ANLY 502 (v2)

Massive Data Fundamentals

Instructors:
Simson Garfinkel
sg1224@georgetown.edu
202-649-0029
Ghaleb Abdulla
Abdulla1@llnl.gov
925-423-5947

Hours: 6:30pm - 9:00pm

Location: Reiss 152

Materials:

Learning Spark, Karau et al., O'Reilly (Feb 2015)

Advanced Analytics with Spark, Ryza et al, O'Reilly (April 2015)

Class Notes, Simson Garfinkel

Assigned papers

Open source software

Milestones

Wed Jan 13
First day of class

Mon Jan 25
Second day of class

Wed Apr 27
Last day of class.
In class presentations.

Wed May 7th
Projects due.

Course Description

"Today's data scientists are commonly faced with huge data sets (Big Data) that may arrive at fantastic rates and in a broad variety of formats. This core course addresses the resulting challenges to data professionals. The course will introduce students to the advantages and limitations of distributed computing and to methods of assessing its impact. Techniques for parallel processing (MapReduce) and their implementation (Hadoop) will be covered, as well as techniques for accessing unstructured data and for handling streaming data. These techniques will be applied to real world examples, using clusters of computational cores and cloud computing. Prerequisite: Good command of R or Python, some knowledge of data structures. Three credits"

Additional Description

The goal of this class is for you to learn the technology, business, science, and social implications of "big data" processing. In recent years there has been an explosion of tools, techniques, and technologies for working with massive data sets. We will start from the beginning, teaching you how to identify, design, and build a system that can analyze massive data. This course emphasizes reading official documentation and building real-world systems using stand-alone Hadoop/Spark environments running in VirtualBox on your personal system, and scalable clusters that you can build on Amazon Web Services. Most classes will include a mix of lecture, class discussion, and live programming/demonstrations.

Students will identify and develop a socially relevant big-data project, acquire data, and write a final report based on data analysis.

Learning Outcomes

At the end of this course, you will be able to:

- Identify technical and social trends in the creation, collection, analysis and storage of massive data.
- Design, cost, and assemble cloud-based computational infrastructure required to perform massive data analysis.
- Perform large-scale data analysis with Python on high-performance workstations using multithreading/multiprocessing and clusters using Hadoop, Map Reduce, Apache Spark, and other advanced technologies.

- Locate, download, "wrangle," and query structured and unstructured data from Internet sources.
- Research and present information about a new "Big Data" tool on the Internet.
- Understand and discuss academic papers about big data technology and related social issues.

Prerequisites and Requirements

Students should have a working knowledge of Python and the Unix command line (tutorials and reviews will be provided for those in need of a refresher). Students should have a laptop with at least 100GB of free disk space, 8GB of RAM, and Oracle's VirtualBox installed. VirtualBox can run on Mac, Windows and Linux machines and runs best on computers with at least 2 cores and an SSD or 7200RPM hard drive. (An external 7200RPM or SSD connected via USB3 is fine; a USB3 thumb drive is not.)

Materials and Student Deliverables

Required Readings are due on the date for which they are appear on the syllabus so that they can be discussed in class. Students are responsible for the content of these readings. Other than the textbook, readings will be made available on Blackboard. **This course expects that you will spend 2-3 hours of preparation time for each hour of class time.**

Homework and Problem Sets are assigned in class and due on a specified date. Problem sets must be turned in via Blackboard. Students are encouraged to turn in their homework on time. Late homework will be accepted for up to 1 week at the cost of 1 point per day, unless arrangements to turn in an assignment late have been made in advance.

Each student is expected to make two *Presentations* during this course. 1) Each student will identify, prepare, and present an article about "massive data" from the open literature. (Articles should be submitted and approved in advance by the instructor.) 2) Each student will identify and test a publicly available data analytics tool, such as an open source tool, a demonstration tool, or an appropriate website. Tools should be tested with publicly available data. Presentations should be 5-10 minutes. Dates for student presentations will be proposed early in the course.

Students will be responsible for a *Final Project.*

References and Additional Materials are optional and provided for students that are interested in delving further. Students are not responsible for the content of these materials.

Grading

Homework assignments, individual presentation, mid-term exam, and group presentations will be combined to create the final grade. Students may collaborate on a problem set or presentation, but all group members will receive the same grade, and each group may only collaborate on a single project.

Grading will be consistent with Section III.A.1 of the GSAS Bulletin, "Grades for Graduate Coursework,"[1] with 100% of all possible credit being equivalent to 4.000 Grade Quality Points and a grade of A+.

Credit in this course will be apportioned as follows:

Student Deliverable	Assigned	Due	Weight
Problem Set 01a: Hadoop Hello!	Jan 13	Jan 29	9

[1] https://sites.google.com/a/georgetown.edu/gsas-graduate-bulletin/iii-academic-regulations-procedures#3a

Problem Set 02a: MapReduce 2	Jan 25	Feb 5	9
Problem Set 03a: AWS Cluster and Joins	Feb 8	Feb ~~19~~ 26	9
Problem Set 04a: Apache Spark and Spark SQL	Feb. ~~22~~ 26	Mar. 18	9
Problem Set 05: TBD	Mar. 21	Apr. 1	9
Midterm		Mar. 21	15
Proposed dates for two presentations		Jan. 15	1
Presentation #1 — An open source big data software tool			4
Presentation #2 — A big data paper			4
Final Project Proposals (2)		Mar. 22	1
Final Project Group Proposal		April 1	1
Final Project Presentation		May 2	5
Final Project Paper		May 5	14
Classroom Participation			10
Total:			100

Note: The lecture order, problem sets, and student deliverables are subject to change. Revisions to this document will be made available on the class Blackboard site and sent by email to students. Students are encouraged to check Blackboard at least once a day.

Syllabus and Schedule

Wed, Jan 13 — L01: What's Massive Data?

Course preview. Massive Data Technology used in this course. Moore's Law and Parallelization. Multithreading. Programming models and challenges for multi-threading. Hadoop introduction. Map/Reduce as the fundamental concept in massive data analytics. Using the Cloudera Quickstart VM.

In Class Lab:

- □ Setting up a laptop with the Cloudera Quick Start VM.

Mon, Jan 18 — Holiday (MLK): Reading and online homework

Mon, Jan 25 —L02a: Map/Reduce

Hadoop introduction. Map/Reduce as the fundamental concept in massive data analytics. Global accumulators.

In Class Lab:

- □ Simple MapReduce with mrjob.

Problem Set Assigned:

- ❑ PS01a: Hadoop Hello World and Word Count using Hadoop with Hadoop Streaming and mrjob. Storing and retrieving files from HDFS. Due January 29th.

Required Readings:

- ☐ MapReduce: Limitations, Optimizations and Open Issues, Kalavri and Vlassov, Trust, Security and Privacy, 2013.

Problem Set Assigned:

- ❏ PS02a: MapReduce2: Parsing Web Logs, Top10, Filters and Joins. Due Feb. 5.

Optional Readings:

- Web Search for the Planet, Barroso, 2003.
- MapReduce: Simplified Data Processing on Large Clusters, Dean & Ghemawt, OSDI 2004.

Fri, Jan 29 — PS01a Due Today!

Mon, Feb 1 — L03a: Filters and Joins with MapReduce, SQL, and Hive.

Filters and Joins in MapReduce. Introducing SQL: SELECT, WHERE, ORDER BY, JOIN.

In Class Lab: Getting to know SQL with sqlite3

- ❏ Select and Join with SQL

Required Readings:

- ☐ DRAM Errors in the Wild: A Large-Scale Field Study, Schroeder et al, SIGMETRICS/Performance '09
- ☐ Amazon EBS documentation, http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/AmazonEBS.html

Optional Readings and Videos:

- ☐ Failure Trends in a Large Disk Drive Population, Pinheiro et al, FAST '07
- ☐ Recommended Lynda Videos: Introduction to Hive and HBase; Understanding Pig
- ☐ Recommended YouTube Videos: Hadoop Introduction to Pig; Hadoop Introduction to Hive

Fri, Feb 5 — PS02a Due Today!

Mon, Feb 8 — L04b: Scaling from One Computer to Thousands (AWS).

Multiprocessing. Grid computing. Virtualization. Cluster computing. Amazon Web Services. Sizing and pricing an on-demand-cluster. Calculating hardware requirements, error rates, and price/performance. HDFS, S3 and Ceph. Tiered storage, cluster storage, massive storage. The Common Rule.

In Class Lab:

- ❏ Creating your First Amazon Cluster

Problem Set Assigned:

- ❏ PS03a: AWS Cluster. Creating a EMR cluster and running big problems. Joins with mrjob. Perform calculations on a large dataset. Downloading, preparing, and computing analytics on a large data set. Pricing Amazon S3. Due Feb. 19

Required Readings:

- ❏ Hadoop 2: What's New, Sanjay Radia and Suresh Srinivas, ;login: vol 39, No. 1, Feb. 2014
- ❏ "Institutional Review Boards and Your Research," Simson Garfinkel and Lorrie Faith Cranor, *Communications of the ACM*, June 2010.
- ❏ The Hadoop Distributed File System, Shvachko et al, MSST 2010, (skim)
- ❏ Ceph as a scalable alternative to the Hadoop Distributed File System, Maltzahn et al, ;Login:, 35:4, August 2010. (skim)

Optional Readings

- Building and installing a Hadoop/MapReduce cluster from commodity components: a case study, Leidner and Berosik, ;login:, Vol 35, No. 1. 2010.
- Apache Hadoop: The Scalability Update, Konstantin V. Shvachko, ;login, June 2011.
- HDFS scalability: the limits to growth, Konstantin V. Shvachko, ;login: Vol. 35, No. 2, April 2010
- Data Availability and Durability with the Hadoop Distributed File System, Robert Chansler, ;login, Vol 37, No. 1.
- Ceph: A Scalable, High-Performance Distributed File System, Weil, Brandt, Miller, Long & Maltzahn, OSDI '06,
- An Analysis of Data Corruption in the Storage Stack, Bairavasundaram et al, Fast '08
- "Were All Those Rainbow Profile Photos Another Facebook Study?," J. Nathan Matias, The Atlantic, June 28, 2015.
- "Introduction to Hadoop Distributed File System Versions 1.0 and 2.0," Manpreet Singh and Arshad Ali, in Big Data Analytics with Microsoft HDInsight, Jan 4, 2016.

Mon Feb 15 — Holiday (President's Day): Reading and online homework

Mon, Feb 22 — L05b: Don Miner, Pig, and Privacy (if we have time)

Special Guest: Donald Miner.

Required Readings:

- ❏ Apache Pig 0.14.0 Overview (This is the version installed on AWS)
- ❏ Apache Pig 0.14.0 Getting Started
- ❏ Apache Pig 0.14.0 Pig Latin Basics
- ❏ Apache Pig 0.8.1 Pig Tutorial (out of date, but still a good explanation)
- ❏ Apache Pig 0.8.1 Pig Latin 1 Reference
- ❏ Apache Pig 0.8.1 Pig Cookbook

Fri, Feb 26 — PS03a Due Today! (new date!)

Mon, Feb 29 — L06: Spark.

Spark's architecture. RDDs. Writing Spark programs in Scala and Python. Internet sources for information about Spark. Word Count in Spark. What's in Web Logs. Log File analysis in Spark. Fair Information Practices.

Required Readings:

- ☐ Learning Spark Chapter 1, “Introduction to Data Analysis with Spark.”
- ☐ Learning Spark Chapter 2, “Downloading Spark and Getting Started.”
- ☐ Learning Spark Chapter 3: “Programming with RDDs.”
- ☐ Learning Spark Chapter 4, “Working with Key/Value Pairs”
- ❑ Learning Spark Chapter 7, “Running on a Cluster”

In Class Lab: Fun with Spark.

Problem Set Assigned: PS04 — Pig and Spark

Fri., Mar 4 — PS04a Due Today!

Mon, Mar 7 — Holiday (Spring Break, Fri Mar 4 — Sun Mar 13)

Mon, Mar 14 — L07: Spark SQL and De-Identification

Creating, transforming, and saving RDDs. File formats. Spark storage options: FS, S3, and HDFS. Brief introduction to SQL. SQLite. Accessing data with Spark SQL. Regular expressions. Ingesting XML and JSON. Extracting data from PDFs. Web Scraping with Spark. De-identification and re-identification. Famous re-identification examples. **Saving and Displaying Your Results.** Getting data out of Hadoop/Spark. Website integration. Displaying information in charts and on maps.

Required Readings:

- ☐ http://spark.apache.org/docs/latest/quick-start.html
- ☐ http://spark.apache.org/docs/latest/programming-guide.html
- ☐ http://spark.apache.org/docs/latest/submitting-applications.html
- ☐ http://spark.apache.org/docs/latest/cluster-overview.html
- ☐ http://minimaxir.com/2015/11/nyc-ggplot2-howto/

Optional Readings:

- Learning Spark Chapter 5, “Loading and Saving Your Data”
- Learning Spark Chapter 9, “Spark SQL” **Text Processing at Scale.**
- Adv. Analytics Chapter 8: “Geospatial & Temporal Analysis on the NYC Taxi Trip Data.”
- http://chriswhong.com/
- Airbnb Blog Entries:
  - “Unboxing the Random Forest Classifier: The Threshold Distributions,” October 1, 2015. http://nerds.airbnb.com/unboxing-the-random-forest-classifier/
  - “Mapping the World,” January 7, 2015: http://nerds.airbnb.com/mapping-world/
  - “When the Cloud Gets Dark: How Amazon’s Outage Affected Airbnb,” April 24, 2011. http://nerds.airbnb.com/when-the-cloud-gets-dark-how-amazons-outage-a/
  - “MySQL in the Cloud at Airbnb,” November 15, 2010. http://nerds.airbnb.com/mysql-in-the-cloud-at-airbnb/
  - “How we partitioned Airbnb’s Main Database in Two Weeks,” October 6, 2015. http://nerds.airbnb.com/how-we-partitioned-airbnbs-main-db/

Special Guest: Jim Koenig

Mon, Mar 21 —Midterm (1 hour)!

Midterm will be administered in class using Blackboard. Open Book. 6:30 - 7:50pm

Mon, Mar 21 — L08: Text and Image Processing at Scale, 8:00 - 9:00pm

Image capture. Privacy issues with Images. bytefish/facerec <facerec@noreply.github.com>

Elasticsearch. Text processing and web mining. Elasticsearch. Document Clustering. Named Entity Extraction, Identity Resolution, and Graph Processing.

Special Guest: Peter Wayner

- Analyzing graph networks with GraphX (Adv. Analytics Chapter 7)

Required Readings:

- Adv. Analytics Chapter 6: "Understanding Wikipedia with Latent Semantic Analysis"

"Making Sense of Found Data," Lecture by danah boyd at the National Science Foundation, Dec 10, 2015 (audio file and slides)

Tue, Mar 22 — Final Project Individual Proposals Due

Mon, Mar 28 — Holiday (Easter Break, Wed Mar 23 — Mon Mar 28)

Fri, April 1 — Final Project Group Proposals Due

Mon, April 4 — L09: LLNL #1: Big Data in High Performance Computing

Guest Lecturers': Todd Gamblin and Abhinav Bhatele

PS 05 Assigned

Mon, April 11 — L10: LLNL #2: Power and Performance data for High Performance Computing

Guest Lecturer: Barry Rountree

Mon, April 18 — L11: LLNL #3: Scientific and Simulation data

Fri, April 22 — PS5 Due

Mon, April 25 — L12: LLNL #4: Scientific Data Analysis Approaches, Architectures, and Workflow Systems

Mon, May 2 — L13: Final Projects.

TBD: Processing streaming data and deploying applications.

Required Readings:

- Learning Spark Chapter 10: "Spark Streaming"

Fri, May 6 — Fri, May 13 — Exam Period.

Fri, May 13 — Final Projects Due

Labs/Homework

Lab/Homework are research-oriented tasks that begin in class and are due the following week. These tasks will typically take between 4 and 10 hours to complete, depending on the skill of the student.

Books and references

You may find the following textbooks useful:

Learning Spark: Lightning-Fast Big Data Analysis, Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, O'Reilly Media, January 2015. $39.99 (Ebook $33.99)

Advanced Analytics with Spark, Sean Owen, Sandy Rzya, Uri Laserson, Josh Wills, O'Reilly Media. $49.99

Spark Cookbook, Rishi Yadav, Packt Publishing, July 2015. $35.99

Mastering Apache Spark, Mike Frampton, Packt Publishing, $54.99

Apache Spark Graph Processing, Rindra Ramamonjison, Packt Publishing, $27.99 (eBook)

http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-spark.html

https://aws.amazon.com/elasticmapreduce/

Academic Integrity

The following materials are from the Georgetown Honor Council website:

> "As a Jesuit, Catholic university, committed to the education of the whole person, Georgetown expects all members of the academic community, students and faculty, to strive for excellence in scholarship and in character.
>
> "To uphold this tradition, the University community has established an honor system for its undergraduate schools, including Georgetown College, the School of Foreign Service, the School of Business, the School of Nursing and Health Studies; for master's degree students except MBA students, and students in the School of Continuing Studies. Students are required to sign a pledge certifying that they understand the provisions of the Honor System and will abide by it.
>
> The Honor Council is the principal administrative body of this system. The Honor Council has two primary responsibilities: to administer the procedures of the Honor System and to educate the faculty and undergraduate student body about the standards of conduct and procedures of the System.
>
> **The Georgetown Student Pledge**
>
> *In pursuit of the high ideals and rigorous standards of academic life I commit myself to respect and to uphold the Georgetown University honor system:*
>
> *To be honest in every academic endeavor, and*
>
> *To conduct myself honorably, as a responsible member of the Georgetown community as we live and work together."*[2]

Academic honesty requires that students perform their own work on problem sets, tests, and final projects. Students must document the references or resources that they use in completing their assignments. Collaboration on problem sets and final projects must be approved in advance. Seeking help from others or collaboration on tests and exams is strictly forbidden. For problem sets and open book exams students may seek assistance from any source, including the books, notes, and on-line sources, provided that they do not involve interaction with a person.

In this class, you are encouraged to collaborate with other students when you study and when you do your homework. Some in-class work will also be in small groups. When working on a homework assignment or a practice exercise, start by yourself, then talk to other students, ask questions, and share your ideas, then complete the work on your own. Do not copy homework from others and do not permit others to copy your work, as this will be considered plagiarism or facilitating plagiarism. Do not use help from outside (e.g. online). Exams are open-book, open-notes, open-computer, but you are not allowed to collaborate with other students or seek any human help during the exam.

Online Materials and Communications

[2] https://honorcouncil.georgetown.edu/

All materials will be accessible through Blackboard, and Blackboard will be used to collect all student assignments. All class announcements will be sent through Blackboard. You are responsible for either having announcements delivered to a mailbox that you check, or monitoring Blackboard for information. We recommend checking Blackboard once per day.

Communications with the professor should take place through Blackboard or the Georgetown email system. Course-related email should use @georgetown.edu addresses. If it is not, the email may not be viewed in a timely fashion.

Attendance Policy

Students are expected to attend each class, to complete all preparatory work (including assigned reading), and to participate actively in lectures, discussions and exercises. Students should bring laptops to class. Students are expected to contact the Instructor in advance for planned absences, and after class as soon as possible in the event of a medical or personal emergency. Work-related absences can be accommodated if the Instructor is notified in advance.