MSBA 6330: Big Data Analytics

Spring 2021 (Remote Instruction)

Instructor	**Dr. De Liu**
Email	deliu@umn.edu
Contact	Slack (http://msba6330spring2021.slack.com) or email
Zoom Meetings	Sec 1: Mondays 8:30-9:45am Sec 2: Mondays 1:00-2:15pm
Office Hours/Live Support	Sec 1: Mondays 9:45-11:00am Sec 2: Mondays 2:15-3:30pm
Course Server	z.umn.edu/csom-atk
Teaching Assistant	TBA

1. INTRODUCTION

This course provides an introduction to big data concepts, environments, processes, and tools from the perspective of business analysts. It uses a hands-on, learning-by-doing approach to develop an understanding of major elements of the big data ecosystems such as Hadoop, MapReduce, Scoop, Hive, and Spark, with a focus on their value propositions and how they can be used to solve data science problems at scale. Through this course, you will also gain an understanding of major concepts in cloud computing, and become familiar with the big data offerings of a leading cloud computing platform -- Amazon Web Services (AWS).

Because the topics covered in this course are nascent and evolving quickly, students are expected to engage in active learning through troubleshooting, experimentation, proactive researching, and sharing (via Slack).

2. LEARNING OBJECTIVES

This course has two main categories of objectives.

1. Develop an understanding of the big data ecosystem, the kinds of problems it aims to solve, the characteristics of big data technologies, and their key advantages and disadvantages. Such an understanding allows you to recognize the opportunities and challenges associated with big data, from both business and technological perspectives, and to guide businesses in adopting and using big data technologies.
2. Develop core competencies in using a variety of essentially big data tools (such as Scoop, Hive, Spark, and Cloud computing) and processes to solve data science problems at scale.

3. PREREQUISITES

This course is built on a sequence of courses offered in the MSBA program including Python programming (MSBA 6310) and database management (MSBA 6320, which is offered in parallel). As prerequisites for this course, students are expected to:

- Have a basic knowledge of computing systems (operating system, file systems, www, etc).

- Have working knowledge of SQL and relational databases.
    - Those who lack knowledge of SQL and relational databases are advised to learn from many available resources, such as the SQL essential training on Lynda.com.
- Have working knowledge of Python.
    - Parts of course will leverage Python scripts, though extensive Python programming is not the focus of this course. If you have not been exposed to these programming languages before, there are many free introductory online courses on Python (e.g. the Python course from Codecademy and Introduction for Python for Data Science from DataCamp).

The following are also helpful to know:

- Basic knowledge of computers (operating systems, file systems, etc).
- Some familiarity with the command-line interface (CLI) and Linux operating systems.
- Some understanding of machine learning methods, e.g., linear regression, logistic regression, and k-means clustering.

4. TEXTBOOK

We do not have a required text for this course. Instead, we rely on our lecture slides, lab notes, additional reading materials (articles & book chapters).

5. SOFTWARE & HARDWARE

Most of the software used for this course is pre-installed on servers that we provide in Amazon's AWS cloud (our first time to use this highly efficient, scalable infrastructure). As a backup, we will also provide an MSBA virtual desktop that runs a Virtual Machine (VM) that has the necessary software and data.

- **Hadoop/Hive**: We use the Cloudera Virtual Machine installed on course provision servers. The servers (depending on the instance) pre-installs big data tools such as Hadoop, MapReduce, Sqoop, Hive, Spark, and some of the datasets used for this course.
- **Spark**: We will use the DataBricks community edition (https://databricks.com/try-databricks) for the Spark portion of the course work. The DataBricks community edition provides a free service for a limited (6GB RAM) virtual machine hosted on AWS. Besides, you may have two other choices, including using the course provisioned server or installing Apache Spark on your computer from https://github.com/paulovn/ml-vm-notebook.
- **Amazon Web Services (AWS)**. (Tentative) Through the amazon academic partner program, you can obtain a **$100 certificate** for use with AWS. This will be needed for the cloud computing part of the labs and assignment (but not for the exam), and possibly for your team project. More details to come. Additionally, you may also try Microsoft Azure/Google Cloud for yourself.

6. GRADING

Your grade will be based on the following activities (tentative):

Activity	Grade %
Team Project	12%
Homework Assignments (7)	21%, 3% each
Self-assessment quizzes	10%
Lab completion	5%
Participation	2%
Midterm Exam	25%
Final Exam	25%

A final letter grade is given according to the following scale.

%	Grade	%	Grade
93.0 or above	A	77.0 - 79.9	C+
90.0 - 92.9	A-	73.0 - 76.9	C
87.0 - 89.9	B+	70.0 - 72.9	C-
83.0 - 86.9	B	60.0 - 69.9	D
80.0 - 82.9	B-	Below 60.0	F

Though it rarely happens, this scale may be adjusted to ensure compliance with MSBA grading policies, which requires the class median GPA of 3.3 +/- 0.2.

6.1 Participation credits

You can earn participation credits by completing surveys, Flipgrid-based participation assignments (e.g. submit a video self introduction video), Zoom participation records, etc. Instructor reserves the rights adjust participation credits based on subjective evaluation of student participation in Zoom classes and on asynchronous channels (e.g. Slack/Flipgrid).

7. HOMEWORK ASSIGNMENTS

7.1 Homework Academic Integrity

All homework assignments are INDIVIDUAL assignments. Plagiarism is prohibited and will be penalized. You may still discuss the homework problems with fellow students. **However, no documents or fragments of homework should be shared between students, physically or digitally (a rule of thumb is that you "typed it with your own hands".** Sharing homework papers is strictly forbidden and is considered plagiarism, and both the one who shared and the one who received can face a penalty, including zero grade for the assignment, letter grade downgrade, and an "F" for the course. All scholarly dishonesty incidence will be reported to the University's Office of Community Standards. In case of difficulty with homework assignments, please reach out to the instructor as early as possible.

Citing someone else's work (including those from the Internet) without accrediting the source properly is also considered plagiarism. **You must precisely point out which part of your answer refer a source and include a reference to the source.** Refer to our submission guidelines for more homework related instructions.

7.2 Late Policy - Flex Days

You have 5 flex days that you may use at your discretion to defer the due date of a homework assignment, self-assessment quiz, or a team project (late lab completion is not accepted). You must use your flex days in the whole day (24 hours) increments. You can defer a single assignment by **a maximum of 2 days**.

You do not need to notify us if you want to use your flex days! Submit your homework when ready and the proper number of flex days will be deducted. Flex days may only be used for homework. You may not use flex days to defer the due date of self-assessment quizzes, active quizzes, your team project, or exams.

Flex days are intended to be used for unforeseen circumstances such as technical difficulties, family emergencies, and personal illness. **After your flex days have been used for the semester, late work will not be accepted and will earn no credit (0 points)**.

7.3 Miscellaneous

If you have not used Canvas assignments before, please watch the tutorial videos here.

You should always save a copy of your submitted work, including homework, team projects, and the exams. Store the copy in a secure location (such as your Google Drive).

Homework assignments are usually graded within one week of the due date. Homework is graded on a scale of 0 to 10. If you get 10, you will receive the full 3% credit. If you get 9, you will receive the 2.7% credit, and so on. The instructor reserves the right to give extra credits for exceptional work.

8. TEAM PROJECT

The team project is evaluated by your peers (50%) and by the instructor (50%). The team project is generally evaluated along the lines of value, presentation, rigor, and creativity. While in general every team member receives the same credit, those who lack contribution (as reflected by peer evaluation within the group) may receive partial to zero credit. For detailed guidelines about team projects, please see the team assignment handout.

9. Instructional Technologies Used in the Course

Slack - Discussion forum

We use the Slack to:

- general purpose interaction/discuss, voice calls (if needed)
- discuss administrative/technical issues
- discuss issues related to labs/homework assignments
    - **If the issue applies to you specifically**, please use a private channel to share with the instructor
    - **If the issue cloud impact many students**, you can share it in public channels (we have separate areas for each Homework)

- students and faculty to share course Internet resources (news, articles, etc)
    - you should include a link to the source and your commentary/expert of the article.

Participation can take many forms; Posting questions, posting replies, and reacting to others' posts are all desirable forms of participation. You're encouraged to give frequent feedback to others (likes and comments) and stay constructive.

FlipGrid - Video Sharing

- We use this for sharing short videos (e.g., self introduction).
- Some participation assignments may be Flipgrid-based.

It is available as a Canvas plug in (accessible from the menu).

VoiceThread - Voice Presentation

We use voice thread for recording a team presentation collaboratively. More details will be forthcoming.

Zoom

We use zoom for synchronous meetings & office hours.

10. OTHER RESOURCES

Getting Help

Slack is the best way to reach the instructor. If you have a technical problem, please attach the scripts, screenshot, error message if any. Appointments are also welcome. For general technical problems concerning user accounts, hardware, or software, you may contact the OIT Help Desk.

Academic Integrity and Scholastic Dishonesty

Academic integrity is essential to a positive teaching and learning environment. All students enrolled in University courses are expected to complete coursework responsibilities with fairness and honesty. Failure to do so by seeking an unfair advantage over others or misrepresenting someone else's work as your own, can result in disciplinary action. The University defines scholastic dishonesty as follows:

Scholastic Dishonesty: Scholastic dishonesty means plagiarizing; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic achievement; acting alone or in cooperation with another to falsify records or to obtain dishonestly grades, honors, awards, or professional endorsement; altering forging, or misusing a University academic record; or fabricating or falsifying data, research procedures, or data analysis.

Within this course, a student responsible for scholastic dishonesty can be assigned a penalty up to and including an F or N for the course. If you have any questions regarding the expectations for a specific assignment or exam, please ask.

Students with Disabilities

The University of Minnesota is committed to providing all students equal access to learning opportunities. Disability Services is the campus office that works with students who have disabilities to provide and/or arrange reasonable accommodations. Students registered with Disability Services, who have a letter requesting accommodations, are encouraged to contact the instructor early in the semester. Students who have, or think they may have, a disability (e.g. psychiatric, attentional, learning, vision, hearing, physical, or systemic), are invited to contact Disability Services for a confidential discussion at 612 626 1333 (V/TTY) or at ds@umn.edu. Additional information is available on the DS website http://ds.umn.edu.

Schedule (tentative and subject to change)

	Date	Subject	Assignments
1	Jan 25	Introduction, Linux Shell, Big Data	
2	Feb 1	Hadoop MapReduce & Ecosystem	**Due: Homework 1**
3	Feb 8	Introduction to Hive	**Due: Homework 2**
4	Feb 15	Hive Optimization & Textual Data Handling	**Due: Homework 3**
5	Feb 22	Introduction to Apache Spark	
6	Mar 1	Spark SQL	
7	Mar 8	**Midterm Exam**	
8	Mar 15	Spark SQL & Spark ML	**Due: Homework 4**
9	Mar 22	Spark ML	**Due: Homework 5**
10	Mar 29	Spark Streaming	**Due: Homework 6**
11	Apr 5	**Spring Break**	
12	Apr 12	Cloud Computing & AWS	**Due: Homework 7**
13	Apr 19	NoSQL	
14	Apr 26	TBD	
15	May 3	**Trend Marketplace**	
	May 8	**Final exam**	

Last revision: January 6, 2021