. **5 points deducted for each day the project is late.** Special circumstances need to be discussed with the instructor ahead of time when possible. I expect each team member to participate on team projects. All team members will receive the same grade for the team projects. However, I reserve the right to make exceptions to that practice as circumstances such as performance imbalance or communication issues warrant. Poor individual contributions to the team project as noted by your team members will result in a poorer individual grade on the team project.

Grading Criteria Rubric and Conversion

Item	Quantity	Points	Total / Percent
Exam 1	**1**	**15**	**15**
Exam 2	**1**	**25**	**25**
Home works	**4**	**10**	**40**
Semester long team project	**1**	**20**	**20**
Total			**100**

Exams, quizzes, assignments and individual term project will receive a numeric score (0-100) each. These scores will be converted to points and totaled to ultimately be converted to letter grade of A, B, C, D, or F as shown in the example below:

	Percent earned by student	**Max points**	**Points to be added to the final grade**
Exam 1	**78%**	**15**	**11.7**
Exam 2	**85%**	**25**	**21.25**
Homework 1	**60%**	**10**	**6**
Homework 2	**75%**	**10**	**7.5**
Homework 3	**80%**	**10**	**8**
Homework 4	**90%**	**10**	**9**
Semester long team project report	**80%**	**20**	**16**
TOTAL		**100**	**79.45**

89.5 – 100 = A
79.5 – 89.49999 = B
69.5 – 79.49999 = C
59.5 – 69.49999 = D
Below 59.5 = F

Grades will not be curved (in the above example, the student with 79.45 will receive a C NOT a B).

Posting of Grades
All student grades will be posted on the Canvas Grade Book and students should monitor their grading status through this tool. Grades will be posted within 10 days after the due date.

COURSE OUTLINE AND CALENDAR

Complete Course Calendar (subjected to change)

Week	Date	Content	Readings	Assignments
1	01/13/2020	Course Welcome, Introduction, and overview of course objectives, and expectations. Introduction to Big Data Analytics, Data Product and Data intensive processing.	Read Syllabus in-depth. Chapter 1 of DAH and Chapter 1 of reference text 2	

2	**01/20/2020**	**Martin Luther King Jr. Day (University Closed)**		
3	01/27/2020	MapReduce	Chapter 2 of HDG and Chapter 2 of reference text 2	
01/29/2020		**Deadline to drop 16-week classes with no record**		
4	02/03/2020	Hadoop Architecture and HDFS	Chapter 2 of DAH and Chapter 3 of HDG	Project proposal available on Canvas.
5	02/10/2020	YARN	Chapter 4 of HDG	
6	02/17/2020	Hadoop Streaming	Chapter 3 of DAH	Project proposal due.
7	02/24/2020	Developing a MapReduce application	Chapter 6 of HDG	
8	03/02/2020	How MapReduce works	Chapter 7 of HDG	
9	**03/09/2020 – 03/12/2020**	**Spring Break (No classes)**		
10	**03/16/2020**	**Exam 1 (based on material covered so far).**		
11	03/23/2020	Spark	Chapter 4 of DAH and Chapter 19 of HDG	
12	03/30/2020	Hive	Chapter 6 of DAH and Chapter 17 of HDG	
13	04/06/2020	HBase	Chapter 6 of DAH and Chapter 20 of HDG	
14	04/13/2020	Data Ingestion with Sqoop	Chapter 7 of DAH, Chapters 14, 15 from HDG	
15	04/20/2020	Data Ingestion with Flume	Chapter 7 of DAH, Chapters 14, 15 from HDG	
15	04/20/2020	Pig	Chapter 8 of DAH and Chapter 16 of HDG	Team Term Project Due.

17	**05/04/2020**	**Exam 2 (based on material covered after exam 1).**

Important University Dates

Refer to University official academic calendar at https://www.tamuct.edu/registrar/academic-calendar.html

TECHNOLOGY REQUIREMENTS AND SUPPORT

I am working on the specific software tools and technologies here and will update the syllabus within the next few days.

Technology Requirements
This course will use the A&M-Central Texas Instructure Canvas learning management system. We strongly recommend the latest versions of Chrome or Firefox browsers. Canvas no longer supports any version of Internet Explorer.

Logon to A&M-Central Texas Canvas [https://tamuct.instructure.com/] or access Canvas through the TAMUCT Online link in myCT [https://tamuct.onecampus.com/]. You will log in through our Microsoft portal.

Username: Your MyCT email address. Password: Your MyCT password

Canvas Support
Use the Canvas Help link, located at the bottom of the left-hand menu, for issues with Canvas. You can select "Chat with Canvas Support," submit a support request through "Report a Problem," or call the Canvas support line: 1-844-757-0953.
For issues related to course content and requirements, contact your instructor.

Other Technology Support
For log-in problems, students should contact Help Desk Central
24 hours a day, 7 days a week
Email: helpdesk@tamu.edu
Phone: (254) 519-5466
Web Chat: [http://hdc.tamu.edu]
Please let the support technician know you are an A&M-Central Texas student.

UNIVERSITY RESOURCES, PROCEDURES, AND GUIDELINES

Drop Policy
If you discover that you need to drop this class, you must complete a Drop Request Form [https://www.tamuct.edu/registrar/docs/Drop_Request_Form.pdf].

Professors cannot drop students; this is always the responsibility of the student. The Registrar's Office will provide a deadline on the Academic Calendar for which the form must be completed, signed and returned. Once you return the signed form to the Registrar's Office, you must go into Warrior Web and confirm that you are no longer enrolled. If you still show as enrolled, FOLLOW-UP with the Registrar's Office immediately. You are to attend class until the procedure is complete to avoid penalty for absence. Should you miss the drop deadline or fail to follow the procedure, you will receive an F in the course, which may affect your financial aid and/or VA educational benefits.

Academic Integrity

Texas A&M University -Central Texas values the integrity of the academic enterprise and strives for the highest standards of academic conduct. A&M-Central Texas expects its students, faculty, and staff to support the adherence to high standards of personal and scholarly conduct to preserve the honor and integrity of the creative community. Academic integrity is defined as a commitment to honesty, trust, fairness, respect, and responsibility. Any deviation by students from this expectation may result in a failing grade for the assignment and potentially a failing grade for the course. Academic misconduct is any act that improperly affects a true and honest evaluation of a student's academic performance and includes, but is not limited to, working with others in an unauthorized manner, cheating on an examination or other academic work, plagiarism and improper citation of sources, using another student's work, collusion, and the abuse of resource materials. All academic misconduct concerns will be referred to the university's Office of Student Conduct. Ignorance of the university's standards and expectations is never an excuse to act with a lack of integrity. When in doubt on collaboration, citation, or any issue, please contact your instructor before taking a course of action.

For more information regarding the Student Conduct process, [https://www.tamuct.edu/student-affairs/student-conduct.html].
If you know of potential honor violations by other students, you may submit a report, [https://cm.maxient.com/reportingform.php?TAMUCentralTexas&layout_id=0].

Academic Accommodations

At Texas A&M University-Central Texas, we value an inclusive learning environment where every student has an equal chance to succeed and has the right to a barrier-free education. The Office of Access and Inclusion is responsible for ensuring that students with a disability receive equal access to the university's programs, services and activities. If you believe you have a disability requiring reasonable accommodations please contact the Office of Access and Inclusion, WH-212; or call (254) 501-5836. Any information you provide is private and confidential and will be treated as such.

For more information please visit our Access & Inclusion Canvas page (log-in required) [https://tamuct.instructure.com/courses/717]

Important information for Pregnant and/or Parenting Students
Texas A&M University-Central Texas supports students who are pregnant and/or parenting. In accordance with requirements of Title IX and related guidance from US Department of Education's Office of Civil Rights, the Dean of Student Affairs' Office can assist students who are pregnant and/or parenting in seeking accommodations related to pregnancy and/or parenting. Students should seek out assistance as early in the pregnancy as possible. For more information, please visit Student Affairs [https://www.tamuct.edu/student-affairs/index.html]. Students may also contact the institution's Title IX Coordinator. If you would like to read more about these requirements and guidelines online, please visit the website [http://www2.ed.gov/about/offices/list/ocr/docs/pregnancy.pdf].

Title IX of the Education Amendments Act of 1972 prohibits discrimination on the basis of sex and gender–including pregnancy, parenting, and all related conditions. A&M-Central Texas is able to provide flexible and individualized reasonable accommodation to pregnant and parenting students. All pregnant and parenting students should contact the Associate Dean in the Division of Student Affairs at (254) 501-5909 to seek out assistance. Students may also contact the University's Title IX Coordinator.

Tutoring
Tutoring is available to all A&M-Central Texas students, both on-campus and online. Subjects tutored on campus include Accounting, Advanced Math, Biology, Finance, Statistics, Mathematics, and Study Skills. Tutors are available at the Tutoring Center in Warrior Hall, Suite 111. Tutor.com tutoring **will not offer writing support** beginning August 1, 2019, but will continue to offer other tutoring support.

If you have a question regarding tutor schedules, need to schedule a tutoring session, are interested in becoming a tutor, or have any other question, contact Academic Support Programs at (254) 501-5836, visit the Office of Student Success at 212F Warrior Hall, or by emailing studentsuccess@tamuct.edu .

Chat live with a tutor 24/7 for almost any subject from on your computer! Tutor.com is an online tutoring platform that enables A&M-Central Texas students to log in and receive online tutoring support at no additional cost. This tool provides tutoring in over 40 subject areas. Access Tutor.com through Canvas.

University Writing Center
Located in Warrior Hall 416, the University Writing Center (UWC) at Texas A&M University–Central Texas (TAMUCT) is a free workspace open to all TAMUCT students from 10:00 a.m.-5:00 p.m. Monday thru Thursday with satellite hours in the University Library Monday thru Thursday from 6:00-9:00 p.m. This semester, the UWC is also offering online only hours from 12:00-3:00 p.m. on Saturdays.

Tutors are prepared to help writers of all levels and abilities at any stage of the writing process.

While tutors will not write, edit, or grade papers, they will assist students in developing more effective composing practices. By providing a practice audience for students' ideas and writing, our tutors highlight the ways in which they read and interpret students' texts, offering guidance and support throughout the various stages of the writing process. In addition, students may work independently in the UWC by checking out a laptop that runs the Microsoft Office suite and connects to WIFI, or by consulting our resources on writing, including all of the relevant style guides. Whether you need help brainstorming ideas, organizing an essay, proofreading, understanding proper citation practices, or just want a quiet place to work, the UWC is here to help!

Students may arrange a one-to-one session with a trained and experienced writing tutor by visiting the UWC during normal operating hours (both half-hour and hour sessions are available) or by making an appointment via WCOnline [https://tamuct.mywconline.com/]. In addition, you can email Dr. Bruce Bowles Jr. at bruce.bowles@tamuct.edu if you have any questions about the UWC and/or need any assistance with scheduling.

University Library

The University Library provides many services in support of research across campus and at a distance. We offer over 200 electronic databases containing approximately 250,000 eBooks and 82,000 journals, in addition to the 85,000 items in our print collection, which can be mailed to students who live more than 50 miles from campus. Research guides for each subject taught at A&M-Central Texas are available through our website to help students navigate these resources. On campus, the library offers technology including cameras, laptops, microphones, webcams, and digital sound recorders.

Research assistance from a librarian is also available 24 hours a day through our online chat service, and at the reference desk when the library is open. Research sessions can be scheduled for more comprehensive assistance, and may take place on Skype or in-person at the library. Assistance may cover many topics, including how to find articles in peer-reviewed journals, how to cite resources, and how to piece together research for written assignments.

Our 27,000-square-foot facility on the A&M-Central Texas main campus includes student lounges, private study rooms, group work spaces, computer labs, family areas suitable for all ages, and many other features. Services such as interlibrary loan, TexShare, binding, and laminating are available. The library frequently offers workshops, tours, readings, and other events. For more information, please visit our Library website [http://tamuct.libguides.com/index].

OPTIONAL POLICY STATEMENTS

A Note about Sexual Violence at A&M-Central Texas

Sexual violence is a serious safety, social justice, and public health issue. The university offers support for anyone struggling with these issues. University faculty are mandated reporters, so if

someone discloses that they were sexually assaulted (or a victim of Domestic/Dating Violence or Stalking) while a student at TAMUCT, faculty members are required to inform the Title IX Office. If you want to discuss any of these issues confidentially, you can do so through Student Counseling (254-501-5955) located on the second floor of Warrior Hall (207L).

Sexual violence can occur on our campus because predators often feel emboldened, and victims often feel silenced or shamed. It is incumbent on ALL of us to find ways to actively create environments that tell predators we don't agree with their behaviors and tell survivors we will support them. Your actions matter. Don't be a bystander; be an agent of change. For additional information on campus policy and resources visit the Title IX webpage [https://www.tamuct.edu/departments/compliance/titleix.php].

Behavioral Intervention

Texas A&M University-Central Texas cares about the safety, health, and well-being of its students, faculty, staff, and community. If you are aware of individuals for whom you have a concern, who are exhibiting concerning behaviors, or individuals causing a significant disruption to our community, please make a referral to the Behavioral Intervention Team. Referring your concern shows you care. You can complete the referral online [https://cm.maxient.com/reportingform.php?TAMUCentralTexas&layout_id=2].

Anonymous referrals are accepted. Please see the Behavioral Intervention Team website for more information [https://www.tamuct.edu/student-affairs/bat.html]. If a person's behavior poses an imminent threat to you or another, contact 911 or A&M-Central Texas University Police at 254-501-5800.

INSTRUCTOR POLICIES

Policies related to absence, grading, etc.

- You are responsible for all class material presented during an absence.
- Home works must be submitted through Canvas and due on the mid night (11:59 PM) of specified due date/time. No email submissions will be accepted except under unforeseen circumstances.
- Late penalties will be applied to items submitted after due dates as per the per day late submit penalty guidelines stated above.
- If you cannot take in-class exam, you must inform me at least one week beforehand and must plan to take the exam during my office hours within 10 days of exam due dates.

My personal statement

- You will receive feedback in the form of graded assignments within 14 days after the due date.

- I want you to read the feedback that I provide to you (your personal grading notes and Canvas emails).
- I am almost always available via email and typically respond within 24 hours except on weekends.
- I prefer email to phone conversations.
- I reserve the right to modify the course syllabus during the semester for the benefit of the students.
- I reserve the right to supplement materials presented in the text with additional course material that may help the students to understand the topic better.
- I reserve the right to modify grading policy rubrics. Any change to grading rubrics will be applied to current and possible future assignments.

Copyright Notice

Students should assume that all course material is copyrighted by the respective author(s). Reproduction of course material is prohibited without consent by the author and/or course instructor. Violation of copyright is against the law and Texas A&M University-Central Texas' Code of Academic Honesty. All alleged violations will be reported to the Office of Student Conduct.