

DS 3022

Systems II: Advanced Data Science Systems

Course description

Data Engineering is the practice of designing, building, and maintaining the infrastructure and systems that collect, store, process, and deliver data reliably at scale for data science, analytics and applications. Data engineers must understand the following:

- **Source Data:** Comfort working with a variety of shapes, types, formats, scales, and dimensions of data. These may be in batches as files, streamed in real-time, or triggered by events.
- **Pipeline Construction:** Methods for designing and building automated workflows (ETL/ELT pipelines) to extract data from various sources, transform it into usable formats, and load it into data warehouses, lakes, or other storage systems.
- **Tool / System / Platform Selection:** Develop a strong grasp over when to use what tool and why to deliver a solution. This includes AI-enabled code generation, testing, A/B strategies, deployment, and monitoring.
- **Infrastructure Management:** Design and maintain the underlying architecture including databases, cloud platforms, streaming systems, and compute resources that handle massive volumes of data efficiently.
- **Data Quality & Reliability:** Ensuring data accuracy, consistency, and availability by implementing monitoring, testing, and error handling mechanisms so downstream users can trust the data they're working with.
- **Enabling Data Science:** Creating the foundation that allows data scientists, analysts, and business users to access clean, organized data for machine learning, and decision-making without worrying about the underlying complexity.

Data engineers are the "plumbers" of the data world, ensuring information flows smoothly from its raw sources to the people and systems that will consume it. It is an essential process for any meaningful data science to emerge from data.

In both education and industry, engineers (software, system, and data) tend to encounter a "dividing line" – once crossed individuals know and understand enough on their own to solve future problems or how to teach themselves new tools, technologies, or approaches. An undercurrent of this course is to bring you closer to, or over that dividing line to a higher level of proficiency and autonomy managing and preparing data for deeper analysis.

Specific blocks of content for this course will be:

Section 1 – Essential Tools & Strategies

This foundational section introduces and reinforces core data engineering concepts that serve as building blocks throughout the course. Students will master SQL fundamentals, explore various data architectures and design patterns, work with diverse data formats (CSV, JSON, Parquet), and develop strategies for intelligently managing and scaling data lakes. These essential skills and concepts will be applied consistently across all subsequent course sections, providing the technical foundation necessary for modern data engineering practices.

Tools & Concepts: data types, data formats, ETL/ELT, API requests, DuckDB, SQL, object storage, synthetic data.

Section 2 – Batch Processing

Batch processing represents the cornerstone of data engineering workflows, built around the concept of repeatable, automated "pipelines" that execute a series of coordinated steps, tasks, and business logic. Students will design and implement these pipelines using Directed Acyclic Graph (DAG) structures, then deploy them in Prefect to achieve robust scheduling, horizontal scaling, comprehensive automation, detailed execution tracing, and enterprise-grade reliability. This approach enables efficient processing of large data volumes during off-peak hours while maintaining data quality and operational oversight.

Data Science Applications: End-of-day financial transaction processing, automated monthly business reporting, comprehensive survey data analysis, web server log aggregation and analysis, customer segmentation workflows, and historical data backfill operations.

Tools & Concepts: Basic pipelining with Prefect, Transform modeling with DBT.

Section 3 – Stream Ingestion

Modern data-driven organizations increasingly require real-time analysis capabilities to extract actionable insights within seconds or minutes of data generation. This section teaches students how to connect to live data streams, efficiently capture and parse individual events, and reliably store them in distributed logging systems such as Apache Kafka. Students will learn to design systems that can handle high-velocity data flows while maintaining data integrity and enabling both immediate processing and future batch analysis of the same data streams.

Data Science Applications: Real-time fraud detection systems, live financial market analysis, payment processing monitoring, IoT sensor data collection, social media sentiment tracking, website clickstream analytics, and wearable device health monitoring.

Tools & Concepts: Time-series data, Apache Kafka, producers and consumers, topics, TTL, windowing and grouping, offsets, ingestion patterns, MQTT/SSE/WSS protocols.

Section 4 – Event-Driven Computing

Event-driven architecture leverages cloud-native design patterns to create highly efficient, cost-effective data processing systems that respond dynamically to incoming data without maintaining idle infrastructure. This serverless approach automatically scales resources up during heavy data bursts and scales down during quiet periods, optimizing both performance and operational costs. Students will learn to design systems that can handle unpredictable event patterns while maintaining responsiveness and reliability.

Data Science Applications: IoT sensor data processing with variable transmission patterns, automated data quality validation triggered by new data arrivals, dynamic model retraining based on data drift detection, real-time recommendation engine updates, automated alert systems for anomaly detection, and customer behavior analysis with immediate personalization responses.

Tools & Concepts: Event patterns, AWS Lambda, API Gateway, SNS/SQS, events & triggers.

Grade scale and breakdown

- **Data Project 1** 20%
- **Data Project 2** 20%
- **Midterm Exam** 10%
- **Data Project 3** 20%
- **Data Project 4** 20%
- **Final Exam** 10%

The exams will cover the material you worked on since the beginning of the course until the week before the exam.

Grades will be awarded according to the UVA grading scale:

A 10-9 (90-100%)	B+ 7,9-7 (70-79%)	C 5,4-5 (50-54%)	F 4,9-0 (Below 50%)
A- 8,9-8 (80-89%)	B 6,9-6 (60-69%)		
	B- 5,9-5,5 (55-59%)		

Materials

Primary Texts, all available for free (see UVA in Valencia platform for links):

- Joe Reis & Matt Housley, Fundamentals of Data Engineering, O'Reilly (2022).
- Matt Palmer, Understanding ETL. O'Reilly (2024).
- Joe Naso & Colton Padden, Data Platform Fundamentals. Dagster (2025).
- Martin Kleppmann & Chris Riccomini, Designing Data Intensive Applications, O'Reilly (2026).
- Mark Needham, Michael Hunger, & Michael Simons, DuckDB in Action. Manning (2024).

Course planning and calendar

Date	Contents
07/09/2026	Course introduction
10/09/2026	Essential Tools & Strategies (Session 1)
14/09/2026	Essential Tools & Strategies (Session 2)
17/09/2026	Hands-on Project 1- Essential Tools & Strategies (Session 3)
21/09/2026	Hands-on Project 1
24/09/2026	Project 1 Due
28/09/2026	Batch Processing (Session 1)
01/10/2026	Batch Processing (Session 2)
05/10/2026	Hands-on Project 2 - Batch Processing (Session 3)
08/10/2026	Hands-on Project 2
12/10/2026	(No class) Día de la Hispanidad
15/10/2026	(No class) Fall Break
19/10/2026	Project 2 Due
22/10/2026	Stream Ingestion (Session 1)
26/10/2026	(MIDTERM EXAMS) *
29/10/2026	(MIDTERM EXAMS) *
02/11/2026	Stream Ingestion (Session 2)
05/11/2026	Hands-on Project 3 - Stream Ingestion (Session 3)
09/11/2026	Hands-on Project 3
12/11/2026	Hands-on Project 3
16/11/2026	Project 3 Due
19/11/2026	Event-Driven Computing (Session 1)
23/11/2026	Event-Driven Computing (Session 2)
26/11/2026	Event-Driven Computing (Session 3)
30/11/2026	Hands-on Project 4 - Event-Driven Computing (Session 4)
03/12/2026	Hands-on Project 4
07/12/2026	(No class) Puente
10/12/2026	Project 4 Due
14/12/2026	(FINAL EXAMS) *

* Specific dates for midterm and final exams will be finalized by the University later this term to ensure a balanced schedule across all courses.

Important information

1) Use of Spanish at the Center

The use of Spanish at the Center is strongly encouraged. Remember that you have come to Spain to learn Spanish and immerse yourself in its culture, and there is no better way to do so than by using Spanish on as many occasions as possible.

2) Class Attendance and Tardiness

2.1) Attendance and punctuality are essential for the successful completion of the course. Consequently, unexcused absences and lateness will negatively affect the final grade.

2.2) Only one unexcused absence is permitted. Please bear this in mind when planning trips or outings.

2.3) Arriving more than 5 minutes late to class will count as half an absence (1/2). It is the student's responsibility to notify the instructor of an absence before the class begins.

2.4) Arriving more than 30 minutes late to class will be considered an absence.

2.5) Four (4) unexcused absences will result in the student not receiving credit for the course.

2.6) Only documented absences will be considered excused: a religious holiday, a family emergency, or a medical appointment. Medical absences must be supported by a doctor's note and submitted at reception on the day the student returns to class after the incident has ended.

3) Academic Integrity

3.1) The exchange of information is not only permitted but is also a way of enriching the knowledge acquired during the course. However, all work submitted for each course must be individual and original in every respect. Any sources and direct quotations used must be properly cited.

3.2) Remember that the use of artificial intelligence to complete assignments or examinations is strictly prohibited unless an instructor expressly states otherwise. Such use constitutes a lack of authenticity, and may amount to direct plagiarism, in addition to undermining the development of critical thinking and writing skills, which are among the primary goals of this program. Failure to comply with this rule will result in a grade of "F" for the course.

3.3) Any work that has received credit for one course will NOT be accepted for credit in another course. There will be NO exceptions to this rule. Reusing such work will result in immediate failure of the second course. If assignments for two different courses are related in any way and the student has questions about whether the second project is permissible, the matter must be discussed with all instructors involved. This will prevent any possible unintentional violation of this rule.

3.4) As with punctuality, assignments must be submitted on time. Any late submission will result in a reduction of the grade for that assignment.

3.5) Failure to comply with the provisions set out above will result in the loss of the corresponding course credits for the course in which these rules have been violated.

3.6) If you have a documented special need that requires academic accommodations (SDAC), it must be certified by your home institution. Please contact the Head of Studies in Valencia, Mari'a Andre's (maria.andres@uvavalencia.org).

4) Activities, Workshops, and Lectures

Throughout the course, various activities, workshops, and lectures are organized. Participation may be voluntary or mandatory if the instructor considers the activity relevant to the course.

5) Classroom Conduct

5.1) It is very important that all students arrive to class on time, prepared, and with the materials needed for the day. During class, students must treat one another and the instructor with respect, demonstrate a positive attitude, and help create a participatory and collaborative environment that supports effective learning. Any violation of these expectations will have direct consequences.

5.2) Electronic devices (laptop, mobile phone, smartwatch) may be used in class only for course-related purposes. If a student uses such technology for personal reasons during class, the student may be penalized with an unexcused absence.

5.3) The use of computer or electronic technology during examinations is prohibited unless the instructor determines otherwise in accordance with the examination format.

6) Course Content

On occasion, and always within the context of the course, certain "sensitive" topics may be addressed both inside and outside the classroom. Therefore, if a student feels offended by such content in a particular class, the student may leave the class, and doing so will not have any negative effect whatsoever on the final grade. Faculty members will always be open to communication and dialogue in order to promote better intercultural understanding.

7) Final Clarification

The instructor may modify or add activities and/or change the dates indicated according to the learning needs of the group. Effective learning requires not only careful planning, but also a certain degree of flexibility.