MSA 8050: Scalable Data Analytics

Dr. Péter Molnár, Fall 2026

Contents

1 Course Details

This course covers essential concepts and tools for large scale data analytics. Topics include 1/ functional and parallel programming paradigms and languages, 2/ ore components of large scale platforms, 3/ scalable machine learning algorithms, and 4/ real-time data analysis with big data tools. Programming projects demonstrate design and implementation of large scale analytics pipelines for structured and unstructured data.

1.1 Location and Time

Wednesday	7:15 PM - 9:45 PM
Location:	Buckhead Center

1.2 Instructor

Instructor:	Dr. Péter Molnár
Email:	pmolnar@gsu.edu
Office Phone:	+1.404.413.7713
Office hours:	by Appointment
Office Buckhead:	Buckhead Center, 6th floor
Office Downtown:	35 Broad Street, 9th floor

During the term, it is highly recommended that you contact the instructor and TAs, in-person or via email. They are available to help you focus your projects, gain access to resources, and answer your questions. Please try to contact them at least once during the term to discuss your project. Your class members are also a good source of help.

1.3 Course Website

Class information will be posted on the class website (https://msa8050.molnar.ai/) and iCollege site (https://icollege.gsu.edu). There will be links to other websites with course related material.

2 Overview

This course is designed for students who are comfortable with the fundamentals of data analysis and machine learning on small datasets and are ready to explore what changes when those same ideas must operate at scale. It integrates algorithmic foundations, distributed data systems, and applied project work to help students move beyond notebook-scale workflows and reason explicitly about computation, memory usage, communication costs, and system design. Over the semester, students study recommendation systems, graph analytics, forecasting, frequent pattern mining, and sequential pattern mining, with a particular focus on how these methods behave in large, parallel environments.

A central objective of the course is to connect theory with implementation. In the theoretical portion of each session, students examine the mathematical and algorithmic foundations of scalable methods such as matrix factorization, FP-Growth, PrefixSpan, PageRank, connected components, triangle counting, and community detection. These discussions emphasize approximation, sparsity, scalability, and the implications of distributed execution. In the technical portion, students implement these ideas using Apache Spark and related tools, allowing them to see how algorithm design interacts with ETL workflows, resource management, partitioning strategies, shuffles, and workflow automation.

The course is intentionally project-centered. Students select from a set of retail analytics project tracks and work in small teams to build an end-to-end solution, starting from raw data and progressing through data preparation, ETL or ELT pipeline design, exploratory analysis, baseline modeling, improved modeling, and quantitative evaluation. Each project requires students to evaluate both system performance (such as runtime and resource utilization) and analytical quality (such as recommendation accuracy, segmentation quality, or forecasting error), and to justify design decisions in terms of these trade-offs.

By the end of the course, students should be able to explain the theoretical foundations of core scalable data science methods, design and implement reproducible Spark-based data and machine learning pipelines, and critically assess how system constraints shape algorithm selection. They should also be able to collaborate effectively on realistic data products, communicate technical decisions clearly, and connect academic readings and research ideas to practical large-scale analytics workflows. The broader goal of the course is not only to teach the use of large-scale tools, but to help students think like data scientists

and engineers who can bridge theory, systems, and real-world applications.

Prerequisite: MSA 8010 - Data Programming for Analytics or IFI 8410 - Introduction to Programming and Predictive Analytics for Business

2.1 Learning objectives

Upon successful completion of this course, you will accomplish the following objectives and outcomes. Students who complete this course will gain "Ready for work" skills, including:

1. Explain the theoretical foundations of core scalable data science methods — including matrix factorization, frequent and sequential pattern mining (FP-Growth, PrefixSpan), graph analytics (PageRank, connected components, triangle counting, community detection), and many-series forecasting — and reason about their behavior with respect to approximation, sparsity, and distributed execution. Students can select an appropriate algorithm for a large-scale problem and articulate why it scales.
2. Design and implement reproducible Apache Spark data and machine-learning pipelines, including ETL/ELT workflows, partitioning strategies, shuffle management, and workflow automation. Students deliver working pipelines that run reliably on a cluster and can be re-executed end-to-end from raw inputs.
3. Evaluate scalable systems along both axes of system performance (runtime, memory, communication cost, resource utilization) and analytical quality (recommendation accuracy, segmentation quality, forecasting error), and use those measurements to drive design decisions. Students produce quantitative baseline-versus-improved comparisons and justify trade-offs with evidence.
4. Build an end-to-end retail analytics solution in a small team, progressing from raw data through preparation, exploratory analysis, baseline modeling, improved modeling, and evaluation. Students ship a complete, deployed data product with a written report and presentation that connects design choices to measured results.
5. Collaborate effectively on realistic data products and communicate technical decisions clearly to both technical and non-technical audiences, connecting academic readings and research to practical large-scale workflows. Students demonstrate professional teamwork artifacts (charter, version control, reproducible code) and present their work in a final showcase.

3 Schedule

The course schedule is shown in the table below. However, the topics may change according to the interests and abilities of the class. Materials may be updated up to 24 hours prior to class.

3.1 Course Schedule

#	Date	Topic	In Class	Project Milestone
1	2026-08-26	Course Overview & Scale	ACT01	
2	2026-09-02	Problem Formulation & Planning	ACT02	M01: Execution plan
3	2026-09-09	Recommenders & Frequent Patterns	ACT03	
4	2026-09-16	Graph Basics & Connectivity	ACT04	M02: Data Exploration
5	2026-09-23	Time Series & Many-Series Forecasting	ACT05	
6	2026-09-30	Factorization & Sequential Patterns	ACT06	M03: Baseline implementation
7	2026-10-07	Communities, Triangles & Structure	ACT07	
8	2026-10-14	Parallel ETL & Distributed Patterns	ACT08	
9	2026-10-21	Model Selection & Large-Scale Experiments	ACT09	M04: Performance Improvements
10	2026-10-28	Optimization & Iterative Algorithms	ACT10	
11	2026-11-04	Robustness, Drift & Monitoring	ACT11	
12	2026-11-11	Algorithm-System Trade-offs & Cases	ACT12	M05: Draft report
13	2026-11-18	Synthesis & Presentation Prep	ACT13	
14	2026-12-02	Last day of class: Presentations		
	2026-12-05			M06: Completed Project

3.2 Schedule of due dates of the project milestones for each section

Milestone	Due
M01: Team roster, project choice, and 1-page execution plan	2026-09-02
M02: Data understanding & initial ETL report (schema, data quality, first ETL run)	2026-09-16
M03: Baseline system implemented and evaluated (ETL+features, metrics, runtime notes)	2026-09-30
M04: Improved system implemented, with comparative technical and quality metrics and initial trade-off analysis	2026-10-21
M05: Draft final report and slides with full baseline vs improved metrics and narrative	2026-11-11
M06: Final reports and video	2026-12-09

4 Required Resources

4.1 Textbooks

Links to selected textbook chapters will be provided. Textbooks are available through the GSU Library subscription to O'Reilly Media at https://go.oreilly.com /georgia-state-university/home/.

4.2 Compute Requirements

Students are required to bring their laptop computer to class for class activities and in-class knowledge checks (quizzes) and discussion. Data processing will be performed on the Analytics Research Cluster (ARC) using open-source software and libraries.

API Access to Hosted GenAI Models: Limited access to LLMs and other GenAI models will be provided. You may also use cloud hosted APIs like OpenAI (https://platform.openai.com/docs/overview) and Google Colab (https://colab.research.google.com/) for GPU supported notebooks and processing environments.

Laptop or desktop computer to complete assignments: ARC provides a web-interface that supports Google Chrome, Firefox and other common web-browsers. A full keyboard and trackpad or mouse are needed to efficiently write code. The browsers on Chrome Books and iPads my not fully support the web-interface.

Virtual Private Network (VPN): VPN access is needed to access ARC and other compute resources. Visit https://technology.gsu.edu/technology-servic

es/cybersecurity/virtual-private-network/ to configure access to the GSU-VPN on your device. (You may seek help from the Technology Service Desk.)

5 Evaluation

Students are evaluated by the deliverables summarized in the table below:

Assignment	Percentage
Class Activities (best 10 out of 13)	20%
Group Project	80%
Total	**100%**

5.1 Grading Scale

A+	A	A-	B+	B	B-	C+	C	C-	D	F
97%–100%	91–96.9	89.5–90.9	87–89.4	83–86.9	79.5–82.9	77–79.4	72–76.9	69.5–71.9	60–69.4	Below 59.9

6 In-class Activities

There will be a graded class activity in most sessions. Students are expected to participate at the time of the activity. These activities may include ad-hoc knowledge checks (quizzes) and discussions. **There are no make-up assignments for missed activities**. The best ten (10) scores of all activities will be used for the final grade. Class activities are individual work unless otherwise stated.

7 Group Project

The group project is a central component of the course and constitutes the majority of the final grade. Teams of up to three (3) students will collaborate to design, implement, and evaluate a scalable data analytics system aligned with the course learning objectives.

Students will select their project from a curated set of instructor-provided retail analytics projects. These projects are designed to ensure consistent scope, data scale, and technical rigor across teams. Students may not propose their own project topics. Project selection will occur early in the semester after teams are formed. Each team must submit a group charter defining member roles and contributions and must use the internal GitLab repository for all project work,

with instructors and TAs granted access. Projects progress from raw data and ETL/ELT pipeline design through baseline and improved modeling, evaluation, and final reporting. Teams are expected to assess both system performance and analytical quality.

The project culminates in a final report and presentation and is intended to reflect realistic, team-based analytics work, resulting in a reproducible technical artifact suitable for professional portfolios.

Teams use the internal GitLab repository (https://git.insight.gsu.edu) to manage their project, and deploy their solutions on the ARC. This assignment not only fosters teamwork but also helps students build practical skills and create a tangible artifact for future professional use.

Group project deliverables are due at the posted date and time (see the milestones table). Late submissions will be penalized with a reduction to 70% of the total score within the first 24 hours, and to 50% thereafter. No submissions will be accepted beyond 72 hours from the original due date.

8 Use of Internet resources and Generative AI

The purpose of assignment is to practice what you learned and verify your understanding of concepts. **The use of Generative AI or any other tools and resources is generally prohibited during quizzes.** You are encouraged to utilize Internet resources (like https://stackoverflow.com/) and GenAI tools (like ChatGPT, Codex, Claude Code, etc.) for your homework and project assignments. If you make use of these tools indicate in your program code where you found (parts of) the solution or the AI coding tool and prompt that produced the code segment.

9 Student Expectations and Class Policies

Students should plan for 2 - 3 hours of work outside of class each week for each course credit hour. Thus, a 3-credit course averages between 6 and 9 hours of student work outside of the classroom, each week. See GSU site for Academic Success: https://success.students.gsu.edu/

9.1 Arbitration

There will be a one-week arbitration period after graded activities are returned. Within that one-week period, you are encouraged to discuss any assumptions and/or misinterpretations that you made on the activity that may have influenced your grade.

9.2 Attendance

This course is designed and delivered **primarily as an in-person class**, and regular on-campus attendance is strongly encouraged and expected. In-class discussions, activities, and collaborative work are an essential part of the learning experience.

Recognizing that some students may have **occasional professional obligations** (such as business travel or work-related commitments), **limited hybrid (remote) participation** is permitted on an as-needed basis. Students should notify the instructor in advance whenever possible when planning to attend remotely.

Hybrid participation is intended to accommodate **temporary and infrequent conflicts and is not a substitute for regular in-person attendance**. Students attending remotely are expected to join synchronously, remain engaged, and participate to the extent possible. Some in-class activities may not be fully replicable in a remote format, and students are responsible for completing any associated requirements.

Students who anticipate frequent inability to attend in person should discuss their circumstances with the instructor early in the semester to determine whether the course format is appropriate for their situation.

9.3 Submission of Deliverables

Unless specific, prior approval is obtained, no deliverable will be accepted after the specified due date.

If you have a legitimate personal emergency (e.g., health problem) that may impair your ability to submit a deliverable on time, you must take the initiative to contact the instructor before the due date/time (or as soon after your emergency as possible) to communicate the situation. Make-up quizzes will not be given.

All assignments must be submitted using the designated mechanism that is specified in the assignment (usually via GitLab, iCollege or ARC). **Assignments via email will not be accepted.**

9.4 Student Behavior

Behavior in class should be always professional. People must treat each other with dignity and respect for scholarship to thrive. Behaviors that are disruptive to learning will not be tolerated and may be referred to the Office of the Dean of Students for disciplinary action.

9.5 Discrimination and harassment

Discrimination and/or harassment will not be tolerated in the classroom. In most cases, discrimination and/or harassment violates Federal and State laws and/or University Policies and Regulations. Intentional discrimination and/or harassment will be referred to the Affirmative Action Office and dealt with in accordance with the appropriate rules and regulations.

Unintentional discrimination and/or harassment is just as damaging to the offended party. But it usually results from people not understanding the impact of their remarks or actions on others, or insensitivity to the feelings of others. We must all strive to work together to create a positive learning environment. This means that everyone should be sensitive to the feelings of others, and tolerant of the remarks and actions of others. If you find the remarks and actions of another individual to be offensive, please bring it to their attention. If you believe those remarks and actions constitute intentional discrimination and/or harassment, please bring it to my attention.

10 Teams of Group Projects

10.1 Team Management

Early in the semester, teams will form. If there are problems during the semester, the following methods will be used:

10.1.1 Terminating team members

As in any organization, there may be people in your group who are not willing or able to perform to the level of excellence demanded by the team. The process used to improve team member performance and/or to terminate a team member's membership in the team will involve the following steps:

- Discuss the poor performance with the individual and the standards he or she is expected to meet. As a team, document the discussion including all members' agreed-upon understanding of the standards of performance and the individual's shortfall from those standards. The document should describe what the individual must do to meet the team's standards and the time frame in which the individual will come up to the standards. This agreement should be signed by all team members, and a copy should be sent to the instructors.
- If the agreement is not met, the team, including the individual in question, will schedule a meeting with the faculty. The team will bring a copy of the contract to the meeting for the faculty and will discuss the individual's performance with the faculty. The individual will be terminated or given a final chance to improve his or her performance during that meeting and within a given time frame.

- If the performance does not improve within the time frame, the individual will be terminated from the team.
- If the individual is terminated, the individual may seek to join another team. Alternatively, he or she must complete all course work in its entirety by himself or herself from that point forward.

10.1.2 Resigning from a team

A student may resign from a team and switch to a different one. The work that was done while a team member is the property of both the team and the individual so all can use the work product. Faculty will facilitate the placement of the resigning person on a different team.

Teams will be allowed for some activities during the term. Please note that unless the activity is explicitly identified as a "team activity", I expect everyone to perform their own work (your hands on the keyboard). For team activities, you will be allowed to work with partners (of your choosing).

- Initial teams must be established by the second week of classes. Established teams may continue working together on subsequent team activities. Team membership may change during the term if problems arise. However, team members must be designated within one week of the due date for the team activity. Exception: you may withdraw from a team at any time and submit an assignment individually.
- Teams will submit one assignment for all team members. In most cases, each member of the team will get the same score. However, an individual's score may be reduced at the discretion of the instructor.
- Each team assignment must include the following:
  - Tasks completed by each member.
  - Percentage of the total work completed by each member.
- Any individual with a low team contribution will be removed from their team.

11 Official Department and University Policies

1. Prerequisites are strictly enforced. Students failing to complete any of the prerequisites with a grade of "C" or higher will be administratively withdrawn from this course with loss of tuition fees. There are no exceptions, except as granted by the instructor with the approval of the department.
2. Students are expected to attend all classes and group meetings, except when precluded by emergencies, religious holidays, or bona fide extenuating circumstances.

3. Students who, for non-academic reasons beyond their control, are unable to meet the full requirements of the course should notify the instructor, by email, as soon as this is known and prior to the class meeting. Incompletes may be given if a student has ONE AND ONLY ONE outstanding assignment.
4. A "W" grade will be assigned if students withdraw before mid-semester if (and only if) they have maintained a passing grade up to the point of withdrawal. Withdrawals after the mid-semester date will result in a grade of "WF". See the GSU catalog or registrar's office for details.
5. Spirited class participation is encouraged and informed discussion in class is expected. This requires completing readings and assignments before class.
6. All knowledge assessments and individual assignments are to be completed by the student alone with no help from any other person.
7. Collaboration within groups is encouraged for project work. However, collaboration between project groups will be considered cheating.
8. Copying work from the Internet without a proper reference is considered plagiarism and subject to disciplinary action as delineated in the GSU Student Handbook.
9. Any non-authorized collaboration will be considered cheating, and the student(s) involved will have an Academic Dishonesty charge completed by the instructor and placed on file in the Dean's office and the CIS Department. All instructors regardless of the type of assignment will apply this Academic Dishonesty policy equally to all students. Abstracted from GSU's Student Handbook Student Code of Conduct "Policy on Academic Honesty and Procedures for Resolving Matters of Academic Honesty" https://codeofconduct.gsu.edu/.

As members of the academic community, students are expected to recognize and uphold standards of intellectual and academic integrity. The University assumes as a basic and minimum standard of conduct in academic matters that students be honest and that they submit for credit only the products of their own efforts. Both the ideals of scholarship and the need for fairness require that all dishonest work be rejected as a basis for academic credit. They also require that students refrain from any and all forms of dishonorable or unethical conduct related to their academic work.

Students are expected to discuss with faculty the expectations regarding course assignments and standards of conduct. Here are some examples and definitions that clarify the standards by which academic honesty and academically honorable conduct are judged at GSU.

11.1 Plagiarism

Plagiarism is presenting work by another person or AI as one's own. Plagiarism includes any paraphrasing or summarizing of the works of another person or AI without acknowledgment, including the submitting of another student's work as one's own. Plagiarism frequently involves a failure to acknowledge in the text, notes, or footnotes the quotation of the paragraphs, sentences, or even a few phrases written or spoken by someone else. The submission of research or completed papers or projects by someone else is plagiarism, as is the unacknowledged use of research sources gathered by someone else when that use is specifically forbidden by the faculty member. Failure to indicate the extent and nature of one's reliance on other sources is also a form of plagiarism. Any work, in whole or part, taken from the Internet or other computer-based resource without properly referencing the source (for example, the URL) is considered plagiarism. A complete reference is required in order that all parties may locate and view the original source. Finally, there may be forms of plagiarism that are unique to an individual discipline or course, examples of which should be provided in advance by the faculty member. The student is responsible for understanding the legitimate use of sources, the appropriate ways of acknowledging academic, scholarly, or creative indebtedness, and the consequences of violating this responsibility.

11.2 Cheating on Examinations

Cheating on examinations involves giving or receiving unauthorized help before, during, or after an examination. Examples of unauthorized help include the use of notes, texts, or "crib sheets" during an examination (unless specifically approved by the faculty member) or sharing information with another student during an examination (unless specifically approved by the faculty member). Other examples include intentionally allowing another student to view one's own examination and collaboration before or after an examination if such collaboration is specifically forbidden by the faculty member. The use of Generative AI tools during knowledge assessments is prohibited.

Unauthorized Collaboration. Submission for academic credit of a work product, or a part thereof, represented as its being one's own effort, which has been developed in substantial collaboration with another person or source or with a computer-based resource is a violation of academic honesty. It is also a violation of academic honesty knowingly to provide such assistance. Collaborative work specifically authorized by a faculty member is allowed.

11.3 Falsification

It is a violation of academic honesty to misrepresent material or fabricate information in an academic exercise, assignment or proceeding (e.g., false, or misleading citation of sources, the falsification of the results of experiments or

of computer data, false or misleading information in an academic context in order to gain an unfair advantage).

11.4 Multiple Submissions

It is a violation of academic honesty to submit substantial portions of the same work for credit more than once without the explicit consent of the faculty member(s) to whom the material is submitted for additional credit. In cases in which there is a natural development of research or knowledge in a sequence of courses, use of prior work may be desirable, even required; however, the student is responsible for indicating in writing, as a part of such use, that the current work submitted for credit is cumulative in nature.