THE UNIVERSITY OF CHICAGO

Master of Science in Applied Data Science

Course Syllabus

BIG DATA PLATFORMS

ADSP 31013/IP02/03

Autumn 2023

Class: Tuesdays/Wednesdays (6-9 PM CST)
Location: NBC Tower

Instructor: **Ashish Pujari**
Instructor Email: apujari@uchicago.edu

TA: **John Kanellopoulos**
TA Email: jkanellopoulos@uchicago.edu

Responsiveness: When contacting your instructor or TA with questions, please allow 24-48 hours for a response. Technical questions about assignments and coursework should be addressed to the TA first with the instructor cc'd on the message if necessary. You are encouraged to attend all TA sessions and office hours to get your questions answered.

COURSE DESCRIPTION

Big data has evolved rapidly in the last decade with applications in healthcare, retail, banking and finance, manufacturing, IoT, and other industries. This course teaches students about processing large datasets on the cloud and applying machine learning to big data problems.

Students will gain an in-depth understanding of the key concepts in big data storage and analytics, as well as parallel and distributed computing. They will learn about applications of big data in data engineering, machine learning, search and recommendation systems, and real-time analysis. They will also explore advanced concepts such as graph computing, real-time data processing and recommendation engines on big data. Students will leverage a combination of open-source and cloud-based services such as AWS Athena, GCP Big Query, Spark, SparkML, Spark NLP, and Airflow, Kafka.

Students will gain hands-on experience on public cloud platforms such as Google Cloud Platform (GCP), Amazon Web Services (AWS) and virtualization technologies such as Docker and Kubernetes. In addition to programming on cloud infrastructure, students will also be exposed to best practices in infrastructure as code, data lake management and governance.

PREREQUISITES

To be successful in the course, the following are the prerequisites.

- Know your computer (Setting environment variables, Using the Mac/PC terminal, traversing applications/folders, updating security preferences)
- Linux Boot Camp
- Programming for Analytics - Python
- Data Engineering Platforms for Analytics

COURSE MATERIALS

Class topics are sourced from multiple sources including the following required/recommended books. While reading assignments will be supplemented, the books jointly cover the class topics, with both construct and methods. Note: The hands-on exercises in class as well as assignments have been custom designed for this course and are based on online data sources.

Suggested:

- Designing Data-Intensive Applications
- Mining of Massive Datasets
- Spark in Action

SOFTWARE & TOOLS

This course will require working in

- Python Anaconda
- Jupyter Notebooks
- Docker Containers
- Apache Spark, Apache Airflow, Apache Kafka
- Google Cloud Platform (GCP) – Google Cloud Storage (GCS), Cloud Data Proc, Big Query
- Amazon Web Services (AWS) – S3, Athena

Note: These software applications work best on PC's/Macs. Ensure the computer you are using provides you with the authority to perform these installs. Some work-related computers may not permit such installations without admin rights.

LEARNING OUTCOMES

After completing this course, students should be able to:

- Design high-level architectures and select technologies for big data analytics projects.
- Leverage big data cloud computing services for data engineering and data science.
- Perform data wrangling, transformation, and exploratory data analysis on big datasets.
- Utilize Apache Spark for distributed in-memory data processing.
- Perform machine learning on both structured and unstructured big datasets.

- Apply big data techniques related to Machine Learning, Natural Language Processing (NLP), Recommendation engines, Association Rules Mining and Graph Computing

METHODS OF INSTRUCTION

This is a 10-week course with pre-class work and in-person activities. We will meet weekly for in-person class sessions, and I will provide online office hours. In-person class sessions will include lecture, discussion, and instructional activities. Students are expected to review the week's pre-class work ***prior*** to each class session in order to fully participate in class. Weekly attendance and participation are expected. In addition, students should plan to log into Canvas at least once a week.

Pre-class work includes a review of course content, completion of weekly assignments, and collaboration with classmates. Pre-class content may include video lectures, readings, supplemental videos, discussions, quizzes, homework, and completion of a final project.

TA sessions will be arranged as a group and held on a weekly basis. The TA will be available to provide additional instruction on course material and provide support for homework.

EVALUATION

Your course grade will be calculated as follows:

- Individual Assignments (4) - 50%
- Class Quizzes (2) - 10%
- Course Project - 40%

COURSE PROJECT

The goal of the group project is to perform exploratory data analysis and apply machine learning techniques to solve a well-defined business problem on a large dataset. Students will be assessed on their understanding of various concepts taught in the course as well as their ability to leverage big data tools and infrastructure. Students will work in teams of 3 or 4 members and collaborate on their Big data class project that would involve data ingestion, processing and machine learning on a large data set. Students will submit the project source code and present their findings in the final class.

GRADING SCALE

A = 93%–100%
A- = 90%–92%
B+ = 87%–89%
B = 83%–86%
B- = 80%–82%
C+ = 77%–79%
C = 73%–76%
C- = 70%–72%
F = 0%–69%

COURSE SCHEDULE

Academic quarters consist of 9 weeks of instruction, with the 10th week for assessment or course rescheduling. Refer to the university's academic calendar for quarterly start and end dates.

Important Note: Changes may occur to the syllabus at the instructor's discretion. When changes are made, students will be notified via email and in-class announcement.

Week	Topic	Assignments Due
Week 1	Introduction to Big Data • Big data overview • Tools and technologies • Parallel and distributed computing	
Week 2	Big Data SQL • Data Lakes – GCS, S3 • Amazon Athena • Google Big Query	
Week 3	Containers and Virtualization • Containers and Virtual Machines • Docker and Kubernetes • Infrastructure as Code	Assignment 1 due
Week 4	Spark Fundamentals • Introduction to Apache Spark • Data Structures, RDD and Data Frames • EDA and Data Transformation	Project checkpoint 1 Reading material Quiz 1
Week 5	Big Data Machine Learning • Introduction to Distributed ML • Feature Engineering • Regression, Classification • Hyperparameter Tuning	Assignment 2 due
Week 6	Natural Language Processing • Introduction to NLP on Big data • Spark NLP • Embeddings, NER, Sentiment Analysis	Project checkpoint 2 Reading material
Week 7	Recommendation Systems • Introduction to Recommendation Systems	Assignment 3 due

	• Association Rules Mining • Collaborative Filtering	
Week 8	Graph Computing • Introduction to big data graph computing • Graph algorithms and applications • Spark Graph Frames	Project checkpoint 3 Reading material Quiz 2
Week 9	Pipelines and Streaming • Data pipelines on Apache Airflow • Real-Time and Streaming on Apache Kafka	Assignment 4 due
Week 10	Final Project Presentation	Project submissions

ATTENDANCE

This course will meet once a week. Your attendance is required and paramount to your success in this class. You are allowed to miss at most two sessions, provided that you make arrangements with the instructor in advance.

In order to allow students to follow quarantine guidelines, I am prepared to offer you the ability to complete your coursework remotely while you self-isolate.

Students who believe they may have been exposed to COVID-19 or who have tested positive for COVID-19 should follow the precautions recommended by the Centers for Disease Control and Prevention (CDC). The University will rely on these guidelines and is discontinuing the previous customized exposure protocols. Please note that anyone with symptoms of respiratory illness should get tested and wear a mask around others until the symptoms are gone.

Those who test positive for COVID-19 must stay at home or in their residence hall and follow CDC guidance.

LATE WORK

All assignments must be submitted to this course's Canvas site on the due dates as per the course calendar. If you turn in an assignment late, 10% credit will be deducted from the total score for each day after the deadline. Assignments turned in more than one week late will not receive credit. In the case of unexpected events, you must contact me before the assignment due date in order to receive a grace period. Students can only receive up to 1 grace periods in the course.

REQUESTING REASONABLE ACCOMODATIONS

The University of Chicago is committed to ensuring equitable access to our academic programs and services. Students with disabilities who have been approved for the use of academic accommodations by Student Disability Services (SDS) and need a reasonable accommodation(s) to participate fully in

this course should follow the procedures established by SDS for using accommodations. Timely notifications are required in order to ensure that your accommodations can be implemented. Please meet with me to discuss your access needs in this class after you have completed the SDS procedures for requesting accommodations.

Phone: (773) 702-6000
Email: disabilities@uchicago.edu

Please follow accommodation implementation instructions provided by the disability liaison in the division after you have completed the SDS procedures for requesting accommodations.

You may want to begin by reading through the information published on the Student Disability Services website.

ACADEMIC HONESTY & PLAGIARISM

It is contrary to justice, academic integrity, and to the spirit of intellectual inquiry to submit another's statements or ideas of work as one's own. To do so is plagiarism or cheating, offenses punishable under the University's disciplinary system. Because these offenses undercut the distinctive moral and intellectual character of the University, we take them very seriously.

Proper acknowledgment of another's ideas, whether by direct quotation or paraphrase, is expected. In particular, if any written or electronic source is consulted and material is used from that source, directly or indirectly, the source should be identified by author, title, and page number, or by website and date accessed. Any doubts about what constitutes "use" should be addressed to the instructor.

Review the University of Chicago Student Manual Academic Honesty and Plagiarism policy.
Review the University of Chicago International Affairs Plagiarism website.

Review the University of Chicago Libraries site on copyright.

EXPECTATIONS FOR STUDENTS

As a student, you will get as much or as little out of this course as the effort you put in. My expectation is that you become involved in the class through participation in discussions, class sessions, group projects, etc. and show respect to all participants in the class. Time management is important to stay focused on assignments and complete them by the required due dates. Please communicate any concerns or issues to the me in order to gain a better understanding of the content and stay focused in the class.

INSTRUCTOR COMMITMENTS

I will provide guidance on the course materials and return graded work with constructive feedback in a timely manner, within two weeks from the assignment due date. Assignments submitted late will receive feedback within 72 hours of their submission. I will make every effort to respond to your questions within 24 hours. I will also be available to meet with you during weekly office hours. As a part-time instructor, I have a day job therefore I request you contact me via email for most situations or questions.

COURSE EVALUATION

Towards the end of the course, you will receive an email from the Office of the University Registrar reminding you to provide feedback on the course. You will receive consistent reminders throughout the period when the evaluation is open, and the reminders will stop once you have completed the evaluation. Please note:

- The evaluation is completely anonymous. When the results are released, instructors and departments will not be able to tell which student provided the individual feedback.
- Because it is anonymous and the results are not released to faculty or departments until after grades have been submitted, the feedback will not impact a student's grade.
- The feedback is important so that the instructor can gain insight in to how to improve their teaching and the department can learn how best to shape the curriculum.

STATEMENT OF INTENT

By remaining in this course, you are agreeing to accept this syllabus as a contract and to abide by the guidelines outlined in the document. You will be notified should there be a necessary change to the syllabus.