This syllabus is effective as of Thursday, January 03, 2019 at 09:44 PM

ANLY502 - Massive Data Fundamentals Georgetown University Spring 2019

Course Information

- **Instructors:** Marck Vaisman (mv559 at georgetown.edu), Irina Vayndiner (iv95 at georgetown.edu)
- **Classroom:** Car Barn 203
- **Time:** Monday 6:30-9:00pm (except 1/9/19 which meets on a Wednesday)
- **TA's:**
  - TBD
- **TA Office Hours:**
  - TBD

Course Description

Data is everywhere! Many times, it's just too big to work with traditional tools. This is a hands-on, practical workshop style course about using cloud computing resources to do analysis and manipulation of datasets that are too large to fit on a single machine and/or analyzed with traditional tools. The course will focus on Spark, MapReduce, the Hadoop Ecosystem and other tools.

You will understand how to acquire and/or ingest the data, and then massage, clean, transform, analyze, and model it within the context of big data analytics. You will be able to think more programmatically and logically about your big data needs, tools and issues.

Credit Hours

This is a 3 credit graduate course. You will spend approximately 3 hours per week in class. It is expected that you will spend approximately 2-3 hours of outside classroom activities (required readings, homework problems, completion of labs, quizzes, etc.) for each hour of class time. You will spend 36 hours in instructional time, and approximately 100 hours in out-of-classroom time.

Course Objectives

- Operate big data tools and cloud infrastructure, including Spark, MapReduce, Hadoop and other tools in the big data ecosystem
- Recognize and 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 Microsoft Azure and Amazon Web Services
- 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, store and present
- Be aware of the responsibilities that are associated with performing analysis of large datasets

Prerequisites

Essential

- Experience with R, Python, and SQL for reading files, manipulating and analyzing data. **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

Optional and Useful

- Experience with the command line and terminal shell to navigate the file system, manipulate files (create, move, delete, etc.)
- Experience with remote computing via ssh
- Understand shell executables

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.

- git - the simple guide
- DataCamp: Introduction to Git for Data Science
- DataCamp: Introduction to Shell for Data Science
- DataCamp: Introduction to SQL for Data Science

Books, Software and Cloud Resources

We have chosen several reference books for this course that cover different parts of the material. There are assigned readings and it is expected that you will read the assigned readings before coming to class. These books are all available on Safari Books Online, and you should be able to access these resources as a Georgetown student. Visit the Georgetown Library e-book information page and click on "Safari Books Online" for more information.

Required Books (for assigned readings)

- 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.

Additional Recommended Books

- Tomasz Drabas, Denny Lee (2017). *Learning Pyspark*. Packt Publishing. ISBN: 9781786463708.
- Ofer Mendelevitch, Casey Stella, Douglas Eadline (2016). *Practical Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale*. Addison-Wesley Professional. ISBN: 9780134024141.
- Krishna Sankar (2016). *Fast Data Processing with Spark 2 - Third Edition*. Packt Publishing. ISBN: 9781784392574.

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.

DataCamp

All students in the course will have access to 6 months of all DataCamp courses. More details will be provided in class.

Required For Windows Users Only

If you have a Windows laptop, please download the Babun Shell before first class. **Do not install, we will do it in class.** Direct download link is here: download link.

Mac Users (Optional)

I recommend 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.

Credit Cards

You will need a credit card (not a debit card) to create an account on Amazon Web Services. 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 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 many 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. The completion of the in-class lab exercises is part of your attendance/participation portion of the grade.

Quizzes

You will take online quizzes about the topics/ideas discussed in class and from the readings. The purpose of the quizzes is to reinforce your knowledge about the tools and platform and also to help you remember the nomenclature and terms used in class. The quizzes are open book and you will take them online through Canvas. You can take them at your convenience within the established time window.

In lieu of a final exam, there will be an in-class, closed book *final* quiz towards the end of the semester.

Assignments

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.

We will be using GitHub Classroom for problem sets and assignment submissions. When an assignment is created, we will email you a link that will clone the assignment and create a private repository for you in your own GitHub account. You will perform your work within the repository and then push back to GitHub for submission. If you do not have a GitHub account, please create one.

The last assignment can be considered a *mini individual project* and will be longer, more difficult and comprehensive. This last assignment will also have a higher weight and will require you to do an end-to-end analytical pipeling using the big data tools.

Group Paper Presentation

We have chosen 12 seminal papers in the big data space that will be presented throughout the semester, usually one per week. Students will be assigned to groups of no more than 5 students, and the papers will be randomly assigned. The group will read the paper, and prepare a presentation that will be given at the beginning of the class session.

Grading

- Assignments: 70% (5 homeworks at 10% each, mini project at 20%)
- Quizzes: 20% (online quizzes 10%, final quiz 10%)
- Attendance/Participation: 5% (attendance, in-class discussion, completion of in-class labs)
- Group Paper Presentation: 5%

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

- A: 93-100
- A-: 90-92
- B+: 87-89
- B: 83-86
- B-: 80-82

Grading Rubric for Homework Assignments

- We will look at the results files and the scripts. If the result files are exactly what is expected, in the proper format, etc., we may run your scripts to make sure they produce the output. If everything works, you will get full credit for the problem.
- If the results files are not what is expected, or the scripts produce something different from what is expected, we will look at code and provide partial credit where possible and applicable.
- Points will be deducted for each/any the following reasons:
  - Instructions are not followed
  - Output is not in expected format (not sorted, missing fields, wrong delimiter, unusual characters in the files, etc.)
  - There are more files in your repository than need to be
  - There are additional lines in the results files (whether empty or not)
  - Files in repository are not the requested filename

Course Calendar

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

Date	Session	Title	Topics	Lab	Reading	Assignment	Quiz
Jan 09	1	Course Overview, Big Data, Cloud Computing	Create cloud accounts, create GitHub account, provide account information to Professors				
Jan 14	2	Cloud Computing providers: AWS and Azure, IAAS, PAAS and SAAS, Linux/Command Line review	Setup environment, create SSH keys, start and connect to an instance in the cloud	paper 1		A1 released (Python script) - due xxxx	
Jan 28	3	Distributed computing, HPC, Hadoop, Distributed filesystems, MapReduce programming model	Start and connect to a cluster, Run built-in Hadoop examples on cluster, Examine the different user interfaces		Bengfort & Kim: Chapter 2		
Feb 04	4	Hadoop Streaming	Run the "Hello World" of Hadoop, the word count using Hadoop Streaming		Bengfort & Kim: Ch. 3	A2 released (Hadoop Streaming) - due xxxx	
Feb 11	5	Overview of scalable database systems, Massively Parallel Processing databases, Neteeza, Greenplum, RedShift					
Feb 25	6	Pig and Hive	Store a dataset in a Hive table, Run and example Pig job		Bengfort & Kim: Ch. 6,8	A3 released (Hive/Pig) - due xxxx	
Mar 11	7	Spark introduction, Resilient Distributed Datasets, PySpark	Start a PySpark session, Create RDDs, Operate on RDDs		Chambers & Zaharia: Ch. 1-3, 12, 32	A3 released - due Mar 21	
Mar 18	8	SQL Review, Intro to SparkSQL	Perform operations on Spark dataframes using SparkSQL		Chambers & Zaharia: Ch. 10		Q4 - due Apr 6
Mar 25	9	SparkML, Issues with ML algorithms on large datasets	Build a model		Chambers & Zaharia: Ch. 24, 25	A4 released - due Apr 11	

Date	Session	Title	Topics	Lab	Reading	Assignment	Quiz
Apr 10	01	Spark Streaming	TBD		Chambers & Zaharia: Ch. 20, 21		
Apr 11	08	NoSQL	Store data in a NoSQL data store				
Apr 12	15	GraphX API for Spark	Analyze a large graph		Chambers & Zaharia: Ch. 30		
Apr 13	29	Apache Drill, other topics	TBD				Q6 - in class

Class will not meet on Jan 21 (MLK Holiday), Feb 18 (President's Day), Mar 04 (Spring Break), Apr 22 (Easter Break).

Policies

General Course Policies

- **Attendance:** Given the technical nature of this course, and the breadth of topics discussed, you must attend every class session. **Attendance will be tracked and is part of your grade.** Please contact us in advance if you are not able to attend class. Excused absences in advance will not affect your attendance grade. **Unexcused absences will be deducted from your attendance grade.**
- **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. Please refrain from other activities. **You may not use your computer while your peers or guest speakers are presenting.**
- **E-mail:** We (Irina and Marck) will try to respond to email within 36 hours. Use email for questions **not** related to homework, material or techincal questions. If you do email us, please email **both** Irina and Marck.
- **Online Discussion Boards:** Please use the discussion board on Canvas for questions about the course, homework assignments, technical issues, etc. TA's will be monitoring them and providing answers. Answering individual questions submitted by email does not scale, and the likelihood of many students having the same question is high. Using the forums is a great resource for everyone.
- **Name Tents:** You will be given a name tent **to be used in every class** so we get to know you better and quicker. TA's will keep the name tents and will be used to track attendance. Grab your name tent on the way in and return it before leaving.
- **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.**
- **Homework Submission:** Homeworks take time, so please do not wait until the last minute to start them. Give yourself a 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**: Each student has 2 (two) "late" days that can be used throughout the semester however they wish. You can submit one assignment two days late, or two assignments one day late. Homework due dates will not be extended, and late homeworks will incur a late penalty.

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 you have any issue please do not wait until the last minute to speak with us. You will find that we are fair, reasonable and flexible and we deeply care about your success.

Academic Integrity

All students are expected to maintain the highest standards of academic and personal integrity in pursuit of their education at Georgetown. Academic dishonesty, including plagiarism, in any form is a serious offense, and students found in violation are subject to academic penalties that include, but are not limited to, failure of the course, termination from the program, and revocation of degrees already conferred. All students are held to the Georgetown University Honor Code. For more information about the Honor Code see http://gervaseprograms.georgetown.edu/honor/

You may collaborate with other students during in-class labs to facilitate collective learning. Regarding assignments, though, only limited collaboration is allowed.

- You may discuss ideas on how to solve the problem with other students
- You may work alongside other students
- You may get help from other students if you are stuck and/or are having technical difficulties on steps to take

BUT

- **You may not share code**
- **All submitted code and scripts written must be your own**
- **You must do all your work on your own cloud resources**

We have a ZERO TOLERANCE POLICY.

Accommodations for students with disabilities

Students with documented disabilities have the right to specific accommodations that do not fundamentally alter the nature of the course. Please alert us should you require accommodations.

Title IX Sexual Misconduct Statement

Please know that as faculty members we are committed to supporting survivors of sexual misconduct, including relationship violence and sexual assault. However, university policy also requires us to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct.

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu