GEORGETOWN UNIVERSITY | McCourt School of Public Policy

Massive Data Fundamentals

(PPOL 5206)

Syllabus

Spring 2025

Summary Information

PROFESSOR: Jeremy Skog

EMAIL: js5497@georgetown.edu

PHONE: (703) 615-8397

OFFICE HOURS: By appointment

MEETING TIME: Thursdays (6:30 PM – 9:00 PM)

MEETING LOCATION: Room 502; 125 E St. NW

Course Description

Data are everywhere, but sometimes they are too massive to analyze with conventional technology—whether it's a book of calculated values or your laptop. This course addresses these challenges by exploring modern big data methods that leverage high-performance computing, distributed computing, and cloud technologies.

In this hands-on, workshop-style course, you will learn to use cloud computing resources to analyze and manipulate datasets that are too large for a single machine or traditional tools. Core skills include data ingestion, cleaning, transformation, analysis, and modeling, all within the context of big data analytics. You will develop a programmatic and logical approach to addressing big data challenges.

The course emphasizes tools such as SQL, the Hadoop Ecosystem, and Spark, with Python as the primary programming language. You will also explore the historical context of these technologies, gaining insight into why they were developed and how to solve similar modern problems. Additionally, the course covers the evolving legal landscape of cloud computing and data operations, equipping you to navigate the complexities of modern data governance.

Learning Objectives

- Setup, operate and manage big data tools and cloud infrastructure, including Spark, MapReduce, DataBricks, Hadoop on Microsoft Azure and Amazon Web Services
- Understand data regulations and how data lives and is used within the legal environment
- Use ancillary tools that support big data processing, including git, IDEs, SDKs, and Command line interfaces
- Execute a big data analytics project from start to finish: ingest, wrangle, clean, analyze, store, and present
- Develop strategies to break down large problems and datasets into manageable pieces
- Identify broad spectrum resources and documentation to remain current with big data tools and developments
- Communicate and interpret the big data analytics results through written, graphical, and verbal methods

Grading Policies

Final Grades

The final grade in this course will be determined by two presentations, lab results, occasional quizzes, and participation.

Assignment	*Points*
Project 1: Data Environment	
Project 1 Presentation	= 40
Project 1 Briefing Report	= 60
Presentation 2: Grant Proposal	
Project 2 Presentation	= 40
Project 2 Website	= 160
Participation in discussions	= 50
Lab Results	= 100
TOTAL	= 450 points possible]

Attendance Policy:

Attendance in class is expected as learning from your peers is an important part of what the school offers. If you are unable to attend, please contact me via email (preferred).

Late Assignment Policy:

Labs and quizzes are due *by the due date and time provided*. Late assignments will be penalized 20% for every day they are overdue.

Participation Policy:

Participation in class and in discussions on Canvas is encouraged. Please be aware that anything you post online will potentially be public. Classes may be recorded and distributed to students.

Canvas Site

A Canvas course site is set up for this course. Each student is expected to check the site throughout the semester as Canvas will be the primary venue for outside classroom communications between the instructors and the students. Students can access the course site at http://canvas.georgetown.edu/. Support for Canvas is available at (833) 476-1171 or canvas-help@georgetown.edu.

Generative AI Policy

Course AI Policy

Artificial Intelligence tools have long been a part of the data science landscape. The introduction of Large Language Models has popularized the use of generative AI and new uses are being developed every day. I expect that you will need to use these tools in your future endeavors.

Using machine learning (ML) in this course is expected. Generative Business Intelligence (BI) may be used as part of the learning process. Generative Artificial Intelligence (AI) may be used to supplement, but not replace, the learning process. Students should document how they used generative AI or other tools as part of a statement when submitting work.

Georgetown's Policy

Georgetown Honor Council's Standards of Conduct Policy (2024) states:

> "The question of how to acknowledge [AI-generated intellectual work], and whether it is to be allowed at all, is answered by individual course policies. It is, as always, the students' responsibility to be sure that they are following the rules laid out by their professors. Note that, as with all source material, this applies both to work taken directly from the AI generator and to work that has been paraphrased before being used in coursework. If you didn't generate the words yourself, say so by quoting and citing the source; if you generated the words but not the content and ideas, say so by citing the source."

Learning Resources

Skills

Pre-requisites

- Experience with Python. **Note:** We will use Python as the primary API interface to use Apache Spark, through PySpark. We will also use Python for examples in data processing and to interact with SQL. Python SDKs will be used to interact with cloud services.

- Familiarity with data concepts

Some tutorials to brush up on these skills:

- git - the simple guide
- Nico Riedmann's Learn git concepts, not commands
- SQLBolt - Learn SQL with simple, interactive exercises
- The Missing Semester of Your CS Education

Books, Software and Cloud Resources

Readings (for assigned readings)

There is no required textbook for the course as a significant amount of material is available online. I have selected specific chapters, articles, and papers from several sources, and these will be provided to you in PDF format. **I recommend completing the assigned readings prior to the lectures.**

Suggested Software

General Environment

- PowerShell
- SSH
- GitHub Desktop

Databases

- DBeaver
- MySQL
- PostgreSQL

Programming Languages and IDEs

- Python
- Spark
- Java
- VSCode
- PyCharm
- Notepad++
- VI
- Anaconda
- JuPyteR

AWS

- CLI
- SDK

Azure

- CLI
- SDK

Cloud Resources

You will use cloud resources on Microsoft Azure and Amazon Web Services. We will discuss how to set up your accounts and environments in class and lab within the first couple of weeks.

Modules

Schedule for the semester PPOL 5206 Spring 2025

This course is divided into 3 phases: Intro to Big Data and the Cloud, SQL & Data Engineering, and Distributed Computing

Module	Date		Notes
Phase 1: Intro to Big Data and the Cloud			
Intro to Big Data and the Cloud	2025-01-09	Course overview	Overview of course structure, policy discussions, intro to big data concepts, setting up computing environments (AWS, Azure, GCP basics). **LAB:** Setting up your computing environment, & GitHub accounts. “Hello World” GitHub Page.
Intro to Big Data and the Cloud	2025-01-16	Cloud Fundamentals	Scaling technologies, cloud services overview (AWS/Azure), metadata. **LAB:** AWS or Azure accounts. “Hello World” basics in the Cloud.
Intro to Big Data and the Cloud	2025-01-23	Cloud Data Storage	Data lakes vs. warehouses, intro to cloud databases, and cloud SDKs for data management.
Phase 2: SQL & Data Engineering			
SQL & Data Engineering	2025-01-30	SQL Fundamentals	History of SQL, data modeling basics, foundational queries, and working with SQLite.
SQL & Data Engineering	2025-02-06	Advanced SQL & Python	Advanced SQL queries, integrating SQL with Python, parallel processing with Dask. Alternative data formats.
SQL & Data Engineering	2025-02-13	Building Pipelines	ETL/ELT basics, APIs for data ingestion, intro to data engineering workflows in the cloud. DuckDB.
Phase 3: Distributed Computing			
Distributed Computing	2025-02-20	Intro to MapReduce	MapReduce concepts, Hadoop fundamentals, and examples of distributed data workflows.

Schedule for the semester PPOL 5206 Spring 2025

This course is divided into 3 phases: Intro to Big Data and the Cloud, SQL & Data Engineering, and Distributed Computing

Module	Date		Notes
Project Presentations 1	2025-02-27	Interlude: Data Ethics & Disaster Recovery	Data ethics, laws (e.g., GDPR), disaster recovery plans, and cloud security basics. **LAB:** Country presentations.
	2025-03-06	**NO CLASS - Spring Break**	
Distributed Computing	2025-03-13	Big Data with Hadoop and Cloud EMR	Practical session on setting up Hadoop and processing large datasets.
Distributed Computing	2025-03-20	Intro to Apache Spark	Spark architecture and RDDs, setting up Spark with AWS SageMaker/EMR.
Distributed Computing	2025-03-27	Spark SQL & DataFrames	Structured data processing with Spark SQL and the Spark DataFrame API.
Distributed Computing	2025-04-03	Machine Learning	Machine Learning with Spark. Current generative AI/BI technologies.
Distributed Computing	2025-04-10	Scaling Machine Learning	Scaling machine learning with Spark. Snowflake. DataBricks.
	2025-04-17	**NO CLASS - Easter Break**	
Project Presentations 2	2025-04-24	Final Project discussion	***Last class for the semester***

Schedule for the semester PPOL 5206 Spring 2025

This course is divided into 3 phases: Intro to Big Data and the Cloud, SQL & Data Engineering, and Distributed Computing

Module	Date	Notes
	2025-05-10	**Last Day of Finals – Any Project Revisions Due**

Warning

IT IS YOUR RESPONSIBILITY TO MANAGE CLOUD CREDITS AND RESOURCES PROVIDED TO YOU. YOU MUST SHUT DOWN YOUR CLOUD RESOURCES WHEN NOT IN USE.

Learning Activities, Communication and Evaluation

This is a hands-on, practical, workshop style course that provides opportunities to use the tools and techniques discussed in class. Although this is not a programming course per se, there is programming involved.

Lectures and Labs

This course is split into a lecture/lab format, where every class session will have a lecture portion, and most sessions will have an in-class lab portion:

- During the lecture, we will discuss the concepts and techniques as well as the history and development of these big data tools and cloud platforms.
- During the lab sessions, you will be completing exercises and following examples which are designed to show you how to implement ideas and concepts with various tools. We will start the labs in class, but we might not finish. It is your responsibility to complete the labs (**which are part of the grade**).

Lectures may not cover all the material, and some topics will be introduced in the lab or through readings/assignments.

Office Hours

Instructors and TAs are available to meet to answer questions, review material, and support your learning.

Readings

On certain weeks, readings will be assigned to prepare you for the lecture material being presented. These readings should take an hour or less per week.

Online Quizzes

Quizzes will be given a few times during the semester during lab or lecture at random intervals and times. Quizzes ensure you are keeping up with the material presented in the class. Quizzes are meant to be brief and low-stress with a time limit of 5-10 minutes. The material will be drawn from lectures, labs, and readings.

Lab Deliverables

Each lab will have a deliverable. It is essential that you learn the skills presented in the labs so that you can effectively complete the assignments and the big data project. The lab deliverables can sometimes be completed during lab; however, it is your responsibility to complete the deliverable as part of your work outside of lecture/lab time.

Projects

There are two projects that are designed to enhance your learning of the data environment and how data is used throughout the world. I am a strong believer that projects and the hands-on work they involve are one of the best ways to learn this material. My hope is that at the end of the course you will have the beginnings of a portfolio that you can share with future colleagues and employers, as well as refer to for your future self.

Legal Data Environment Project

The legal and physical environments for data vary significantly across the world and are constantly evolving. For this project, students will select a specific country to research and analyze its data infrastructure and legal landscape.

Students will investigate key areas including the country's privacy and data protection laws, presence and operations of major cloud providers, and data center and internet infrastructure. The analysis should also consider market penetration of digital services, regulatory compliance requirements, and the country's readiness for advanced data-driven technologies.

Students will examine how historical context, regional influences, and cultural factors shape the country's approach to data governance. This includes understanding international agreements, economic implications, and the evolving relationship between technology adoption and policy development.

The project culminates in two deliverables: a written briefing report analyzing these elements and their interconnections, and a concise presentation highlighting key findings, critical factors, and future outlook for data systems in the chosen country. Students are encouraged to include relevant comparisons with peer nations to provide context for their analysis.

Big Data Analytics Project

The final project challenges students to demonstrate mastery of big data analytics applied to public policy through hands-on work with a large dataset. Working in teams of up to four members, students will pick a policy area of interest and tackle datasets that could be too large for a standard data analysis process.

The goal for this assignment is to conduct a preliminary analysis of a dataset, laying the groundwork for future research. The final product--a project website and presentation--will be a proof-of-concept presented to a "grant funder" explaining the importance of your project, your initial analysis, and your future goals.

Teams will develop an analysis plan and conduct an initial investigation that demonstrate the core skills developed throughout the semester, including data ingestion and transformation, exploratory analysis, and the application of appropriate statistical techniques. The project emphasizes creating a convincing research proposal that demonstrates both technical capabilities and policy relevance.

The project follows an iterative development process with structured intermediate assignments. These stepping stones focus on specific components – from initial data exploration to policy implications –

while building toward the final deliverable. Each stage incorporates feedback from peers, teaching assistants, and professors to help refine both technical approaches and policy arguments.

Students will create deliverables suitable for both technical and non-technical audiences, including a project website, presentation materials, and technical documentation. The final presentation will simulate a grant funding pitch, where teams must justify their research approach, demonstrate preliminary findings, and outline a compelling path for future investigation. Project timelines and specific deliverable requirements will be detailed during class sessions.

University Policies

Mental Health Resources

Mental health is a serious and growing problem across the world and in universities in particular. Georgetown offers counseling and other services. More information is available here: https://studenthealth.georgetown.edu/mental-health/

Disability

If you believe you have a disability, then you should contact the Academic Resource Center (arc@georgetown.edu) for further information. The Center is located in the Leavey Center, Suite 335 (202-687-8354). The Academic Resource Center is the campus office responsible for reviewing documentation provided by students with disabilities and for determining reasonable accommodations in accordance with the Americans with Disabilities Act (ADA) and University policies. For more information, go to https://academicsupport.georgetown.edu/disability/.

Important Academic Policies and Academic Integrity

McCourt School students are expected to uphold the academic policies set forth by Georgetown University and the Graduate School of Arts and Sciences. Students should therefore familiarize themselves with all the rules, regulations, and procedures relevant to their pursuit of a Graduate School degree. The policies are located at: https://grad.georgetown.edu/academics/current-students/

Warning

I have a ZERO TOLERANCE POLICY and if students are found to be in violation there will be consequences, and the students will be penalized and reported.

Rules

- **In-class labs:** you may collaborate with other students during in-class labs to facilitate collective learning.
- **Group project:** by nature, it is a group project, and collaboration is to be confined within groups. **You may not collaborate across groups.**
- The following collaboration and communication is allowed
    - The discussion of ideas and approaches to problems
    - The discussion and resolution of any technical issue
    - Sharing code snippets (a few lines) for a very specific things

Statement on Sexual Misconduct

Georgetown University and its faculty are committed to supporting survivors and those impacted by sexual misconduct, which includes sexual assault, sexual harassment, relationship violence, and stalking. Georgetown requires faculty members, unless otherwise designated as confidential, to report all disclosures of sexual misconduct to the University Title IX Coordinator or a Deputy Title IX Coordinator. If you disclose an incident of sexual misconduct to a professor in or outside of the classroom (with the exception of disclosures in papers), that faculty member must report the incident to the Title IX Coordinator, or Deputy Title IX Coordinator. The coordinator will, in turn, reach out to the student to provide support, resources, and the option to meet. [Please note that the student is not required to meet with the Title IX coordinator.]. More information about reporting options and resources can be found on the Sexual Misconduct Website: https://sexualassault.georgetown.edu/resourcecenter.

If you would prefer to speak to someone confidentially, Georgetown has a number of fully confidential professional resources that can provide support and assistance. These resources include:

Health Education Services for Sexual Assault Response and Prevention: confidential email sarp@georgetown.edu

Counseling and Psychiatric Services (CAPS): 202.687.6985.

More information about reporting options and resources can be found on the Sexual Misconduct Website.

Provost’s Policy Accommodating Students’ Religious Observances

Georgetown University promotes respect for all religions. Any student who is unable to attend classes or to participate in any examination, presentation, or assignment on a given day because of the observance of a major religious holiday (see below) or related travel shall be excused and provided with the opportunity to make up, without unreasonable burden, any work that has been missed for this reason and shall not in any other way be penalized for the absence or rescheduled work. Students will remain responsible for all assigned work. Students should notify professors in writing at the beginning of the semester of religious observances that conflict with their classes. The Office of the Provost, in consultation with Campus Ministry and the Registrar, will publish, before classes begin for a given term, a list of major religious holidays likely to affect Georgetown students. The Provost and the Main Campus Executive Faculty encourage faculty to accommodate students whose bona fide religious observances in other ways impede normal participation in a course. **Students who cannot be accommodated should discuss the matter with an advising dean.**

The McCourt Policy-Writing Center (MPWC)

The McCourt Policy-Writing Center (MPWC) offers services and resources for students to help improve policy-writing skills. These include workshops, peer review facilitation, open houses, lectures, **online modules**, and **individual appointments**. Click **here** for the full calendar of events, and email mpwc@georgetown.edu for general inquiries.