MET CS 777

Big Data Analytics
2026 Summer 2 (June 30–August 17)
4 credits

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 and Apache Spark.

Students will implement data mining algorithms and execute them on real cloud systems like Google Cloud, Amazon AWS, 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, which make the course very compact for a six-week online course.

To implement the assignments, students need to have excellent knowledge of 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 successfully 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

Instructional Team

Instructor: Dimitar Trajanov, PhD
Computer Science Department
Metropolitan College
Boston University
dtrajano@bu.edu

Prof. Dimitar Trajanov, Ph.D., is a Visiting Research Professor at Boston University and full professor 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, and in his tenure, the Faculty became the largest technical Faculty in Macedonia. Dimitar Trajanov is the leader of the Regional Social Innovation Hub, which was established in 2013 as a cooperation between UNDP and the Faculty of Computer Science and Engineering.

His professional experience includes working as a Senior 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 200 journal and conference papers and seven books. He has been involved in more than 80 research and industry projects

Materials

Required Book

There is no required textbook for the class. All class material will be conveyed during the lecture. The following recommended books and materials are available online.

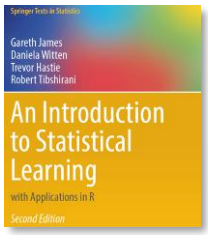
Materials and Digital Learning Assets

Online Materials

The course adopts a "learning by example" approach, offering a hands-on approach through more than 50 Python notebooks and scripts. This educational strategy fosters real-world relevance, deepens theoretical understanding, and hones both hard and soft skills essential to the field.

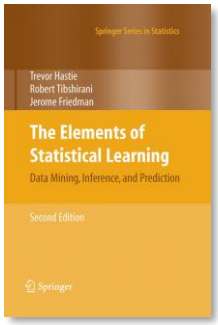
Please download the CS 777 Learning by Example Guide (PDF available in Syllabus and Resources folder on course Blackboard site).

Spark Online Guides



The official PySpark documentation is maintained by the developers themselves, ensuring that it is always up-to-date and accurate with the latest changes in the platform.

The following guides are highly recommended for anyone learning or working with PySpark:



- PySpark Getting Started: This is the Getting Started guide that summarizes the basic steps required to setup and get started with PySpark.

- RDD Programming Guide: This guide provides an overview of Spark basics, including RDDs (the core but older API), accumulators, and broadcast variables.

- Spark SQL, Datasets, and DataFrames: This guide is focused on processing structured data with relational queries using a newer API than RDDs.

- MLlib: This guide provides detailed information on how to apply machine learning algorithms in PySpark.

- Spark Python API. This is the PySpark API documentation, which provides detailed information on the PySpark API, including its modules, classes, and methods.

- Structured Streaming Programming Guide. This guide is focused on Structured Streaming which is a scalable and fault-tolerant stream processing engine built on the Spark SQL engine.

Books

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2021)
An introduction to statistical learning (2nd ed.)
Springer-Verlag

This book is available for PDF download

Hastie, T. and Tibshirani, R. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.).
Springer-Verlag.

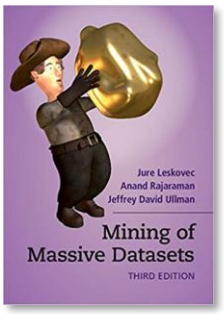
ISBN-13: 978-0-387-84858-7

This book is available for PDF download.

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.

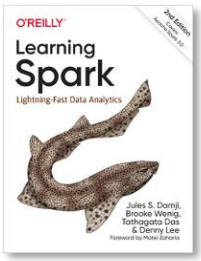
Missing details



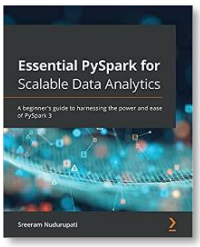
Other Materials and Resources

Spark Programming

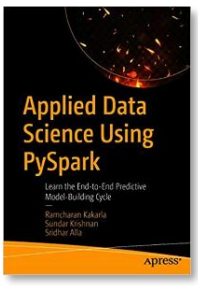
Spark in Action
Perrin, J. (2020). Spark in action (2nd ed.). (Covers Apache Spark 3 with examples in Java, Python, and Scala)
O'Reilly Media Inc.



Learning Spark
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** with all of the course code examples.

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.

Study Guide

Module 1 Study Guide and Deliverables

(June 30-July 6)
MapReduce Data Processing Pattern

Readings	Online lecture material topics: • Definition of Big Data and its five V's (Volume, Velocity, Variety, Veracity, Value) • Overview of industry applications and emerging trends • Fundamental challenges in handling massive, fast-moving datasets. • Introduction to Apache Hadoop and MapReduce • Apache Spark • Spark Programming (Python and PySpark). PySpark advantages for data scientists and Python developers. • The concept of Resilient Distributed Dataset (RDDs)
Assignments	Assignment 1 due Sunday, July 12 at 6:00 AM ET
Assessments	Quiz 1 due Friday, July 10 at 6:00 AM ET

Live Classroom	• Tuesday, June 30 from 5:30–7:00 PM ET • Saturday July 4 from 11:00–12:30 PM ET • Live Office (facilitator sessions): to be scheduled each weekend

Module 2 Study Guide and Deliverables

(July 7-July 13)

Large-Scale Data Processing and Storage

Readings	Online lecture material topics: • Setting up PySpark Environment: Installation and basic configuration (Local mode). • Running Spark programs on the Cloud • Spark - RDDs, DataFrames, Spark SQL • PySpark + NumPy • Code Optimization, Cluster Configurations • Distributed File Storage Systems
Assignments	Assignment 2 due Wednesday, July 15 at 6:00 AM ET
Assessments	Quiz 2 due Tuesday, July 14 at 6:00 AM ET
Live Classroom	• Tuesday, July 7 from 5:30–7:00 PM ET • Saturday July 11 from 11:00–12:30 PM ET • Live Office (facilitator sessions): to be scheduled each weekend

Module 3 Study Guide and Deliverables

(July 14-July 20)

Data Modeling and Optimization Problems

Readings	Online lecture material topics: • Introduction to Modeling: Numerical vs. Probabilistic • Introduction to Optimization Problems • Batch and Stochastic Gradient Descent • Processing real-time data: Spark Structured Streaming
Assignments	Assignment 3 due Wednesday, July 22 at 6:00 AM ET
Assessments	Quiz 3 due Tuesday, July 21 at 6:00 AM ET
Live Classroom	• Tuesday, July 14 from 5:30–7:00 PM ET • Saturday July 18 from 11:00–12:30 PM ET

	• Live Office (facilitator sessions): to be scheduled each weekend

Module 4 Study Guide and Deliverables

(July 21-July 27)

Large-Scale Supervised Learning

Readings	Online lecture material topics: • Introduction to Supervised Learning • Generalized Linear Models and Logistic Regression • Developing a Distributed ML Regression Model • Overfitting and Regularization • Dealing with unbalanced data • Spark ML library
Assignments	Assignment 4 due Wednesday, July 29 at 6:00 AM ET
Assessments	Quiz 4 due Tuesday, July 28 at 6:00 AM ET
Live Classroom	• Tuesday, July 21 from 5:30–7:00 PM ET • Saturday July 25 from 11:00–12:30 PM ET • Live Office (facilitator sessions): to be scheduled each weekend

Module 5 Study Guide and Deliverables

(July 28-Aug 3)

Unsupervised Learning on Large-Scale Data

Readings	Online lecture material topics: • Introduction to Unsupervised learning • Clustering: K-means, Hierarchical Clustering and Gaussian Mixture Models • Recommender Systems • Matrix Factorization • Dimensionality Reduction • Latent variables •
Assignments	• Term Project Proposal due Tuesday, Aug 4 at 6:00 AM ET • Assignment 5 due Wednesday, Aug 5 at 6:00 AM ET
Assessments	Quiz 5 due Tuesday, Aug 4 at 6:00 AM ET
Live Classroom	• Tuesday, July 28 from 5:30–7:00 PM ET • Saturday Aug 1 from 11:00–12:30 PM ET • Live Office (facilitator sessions): to be scheduled each weekend

Module 6 Study Guide and Deliverables

(Aug 4-Aug 10)

Text Mining

Readings	Online lecture material topics: • Natural Language Processing (NLP) • Using PySpark for distributed text processing. • Topic Modeling: Latent Dirichlet Allocation • Large Language Models (LLM) • Using ChatGPT API in Spark • AI Agents in SparkCode Generation for Data Analysis •
Assignments	Term Project due Tuesday, August 11 at 6:00 AM ET
Assessments	None
Live Classroom	• Tuesday, Aug 4 from 5:30–7:00 PM ET • Saturday Aug 8 from 11:00–12:30 PM ET • Live Office (facilitator sessions): to be scheduled each weekend • Final Exam Prep session: to be scheduled
Course Evaluation	Please complete the course evaluation once you receive an email or Blackboard notification indicating the evaluation is open. Your feedback is important to MET, as it helps us make improvements to the program and the course for future students.

Final Exam Details

The Final Exam is a proctored exam available from Wednesday, August 12, at 6:00 AM ET to Saturday, August 15, at 11:59 PM ET. The Computer Science department requires that all final exams be administered using an online proctoring service that you will access via your course in Blackboard. Additional information regarding your proctored exam will be forthcoming from the Assessment Administrator.

The Final Exam will be open book/open notes and is accessible only during the final exam period. Your proctor will enter the password to start the exam.

Final Exam duration: three hours.

The exam features a combination of multiple choice, essay, and file response questions.

Grading Information

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:

Overall Grading Percentages	
5 × Homework Assignments	40%
5 × 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 implementation of 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.

Weekly Quizzes

Quizzes will evaluate students understanding of concepts presented in the corresponding week's module. Students should ensure adequate preparation before starting the quiz. It will not be possible to do well on the quiz without first reviewing

the course material in depth and attempting to understand all examples and test yourself questions. It is recommended that you complete the quiz after you feel comfortable with the material and have asked any questions that you may have had.

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, analyze and train machine learning algorithms. You will present your project in the form of a 15-minutes online presentation. Clear project development guidelines will be provided in the course content in the "Assignment" section.

Final Exam

The Computer Science department requires that all final exams be administered using an online proctoring service that you will access via your course in Blackboard. Additional information regarding your proctored exam will be forthcoming from the Assessment Administrator.

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

* Grading scale may be adjusted at the discretion of the course instructor.

Lateness

We recognize that emergencies occur in professional and personal lives. If an emergency occurs that prevents your completion of homework by a deadline, please notify your facilitator or instructor. This must be done in advance of 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 Policy

Please visit Metropolitan College's website for the full text of the department's *Academic Conduct Code*.

A Definition of Plagiarism

“The academic counterpart of the bank embezzler and of the manufacturer who mislabels products is the plagiarist: the student or scholar who leads readers to believe that what they are reading is the original work of the writer when it is not. If it could be assumed that the distinction between plagiarism and honest use of sources is perfectly clear in everyone’s mind, there would be no need for the explanation that follows; merely the warning with which this definition concludes would be enough. But it is apparent that sometimes people of goodwill draw the suspicion of guilt upon themselves (and, indeed, are guilty) simply because they are not aware of the illegitimacy of certain kinds of “borrowing” and of the procedures for correct

identification of materials other than those gained through independent research and reflection."

"The spectrum is a wide one. At one end there is a word-for-word copying of another's writing without enclosing the copied passage in quotation marks and identifying it in a footnote, both of which are necessary. (This includes, of course, the copying of all or any part of another student's paper.) It hardly seems possible that anyone of college age or more could do that without clear intent to deceive. At the other end there is the almost casual slipping in of a particularly apt term which one has come across in reading and which so aptly expresses one's opinion that one is tempted to make it personal property."

"Between these poles there are degrees and degrees, but they may be roughly placed in two groups. Close to outright and blatant deceit-but more the result, perhaps, of laziness than of bad intent-is the patching together of random jottings made in the course of reading, generally without careful identification of their source, and then woven into the text, so that the result is a mosaic of other people's ideas and words, the writer's sole contribution being the cement to hold the pieces together. Indicative of more effort and, for that reason, somewhat closer to honest, though still dishonest, is the paraphrase, and abbreviated (and often skillfully prepared) restatement of someone else's analysis or conclusion, without acknowledgment that another person's text has been the basis for the recapitulation."

The paragraphs above are from H. Martin and R. Ohmann, *The Logic and Rhetoric of Exposition, Revised Edition.* Copyright 1963, Holt, Rinehart and Winston.

Academic Conduct Code

I. **Philosophy of Discipline**

The objective of Boston University in enforcing academic rules is to promote a community atmosphere in which learning can best take place. Such an atmosphere can be maintained only so long as every student believes that his or her academic competence is being judged fairly and that he or she will not be put at a disadvantage because of someone else's dishonesty. Penalties should be carefully determined so as to be no more and no less than required to maintain the desired atmosphere. In defining violations of this code, the intent is to protect the integrity of the educational process.

II. **Academic Misconduct**

Academic misconduct is conduct by which a student misrepresents his or her academic accomplishments, or impedes other students' opportunities of being judged fairly for their academic work. Knowingly allowing others to represent your work as their own is as serious an offense as submitting another's work as your own.

III. **Violations of this Code**

Violations of this code comprise attempts to be dishonest or deceptive in the performance of academic work in or out of the classroom, alterations of academic records, alterations of official data on paper or electronic resumes, or unauthorized collaboration with another student or students. Violations include, but are not limited to:

A. **Cheating on examination**. Any attempt by a student to alter his or her performance on an examination in violation of that examination's stated or commonly understood ground rules.
B. **Plagiarism.** Representing the work of another as one's own. Plagiarism includes but is not limited to the following: copying the answers of another student on an examination, copying or restating the work or ideas of another person or persons in any oral or written work (printed or electronic) without citing the appropriate source, and collaborating with someone else in an academic endeavor without acknowledging his or her contribution. Plagiarism can consist of acts of commission-appropriating the words or ideas of another-or omission failing to acknowledge/document/credit the source or creator of words or ideas (see below for a detailed definition of plagiarism). It also includes colluding with someone else in an academic endeavor without acknowledging his or her contribution, using audio or video footage that comes from another source (including work done by another student) without permission and acknowledgement of that source.
C. **Misrepresentation or falsification of data** presented for surveys, experiments, reports, etc., which includes but is not limited to: citing authors that do not exist; citing interviews that never took place, or field work that was not completed.
D. **Theft of an examination**. Stealing or otherwise discovering and/or making known to others the contents of an examination that has not yet been administered.
E. **Unauthorized communication during examinations**. Any unauthorized communication may be considered prima facie evidence of cheating.
F. **Knowingly allowing another student to represent your work as his or her own**. This includes providing a copy of your paper or laboratory report to another student without the explicit permission of the instructor(s).
G. **Forgery, alteration, or knowing misuse of graded examinations, quizzes, grade lists, or official records of documents**, including but not limited to transcripts from any institution, letters of recommendation, degree certificates, examinations, quizzes, or other work after submission.
H. **Theft or destruction of examinations or papers** after submission.
I. **Submitting the same work in more than one course** without the consent of instructors.
J. **Altering or destroying another student's work or records**, altering records of any kind, removing materials from libraries or offices without consent, or in any way interfering with the work of others so as to impede their academic performance.
K. **Violation of the rules governing teamwork**. Unless the instructor of a course otherwise specifically provides instructions to the contrary, the following rules

apply to teamwork: 1. No team member shall intentionally restrict or inhibit another team member's access to team meetings, team work-in-progress, or other team activities without the express authorization of the instructor. 2. All team members shall be held responsible for the content of all teamwork submitted for evaluation as if each team member had individually submitted the entire work product of their team as their own work.

L. **Failure to sit in a specifically assigned seat during examinations.**
M. **Conduct in a professional field assignment that violates the policies and regulations of the host school or agency.**
N. **Conduct in violation of public law occurring outside the University that directly affects the academic and professional status of the student, after civil authorities have imposed sanctions.**
O. **Attempting improperly to influence the award of any credit, grade, or honor.**
P. **Intentionally making false statements to the Academic Conduct Committee or intentionally presenting false information to the Committee.**
Q. **Failure to comply with the sanctions imposed under the authority of this code.**

Important Message on Final Exams

Dear Boston University Computer Science Online Student,

As part of our ongoing efforts to maintain the high academic standard of all Boston University programs, including our online MSCIS degree program, the Computer Science Department at Boston University's Metropolitan College requires that each of the online courses includes a proctored final examination.

By requiring proctored finals, we are ensuring the excellence and fairness of our program. The final exam is administered online.

Specific information regarding final-exam scheduling will be provided approximately two weeks into the course. This early notification is being given so that you will have enough time to plan for where you will take the final exam.

I know that you recognize the value of your Boston University degree and that you will support the efforts of the University to maintain the highest standards in our online degree program.

Thank you very much for your support with this important issue.

Regards,

Professor Lou Chitkushev, Ph.D.
Associate Dean for Academic Affairs
Boston University Metropolitan College

Who's Who: Roles and Responsibilities

You will meet many BU people in this course and program. Some of these people you will meet online, and some you will communicate with by email and telephone. There are many people behind the scenes, too, including instructional designers, faculty who assist with course preparation, and video and animation specialists.

People in Your Online Course in Addition to Your Fellow Students

Your Facilitator. Our classes are divided into small groups, and each group has its own facilitator. We carefully select and train our facilitators for their expertise in the subject matter and their excellence in teaching. Your facilitator is responsible for stimulating discussions in pedagogically useful areas, for answering your questions, and for grading homework assignments, discussions, term projects, and any manually graded quiz or final-exam questions. If you ask your facilitator a question by email, you should get a response within 24 hours, and usually faster. If you need a question answered urgently, post your question to one of the urgent help topics, where everyone can see it and answer it.

Your Professor. The professor for your course has primary responsibility for the course. If you have any questions that your facilitator doesn't answer quickly and to your satisfaction, then send your professor an email in the course, with a cc to your facilitator so that your facilitator is aware of your question and your professor's response.

Your Lead Faculty and Student Support Administrator, Jen Sullivan. Jen is here to ensure you have a positive online experience. You will receive emails and announcements from Jen throughout the semester. Rachel represents Boston University's university services and works for BU Virtual. She prepares students for milestones such as course launch, final exams, and wrapping up a course and preparing for the next launch. She is a resource to both students and faculty. For example, Jen can direct your university questions and concerns to the appropriate party. She also handles general questions regarding Online Campus functionality for students, faculty, and facilitators, but she does not provide tech support. She is enrolled in all classes and can be contacted within the course through Online Campus email as it is running. You can also contact her by external email at jensul@bu.edu.

People Not in Your Online Course

Although you will not normally encounter the following people in your online course, they are central to the program. You may receive emails or phone calls from them, and you should feel free to contact them.

Your Computer Science Department Online Program Coordinator, Michelle Younger. Michelle administers the academic aspects of the program, including admissions and registration. You

can ask her questions about the program, registration, course offerings, graduation, or any other program-related topic. She can be reached at metcsol@bu.edu or (617) 353-2566.

Professor Guanglan Zhang, Computer Science Department Chairman. You can reach Professor Zhang at guanglan@bu.edu or at 617-358-5688.

Professor Lou T. Chitkushev, Associate Dean for Academic Affairs, Metropolitan College. Dr. Chitkushev is responsible for the academic programs of Metropolitan College. Contact Professor Chitkushev with any issues that you feel have not been addressed adequately. The customary issue-escalation sequence after your course facilitator and course faculty is Professor Zhang, and then Professor Chitkushev.

Professor Tanya Zlateva, Metropolitan College Dean. Dr. Zlateva is responsible for the quality of all the academic programs at Boston University Metropolitan College.