CSCI 49395: Big Data Analytics and Applied Privacy

Fall 2026: Tuesday/Thursday 1:00 - 2:15 PM, Room: North Building 1001F

Instructor:	Mohammad Hossein Manshaei, mohammad.manshaei@hunter.cuny.edu
WWW:	Course related material can be found at **Brightspace** and **www.bdap-hunter.com** Please check both regularly for course updates!
Office:	E1004, Hunter East
Office Hours:	Tuesday/Thursday 3 – 5 PM or by appointment

Textbook

No required textbook. All readings will be provided via Brightspace.

Some Recommended Resources:
- "Privacy and Machine Learning" – Google Research (available free online)
- "The Algorithmic Foundations of Differential Privacy" – Dwork & Roth (available free online)

Additional readings and research papers will be assigned throughout the semester.

Course Content and Objectives

This course introduces the core principles and technologies of big data systems, with a strong emphasis on artificial intelligence (AI) and privacy protection at scale. Students will learn to design scalable data pipelines, apply AI and analytics to large-scale datasets, and implement privacy-preserving techniques in distributed environments. The course is organized around the theme of building **trustworthy big data systems**, platforms that are **scalable**, **intelligent**, and **privacy-aware**. Rather than treating Big Data, AI, and Privacy as separate topics, we integrate them around real-world challenges. We start with privacy foundations to understand the threats and regulations that motivate the need for trustworthy systems, then explore the tools and techniques needed to build them. Students gain hands-on experience with real-world big data tools. By the end of the course, students will be prepared for roles at the intersection of big data, AI, and privacy, including Privacy Engineer, AI Ethics Specialist, and Compliance Analyst. **The course is designed to meet employer demand identified through the CUNY Tech Futures industry partnerships, bridging the gap between technical AI skills and privacy expertise.**

Prerequisites

CSCI 335 (Data Structures and Algorithms) or equivalent.
Students should have:

- Proficiency in Python programming
- Familiarity with basic SQL and data processing
- Understanding of fundamental machine learning concepts

Learning Outcomes

By the end of this course, students will be able to:

1. **Design and build** scalable big data pipelines using distributed frameworks (Apache Spark, streaming, NoSQL, cloud) for large-scale data processing and analytics.
2. **Apply AI and machine learning techniques** (LLMs and distributed training) to extract insights from large-scale datasets.
3. **Identify and assess privacy risks** in big data and AI systems, including re-identification, membership inference, and regulatory compliance (GDPR, CCPA, NYC AI Law).
4. **Implement privacy-preserving techniques** including anonymization (k-anonymity, l-diversity), differential privacy, federated learning, and cryptographic methods in big data environments.
5. **Integrate and communicate** big data, AI, and privacy knowledge to design a complete **trustworthy big data system** that is scalable, intelligent, and privacy-aware.

Syllabus and Readings

The syllabus provides an overview of the course content and major topics. All required readings and supplementary materials will be posted on Brightspace prior to each session. Please review assigned readings before class.

For enrollment verification, please complete the first discussion that will be posted in Brightspace during the first week.

Grading

Lab Assignments: 35%
Final Research Project: 25%
Assignments (Problem Sets): 10%
Final Exam: 30%

Lab Assignments

Labs are hands-on coding exercises that follow a step-by-step learning approach. Students will work with real-world datasets and implement practical tasks covering Big Data systems, AI analytics, and privacy risks. We expect to have **7 labs** throughout the semester. More details will be provided during the semester.

Final Research Project

The final project is a **group-based research project** focused on designing a trustworthy Big Data system. Teams will explore recent work published in renowned conferences, submit a proposal, and conduct experiments to test, learn, and understand the concepts and techniques. Projects will be evaluated based on a final presentation and report. Detailed guidelines and expectations will be provided during the semester.

Assignments (Problem Sets)

Problem sets consist of **theoretical exercises** covering key concepts from different areas of the course. They are designed to reinforce the mathematical and conceptual foundations that will also be assessed in the written exam.

Exam

This course emphasizes hands-on learning and practical application over traditional testing. Accordingly, there will be one written exam assessing the theoretical and mathematical concepts covered in class, with questions similar to those in the problem sets and review materials.

All assignments are due at 11:59 PM on the specified date. Late submissions will be penalized 10% per day for 3 days maximum.

Course Material and Communication

All course materials, including slides, readings and research papers, course documents, assignments, and announcements, will be posted on Brightspace. An accompanying course website will also be available to enrolled students at www.BDAP-Hunter.com. The website will require authentication. Certain lab assignments and project deliverables will be completed or submitted through this website. Additional instructions will be provided during the semester. We will use the Brightspace Discussion Forum for general course discussions and questions. Before emailing the teaching staff about course content, assignments, or other general questions, please check the discussion board. If your question has not already been answered, post it there so that the response is available to the entire class. Students are also encouraged to help answer their classmates' questions. Do not post code solutions on the discussion board. For personal or individual matters, you may contact the instructor directly by email. Please include CSCI 49395: at the beginning of the subject line.

Hunter College Official Policies

Hunter College regards acts of academic dishonesty (e.g., plagiarism, cheating on examinations, obtaining unfair advantage, and falsification of records and official documents) as serious offenses against the values of intellectual honesty. The College is committed to enforcing the CUNY Policy on Academic Integrity and will pursue cases of academic dishonesty according to the Hunter College Academic Integrity Procedures.

Hunter College Policy on Sexual Misconduct

In compliance with the CUNY Policy on Sexual Misconduct, Hunter College reaffirms the prohibition of any sexual misconduct, which includes sexual violence, sexual harassment, and gender-based harassment retaliation against students, employees, or visitors, as well as certain intimate relationships. Students who have experienced any form of sexual violence on or off campus (including CUNY- sponsored trips and events) are entitled to the rights outlined in the Bill of Rights for Hunter College.
a. Sexual Violence: Students are strongly encouraged to immediately report the incident by calling 911, contacting NYPD Special Victims Division Hotline (646-610-7272) or their local police precinct, or contacting the College's Public Safety Office (212-772-4444).
b. All Other Forms of Sexual Misconduct: Students are also encouraged to contact the Director of the Office of Student Conduct (https://hunter.cuny.edu/students/student-affairs/office-of-student-conduct/contact/) CUNY Policy on Sexual Misconduct Link:
https://www.cuny.edu/wp-content/uploads/sites/4/page-assets/about/administration/offices/ovsa/policies/Sexual-misconduct-8.30.18-PSM-2018-005.pdf

In compliance with the ADA and with Section 504 of the Rehabilitation Act, Hunter College is committed to ensuring educational access and accommodations for all its registered students. Hunter College's students with disabilities and medical conditions are encouraged to register with the Office of AccessABILITY for assistance and accommodation. For information and appointment contact the Office of AccessABILITY located in Room E1214 or call (212) 772-4857 /or VRS (646) 755-3129.

Bullying, cyberbullying, online hate, intimidation, threats, harassment, and pressure to share schoolwork are all forms of violence. CUNY holds a zero tolerance stance towards all such acts. The University is committed to prevention of any form of bullying, will respond promptly to threats and/or acts, and will protect victims of bullying from retaliation. As a criminal matter, the New York Attorney General defines cyberbullying as the use of email, websites, instant messaging, chat rooms, text messaging and digital cameras to antagonize and intimidate others. Disrupting a teleconferencing platform (such as Zoom/Skype/Brightspace Collaborate Ultra) is a federal crime.

Note that details of this document are subject to change if the need arises.

Course Schedule

	Topic	Brief Description
1	Course Introduction: Trustworthy Big Data Systems	Introduces the course and the integration of big data, AI, and privacy in building trustworthy data systems.
2	Privacy Fundamentals and PII	Covers fundamental privacy concepts, personally identifiable information (PII), and privacy risks in large-scale data systems.
3	Privacy Attacks and Data Anonymization	Examines re-identification and linkage attacks and introduces common data anonymization techniques and their limitations.
4	Differential Privacy: Foundations and Applications	Introduces the principles of differential privacy and its mathematical foundations for providing rigorous privacy guarantees and explores practical applications.
5	Privacy Regulations and Compliance	Provides an overview of major privacy regulations and their implications for the design and operation of data-driven systems.
6	Apache Spark and Big Data Pipelines	Introduces Apache Spark and scalable data processing.
7	Streaming and NoSQL Systems	Introduces real-time data processing and NoSQL databases.
8	Cloud Computing and Data Lakes	Covers cloud-based data processing, data lakes, and modern architectures for storing and analyzing large-scale datasets.
9	Large Language Models	Introduces LLMs and examines risks associated with their training, deployment, and use.
10	Distributed AI and Large-Scale Model Training	Introduces distributed approaches for training large AI models and discusses their scalability and security challenges.
11	Privacy-Preserving Machine Learning	Examines privacy attacks against machine learning models and techniques for privacy-preserving model training and evaluation.
12	Federated Learning and Decentralized AI	Introduces federated learning and related approaches for collaborative and privacy-preserving learning across distributed data sources.
13	Cryptography for Big Data Security	Introduces key cryptographic techniques for privacy-preserving computation, including homomorphic encryption, MPC, TEEs, and secure aggregation.
14	Final Project Presentations	Students present their final projects and reflect on the integration of big data, AI, and privacy concepts covered throughout the course.