CSE 398/498 BIG DATA ANALYTICS

Fall 2020 • 2:05 pm – 3:20 pm MW • Location TBD[1]

Instructor **Professor Daniel Lopresti**
Email dal9@lehigh.edu ~ Ext 85782 ~ Office Hours TBD

Texts *Advanced Analytics with Spark: Patterns for Learning from Data at Scale (2nd Edition)* by Sandy Ryza, Uri Laserson, Sean Owen, and Josh Wills, O'Reilly Media, 2015, ISBN-13: 978-1491972953. (primary)

Others TBD

CourseSite Materials and discussion forums will be available @ http://coursesite.lehigh.edu/

Grading

● Homework assignments	50 points (25%)
● Class participation	50 points (25%)
● Final project, presentation, and write-up	100 points (50%)

Notes This 3-credit project course gives a practical working knowledge of large-scale data analysis using the popular open source Apache Spark framework. Spark provides a powerful model for distributing programs across clusters of machines and elegantly supports patterns that are commonly employed in big data analytics, including classification, collaborative filtering, and anomaly detection, among others.

Enrollment in this course is limited and requires permission of the instructor. Please note that this is not a basic course on data mining, machine learning, or distributed computing; it assumes you already know something about these topics and/or you can pick up the necessary details on your own. The course also assumes you already have substantial programming experience in one or more high-level languages.

Accommodations for Students with Disabilities If you have a disability for which you are or may be requesting accommodations, please contact both your instructor and the Office of Academic Support Services, Williams Hall, Suite 301 (610-758-4152) as early as possible in the semester. You must have documentation from the Academic Support Services office before accommodations can be granted.

Principles of Our Equitable Community Lehigh University endorses The Principles of Our Equitable Community[2]. We expect each member of this class to acknowledge and practice these Principles. Respect for each other and for differing viewpoints is a vital component of the learning environment inside and outside the classroom.

[1] This hands-on project course will be taught using live (synchronous) Zoom video conferencing. It will also employ online resources that are accessible remotely from anywhere in the world, including access to the CSE Department computing network and Lehigh's CourseSite system. The time for class meetings will be determined by polling the students who are registered in the course to avoid conflicts, accommodate time zone differences, and identify times that are convenient for everyone in the class. This class can be taken entirely remotely, but if permitted, optional in-person meetings may also be arranged on campus for students local to Lehigh. We will always follow all federal, state, and local safety guidelines.

[2] http://www.lehigh.edu/~inprv/initiatives/PrinciplesEquity_Sheet_v2_032212.pdf

CSE 398/498 BIG DATA ANALYTICS

Fall 2020 • 2:05 pm – 3:20 pm MW • Location TBD

Academic Integrity

The work you submit in CSE 398/498 must be entirely your own. While I encourage you to discuss basic concepts with others, plagiarism is never acceptable. Such cases will be referred to the University Committee on Discipline and, if you are found guilty, you may be given the failing grade WF in the course. If you have questions about this policy at any point, ask me. It is far better to be safe than sorry when your academic career may be on the line.

Tentative Course Schedule

Date	Topic	Readings*	Note
Week 1	Course Intro		
Week 2	Hands-on Intro to Scala and Spark	Chapters 1, 2	
Week 3	Recommending Music and the Audioscrobbler Data Set	Chapter 3	
Week 4	Predicting Forest Cover with Decision Trees	Chapter 4	
Week 5	Anomaly Detection in Network Traffic with K-means Clustering	Chapter 5	
Week 6	Understanding Wikipedia with Latent Semantic Analysis	Chapter 6	
Week 7	Analyzing Co-occurrence Networks with GraphX	Chapter 7	
Week 8	Geospatial and Temporal Data Analysis on the NYC Taxi Trip Data	Chapter 8	
Week 9	Estimating Financial Risk through Monte Carol Simulation	Chapter 9	
Week 10	Analyzing Genomics Data and the BDG Project	Chapter 10	
Week 11	Analyzing Neuroimaging Data with PySpark and Thunder	Chapter 11	
Week 12	Special Topics		
Week 13	Special Topics		
Week 14	Final Project Presentations		
Week 15	Course Wrap-Up		

* All readings are taken from *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*. In addition to reading each chapter along with the associated supplemental materials assigned throughout the course, you will work through the programming examples in the book on your own. After you have programmed what you find in the book, you will implement your own extensions and enhancements which you will be asked to demonstrate for the class. Creativity will be rewarded! Class participation will constitute a significant portion of your grade.