

Big Data Analytics

MET CS 777 SUM2 On-Campus

Classes: Tuesdays 6-9:30 pm and Thursdays 6-9:30 pm classroom CAS 208
(from 7/5/22 to 8/11/22)

Prof. Dimitar Trajanov, Ph.D.

Mail: dtrajano@bu.edu

Office hours: Thursday 3-5 pm on my Zoom link or by appointment

Course Description

This course is an introduction to large-scale data analytics. Big Data analytics is the study of how to extract actionable, non-trivial knowledge from a massive number of data sets. This class will focus both on the cluster computing software tools and programming techniques used by data scientists and the important mathematical and statistical models used in learning from large-scale data processing. On the tool's side, we will cover the basic systems and techniques to store large volumes of data and modern systems for cluster computing based on MapReduce patterns such as Hadoop MapReduce, Apache Spark, and Flink.

Students will implement data mining algorithms and execute them on real cloud systems like Amazon AWS, Google Cloud, or Microsoft Azure by using educational accounts. On the data mining models side, this course will cover the main standard supervised and unsupervised models and will introduce improvement techniques on the model side.

Course Prerequisites

- We expect you to have a solid background in Python programming and understand basic statistics and machine learning. The following classes are required/recommended: MET CS 521, MET CS 544 and MET CS 555, or MET CS 677.
- If you do not have the required/recommended courses, you need the instructor's consent.
- This class includes topics from Cloud Computing, Parallel Processing, and Machine Learning, making the course very compact for a six-week online course.
- To implement the assignments, students need to have excellent knowledge of the Python programming language and some basic Linux knowledge. Assignments are very time-consuming, and you should take this course when you have at least 20 hours per week.

Learning Objectives

By completing this course, you will be able to:

- Explain the main challenges of Big Data Processing.
- Run a Big Data Processing pipeline on Google Cloud (or Amazon AWS).
- Implement Big Data code in Apache Spark (in PySpark).
- Run Supervised and Unsupervised machine learning on Large-Scale Data.

Laptop Requirement

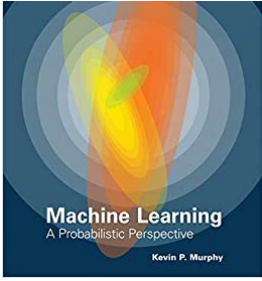
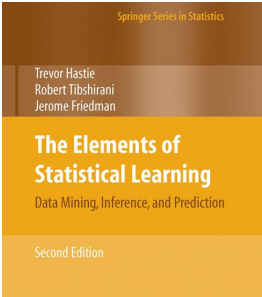
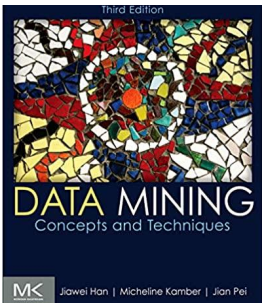
Students should have a personal laptop. We will use laptops to write Python programs and do the quizzes in the classroom. Also, for the Final exam, Laptops are required.

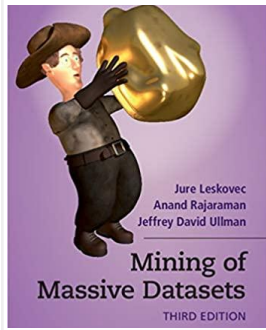
Materials

Required Book

There is no required textbook for the class. There are detailed lecture notes, and all class material will be conveyed during the lecture.

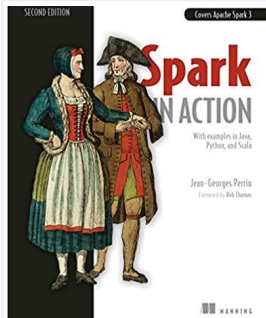
Recommended Books

	<p>Murphy, K. (2012). Machine learning: a probabilistic perspective The MIT Press</p> <p>ISBN-13: 978-0262018029</p>
	<p>Hastie, T. and Tibshirani, R. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer-Verlag.</p> <p>ISBN-13: 978-0-387-84858-7</p> <p>This book is available for PDF download.</p>
	<p>Han, J., Kamber, M., Pei, J. (2009). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.</p> <p>ISBN-13: 978-9380931913</p>

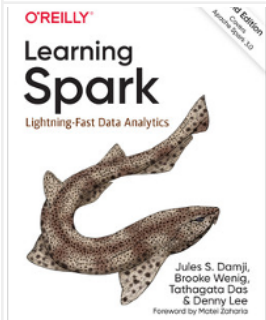


Leskovec, J. Rajaraman, A., Ullman, J. (2014). Mining of massive datasets.
Cambridge University Press.
By agreement with the publisher, you can [download the book](#) for free from this page

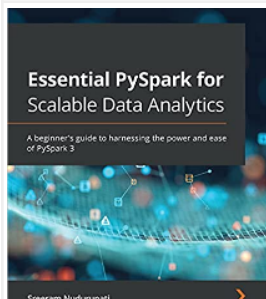
Other Materials and Resources



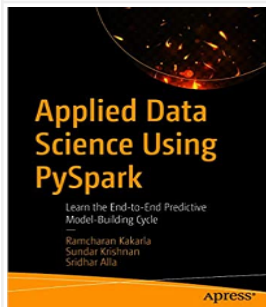
Perrin, J. (2020). Spark in action (2nd ed.). (Covers Apache Spark 3 with examples in Java, Python, and Scala)
O'Reilly Media Inc.




Damji, J., Wenig, B., Das, T., Lee, D. (2020). Learning spark (2nd ed.)
O'Reilly Media Inc.



Nudurupati, S. (2021). Essential PySpark for scalable data analytics: A beginner's guide to harnessing the power and ease of PySpark 3
Packt Publishing



Ramcharan, K., Sundar, K., Alla, S. (2020). Applied data science using PySpark: Learn the end-to-end predictive model-building cycle
Apress

	Main Apache Spark documentation website
---	---

GitHub

This course has a **GitHub repository** (<https://github.com/trajanov/BigDataAnalytics>) with all of the course code examples.

Course website

This course will use the Blackboard Learn site. Students are required to have a BU ID and password to log in. If you do not have a BU ID yet, note that this takes some time so be sure to start this process well before class starts. The BlackBoard site is <https://learn.bu.edu>

Usage of Cloud Machines

In this class, we use real-world cloud systems existing on Google Cloud (or Amazon AWS). You will receive educational credit coupons or credited access to such cloud systems. You should never use your private account or use your credit card for this class assignment. You will receive enough education credits to run successful assignments on Google Cloud.

The credit amount is 50 USD for Google Cloud. You should use only this amount to finish your assignments. This would be more than enough to finish the assignments, learn how Google Cloud (or AWS) works, and have your first enjoyable experience with it. You can choose different numbers of Machines and different configurations of those machines. And each will cost you differently!

Since this is real money, it makes sense to develop your code and run your jobs locally, on your laptop, using the small data set (we will provide two types of the same data set, small and big). Once things are working, you'll then move to Amazon AWS or Google Cloud. We will ask you to run your Spark jobs over the "real" data using a set of cluster machines.

Class Policies

Assignment Completion & Late Work

All assignments should be submitted on time.

Work submitted late without any reason provided will result in a grade deduction:

- 10% penalty for 24 hours late
- 20% penalty for 48 hours late
- After 48 hours, the assignments are not accepted

Attendance & Absences

We recognize that emergencies occur in professional and personal lives. If an emergency prevents your completion of homework by a deadline, please notify your instructor. This must be done before the deadline (unless the emergency makes this impossible, of course). Additional documentation may be requested. Work submitted late without any reason provided will result in a grade deduction: we want to be fair to everyone in this process, including the vast majority of you who sacrifice so much to submit your homework on time in this demanding schedule.

Academic Conduct Code

Cheating and plagiarism will not be tolerated in any Metropolitan College course. They will result in no credit for the assignment or examination and may lead to disciplinary actions.

Please take the time to review the Student Academic Conduct Code:

http://www.bu.edu/met/metropolitan_college_people/student/resources/conduct/cod_e.html.

This should not be understood as discouragement from discussing the material or your particular approach to a problem with other students in the class. On the contrary – you should share your thoughts, questions, and solutions. Naturally, if you choose to work in a group, you will be expected to come up with more than one highly original solution rather than the same mistakes.

Academic Misconduct Regarding Programming

In a programming class like ours, there is sometimes a very fine line between “cheating” and acceptable and beneficial interaction between peers. Thus, it is essential that you fully understand what is and what is not allowed in collaboration with your classmates. We want to be 100% precise, so there can be no confusion.

The rule on collaboration and communication with your classmates is very simple: you cannot transmit or receive code from or to anyone in the class in any way—visually (by showing someone your code), electronically (by emailing, posting, or otherwise sending someone your code), verbally (by reading code to someone) or in any other way we have not yet imagined. Any other collaboration is acceptable.

The rule on collaboration and communication with people who are not your classmates (or your TAs or instructor) is also very simple: it is not allowed in any way, period. This disallows (for example) posting any questions of any nature to programming forums such as StackOverflow. As far as going to the web and using Google, we will apply the “two-line rule.” Go to any web page you like and do any search you like. But you cannot take more than two lines of code from an external resource and include it in your assignment in any form. Note that changing variable names or otherwise transforming or obfuscating code you found on the web does not render the “two-line rule” inapplicable. It is still a violation to obtain more than two lines of code from an external resource and turn it in, whatever you do to those two lines after you first obtain them.



Furthermore, you should cite your sources. Add a comment to your code that includes the URL(s) you consulted when constructing your solution. This turns out to be very helpful when you're looking at something you wrote a while ago, and you need to remind yourself what you were thinking.

Grading Criteria

Please check the Study Guide in the syllabus for Live Classroom dates and specific due dates for assignments and assessments.

Grading Structure and Distribution

The grade for the course is determined by the following:

Activity	Percentages
5 x Homework Assignments	40%
5 x Weekly Quizzes	20%
Term Project and Presentation	10%
Final Exam	30%

Assignments

Homework assignments are focused on applying theory learned in the week's module to a set of data and analyzing that data in PySpark. Weekly homework assignments will focus on implementing data processing and machine learning algorithms in Apache Spark (PySpark). You will use Google Cloud to run your Spark code on large data sets. Free of charge usage credits for Google Cloud will be provided through Education accounts.

Due Time: At the end of each module (Please check the Study Guide or the Syllabus for the specific due date).

Where to submit: The "Assignments" section in the left-hand course menu.

Weekly Quizzes

Quizzes will evaluate students' understanding of concepts presented in the previous week's module. Students should ensure adequate preparation. Doing well on the quiz will not be possible without first reviewing the course material in-depth, attempting to understand all examples, and testing yourself. There are five quizzes.

Term Project and Presentation

At the end of this course, you will work on your own Big Data project. You will work on a large data set and analyze and train machine learning algorithms. You will present your project in the form of a 10 minutes presentation. Clear project development guidelines will be provided in the course content in the "Assignment" section.

Final Exam

The Final Exam will be an open book/open notes, and its duration is three hours. The exam features a combination of multiple-choice, essay, and coding tasks.

Translation between letter grades and percentages.

A (Excellent)	95-100
A- (Excellent; minor improvement needed)	90-94.99
B+ (Very good)	87-89.99
B (Good)	83–86.99
B- (Good; some improvements needed)	80-82.99
C+ (Satisfactory; some significant improvements needed)	77-79.99
C (Satisfactory; significant improvements needed)	73–86.99
C- (Satisfactory; significant improvements required)	70-82.99
D (Many significant improvements required)	65
Unacceptable	0

Important Dates: Add/drop

Standard six-week course in Summer Session 2 (SUM2)

- Course start date: Tuesday, July 5, 2022,
- Last day to add: Monday, July 11
- Last day to drop without a “W” grade: Monday, July 11.
- Last day to drop with a “W” grade: Thursday, July 28.
- Course end date: Thursday, August 11.

Class Meetings, Lectures & Assignments

Date	Module	Topics	Quiz date	Assignment due date
5 Jun 2022	Module 1 Introduction to Big Data Processing	<ul style="list-style-type: none"> • Introduction to Big Data Analytics. What is Big Data? What are the challenges? • Introduction to Apache Hadoop and MapReduce. Apache Spark. • Spark programming. (Python and PySpark) • Spark - Resilient Distributed Dataset (RDDs). 	12 Jun 2022	14 Jun 2022
7 Jun 2022				
12 Jun 2022	Module 2 Large-Scale Data Processing With PySpark	<ul style="list-style-type: none"> • Spark - RDDs, DataFrames, Spark SQL • PySpark + NumPy + SciPy, Code Optimization, Cluster Configurations • Linear Algebra Computation in Large Scale. • Distributed File Storage Systems 	19 Jun 2022	21 Jun 2022
14 Jun 2022				
19 Jun 2022	Module 3 Data Modeling and Optimization Problems	<ul style="list-style-type: none"> • Introduction to modeling: numerical vs. probabilistic vs. Bayesian • Introduction to Optimization Problems • Batch and stochastic Gradient Descent • Newton's Method • Expectation-Maximization, • Markov Chain Monte Carlo (MCMC) 	26 Jun 2022	28 Jun 2022
21 Jun 2022				
26 Jun 2022	Module 4 Large-Scale Supervised Learning	<ul style="list-style-type: none"> • Introduction to Supervised learning • Generalized Linear Models and Logistic Regression • Regularization • Support Vector Machine (SVM) and the kernel trick • Outlier Detection • Spark ML library 	2 Aug 2022	3 Aug 2022
28 Jun 2022				
2 Aug 2022	Module 5 Large-Scale Unsupervised Learning	<ul style="list-style-type: none"> • Introduction to Unsupervised learning • K-means / K-medoids • Gaussian Mixture Models • Dimensionality Reduction • Spark MLlib for Unsupervised Learning 	4 Aug 2022	6 Aug 2022
4 Aug 2022	Module 6 Large Scale Text Mining	<ul style="list-style-type: none"> • Latent Semantic Indexing • Topic models • Latent Dirichlet Allocation • Spark ML library for NLP 	No quiz	No assignment
9 Aug 2022		Team Project Presentations		
11 Aug 2022		Final Exam		

**Lectures, Readings, and Assignments are subject to change and will be announced in class as applicable within a reasonable time frame.*

Instructor Biography



Prof. Dimitar Trajanov, Ph.D. is Visiting Research Professor at Boston University and Head of the Department of Information systems and network technologies at the Faculty of Computer Science and Engineering - ss. Cyril and Methodius University—Skopje. From March 2011 until September 2015, he was the founding Dean of the Faculty of Computer Science and Engineering. In his tenure, the Faculty has become the largest technical Faculty in Macedonia. Dimitar Trajanov is the leader of the Regional Social Innovation Hub, established in 2013 in cooperation with the United Nations

Development Programme. His professional experience includes working as a Data Science Consultant for one of the largest Pharmaceutical companies, a Data Science consultant for UNDP in North Macedonia, and a software architect in a couple of startups. Dimitar Trajanov is the author of more than 170 journal and conference papers and seven books. He has been involved in more than 70 research and industry projects.