This syllabus is effective as of Sunday, May 03, 2020 at 07:58 PM

DNSC 6290 Working with Large Datasets the George Washington University Summer 2020

Course Information

- **Instructors:** Marck Vaisman (marck at gwu.edu)
- **Classroom:** Online
- **Time:** Wednesdays 6:10-8:40pm (actual time may be adjusted)
- **TA's:**
- **Course Description:** Learn to work with datasets that will not fit on a single machine because of storage, memory or processing constraints. This practical workshop-style course teaches students cloud computing and open-source tools to work with large datasets. Topics include: creating and setting up a cluster in the cloud, parallelization techniques, Hadoop, MapReduce, and Apache Spark.

Learning Objectives

- Operate big data tools and cloud infrastructure, including Spark, MapReduce, Hadoop
- Use ancillary tools that support big data processing, including git and the Linux command line
- Setup and manage big data infrastructure and tools in the cloud on Amazon Web Services and Microsoft Azure
- Identify broad spectrum resources and documentation to remain current with big data tools and developments
- Execute a big data analytics exercise from start to finish: ingest, wrangle, clean, analyze and store

Pre-requisites

- Experience with Python and SQL. **Note:** We will use Python as the primary interface to Apache Spark, through PySpark
- Understand programming concepts (flow control, input/output, variable assignment.)
- Experience with git and GitHub

Refresher Tutorials

It is highly recommended that you go through the following tutorials if you need a refresher or are new to the topics of git, the command line, and SQL. If these topics are new to you, you must take these online courses before the course begins.

- git - the simple guide
- codeacademy - Learn the Command Line
- codeacademy - Learn SQL
- Nico Riedmann's Learn git concepts, not commands

Books, Software and Cloud Resources

Readings (for assigned readings)

There is no required textbook for the course. We have selected specific chapters from several sources as well as several seminal papers in the big data space, and these will be provided to you in PDF format. We may also provide supplemental materials (articles, links, videos, etc.) to complement the readings. **You must read assigned readings prior to the lectures.**

Chapter readings are from the following books:

- Benjamin Bengfort, Jenny Kim (2016). Data Analytics with Hadoop: An Introduction for Data Scientists. O'Reilly Media. ISBN: 9781491913703.
- Bill Chambers, Matei Zaharia (2018). Spark: The Definitive Guide. O'Reilly Media. ISBN: 9781491912218.

- JT Wolohan (2019). Mastering Large Datasets with Python. Manning.

Cloud Resources

You will be using cloud resources on Microsoft Azure and Amazon Web Services. We will discuss how to setup your account and environment in the first class session.

Terminal and Command Line

You will be using the command line and terminal heavily in this course. Please take a look at the following depending on your laptop's operating system

- **For Windows users:** Students with Windows machines will be using PowerShell as the terminal for command line operations. Please read this article to enable it on your machine. **You will need the latest version of Windows 10, make sure to update your machine!** Even if you have PuTTY or another ssh client installed, we will be using PowerShell as the terminal in this course and will not support anything else.
- **For Mac users:** Macs have a built in Terminal. However, I prefer using iTerm as another Terminal application for your Mac. I've been using it for years and I love it. This is not required, but truly recommended.
- **Linux users:** nothing to do.

Credit Cards

You will need a credit card or a debit card to be able to create an account on Amazon Web Services. **The card will not be charged unless you use up all of the course credits.** If you do not have a credit card, you may consider getting a pre-paid VISA card which you can use as the credit card when you create the account.

Learning Activities, Communication and Evaluation

This is a hands-on, practical, workshop style course that provides opportunities to use the tools and techniques discussed in class. Although this is not a programming course per se, there is programming involved.

Lectures and In-Class Labs

Every class session will have a lecture portion and most sessions will have an in-class lab portion. Lab exercises are designed to get you familiar with the tools discussed in class. In these labs, we will work through simple examples.

Online Quizzes

Weekly quizzes will ensure students are keeping up with the material presented in the class. The material for the quizzes will be drawn from lectures, labs, and readings. These quizzes are meant to be brief and will have a time limit of 15 minutes. Students will be responsible for providing definitions for terminology, describing concepts, reading computer results, answering basic analysis questions and, perhaps very straight-forward computations. These quizzes will generally be in multiple choice or short answer format. The quizzes will be open book, and are individual assignments.

Problem Sets

You will be given problem sets as homework assignments. The goal of these problem sets is to use the big data tools to answer some questions about large datasets. The problem sets will build on the labs and will be much more elaborate. Deliverables from the problem sets will usually include code written for your programs and the output produced.

Important Note: We reuse problem set questions, we expect students not to copy, refer to, or look at the solutions in preparing their answers. Since this is a graduate-level class, we expect students to want to learn and not search online for answers.

Group Project

Students will assemble into groups of 3-4 students and propose, perform, and write up an analysis of a *big* dataset using the tools learned in class. *Big* is defined as "a dataset that is so large that you cannot work with it on a laptop."

Details for the project can be found here.

Grading

- Completion of weekly labs: 10%
- Weekly quizzes: 20%
- Problem Sets: 30%
- Group Project: 40%

Total is 100%. We have no plans to curve the final grade, so the letter grade will follow standard guidelines:

- A: >= 93
- A-: >= 90, < 93
- B+: >= 87, < 90
- B: >= 83, < 87
- B-: >= 80, < 83
- C: < 80

Course Calendar

This calendar is subject to change. We will make make any changes known in advance.

Class	Date	Topics	Lab	Readings	Notes
1	Wed May-20	Course Overview. Big Data concepts. Cloud computing and evolution of cloud technologies	Setting Up		
2	Wed May-27	Scaling up on a single machine. Functional programming. Map and reduce functions. Parallelization.	Parallelization		
3	Wed Jun-03	MapReduce. Hadoop and Hadoop Streaming. Distributed filesystems.	Running a Hadoop job		
4	Wed Jun-10	Spark 1: Intro to Spark and PySpark. RDDs. Spark DataFrames. SparkSQL	Intro to Spark, working with RDDs		
5	Wed Jun-17	Spark 2: Marchine Learning with Spark	Machine Learning with Spark		
6	Wed Jun-24	Spark 3: Spark Streaming. Other big data tools. Course wrapup.	Spark Streaming		

Policies & Expectations

General Policies

- **Attendance and punctuality:** Attendance is mandatory. Given the technical nature of this course, and the breadth of topics discussed, you are expected to attend each class, to complete all readings, and to participate actively in lectures, discussions and exercises. We understand there may be times you may need to miss class, please inform us in advance if you are not able to attend class for any reason.
- **Participation:** We love participation. Read. Raise your hand. Ask questions. Make comments. Challenge us. Acknowledge us. If we speak for three hours to a silent classroom, it is a lot more boring and tiring for everyone.
- **Computer Usage:** You must bring your laptop to class to work on **labs**. No phone or computer use (i.e browsing, social media, shopping, etc.) is allowed during lecture.
- **Email and Online Discussion Boards:** Please use the discussion boards on for questions about the course, homework assignments, technical issues, etc. Staff will be monitoring them and providing answers on a regular basis. Individual emails will not be answered except for special circumstances.
- **Cloud Resources:** You will create cloud accounts on Amazon Web Services and Microsoft Azure. You will get credits on both platforms that will be enough to support your coursework throughout the semester. **It is your responsibility to manage the credits and resources yourself. If you run out of credits because you do not shut down your resources, we cannot help you.**
- **Allocate Time for Assignments:** Homeworks take time, so please do not wait until the last minute to start them. Give yourself several days to work on problem sets. While the tools have matured a lot over the years, there are cases where you will run into unforseen technical difficulties. All homework assignments have been thoroughly tested using the technical configuration provided in the assignment and they work. *"It didn't work for me"* is not an excuse. *"I lost my code because I didn't push to github"* is not an excuse. *"It took me too long because it was the first time I'm doing it"* is not an excuse.
- **Late Policy:** Due dates will not be extended, and late submissions will incur a late penalty of 25% per day.
- **Class materials are for class use only!:** Please refrain from making your private GitHub repositories or any other class materials public.

Open Door Policy

Please approach or get in touch with us if something is not working for you regarding the class, methods, etc. Our pledge to you is to provide the best learning experience possible. If any issue comes up, please do not wait until the last minute to bring it up. We will work with you individually to try to resolve in a timely manner.

Academic Integrity

The code of academic integrity applies to all courses in the George Washington School of Business. Please become familiar with the code. All students are expected to maintain the highest level of academic integrity throughout the course of the semester. Please note that acts of academic dishonesty during the course will be prosecuted and harsh penalties may be sought for such acts. Students are responsible for knowing what acts constitute academic dishonesty. The code may be found at http://www.gwu.edu/~ntegrity/code.html

Collaboration Policy

- **In-class labs:** you may collaborate with other students during in-class labs to facilitate collective learning.
- **Group project:** by nature, it is a group project and collaboration is to be confined within groups. **You may not collaborate across groups.**
- **Online quizzes:** these are individual assignments and must be completed individually.

- **Problem sets:** students may work on homework assignments together. However, students must submit their own work, and work must be done in the students' own cloud resources. Assignments that do not represent the student's own work will be awarded a score of 0. This includes copying and pasting other students' code or not doing the work on your own cloud resources. You are encouraged to learn from each other, however, you must write your own code.

We have a ZERO TOLERANCE POLICY.

University Policies and Support Services

Religious Accommodation

Students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance. Faculty should extend to these students the courtesy of absence without penalty on such occasions, including permission to make up examinations. Faculty who intend to observe a religious holiday should arrange at the beginning of the semester to reschedule missed classes or to make other provisions for their course-related activities.

Disability Support Services (DSS)

Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Rome Hall, Suite 102, to establish eligibility and to coordinate reasonable accommodations. For additional information please refer to http://gwired.gwu.edu/dss/

Mental Health Services 202-994-5300

The University's Mental Health Services offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include: crisis and emergency mental health consultations confidential assessment, counseling services (individual and small group), and referrals. http://counselingcenter.gwu.edu