UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

Department of Decision, Operations and Information Technologies
University of Maryland

BUDT 758B-0501/0502
Big Data: Strategy and Analytics
Spring 2016

Instructors: **Guodong (Gordon) Gao and Anand Gopal**
4325 / 4307 Van Munching Hall
301-405-2218/ 301-405-9681, {ggao, agopal}@rhsmith.umd.edu

Lab Session Lead: **Ujjwal Goel <ujjwal@cs.umd.edu>**

Class Meets: 0501: M/W 11:00 to 12:15 PM, VMH 1518
0601: M/W 12:30 to 1:45 PM, VMH 1518

Office Hours: TBA

Course Introduction

Digitization is occurring in every aspect of business and our daily life. As a result, huge amount of data is being generated. Big data represents unprecedented opportunities for companies to generate insights and create wealth. At the same time, much of the big data is unstructured, in real time and only loosely connected. It defies the traditional ways of managing databases. This creates challenges even to tech-savvy companies on how to leverage the big data to gain competitive advantages.

This course uses a hands-on, learning-by-doing approach to understanding the concepts behind Big Data, some of the key technologies used within this ecosystem, the strategic drivers of Big Data and the value propositions that these technologies provide to industries. In addition, the course will also serve as an introduction to some of the key technologies within this ecosystem, such as Hadoop, AWS, Pig, Hive, Impala and Spark. The focus is on creating awareness of the technologies, allowing some level of familiarity with them through assignments, and enabling some strategic thinking around the use of these in business.

Since the course is designed to be experiential, the second half of the semester will be spent on company projects in the domain of big data and analytics. The specific content of these projects will be determined over the course of the semester in consultation with the corporate partners. These projects will entail not only hand-on work on datasets and big data technologies but also in understanding how big data can add value to the companies.

The technology of Big Data is still evolving very rapidly. Therefore, there is a level of experimentation with new material that will take place during the semester. Students are required to be flexible as and when topics or material in class are revised or modified. We will do our best to ensure that no undue burden is placed on students.

Learning Objectives
The course has two primary objectives:

1. To allow students to have working knowledge and exposure to key elements of a big data technology platform
2. To allow students to understand critical business and strategic issues around the use of big data in organizations and to help guide the successful design and implementation of big data strategy

Though mastery of this content requires more than one course, an introductory course, such as this, is useful in allowing students to gain much-needed familiarity with these technologies and concepts. That is the objective of this course.

Prerequisites

1. Databases, specifically working knowledge of relational databases and SQL
2. Working knowledge of Linux/Unix is useful

Software Needed
Much of the software needed for big data applications tend to be open source. Therefore, the source programs are free. However, for the purposes of the course, we will be using versions provided by the Cloudera Academic Partnership.

1. Cloudera CDH VM – provided by instructors (Hadoop, Pig, Hive, Impala)
2. Amazon Web Services – provided through the instructors

Required Reading Material
A significant proportion of the reading material for this course is available online and is free. When necessary, additional reading material will be posted on Canvas/ELMS.

Optional useful sources are listed below; these are not required but are good reference material.

1. Hadoop: The Definitive Guide, by Tom White (http://it-ebooks.info/book/5629/)
2. Big Data: A Revolution That Will Transform How We Live, Work, and Think, by Viktor Mayer-Schonberger and Kenneth Cukier (http://www.big-data-book.com/)
3. Mining of Massive Datasets. Hardcopy: Amazon.com E-version: Free available here

Course Format and Grading

Classes
We meet twice each week. Mondays will be mostly lecturing, and Wednesdays will be mostly in-class exercise. Given the diversity of the topic, there will be several guest lectures in class, with visitors from industry presenting their own perspectives on the value of big data.

Assignments

We have 3 homework assignments. These assignments are mainly from the lectures. These assignments will help you understand concepts and ideas you've learned from the class.

Class project

There is a class project for each group. The size of each group is 5, and the group will be assigned by TA randomly. Two types of formats are acceptable: a consulting case study or a runnable system (frontend + backend). For the case study, each group will be assigned a case (mostly, they are real data and problems in industry). For the system, you can use some existing online datasets or download your own datasets from online resources, like Facebook, Twitter, Yelp, Amazon.com, Yahoo financial news, etc. Then run existing big data analysis algorithms to show some interesting results.

If you miss your project presentation without an extremely good excuse, you will receive a grade of **ZERO** for that. If you think you have an excuse for missing your presentation, please discuss it with me, in advance if possible. If I judge that your excuse is reasonable, I will -- depending on the circumstances -- either give you a make-up presentation, or I will average your other grades so that the missing grade does not count against you.

Grading

Your final grade for the course will be composed from the following items:

Class participation:	10%*1=10%
Class project:	35%*1=35%
Lab report:	5%*5=25%
Assignments:	10%*3 =30%

Academic Integrity

The Robert H. Smith School of Business recognizes honesty and integrity as necessary cornerstones to the pursuit of excellence in academic and professional business activities. The University's *Code of Academic Integrity* is designed to ensure that the principles of academic honesty and integrity are upheld. All students are expected to adhere to this Code. The Smith School does not tolerate academic dishonesty. All acts of academic dishonesty will be dealt with in accordance with the provisions of this code. Please visit the following website for more information on the University's *Code of Academic Integrity:* http://www.inform.umd.edu/CampusInfo/Departments/JPO/AcInteg/code_acinteg2a.html

Plagiarism Policy: Inevitably in a programming course, it seems that a few people will turn in work that is not their own. You should understand that it is usually easy to detect copying of programs -- even when a program is modified to try to disguise its source. Copying a program, or letting someone else copy your program, is a form of academic dishonesty and the penalties can be found ***here***.

Schedule (subject to change)

Session	Topics	Lab	Assignment Due
1/14/2016	No class -rescheduled		
1/20/2016	Introduction. Big Data – Why business should care		
1/25/2016	Business Value of Big Data - Frameworks		
1/27/2016		Case Analysis / Write-up	Case Write-up Due
2/1/2016	Overview of the Hadoop Ecosystem		
2/3/2016	HDFS and Hue	Lab 2: Set up Cloudera Training Virtual Machine, Lab 3	
2/8/2016	The MapReduce Framework		
2/10/2016	Hadoop Installation and configuration	How MR is applied to various contexts.	
2/15/2016	Sqoop and Pig		
2/17/2016		Lab: Sqoop and Pig	
2/22/2016	Impala and Hive		
2/24/2016		Lab 6	
2/29/2016	Flume / NoSQL		
3/2/2016		Lab 9	
3/7/2016	AWS Basics		
3/9/2016		Lab on AWS	
3/14/2016	Spring break		
3/16/2016	Spring break		
3/21/2016	Guest speaker 1		
3/23/2016	Go over assignment #1(AWS). Project Discussion		#AWS assignment due
3/28/2016	Spark basics		
3/30/2016	Spark basics - continue	Lab 10	
4/4/2016	Spark Application Development	Lab 13	
4/6/2016	Big Data ML Applications: Clustering	Lab 16 – K-means Clustering	
4/11/2016	Association Rules		#Spark assignment due
4/13/2016	Collaborative Filtering		
4/18/2016		Vector Similarity	
4/20/2016	Topic Modeling (LDA)		
4/25/2016		Sentiment Analysis	
4/27/2016	Guest speaker 2		
5/2/2016	Group Project Presentation		
5/4/2016	Group Project Presentation		Final Group Project Report Due