Raymond A. Mason School of Business
WILLIAM & MARY

BUAD 5722 BIG DATA

Course Syllabus, Spring 2022

Last revised 1/21/2022

Instructor:	Arturo A Castellanos Bueso
Office:	Miller Hall 3067D
Phone:	757-221-1706 (office)
E-mail:	Instructor: aacastellanosb@wm.edu TA: Ahmed Mustapha aomustapha@email.wm.edu (Please put BUAD5722 in the subject line)
Class Hours	Mondays, Wednesdays & (some) Fridays See schedule for section time slot
Location	Miller 1027
Midterm Exam	TBA on Blackboard
Office Hours	Ahmed (TA): Mondays 9-10:00 am Professor Castellanos: Tuesdays 3-4:00 p.m (send email to notify instructor)

COURSE OVERVIEW & OBJECTIVES

This course focuses on understanding the concepts of Big Data, why it is being used, and examples of how it is driving business decisions. The data storage and retrieval techniques that have served the information processing industry for decades have proven inadequate in the face of the massive collections of data presently being created by organizations and user-generated content, Internet, and the so-called "Internet of Things." Businesses are requiring a new set of technologies that are specifically designed to deal with these huge data sets. In this course, we will introduce methods to process large-scale data sets.

COURSE MATERIALS

Recommended Textbooks:

- Big Data Science & Analytics by Arshdeep Bahga , Vijay Madisetti
- Cloud computing Solutions Architect by Arshdeep Bahga , Vijay Madisetti

Software:

- Python / Anaconda
- We will be using Amazon Web Services (AWS). Some of the services do require a fee or payment for usage (i.e., < $1 per hour of usage).
- Cloudera Quickstart VM
- Docker
- Datacamp

OFFICE HOURS AND COMMUNICATIONS

Some of the scheduled class times may be used for holding office hours to support the comprehension and application of the material. Such office hours will be either general -open to everyone-, or appointment based depending on the needs of the class or expectations for the corresponding week. Office hours during the class times will be announced before the start of the week and added to Calendly on a weekly basis.

The best way to reach me is via email. I strive to answer all emails within 24-48 hours of receipt during the week and I hope to do the same (with no guarantees) on the weekend. Please plan accordingly and reach out early with any questions or concerns you have. If you send an email close to a deadline, there is a chance I will not reply on time so plan ahead.

COURSE APPROACH & ATTENDANCE & PARTICIPATION

Students are expected to attend lectures. You are allowed 2 unexcused absences throughout the semester. These absences could be for anything including but not limited to personal, medical, professional and job-related reasons. You don't need to ask permission or explain. Every subsequent unexcused absence (regardless of cause) will diminish your attendance grade. Only a note from the Dean of Students or the Athletic Director may excuse an absence.

ASSESSMENT AND EVALUATION METHODS

Assignments:

There will be multiple individual assignments throughout the semester. Assignments will be posted on Blackboard and will be submitted via Blackboard.

It is important that you read assignment descriptions/instructions carefully, and submit your work by the due date/time, using the format instructed. Failure to do so may result in loss of points.

You should take assignment deadlines seriously and plan in advance to allocate sufficient time to meet deadlines. For late submissions the penalty will be a 20% reduction in the grade for that assignment for each day that the assignment is late. Assignments are not accepted 3 days after the due date.

Midterm Exam:

There will be a midterm exam mid-semester. In case classes are moved to an online format, the exam may be proctored by a remote proctoring system. Details for the midterm exam and remote proctoring will be provided.

COURSE GRADING

Final course grade will be computed as follows:

- Assignments : 35%
- Midterm Exam : 30%
- Group project : 25%
- Attendance & Participation : 10%

A	94-100	B	83-86.99	C	73-76.99
A-	90-93.99	B-	80-82.99	C-	70-72.99
B+	87-89.99	C+	77-79.99	F	0-69.99

HONOR CODE & ACADEMIC INTEGRITY

"As a member of the William and Mary community, I pledge on my honor not to lie, cheat, or steal, either in my academic or personal life. I understand that such acts violate the Honor Code and undermine the community of trust, of which we are all stewards."

Academic integrity is an integral component of the College of William and Mary learning experience and any breach of this integrity is very serious and not in keeping with the overall intellectual and ethical foundations of our University. Students are expected to adhere to the College of William and Mary Honor Code and to the general principles of academic honesty. These principles include and incorporate the concept of respect for the intellectual property of others, the expectation that assignments will be submitted according to guidelines specified by the instructor, and that plagiarism of any type is unacceptable.

Notice of Copyright

All course materials, including the syllabus, lectures, presentations, recordings, quizzes, assessments, tests, exams, outlines, assignments, electronic files, and similar materials, for this course are protected by copyright and are the sole property of the course instructor. You may use these materials for your personal, non-commercial educational use. You may not, nor may you knowingly allow others, to reproduce or distribute any course materials publicly without the express written consent of the instructor. This includes providing materials to commercial course material suppliers such as CourseHero and other similar services. To do so is both a copyright violation and a violation of W&M's Honor Code.

Mason School Assignment Codes

Within the Mason School of Business we have developed an Assignment Code and every assignment given in this course will be identified by a code letter. If there are questions related to how an assignment is to be carried out, it is essential that you ask your professor for clarification.

The code is as follows:

CATEGORY A: This is an individual assignment. You may not receive help from anyone on this assignment. It must be 100% your own work. All questions concerning this assignment should be addressed to your professor. It is an honor code offense to give or receive assistance on this assignment.

CATEGORY B: This is a group assignment. Your group may not receive help from anyone outside your group. While your group may choose to delegate the work among the group members, everyone in the group is expected to be prepared to discuss the entire assignment in class. All questions concerning this assignment should be addressed to your professor. It is an honor code offense to give help to other groups and individuals or receive assistance from other groups and individuals.

CATEGORY C: This is an individual assignment. You may work with others or receive help from a tutor on this assignment. You must, however, turn in your own paper. You may not divide the work with others or copy another student's paper; it is an honor code offense to do so.

CATEGORY D: This is a group assignment. You may share information, discuss general concepts and approaches to the assignment with other groups. Everyone in the group should be prepared to discuss the entire assignment in class. Each group must turn in their own work. You may not copy another group's work; it is an honor code offense to do so.

CATEGORY E: This is a timed assignment. You are given a specific length of time within which the work must be completed. It is an honor code offense to violate this time restriction unless you have received permission from your professor.

A note on Category C Assignments: You may work with your classmates on the individual Category C assignments, but all the work that you turn in must be your own. In other words, two people cannot work together to generate one version of the homework solution and then simply turn in two copies of that one work product.

ADA ACCOMMODATION

William & Mary accommodates students with disabilities in accordance with federal laws and university policy. Any student who feels s/he may need an accommodation based on the impact of a learning, psychiatric, physical, or chronic health diagnosis should contact Student

Accessibility Services staff at 757-221-2509 or at sas@wm.edu to determine if accommodations are warranted and to obtain an official letter of accommodation. For more information, please see www.wm.edu/sas.

DIVERSITY & INCLUSION STATEMENT

William and Mary welcomes students from around the country and around the world, and their unique perspectives enrich our learning community. It is our collective responsibility to create and foster an environment that is inclusive and respectful for all. To do this we must demonstrate:

- Respect and responsibility for self and others
- A spirit of generosity
- A life dedicated to inquisitive learning and development
- An acknowledgement that an individual's own words, actions, and relationships show a commitment to these values

ADDITIONAL POLICIES AND STATEMENTs

Class Recordings Made by Instructor

Meetings of this course will be recorded. Recordings will be available only to students registered for this class (as required by FERPA). This is intended to supplement the classroom experience. Students are expected to follow appropriate university policies and maintain the security of passwords used to access recorded lectures. Recordings may not be reproduced, shared with those not in the class, or uploaded to other online environments; violations may be subject to disciplinary action. If the instructor or a William & Mary office plan any other uses for the recordings, beyond this class, students identifiable in the recordings will be notified to request consent prior to such use.

Recording Class Sessions by Students

Recordings of the synchronous class sessions provided by the instructor are protected by both copyright and the Family Education Rights and Privacy Act (FERPA) and may not be shared or redistributed to anyone at any time now or in the future. To do so is both a violation of law and of W&M's Honor Code.

Zoom Protocol

Before connecting with a Zoom session, log into Zoom either on the W&M Zoom website (cwm.zoom.us) or using the Zoom desktop app. This permits the verification of approved Zoom participation via your WM email address and the use of Zoom breakout rooms. When joining a

Zoom session, please change your Zoom screen name to your preferred name (i.e. the name you wish me to use to address you in class) if different from the default name shown in Zoom.

Zoom Etiquette

Please be on time when entering a Zoom session, remain muted unless you have a question, and keep your video on during the entire class session. In order to replicate the in person experience for all students, you are required to show your video during the class session. You may use a virtual background if you wish, but all participants should be able to see you as if you were in the physical classroom.

Syllabus Changes

This is a dynamic syllabus, meaning it may undergo change. This class will follow the rules and guidelines outlined by the university regarding best practices surrounding Covid-19. There is a chance that recommended protocols for delivering class material will change based on new information. In the event this happens, I will update our syllabus on Blackboard and inform you in class or via email. It is the student's responsibility for reviewing the syllabus for changes each week on Blackboard.

Week	Date	Topic	Reading/Training to Complete Before Class
1	Monday 01-24-2022 (1) 11:00-12:20 (2) 12:30-13:50	Expectations Course Introduction	• Introduction • Syllabus Overview • Course Overview
	Wednesday 01-26-2022 (1) 11:00-12:20 (2) 12:30-13:50	Introduction to Big Data	• Big data Fundamentals • VM vs. Cloud
	Friday 01-28-2022 (1) 11:00-12:20 (2) 12:30-13:50	Setting up the lab Environment	• Intro to Sandboxes • Docker Containers • Cloudera's quickstart VM • Cloudera's HDP • Cloudera HDF
2	Monday 01-31-2022 (1) 11:00-12:20 (2) 12:30-13:50	Introduction to Big Data	• Big data Fundamentals **Assignment 1 - Category C** **(SetUp Docker and Cloudera Quickstart VM)**
	Wednesday 02-02-2022 (1) 11:00-12:20 (2) 12:30-13:50	Big Data Ecosystem	• Types of analytics • Big Data characteristics • Data analysis flow • Analytics patterns
	Friday 02-04-2022 (1) 11:00-12:20 (2) 12:30-13:50	Big data examples, applications & case studies	• Big data examples, applications & case studies • Hands-on Sandbox • HDFS
3	Monday 02-07-2022 (1) 11:00-12:20 (2) 12:30-13:50	Batch Processing	HDFS, YARN MapReduce • Programming model • Examples • MapReduce pattern **Assignment 2 - Category A** **(Datacamp Regular Expressions with Python)**

	Wednesday 02-09-2022 (1) 11:00-12:20 (2) 12:30-13:50	Batch Processing	HDFS, YARN MapReduce • Programming model • Examples • MapReduce pattern
4	Monday 02-14-2022 (1) 11:00-12:20 (2) 12:30-13:50	Batch Processing	• Regular expressions • Examples • MapReduce pattern **Assignment 3 - Category A** **(Map Reduce)**
	Wednesday 02-16-2022 (1) 11:00-12:20 (2) 12:30-13:50	Batch Data analysis	• RDBMS • Sqoop • HIVE and Impala
5	Monday 02-21-2022 (1) 11:00-12:20 (2) 12:30-13:50	Batch Data analysis	• Sqoop, Hive, and Impala **Assignment 4 - Category A** **(Impala / Hive Querying)**
	Wednesday 02-23-2022 (1) 11:00-12:20 (2) 12:30-13:50	Data acquisition and storage	• Amazon S3 (AWS Boto) • Hbase • Interacting with APIs • NoSQL Databases
6	Monday 02-28-2022 (1) 11:00-12:20 (2) 12:30-13:50	Data acquisition and storage	• Data in motion • Apache Nifi • Apache Kafka **Assignment 5 - Category A** **(Apache Nifi)**
	Wednesday 03-02-2022 (1) 11:00-12:20 (2) 12:30-13:50	Big Data acquisition and storage	• Data in motion - continued • Midterm logistics
7	Monday 03-07-2022	MIDTERM EXAM (Theory)	

	(1) 11:00-12:20 (2) 12:30-13:50		
	Wednesday 03-09-2022 (1) 11:00-12:20 (2) 12:30-13:50	MIDTERM EXAM (Hands-on)	
8	SPRING BREAK		
9	Monday 03-21-2022 (1) 11:00-12:20 (2) 12:30-13:50	Real-time analytics	Apache Spark Intro to Group project
	Wednesday 03-23-2022 (1) 11:00-12:20 (2) 12:30-13:50	Real-time analytics	Apache Spark
	Friday 03-25-2022 1- 08:00-09:20 2- 09:30-10:50	Real-time analytics	Apache Spark
10	Monday 03-28-2022 (1) 11:00-12:20 (2) 12:30-13:50	Real-time analytics	• Stream processing with Storm • In-memory processing with Spark Streaming • Real-time analysis examples & case studies **Assignment 6 -Category A** **(Datacamp - Big Data fundamentals with Pyspark)**
	Wednesday 03-30-2022 (1) 11:00-12:20 (2) 12:30-13:50	Introduction To Cloud Computing	Amazon AWS
	Friday 04-01-2022 1- 08:00-09:20 2- 09:30-10:50	AWS	Amazon Elastic Map Reduce (EMR)
11	Monday 04-04-2022 (1) 11:00-12:20	AWS	Streaming Data with AWS **Phase I (Category B)**

	(2) 12:30-13:50		
	Wednesday 04-06-2022 (1) 11:00-12:20 (2) 12:30-13:50	AWS	Machine learning
12	Monday 04-11-2022 (1) 11:00-12:20 (2) 12:30-13:50	AWS	• SparkMLib •H2O • Clustering algorithms • Classification algorithms • Regression algorithms **Assignment 7 - Category A** **(Datacamp - Streaming data with AWS)**
	Wednesday 04-13-2022 (1) 11:00-12:20 (2) 12:30-13:50	AWS	• Recommendation systems Big Data analytics case studies with implementations
13	Monday 04-18-2022 (1) 11:00-12:20 (2) 12:30-13:50	Machine Learning on AWS	Project presentations 1 **Phase II (Category B)**
	Wednesday 04-20-2022 (1) 11:00-12:20 (2) 12:30-13:50	Big Data analytics algorithms	Project presentations 2
			Classes End April 22nd

DATACAMP ASSIGNMENTS

During the next six months you will have access to a wonderful online learning platform called DataCamp. You will receive an invitation to join our course on DataCamp in your W&M email. Assignments generally consist of an entire 4-5 hour long course on DataCamp. Some weeks may have more than a single assignment, so be sure to plan your time wisely. All of these assignments are completed on DataCamp's website. All due dates are firm. Under no circumstances will late submissions be accepted. All of the DataCamp assignments are due by 11:59PM (ET) on Sundays. While these assignments are individual endeavors, you may receive help from others as long as you are the person completing the work and the work you hand in is yours. (Category C).

S/N	Title	Link	Completion Deadline
1	Regular Expressions In Python	https://learn.datacamp.com/courses/regular-expressions-in-python	Feb 6, 11:59PM EST
2	Datacamp for Big Data Fundamentals with Pyspark	https://app.datacamp.com/learn/courses/big-data-fundamentals-with-pyspark	Mar 27 11:59PM EST
~~3~~	Streaming Data with Aws Kinesis and Lambda	https://learn.datacamp.com/courses/streaming-data-with-aws-kinesis-and-lambda	Apr 10 11:59PM EST

TEAM PROJECT

Details of the team project can be found on Blackboard. The due date is firm. Under no circumstances will late submissions be accepted. The team project is a group assignment (Category B). Group members will evaluate all members of the group (including themselves) on each individual's value-add to the final product. Leadership and effort will be viewed favorably while the lack of adequate participation will adversely impact your grade. Deliverables include the programming script(s) that performs the analyses, data files that are required to run the script(s), and a recorded presentation of your project, along with the supporting slide deck.

ASSIGNMENTS

As indicated above, DataCamp assignments (Category A). It also includes one group assignment (Category B). Additionally, students are expected to complete all required reading and video instruction prior to the appropriate class to help ensure a high level of active, engaged, and thoughtful discourse during the class.