BUAN 6346.502
Course Syllabus – Fall 2023

Naveen Jindal School of Management
University of Texas at Dallas

Course Information

Course Prefix, Number, Section	BUAN 6346.502
Course Title	Big Data
Term	Fall 2023

Professor Contact Information

Professor	Farzad Kamalzadeh
Email Address	farzad@utdallas.edu
Office Hours	By Appointment
Teaching Assistant (TA)	Parveen Kumar
TA Email Address	Parveen.Kumar@UTDallas.edu

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

Prerequisite: BUAN 6320 or MIS 6320 or MIS 6326.

Please refer to the JSOM's Cirruculum for pre-requisites, co-requisites and/or restrictions.

Course Description

This course covers topics including (1) understanding of big data concepts, (2) manipulation of big data with popular tools, and (3) distributed analytics programming. It is a project-oriented course; thus, students will be required to establish a big data environment, perform various analytics, and report findings in their projects. Though concepts and theoretical aspects are addressed, more emphasis will be on actual operations of a big data system. Students will not only manipulate the basic big data software/system, but also use various dedicated big-data tools and perform distributed analytics programming with popular computer languages.

Student Learning Objectives/Outcomes

1. **Comprehend Big Data Fundamentals:** Understand the characteristics and significance of Big Data in modern contexts.
2. **Master Big Data Technologies:** Gain proficiency in utilizing tools within Hadoop Ecosystem including HDFS, Hive, Pig, HBase and Spark.
3. **Process Big Data in Hadoop:** Become proficient in end-to-end data processing in a Hadoop environment including data import, data processing, data analysis, data visualization.
4. **Apply Analytics and Machine Learning:** Explore machine learning integration, analyze large datasets, and communicate insights effectively using Spark and its capacilities.

Required Textbooks and Materials

Lecture Notes as provided in powerpoint slides

Suggested Course Materials

Hadoop Ecosystem Tools

- Big Data Now: Current Perspectives from O'Reilly Radar by *O'Reilly Radar Team* · 2011, O'Reilly Media, ISBN: 9781449315214
- Hadoop: The Definitive Guide by *Tom White* · 2010, O'Reilly Media, ISBN: 9781449396893
- Hadoop Application Architectures: Designing Real-World Big Data Applications by *Mark Grover, Ted Malaska, Jonathan Seidman, Gwen Shapira* · 2015, O'Reilly Media, ISBN: 9781491900055
- Apache Sqoop Cookbook: Unlocking Hadoop for Your Relational Database by *Kathleen Ting, Jarek Jarcec Cecho* · 2013, O'Reilly Media, ISBN: 9781449364601
- Architecting HBase Applications: A Guidebook for Successful Development and Design by *Jean-Marc Spaggiari, Kevin O'Dell* · 2016, O'Reilly Media, ISBN: 9781491916117
- Getting Started with Impala Interactive SQL for Apache Hadoop by *John Russell* · 2014, O'Reilly Media, ISBN: 9781491905722
- HBase: The Definitive Guide by *Lars George* · 2011, O'Reilly Media, ISBN: 9781449396107
- Programming Pig by *Alan Gates* · 2011, O'Reilly Media, ISBN: 9781449302641
- Using Flume: Flexible, Scalable, and Reliable Data Streaming by *Hari Shreedharan* · 2014, O'Reilly Media, ISBN: 9781491905333
- Spark: The Definitive Guide: Big Data Processing Made Simple by *Bill Chambers, Matei Zaharia* · 2018, O'Reilly Media, ISBN: 9781491912300
- Learning Spark: Lightning-Fast Big Data Analysis by *Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia* · 2015, O'Reilly Media, ISBN: 9781449359065

Software Tools

Below is a list of software tools required to do the assignments for this course. Instructions on how to find, install and use these tools will be later provided in another document.

We will be using tools from Hadoop Ecosystem including HDFS, Sqoop, Hive, Impala, Flume, Pig, HBase, and Spark.

Virtual Machine Emulator

You need a Virtual Machine Emulator Software such as VMware Workstation Player or Oracle VM VirtualBox (or any other emulators you prefer) in order to virtually run a Linux machine.

VM Cloudera Sandbox Image

You also need the Cloudera Sandbox image file which is an image of a Linux Virtual machine with Cloudera Distriibution of Hadoop already installed on it. This image has all the tools you need to do your homework assignments and projects. The file will be provided later.

Assignments & Grading

Assignments

There will be a total of 10 individual assignments. Each assignment will be posted on the Canvas. Assignments must be submitted before the due date. Late submission of assignments will be penalized by the reduction of the grade by 25% for each day (Please note that a couple of minutes up to a couple of hours of late submission is not considered late submission as long as it is not submitted after 8am the next day of deadline.)

- Assignments are individual work and should not be done in groups.
- Details on how to submit the assignments will be provided in the class.

Projects

There will be a total of 2 projects. Students are required to do these projects in groups of 5 members. The details of these projects will be provided in the class.

Students are required to form their groups by the time the first project is announced. Students should enroll in groups defined in eLearning. In case of any questions or problems with grous in eLearning students can send an email to TA.

Projects follow the same late submission policy as assignments.

Grading Policy

Grades will be calculated according to the following weights:

Item	Individual Weight	Total Weight
Assignments (10)	4%	40%
Projects (2)	P1: 30%, P2: 30%	60%
Total		100%

* Bonus points: 3 pts – for those who submit in-class assignments during the class.

** Peer-Review: for each project, team mates will evaluate each other anonymously with “Cooperated/Did not cooperate”. If the majority agrees you did not cooperate, you will lose 20% of each project.

Total weighted score then will be mapped to the letter grade based on the following table (subject to change):

Weighted Score	Letter Grade
[97-100]	A
[93-97)	A-
[90-93)	B+
[85-90)	B
[80-85)	B-
[70-80)	C+
[60-70)	C
[50-60)	C-
[0-50)	F

Academic Calendar

Please find below the academic calendar for this course and the due dates for the assignments and projects.

Session /Week	Date	Topics	Assignment/ Project	Due Date / Notes
1	8/24	Introduction to Big Data	-	-
2	8/31	Big Data Infrastructure	-	-
3	9/7	Introduction to Hadoop Ecosystem	Assignment 1	Next Week 11:59pm
4	9/14	Hadoop Storage Unit (HDFS)	Assignment 2 **Project 1**	**10/19** 11:59pm
5	9/21	Hadoop Processing Unit (MapReduce and YARN)	Assignment 3	Next Week 11:59pm
6	9/28	Sqoop and Flume	Assignment 4	Next Week 11:59pm
7	10/5	Sqoop and Flume	Assignment 5	Next Week 11:59pm
8	10/12	Hive and Impala	Assignment 6	Next Week 11:59pm
9	10/19	Hive and Impala	Assignment 7	Next Week 11:59pm
10	10/26	Pig	Assignment 8 **Project 2**	**12/10** 11:59pm
11	11/2	Pig	Assignment 9	Next Week 11:59pm
12	11/9	HBase	Assignment 10	Next Week 11:59pm
13	11/16	HBase	-	Next Week 11:59pm
-	11/23	**Fall Break – No Class**		
14	11/30	Spark	-	-
15	12/7	Spark	-	-
-	12/10	**Project 3 Due Date**		
-	12/20	**Final Grading**		

Course & Instructor Policies

Attendence Policy

Attendence is not required but highly encouraged. Students who submit all in-class assignments will get 3 additional points.

Syllabus Policy

The descriptions and timelines contained in this syllabus are subject to change at the discretion of the Professor.

Class Materials

The instructor may provide class materials that will be made available to all students registered for this class as they are intended to supplement the classroom experience. These materials may be downloaded during the course, however, these materials are for registered students' use only. Classroom materials may not be reproduced or shared with those not in class, or uploaded to other online environments except to implement an approved Office of Student AccessAbility accommodation. Failure to comply with these University requirements is a violation of the Student Code of Conduct.

Classroom Conduct Requirements Related to Public Health Measures

UT Dallas will follow the public health and safety guidelines put forth by the Centers for Disease Control and Prevention (CDC), the Texas Department of State Health Services (DSHS), and local public health agencies that are in effect at that time during the Spring 2022 semester to the extent allowed by state governance. Texas Governor Greg Abbott's Executive Order GA-38 prohibits us from mandating vaccines and face coverings for UT Dallas employees, students, and members of the public on campus. However, we strongly encourage all Comets to get vaccinated and wear face coverings as recommended by the CDC. Check the Comets United: Latest Updates webpage for the latest guidance on the University's public health measures. Comets are expected to carry out Student Safety protocols in adherence to the Comet Commitment. Unvaccinated Comets will be expected to complete the Required Daily Health Screening. Those students who do not comply will be referred to the Office of Community Standards and Conduct for disciplinary action under the Student Code of Conduct – UTSP5003.

Class Attendance

The University's attendance policy requirement is that individual faculty set their course attendance requirements. Regular and punctual class attendance is expected. Students who fail to attend class regularly are inviting scholastic difficulty. In some courses, instructors may have special attendance requirements; these should be made known to students during the first week of classes. Faculty have the discretion to set an attendance policy for their in-person meetings, but the absences due to COVID-19 cannot be counted against a quarantined student.

Class Participation

Regular class participation is expected. Students who fail to participate in class regularly are inviting scholastic difficulty. A portion of the grade for this course is directly tied to your participation in this class. It also includes engaging in group or other activities during

class that solicit your feedback on homework assignments, readings, or materials covered in the lectures (and/or labs). Class participation is documented by faculty. Successful participation is defined as consistently adhering to University requirements, as presented in this syllabus. Failure to comply with these University requirements is a violation of the Student Code of Conduct.

Class Recordings

Students are expected to follow appropriate University policies and maintain the security of passwords used to access recorded lectures. Unless the Office of Student AccessAbility has approved the student to record the instruction, students are expressly prohibited from recording any part of this course. Recordings may not be published, reproduced, or shared with those not in the class, or uploaded to other online environments except to implement an approved Office of Student AccessAbility accommodation. Failure to comply with these University requirements is a violation of the Student Code of Conduct.

NOTE: if the instructor records any part of the course, then the instructor will need to add the following syllabus statement:

The instructor may record meetings of this course. These recordings will be made available to all students registered for this class if the intent is to supplement the classroom experience. If the instructor or a UTD school/department/office plans any other uses for the recordings, consent of the students identifiable in the recordings is required prior to such use unless an exception is allowed by law.

Off-campus Instruction and Course Activities

N/A

Comet Creed

This creed was voted on by the UT Dallas student body in 2014. It is a standard that Comets choose to live by and encourage others to do the same:

"As a Comet, I pledge honesty, integrity, and service in all that I do."

Academic Support Resources

The information contained in the following link lists the University's academic support resources for all students.

Please see http://go.utdallas.edu/academic-support-resources.

UT Dallas Syllabus Policies and Procedures

The information contained in the following link constitutes the University's policies and procedures segment of the course syllabus. Please review the catalog sections regarding the credit/no credit or pass/fail grading option and withdrawal from class.

Please go to http://go.utdallas.edu/syllabus-policies for these policies.