DATA 503

Fundamentals of Data Engineering

Tues, 200 Market, Portland, OR
Thur, Ford 102, Salem, OR
Spring 2024

Jed Rembold, PhD
jjrembold@willamette.edu
http://willamette.edu/~jjrembold/classes/data503
Ford 214
Office Hours: MW 2:00-4:00, TTh 3:00-4:30, or catch me anytime online!
Office Phone: (503) 370-6860

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

Course Description: As "big data" more and more becomes a common facet of everyday life, the bulk of attention has been focused on the analysis and usage of this information. Such a focus ignores the vital fact that, no matter how much data is gathered, little analysis is possible unless that data has been stored and organized in such a way as to be easily accessible. As more depends on huge repositories of data, the importance of data integrity and scalability continues to increase. This course focuses on the basic skills of a data engineer tasked with acquiring, storing, and maintaining such repositories of information. To this end the course is broken roughly into two parts.

Data engineering is largely structured around the act of acquiring data from various sources, transformin git as necessary, and the storing it and making it accessible. This pipeline has various components and sees huge variation in the software used to build and maintain it. In this class we will endeavour to investigate and showcase each step of the pipeline with at least one viable software candidate. On the storage front, relational databases are one of the industry standards in storing and organizing information, and thus the other half of this class will revolve around learning how to create, manipulate, query, and maintain such relational databases using SQL. In particular, this class will focus on the open-source Postgresql variant of SQL, but the majority of learned techniques will be readily applicable to any other SQL variant. Students will leave the course having a broad theoretical foundation about the questions and trade-offs that underpin modern data storage, as well as feeling comfortable creating and utilizing a relational database to both store and query information.

Prerequisite(s): None
Note: A minimum grade of C- is required for this course to count toward university credit.
Credits: 4.0

Textbooks:
Only the SQL portion of this class has an accompanying text, listed below.

> **Text:** *Practical SQL: A Beginner's Guide to Storytelling with Data* (2nd edition)
> **Author:** Anthony DeBarros
> **ISBN-13:** 9781593278274
> **Comments:** This is the main book I'd suggest acquiring, especially if you haven't done much with SQL in the past. It is fairly cheap, and the older 1st edition will still work fine, the chapter numbers are just different.

If you are looking to go deeper, there are a few supplementary textbooks that you might find interesting. We will pull some ideas from these over the course of the semester, but they are absolutely not mandatory. I list them here only if you have an interest.

> **Optional Supplementary Text 1:** *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems* (1st edition)
> **Author:** Martin Kleppmann
> **ISBN-13:** 978-1449373320
> **Comments:** This is the text I used to teach the non-SQL content the first year I taught this course. 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. Hence why I've pivoted away from it in recent semesters, but it is still excellent for a broad theoretical background on how data is stored.

> **Optional Supplementary Text 2:** *Fundamentals of Data Engineering* (1st edition)
> **Authors:** Joe Reis and Matt Housley
> **ISBN-13:** 9781098108304
> **Comments:** This book *just* came out last summer, and I still haven't gotten a full chance to delve deep into it yet. By all metrics though, it seems like an excellent high level overview of the field of data engineering. 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 or approaches.

> **Optional Supplementary Text 3:** *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. Were I ever to teach a follow-up course to this class, we would spend a considerable amount of time with this text, and if you are strongly interested in going on in this field, it is an easy recommendation.

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

1. The basic tasks of a data engineer
2. The role a data engineer plays amidst organizational and analysis pipelines
3. The fundamentals of working with and querying a relational database using SQL
4. The key factors that drive database design choices
5. More advanced database queries involving text mining or spatial relationships
6. Alternatives to relational databases and future/current developments in the field

Grade Weighting:

Homework	55%
Project	30%
Midterm Video	15%

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

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

- Describe what a relational database is and what advantages or disadvantages they have over other forms of storing data.
- Design and implement a relational database, including creating multiple tables, parsing and inserting data into those tables, and including relationships between table columns.
- Query data from a database, including using advanced filters, descriptive statistics, and joins to combine information from multiple tables.
- Use SQL for analyzing more complicated types of information, including parsing text using regular expressions and analyzing spatial geometric information.
- Describe the role that a data engineer fits within a data analysis team and within a larger organization.
- Implement all aspects of the data engineering pipeline, including ingestion, transformation, and serving.

Course Assessment:

- **Homework**
  - There will be weekly homework sets which will be due Monday or Wednesday nights at 11:59pm. Homework sets will have an accompanying template where solutions can be added, and then be submitted through GitHub Classroom before the deadline. Assignments will be posted on the class webpage each week and the provided link at the top of the assignment should be followed to accept the assignment and to download any possible extra materials for that week's assignment. A subset of the problems will be randomly chosen to be graded, and solution sets provided that students can check their other problems against if they desire.
- **Project**
  - There will be a partner project that will be ongoing throughout much of the semester, but will culminate in a final presentation. The project will give students a chance to interact and explore several pieces of the pipeline of modern data engineering. Deliverables from the project will involve a presentation describing what was done, why certain design choices were made, how the database was organized, and what basic analysis was done to answer an interesting question. Some weekly homework problems will relate to the project as well.
- **Midterm Video**
  - There are currently no exams in this course. Historically, with smaller classes, I've been able to do an oral midterm, which I think has been very educational and useful to students, and helps prepare them for technical interviews. But that is impossible with 40-50 students. So instead, you will have one mid-semester assignment in which you will be tasked with showcasing a particular set of SQL objectives through an explanatory recording. You'll be expected to submit a 10-15 min screen recording of you achieving the stated objectives, explaining what you are doing and why you are doing it along the way. This will be done on your own time, so you can practice however many times you like or slice the video together in separate takes, but the final output should be an educational video similar to some you'd find posted on YouTube walking someone through the various tasks and objectives that you were given.

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. However, there also comes a point when you get too far behind to realistically keep up with the class. In an effort to compromise between the two, my late policy allots you 3 cumulative days (72 hours) of unpenalized late work throughout the entire semester. So you can turn 3 assignments in one day late, 6 assignments in 12 hours late, etc. without penalty. Once you have used up your 3 days (72 hours), assignments will drop in worth by 20% per day. If you are approaching the point where an assignment is heavily penalized, consider just turning in what you have, so that you can move on and keep up with the class. 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	Chapter	Description	Due
1	Tue, Jan 23	SQL: Ch 2-3	Data Engineering Described Tables and SELECT	
2	Mon, Jan 29 Tue, Jan 30	SQL: Ch 4-5	Data Sources Data Types and I/O	HW 1
3	Mon, Feb 05 Tue, Feb 06	SQL: Ch 6	The Shell Calculations with SQL	HW 2
4	Mon, Feb 12 Tue, Feb 13	SQL: Ch 6-7	Remote Connections Joining Tables	HW 3
5	Mon, Feb 19 Tue, Feb 20	SQL: Ch 7	Docker Containers Constraining Tables	HW 4
6	Mon, Feb 26 Tue, Feb 27	SQL: Ch 8-9	Scraping Grouping Showdown	Midterm Recording
7	Mon, Mar 04 Tue, Mar 05	SQL: Ch 12-13	Modeling and Normalization JSON and Date-Time	HW 5
8	Mon, Mar 11 Tue, Mar 12	SQL: Ch 10	Automating Transforms Inspecting and Modifying Data	HW 6
9	Mon, Mar 18 Tue, Mar 19		Subqueries, Crosstabs, and Window Functions	HW 7
10	Tue, Mar 26		*Spring Break*	
11	Tue, Apr 02		**Jed in Mexico for Eclipse**	
12	Mon, Apr 08 Tue, Apr 09	SQL: Ch 14	Regular Expressions Mining Text	HW 8
13	Mon, Apr 15 Tue, Apr 16	SQL: Ch 15	Serving Data Spatial Data with POSTGIS	HW 9
14	Mon, Apr 22 Tue, Apr 23	Ch 16	Views, Functions, and Triggers	HW 10
15	Tue, Apr 30		**Project Presentations**	

Thursday Classes:

Week	Date	Chapter	Description	Due
1	Thu, Jan 18	SQL: Ch 2-3	Data Engineering Described Tables and SELECT	
2	Wed, Jan 24			HW 1
	Thu, Jan 25	SQL: Ch 4-5	Data Sources Data Types and I/O	
3	Wed, Jan 31			HW 2
	Thu, Feb 01	SQL: Ch 6	The Shell Calculations with SQL	
4	Wed, Feb 07			HW 3
	Thu, Feb 08	SQL: Ch 6-7	Remote Connections Joining Tables	
5	Wed, Feb 14			HW 4
	Thu, Feb 15	SQL: Ch 7	Docker Containers Constraining Tables	
6	Wed, Feb 21			HW 5
	Thu, Feb 22	SQL: Ch 8-9	Scraping Grouping Showdown	
7	Wed, Feb 28			Midterm Recording
	Thu, Feb 29	SQL: Ch 12-13	Modeling and Normalization JSON and Date-Time	
8	Wed, Mar 06			HW 6
	Thu, Mar 07	SQL: Ch 10	Automating Transforms Inspecting and Modifying Data	
9	Wed, Mar 13			HW 7
	Thu, Mar 14		Subqueries, Crosstabs, and Window Functions	
10	Wed, Mar 20			HW 8
	Thu, Mar 21	SQL: Ch 14	Regular Expressions Mining Text	
11	Thu, Mar 28		*Spring Break*	
12	Wed, Apr 03			HW 9
	Thu, Apr 04	SQL: Ch 15	Serving Data Spatial Data with POSTGIS	
13	Wed, Apr 10			HW 10
	Thu, Apr 11		**Jed in Mexico for Eclipse**	
14	Thu, Apr 18	Ch 16	Views, Functions, and Triggers	
15	Thu, Apr 25		**Project Presentations**	