DATA 513

Advanced Data Engineering

Tues, GPC 213, Portland, OR
Wed, Ford 102, Salem, OR
Summer 2026

Jed Rembold, PhD
jjrembold@willamette.edu
http://willamette.edu/~jjrembold/classes/data599
Ford 214
Office Hours: M,Th,F afternoons virtually by appointment
Office Phone: (503) 370-6860

This syllabus is subject to change or adaptation as the semester progresses.

Course Description: Data engineers design, implement, and maintain the increasingly large and complex data pipelines that power much of modern technology and society. Working with these pipelines requires a multifaceted combination of network savy, technical scripting, data modeling, and documentation. This course is constructed to plunge students into a semester-long project that has the necessary complexity to demonstrate and encourage good practices in each of these areas. Students should leave the course having strong technical and communication skills that could directly apply to a modern data engineering position, as well as an appreciation or acknowledgment of the complexity of all the work that goes into prepping data for later analysis.

Prerequisite(s): DATA 503
Credits: 4.0

Grade Weighting:

Milestones	35%
Documentation	35%
Reflections	20%
Participation	10%

Letter Grade Distribution:

>= 92.00	A	72.00 - 77.99	C
90.00 - 91.99	A-	70.00 - 71.99	C-
88.00 - 89.99	B+	68.00 - 69.99	D+
82.00 - 87.99	B	62.00 - 67.99	D
80.00 - 81.99	B-	60.00 - 61.99	D-
78.00 - 79.99	C+	<= 59.99	F

Textbooks:
This class has no required textbook. However, for those who are interested in pursing a data engineering position in the future, the following texts are highly recommended:

Text: *Fundamentals of Data Engineering* (1st edition)

Author: Joe Reis and Matt Housley

ISBN-13: 978-1098108304

Comments: This is the most recent book of the collection, and generally offers an excellent high-level overview of the field. Not much on specific applications, but a lot of excellent discussion about the types of tasks expected of data engineers and listings of common tools and practices.

Text: *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems* (1st edition)

Author: Martin Kleppmann

ISBN-13: 978-1449373320

Comments: It is an excellent resource, and really delves into the minutia of different storage systems. It is not perhaps the most applicable if you don't end up needing to do a ton of data architecture, and generally such big picture that it is difficult to put into practice. Nevertheless, it is still excellent for a broad theoretical background on how data is stored, especially with respect to replication and sharding: features that we will *not* be covering this semester.

Text: *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd edition)

Authors: Ralph Kimball and Margy Ross

ISBN-13: 978-1118530801

Comments: Data storage is often all about getting data in the best form for storage, where "best" can have many interpretations. This is the text that everyone talks about when thinking about dimensional modeling, and it has been around since the late 20th century.

Course Objectives:
Over the semester, students will gain working knowledge in:

- Understanding and implementing a modern data pipeline architecture, including data lake, data warehouse, and the orchestration binding them all together.
- Working productively on remote networks.
- Documenting data pipelines, including lineage, data dictionaries, and architectural design records.
- Adjusting to existing pipelines and building atop infrastructure build by others.
- Ingesting data from a wide variety of sources, including relational, graph, and document databases, logs, and event queues.
- Building architecture and dashboards to meet and inform evolving business and technological requirements.

Student Learning Objectives (SLO):
Upon completion of the course, students should be able to:

- Access remote systems through SSH and tunnel out important resources
- Deploy complex containers through use of Docker Compose
- Write and schedule DAGs in Airflow to move data through a pipeline
- Utilize S3 storage as a data lake, with daily snapshots
- Design and model fact and dimension tables tailored to specific business decisions
- Populate fact and dimension tables from snapshot data, including dimension tables with slowly changing dimensions
- Query data from document and graph databases
- Ingest data from paginated API endpoints and text logs
- Use streaming events to enrich data
- Write documentation that is useful for future pipeline maintainers
- Quickly learn an existing pipeline and meaningfully build new features atop it
- Work cohesively as a team to manage a complex and demanding technical system

Course Assessment:

- **Milestones**
  - There will be six milestones spaced across the semester, generally occurring every 1-2 weeks. Each of these milestones will contain a business decision, which will describe the real-world situation in which stakeholders would like to be able to make informed decisions. Students will be responsible for ingesting data from all possible sources, organizing it, and then modeling it in dimension and fact tables for later access from a Grafana dashboard. The dashboard is the end evaluated product here, and will be scored according to the insights it offers to the business decision, the accuracy of those insights, and the effectiveness with which it communicates those insights. When assigned, dashboards will be due at midnight at the end of Wednesdays.
- **Documentation**
  - Projects will be rotating between groups at several points during the semester, which will underscore the importance of excellent documentation. Before each rotation (of which there will be two) a group's documentation should be complete and uploaded to the server. This documentation should include not only used data models and data dictionaries, but the lineage of the data and a record of what and why important architectural decisions were made. The documentation will be evaluated by faculty, though some weight will be given to the experience and perspectives of the group which inherited the documentation. Documentation will be due at the end of the day preceding a group rotation.
- **Reflections**
  - As is the case with most large projects, there are a lot of intangibles that students will be contending with over the course of the semester. In order to get the most out of this project, students will be required to maintain reflection journals, in which they respond to weekly prompts. These weekly prompts will center around what struggles they had over the past week and how they overcame them, what they appreciated about a group member, what they could improve going forwards, and what outstanding questions they might still have. These are scored on a credit/no-credit basis, and should be submitted each week by the end of Wednesday.
- **Participation**
  - This is a highly collaborative course, as almost all work is done through groups. The expectation is that each group will have around 20-25 hours of work to be done each week, through the construction of new DAGs to ingest or model data, creating dashboards to deliver insight to business decisions, or keep up with the documentation. If someone is not pulling their weight, that unfairly piles more work on everyone else. Weekly variation is ok, as everyone has certain weeks when they will be more or less busy, but over the course of the semester all group members should plan to contribute evenly. These participation points are by default fully granted to everyone. The only time they might be deducted is if repeated instances of poor communication or group work sharing become apparent. Students will be first given a warning, and then points will be deducted on repeated infractions. This should be free points, provided you work well with your group.

Course Policies:

Late Work Policy

I understand that sometimes things come up where you are unable to get an assignment in on time, and I strive to be incredibly flexible and accepting of late work. In general, for milestones or reflection journals, you can turn things in up to 48 hours late and still receive full credit. Between 2-4 days late would be 50% credit, and anything past that will not be accepted. For deadlines before a project swap though (and thus for all documentation deadlines) you *must* have things turned in on time. Projects will be rotated the day after the deadline to give the new groups a chance to examine what they inherited in class that night, and so you will lose access to your old project. In the case of extenuating circumstances, please just come talk to me. We'll figure out what can be done.

Incomplete Policy

An incomplete grade will only be granted in the case of prolonged illness or family emergencies that remove the student from the learning environment for an extended time period during the semester. Under no situations will an incomplete be granted due to a student falling behind through lack of motivation, understanding, or time management skills. If you are concerned about your progress and how you are doing in the class, please come visit me! We can sort out where you are struggling and work out a plan to get you back on track.

Classroom Conduct

As an educational institution, Willamette is committed to support the ideals and standards that help create a constructive and healthy learning community. That requires, among other things, encouraging positive classroom behaviors, discouraging disruptive classroom behaviors, and setting clear standards for both of those things.

To that end, constructive classroom behaviors are those that support learners and teachers in an environment that promotes trust, respect, and collaborative learning.

Disruptive classroom behaviors are those that undermine or interfere with the abilities to learn and teach. Clear examples of disruptive behaviors include, but are not limited to:

- Interrupting others or persistently speaking out of turn
- Distracting the class from the subject-matter or discussion at hand
- Making unauthorized recordings or photos of a class meeting or discussion (except as permitted as part of an Accessible Education Services-mandated accommodation)
- Any physical threat, physical, psychological, or sexual harassment, ridicule, or abusive act towards a student, staff member, or instructor in a classroom or related setting.

Willamette Policies:

Academic Honesty

Cheating is defined as any form of intellectual dishonesty or misrepresentation of one's knowledge. Plagiarism, a form of cheating, consists of intentionally or unintentionally representing someone else's work as one's own. Integrity is of prime importance in a college setting, and thus cheating, plagiarism, theft, or assisting another to perform any of the previously listed acts is strictly

prohibited. I may impose penalties for plagiarism or cheating ranging from a grade reduction on an assignment or exam to failing the course. I can also involve the Office of the Dean for further action. For further information, visit: http://www.willamette.edu/cla/catalog/resources/policies/plagiarism_cheating.php.

Time Commitments

Willamette's Credit Hour Policy holds that for every hour of class time there is an expectation of 2-3 hours work outside of class. Thus, for a class meeting three hours a week, you should anticipate spending 6-9 hours outside of class engaged in course-related activities. Examples include study time, reading and homework, assignments, research projects, and group work.

Diversity and Disability

Willamette University values diversity and inclusion; we are committed to a climate of mutual respect and full participation. Our goal is to create learning environments that are usable, equitable, inclusive and welcoming. If there are aspects of the instruction or design of this course that result in barriers to your inclusion or accurate assessment or achievement, please notify me as soon as possible. Students with disabilities are also encouraged to contact the Accessible Education Services office in Smullin 155 at 503-370-6737 or accessible-info@willamette.edu to discuss a range of options to removing barriers in the course, including accommodations.

Tentative Course Outline:

The weekly coverage will almost certainly change as it depends on the progress of the class. However, this should serve as a rough guide.

Tuesday Classes:

Week	Date	Description	Due
1	Tue, May 12	Big Picture and Refreshers	
2	Tue, May 19	Tooling: SSH, Docker Compose, Parquet, DuckDB	
3	Tue, May 26	Introducing Airflow and the Project	
4	Tue, Jun 02 Wed, Jun 03	Airflow Connections, TaskFlow, and Intro to Warehousing	 Airflow Access 1
5	Tue, Jun 09 Wed, Jun 10	SCDs and Dashboarding	 Snapshots 1
6	Tue, Jun 16 Wed, Jun 17	APIs and Reliability	 Milestone 1
7	Tue, Jun 23 Wed, Jun 24 Thu, Jun 25	Document Databases	 Milestone 2 Project Rotations 1 (Documentation Due)
8	Tue, Jun 30	Document Databases	
9	Tue, Jul 07 Wed, Jul 08	Graph Databases	 Milestone 3
10	Tue, Jul 14	Graph Databases	
11	Tue, Jul 21 Wed, Jul 22 Thu, Jul 23	Kafka Events	 Milestone 4 Project Rotations 2 (Documentation Due)
12	Tue, Jul 28	Kafka Events	
13	Tue, Aug 04 Wed, Aug 05	Metrics and Logs	 Milestone 5
14	Tue, Aug 11 Wed, Aug 12 Thu, Aug 13	Sharing and Reflection Activities	 Milestone 6 Final Documentation Due

Wednesday Classes:

Week	Date	Description	Due
1	Wed, May 13	Big Picture and Refreshers	
2	Wed, May 20	Tooling: SSH, Docker Compose, Parquet, DuckDB	
3	Wed, May 27	Introducing Airflow and the Project	
4	Wed, Jun 03	Airflow Connections, TaskFlow, and Intro to Warehousing	Airflow Access 1
5	Wed, Jun 10	SCDs and Dashboarding	Snapshots 1
6	Wed, Jun 17	APIs and Reliability	Milestone 1
7	Wed, Jun 24 Thu, Jun 25	Document Databases	Milestone 2 Project Rotations 1 (Documentation Due)
8	Wed, Jul 01	Document Databases	

Week	Date	Description	Due
9	Wed, Jul 08	Graph Databases	Milestone 3
10	Wed, Jul 15	Graph Databases	
11	Wed, Jul 22 Thu, Jul 23	Kafka Events	Milestone 4 Project Rotations 2 (Documentation Due)
12	Wed, Jul 29	Kafka Events	
13	Wed, Aug 05	Metrics and Logs	Milestone 5
14	Wed, Aug 12 Thu, Aug 13	Sharing and Reflection Activities	Milestone 6 Final Documentation Due