Big Data Systems

Last updated: August 9, 2022

Overview

Instructor Name and Contact Information:

Adam Tashman

Email: apt4c@virginia.edu

Subject Area and Catalog Number: Data Science DS 5110

Year and Term: Fall 2022

Class Title: Big Data Systems

Level: Graduate

Credit Type: Graded

Class Description

Increasingly, data scientists , data engineers and data analysts are working with datasets that exceed the memory of a single machine. This motivates the need for a different paradigm of computing and a different toolset. This course will prepare you for this use case.

The focus of the course is learning Spark, an open-source, general-purpose computing framework that is scalable and blazingly fast. The fundamental data types and concepts will be covered (e.g., resilient distributed datasets, DataFrames). You will learn how to use Spark for large-scale analytics and machine learning, among other topics. Tools for data storage and retrieval will be covered, including Amazon Web Services (AWS) and the Hadoop ecosystem.

A team project is a large component of the course, whereby you will conduct an end-to-end data science project. This simulates the workflow of a professional data scientist, from developing a hypothesis to communicating with stakeholders.

After completing this course, you will have developed valuable data science skills and experience working with big data frameworks.

Required Text

Jules Damji, Brooke Wenig, Tathagata Das, and Denny Lee. 2020. *Learning Spark: Lightning-Fast Big Data Analytics*. **2nd edition**. Sebastopol: O'Reilly Media, Inc.

Tomasz Drabas, Denny Lee. 2017. *Learning PySpark: Build data-intensive applications locally and deploy at scale using the combined powers of Python and Spark 2.0*. Birmingham: Packt Publishing.

Learning Outcomes

Upon successful completion of this course, you will be able to:

1. Execute distributed computing frameworks using Spark
2. Demonstrate knowledge of applications for big data storage, retrieval, processing, and modeling using Amazon AWS, Hive, and others from the Hadoop ecosystem
3. Implement PySpark for prevalent data science tasks, including data analysis and machine learning
4. Execute an end-to-end predictive modeling project using a large dataset

Delivery Mode Expectations

Web-based with weekly live meetings

Required Technical Resources and Technical Components

- VPN app: Cisco AnyConnect

Class Specific Information

Class Instruction and Activities

The topics covered in this course include the following:

- Map Reduce Framework
- Getting started in Spark
- Fundamental objects in Spark: RDDs, Key Value Pairs, DataFrames
- Running on a cluster
- Machine Learning with MLlib Library

    - Model tuning, training, validation
    - Data preprocessing
    - Pipelines
    - Classical problems: classification, regression, clustering, recommendation
- HDFS for distributed data storage
- Hive for querying against big data
- Amazon AWS tools for computing, storage, and retrieval
- Streaming systems including Spark Streaming
- Delta Lakes
- GraphX

Class Requirements

Prior to taking this course, you should meet the following prerequisites:

- At least one programming course
- Regression Analysis
- Machine Learning or Data Mining

The following are strongly recommended:

- Programming in Python (since PySpark will be used in this course)
- At least one course in Probability

Evaluation Standards and Assessments

Quizzes	Quizzes will assess student knowledge and application of topics covered in reading assignments and modules.
Attendance	Student attendance is required.
Journaling	Students track their progress and learning in a Journal that is submitted to the instructor.
Programming Assignments	Programming assignments will be implemented in Jupyter Notebooks and provide hands-on experience writing/modifying Spark code, while working with various datasets.
Final Project	The final project is a large component of the course and it includes data collection, modeling, visualization, and presentation.

Your final letter grade will be determined by the following scale:

A+	100	98.0
A	97.999	93.0
A-	92.999	90.0
B+	89.999	87.0
B	86.999	83.0
B-	82.999	80.0
C+	79.999	77.0
C	76.999	73.0
C-	72.999	70.0
D+	69.999	67.0
D	66.999	63.0
D-	62.999	60.0
F	59.999	0

Class Schedule

Tuesday 8:30pm - 9:30pm Eastern Time

Communication & Student Response Time

Discussion boards are set up in each module and designed to be a place where students can reach out to peers and instructors and ask questions related to content and technology. Students are encouraged to check the discussion boards daily for updates and correspondence. Specific queries regarding your progress should be addressed to me via email and you will receive a response within 24 hours. Throughout our time together, the sooner you inform me of any problem (personal or academic) that may affect your attendance or performance, the better the chance we have of solving it together.

Assignments

Quizzes (20% of grade)

All quizzes are multiple choice, with full points awarded for a correct answer, and no points awarded for an incorrect answer. **They are closed book.**

Attendance (5% of grade)

Student attendance is required. If a student cannot attend a class, the instructor needs to be notified in advance.

Journaling (15% of grade)

The purpose of journaling is to track student learning and growth throughout this course.
The journal will be submitted over the course of the term, appending new entries.

Programming Assignments (30% of grade)

Programming assignments will include exercises in data analysis, pipeline development, and machine learning. The outline of the exercise will be sketched out by the instructor, and students will fill in the missing pieces, as well as modify and run the code. **You may collaborate with your classmates but your code must be your own.**

Final Project (30% of grade)

The final group project will include forming a hypothesis, data acquisition and analysis, programming in PySpark, writing a report, and presenting to the class. There will be an ungraded assignment in each module to help build toward the final project.

The final project will consist of three components, each worth one-third of the final project grade:

1. Code
2. Presentation
3. Paper

Spirit of the Course

Students must attend weekly live sessions and complete the final project as a team. I encourage you to post in the Teams channel and exchange ideas. For the programming assignments and quizzes, you must submit your own work.

Electronic Submission of Assignments

All assignments must be submitted electronically through Collab by the specified due dates and times. It is crucial to complete all assigned work—failure to do so will likely result in failing the class. For late assignments, 10% of the total grade will be deducted per day, where the day means 11:59 p.m. Eastern time cutoff. After five days late, it will be marked as 0 points.

Technical Support

Technical Specifications: Computer Hardware

Operating system: Microsoft Windows 8.1 (64-bit) or Mac OS X 10.10 Minimum hard drive free space: 100 GB, SSD recommended Minimum processor speed: Intel 4th Gen Core i5 or faster
Minimum RAM: 4 GB

Technical Support Contacts

UVaCollab: collab-support@virginia.edu

UVA Policies

SDS Grading Policies

The standing of a graduate student in each course is indicated by one of the following grades:
A+, A, A-; B+,B, B-; C+, C, C-; D+, D, D-; F. B- is the lowest satisfactory grade for graduate credit.

Attendance

Students are expected to attend all class sessions. Instructors establish attendance and participation requirements for each of their courses. Class requirements, regardless of delivery mode, are not waived due to a student's absence from class. Instructors will require students to make up any missed coursework and may deny credit to any student whose absences are excessive. Instructors must keep an attendance record for each student enrolled in the course to document attendance and participation in the class.

University Email Policies

Students are expected to check their official UVA email addresses on a frequent and consistent basis to remain informed of University communications, as certain communications may be time sensitive. Students who fail to check their email on a regular basis are responsible for any resulting consequences.

Mid-Term and End-of-Class Evaluations

Students may be expected to participate in an online mid-term evaluation. Students are expected to complete the online end-of-class evaluation. As the semester comes to a close, students will receive an email with instructions for completing this. Student feedback will be very valuable to the school, the instructor, and future students. We ask that all students please complete these evaluations in a timely manner. Please be assured that the information you submit online will be anonymous and kept confidential.

University of Virginia Honor System

All work should be pledged in the spirit of the Honor System at the University of Virginia. The instructor will indicate which assignments and activities are to be done individually and which permit collaboration. The following pledge should be written out at the end of all quizzes, examinations, individual assignments and papers: "I pledge that I have neither given nor received help on this examination (quiz, assignment, etc.)." The pledge must be signed by the student. For more information, visit www.virginia.edu/honor.

Special Needs

It is my goal to create a learning experience that is as accessible as possible. If you anticipate any issues related to the format, materials, or requirements of this course, please meet with me outside of class so we can explore potential options. Students with disabilities may also wish to work with the Student Disability Access Center to discuss a range of options to removing barriers in this course, including official accommodations. Please visit their website for information on this process and to apply for services online: sdac.studenthealth.virginia.edu. If you have already been approved for accommodations through SDAC, please send me your accommodation letter and meet with me so we can develop an implementation plan together.