Data Engineering — Syllabus

ECE590D—Spring 2026

INSTRUCTOR

Javier Pastorino, Ph.D.
211 Hudson Hall
javier.pastorino@duke.edu
Make and Appointment.

Teaching Assistants

- Jay Ding
- Conrad Li
- Hung Anh Vu
- Vivian Zhang *(Head TA)*

Office Hours

Instructor and TA office hours will be posted in the course website.

Main Sessions

- Section 001:
  Tuesday/Thursday 10:05AM-11:20AM
- Section 003:
  Tuesday/Thursday 8:30AM-9:45AM
- *Both in Wilkinson 132*

Discussions

- Section 011: Friday 10:05AM, 207HH
- Section 012: Friday 10:05AM, 208HH
- Section 031: Friday 8:30AM, 207HH
- Section 032: Friday 8:30AM, 208HH

Course Website

https://duke.is/de

This course introduces the foundations of data engineering and data governance, integrating them with DataOps, and modern data storage systems. It covers core data governance activities, including requirements gathering, data cleansing, extraction, transformation, and loading (ETL). The course introduces networking and data storage concepts beyond individual files, encompassing relational and NoSQL data models deployed on single-server systems, on-premises clusters, and cloud-based platforms. DataOps best practices are applied to design, orchestrate, and operate data pipelines that serve as the backbone of machine learning workflows.

What background do you need to take this course?

This class requires the fluency in Python and Data Structures provided by ECE590D-P4ML — *Programming and Data Structures for Machine Learning*, and it is an enforced prerequisite. As an alternative, students who had taken ECE551, can take this course, though they are responsible for self-learning Python programming. Students who did not take either of these courses, can meet with the instructor to request a permission for the class.

What will you get out of this course?

By the end of this course, you will be equipped to analyze, design, and implement data architectures that solve real-world problems. You will learn how to process and interpret data, interact effectively with users, and create well-structured pipelines that integrate machine learning models or other computational systems to support decision-making.

The course learning objectives are for the student to

- ***Comprehend*** the basis of data engineering.
- ***Understand*** different data models and storage systems that encompass those models.
- ***Analyze*** and ***identify*** data requirements to build an ETL process to support those requirements.
- ***Understand*** the basis of Enterprise hardware and Networks.
- ***Deploy*** and ***use*** storage systems on-premises and on the cloud.
- ***Design*** and ***implement*** Data Architectures to support the business operations (requirements).

How can you be successful in this course?

Course Information

Textbooks

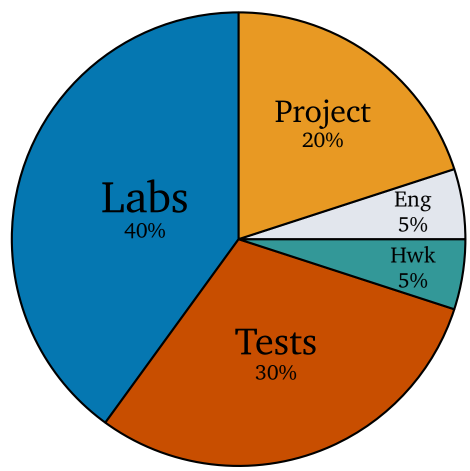
No textbook is required as necessary readings will be provided during the course. However, the following textbooks will be referred in many occasions.

Fundamentals of Data Engineering
Joe Reis and Matt Housley.
O'Reilly Media
ISBN 13: 978-1098108304
The full text is available online at Duke Library.

Database System Concepts
A. Silberschatz, H. Korth, S. Sudarshan
McGraw Hill
ISBN 13: 978-1260084504
The full text is available online at Duke Library.

Course Grading

There are several type of assessments in this course designed to assess your mastery in the concepts and their application to realistic scenarios.



Labs (40%) Labs are hands-on, in-class assessments that focus on the application of the concepts being covered in both readings and lectures. Attendance is required for lab sessions to receive credit for the lab —two tardy arrivals will be excused.

Tests (30%) There are three equally weighted tests. All test will be on paper and include problem-solving and critical-thinking exercises which may include coding activities.

Project (20%) The project constitutes a comprehensive assessment, bringing together all the techniques discussed within a single case study. It will be completed in teams of four students and will include activities such as requirement gathering, data architecture design, and implementation. During the final week of classes, students will **showcase and defend** their project, demonstrating both their technical implementation and their understanding of the underlying concepts.

Homework (5%) Homework is a formative form of assessment designed to support students in the preparation and coverage of course content. Homework assignments may include activities such as coding exercises, research tasks, and case study evaluations. All homework is to be completed individually.

Engagement (5%) Engagement is assessed through student attendance, participation, and timely submission of in-class activities. Attendance is recorded for all sessions (including discussions) and scored as follows: students arriving on time receive 1 point, arriving tardy (within the first ten minutes) receive 0.7 points, and arriving late (more than ten minutes after) receive no points for that session. Engagement activities may include preparation tasks, reading quizzes, reflections, and in-class work, with varying point values (typically 1-3 points). Students who earn 90% or more of the available engagement points receive full marks; otherwise, engagement scores are prorated.

Extensions and Extra Credit Extensions may be granted for assessments under special circumstances. The student must meet with the instructor **before the assessment is due**. There will be **no** extra credit assignments in this course.

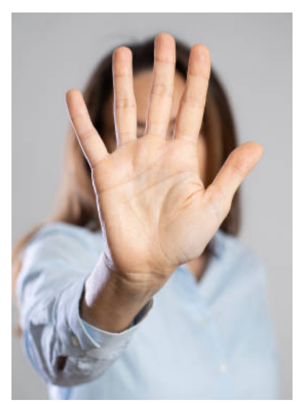
Makeup assessments Make-up opportunities for tests and labs are granted only in cases of documented illness. Students must notify the instructor prior to the scheduled assessment, or as soon as reasonably possible in the event of an emergency. Approved make-up tests and laboratory assignments must be completed within one week of the original assessment date and before grades for the assessment are released (typically within one week).

Regrades If you believe an assessment has been graded incorrectly, you must first meet with the Head TA **within one week of the grade being posted**. If, after this meeting, you remain dissatisfied with the outcome, you must then meet with the instructor within one additional week to discuss the matter further.

Final Letter Grade At the end of the semester, the total weighted grade from each assessment category will be used to compute your final letter grade according to the following scale. Final letter grades may include plus or minus modifiers informed by a student's overall performance throughout the semester.

Course Policies

Use of AI



We live in a time in which AI systems, most notably large language models, are capable of performing many tasks faster than humans and, in many cases, with a high degree of correctness. However, there are also many situations in which the responses *appear* correct but are, in fact, incorrect. In this course, we encourage students to **ethically use AI in the role of a tutor**. *What does this mean?* A tutor is a person (or entity) that supports learning and helps with understanding concepts, but **does not provide direct answers** to assigned problems.

For example, if you need to list the contents of a directory sorted by time in `bash`, using AI as a tutor would involve asking something like: *"Explain how the `ls` command works."* In contrast, asking *"How do I use `ls` to list the contents of directory x sorted by time?"* **is not appropriate**, as this directly asks the AI to solve the assigned problem.

In this course, using AI as a tutor is acceptable. However, asking AI for direct solutions to course problems (as in the second example) is not permitted and is considered academic dishonesty (see below), as course assessments are explicitly designed with this distinction in mind.

Communications

Announcements The course staff (instructors or TAs) will post all announcements in ED Discussions. Please make sure you have notifications on so you don't miss them.

Q&A with TAs/Instructors We will use ED Discussion for questions and answers with the TAs. Questions that relate to your specific solutions should be marked as private, so they are visible only to you and the course staff. All other questions will be public and visible to the entire course.

To ensure timely and consistent support for everyone in the course, students are encouraged to post course-related questions on **ED Discussion**. This allows TAs and instructors to respond efficiently and ensures that answers are visible to the entire class. If such questions are sent directly via email, TAs may ask you to repost them on ED so that all students can benefit from the discussion.

Contact your Instructor In some cases, you may need to contact the instructor to discuss sensitive or personal matters. Outside of ED Discussion, I communicate exclusively through the University email system (Outlook).

If you need to reach me, please send an email rather than using Canvas messages. To help ensure a prompt response, please prefix your email subject line with `DataEng`. I typically respond to emails within 24-48 hours. My email address is listed at the beginning of this document.

In addition, you are always welcome to come during office hours or stop by my office. I will do my best to accommodate meetings outside of office hours when possible.

Academic Integrity

All members of the Duke community, myself included, are expected to abide by the Duke Community Standard (DCS). https://students.duke.edu/get-assistance/community-standard/

> *Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and nonacademic endeavors, and to protect and promote a culture of integrity.*
>
> *To uphold the Duke Community Standard:*
>
> - *I will not lie, cheat, or steal in my academic endeavors;*
> - *I will conduct myself honorably in all my endeavors; and*
> - *I will act if the Standard is compromised.*

What are my expectations with respect to academic integrity? **I expect that all work you submit for a grade is your own, individual work.** It is **not acceptable** to submit someone else's intellectual work product, **including generative AI–produced work**, as your own. This policy also applies to team-based assessments, where all submitted work must represent the collective effort of the group.

In particular, it is not acceptable to obtain, access, consult, or reference solutions to graded assignments prepared by others. This includes, but is not limited to, solutions authored by other current or former students, textbook solution manuals, solutions provided to other current or previous sections, materials for other courses at Duke, materials for other courses at other universities, websites (*e.g.*, CourseHero.com), online services (*e.g.*, 24HourAnswers.com), tutors (*e.g.*, Chegg.com), generative AI services (*e.g.*, ChatGPT), online repositories (*e.g.*, a shared Google Drive), or other materials found online or otherwise not generated by or originated by you.

It is expected that students **will not share our course materials with others**, whether to people as individuals or to groups, organizations, online services, or repositories.

What should you do if you're unsure if an action is allowable? You may encounter situations where whether or not an action would be considered a violation of the DCS is not entirely clear to you. It is ok to be uncertain; there are a lot of situations that fall into gray zones where perspective can shift the evaluation of "ok" vs "not ok," or "right" vs "wrong".

> **If you are unsure whether an action would be considered a violation of the Duke Community Standard, please talk with me *before* you do it.**

A guiding principle is, if you wouldn't want anyone, or everyone, to know you did it, that is likely a strong indicator that you shouldn't do it.

How can you avoid academically dishonest actions? Students most often experience lapses in judgment when they find themselves under time pressure to complete assignments. You can help yourself by starting your assignments well in advance of when they are due. It is never to your advantage to take academically dishonest shortcuts, as doing so will invariably lead to an outcome worse than either submitting an incomplete assignment or not submitting it at all.

Collaboration

Collaboration on exams and homework is not permitted. In other group assessments, like the project, is permitted as far as you discuss only within your group. **You are not permitted to share solutions or code**. *It is not acceptable to share your code; it is not acceptable to receive code.*

Absences, Incapacitation, and Religious Observances

If circumstances prevent you from attending class, reach out to a classmate for notes, review posted notes or come to office hours if you have questions about the material. As you only need 90% of engagement (including attendance) to receive full marks, these are not excused.

In case of a *short-term illness incapacitation* that prevents attending an assessment (e.g, a lab or test), email me as soon as possible. Attach documentation (usually a letter of excuse from your doctor). Per university policy, "**it is your responsibility to meet or otherwise communicate with the instructor of the course in question within 48 hours** to discuss your missed work and how it will be treated in accordance with the instructor's policy. Otherwise, the instructor is under no obligation to accommodate your illness and can treat your missed work as unexcused. "

If a religious observance prevents you from submitting an assignment by the due date complete the form with **at least two weeks before** the date.

Tentative Schedule

The following schedule outlines our plan for this semester. An up-to-date schedule is regularly updated on the course website. While we aim to follow it closely, adjustments may be made as the course progresses.

Week	Topic	
Week 1	Course Introduction and Responsible use of AI Lab 1	
Week 2	DE Lifecycle and Undercurrents. Software Engineering Lab 2	
Week 3	Networks Basics VLans and Firewalls.	Due Homework 1
Week 4	Clusters and Virtualization Lab 3	
Week 5	Conceptual Modeling **Test 1**: *Feb-06*	
Week 6	Relational Model — SQL Lab 4.	Due Homework 2
Week 7	Relational Model and Enterprise Storage Lab 5	
Week 8	Transactions, Concurrency and NoSQL Lab 6.	Due Homework 3
Week 9	Document Stores Lab 7	
Week 10	Spring Break No Class	
Week 11	Pipelines **Test 2**: *Mar-20*	
Week 12	Streams and Orchestration Lab 8	
Week 13	Distributed Processing Lab 9.	Due Homework 4
Week 14	Cloud Systems **Test 3**: *Apr-10*	
Week 15	Lab 10 *Reading Period Starts*	
Week 16	*Reading Period*	Project Due Friday
Finals Week	***Project Defenses***	