Carnegie Mellon University

Data Warehousing 95-797

Meeting Days, Times, Location: Wednesdays 7:30-8:30 EDT on Zoom

Semester: Mini-6**, Year:** 2023

Instructor information

Name	Pete Fein, adjunct professor
Contact Info	pfein@andrew.cmu.edu
Office hours	Online, schedule via https://calendly.com/pete-fein/cmu-office-hours

TA Information

TA name	Dibyanshu Patnaik
TA Contact Info	dpatnaik@andrew.cmu.edu
Office hours	Online, schedule via email

Course Description

Data warehouses are the central component of a modern data stack. They have solved the problem of analyzing massive amounts of structured & semi-structured data and are cost-effective, performant and easy to use. Data warehouses are the foundation for reporting, ad hoc analysis, business intelligence and machine learning, and enable collaboration among a diversity of users and stakeholders across organizations of all sizes.

This class will provide students with the conceptual background and hands on keyboard skills needed to utilize a data warehouse effectively. Throughout the course, students will work on an end-to-end development project, building a working data platform for New York City transit data. Using actual taxi, rideshare, bike share and weather data, students will answer real-world analytics questions, such as "How does location and time of day affect trip length?" and "How does weather affect transit preferences?". By the end, students will be empowered with the tools and techniques needed to take a real-world data project from problem statement to prototype to production.

Prerequisites

95–703 A: Database Management. Basic knowledge of programming (Python strongly recommended) and UNIX shell. Basic data analysis.

Learning Objectives

- Implement data ingest techniques (ETL)
- Transform data using dbt
- Maintain data quality
- Compare modern and classic strategies of data modeling
- Write advanced SQL for data analytics, including geographic and time series
- Understand data warehouse architecture
- Create reports, analysis & visualizations

Students will develop secondary skills in software engineering, testing, benchmarking, version control, reading technical resources and technology selection. They will learn tools to think with as well as tools to code with, ensuring future career growth as the technological landscape changes.

Weekly Course Schedule

- Saturday morning: week N (current) video lectures posted; reading list finalized
- Sunday 9 AM EDT: deadline to request extension for week N-1(previous) homework
- Sunday midnight EDT: week N-1(previous) homework due
- Wednesday afternoon: week N-1(previous) homework graded & canonical solution posted
- Wednesday 7:30 PM EDT: week N (current) quiz due
- Wednesday 7:30 – 8:30 PM EDT: Live recitation on Zoom. Review week N-1(previous) homework & homework kickoff for week N (current)

You should spend Sun/Mon/Tues/Wed watching lectures, doing readings & reviewing software documentation **before** the live recitation Wednesday night. Spend Thurs/Fri/Sat/Sun working on homework.

Coding Assignments

Each weekly coding assignment will be explained in a dedicated video lecture segment; see the corresponding slides for links to tools and software documentation you should use.

- Grading is based separately on the work you do each week.
- Initial setup: create a new repo in your GitHub account (`you`) from template
- Weekly procedure
  1. `git tag week-1-start`
  2. <hack, hack, hack>
  3. `git tag week-1-end`
  4. In Canvas, submit URL to GitHub diff view:
     `https://github.com/you/cmu-95797/compare/week-1-start...week-1-end`
- If you mess up the previous week, you can either:
  - manually integrate homework solution into your existing code base (be sure to tag `week-2-start` **afterwards, not before**)
  - Start over by creating a new repo from the template (no penalty!)

Live recitations on Zoom will be used to discuss previous week's homework solutions and kick off The current week's coding assignment. **Come prepared** having already reviewed software documentation.

Assessments

The final course grade will be calculated using the following categories:

Assessment	Percentage of Final Grade
Weekly quiz	10%
Participation	10%
Short answer questionsWill be used	30%
Coding assignments	50%

- Weekly quizzes: “did you watch it?” - one question per recorded lecture segment
- Participation: “did you show up?” - 0 or 1 each week, full credit earned for *any* participation in weekly recitations or on Canvas forum
- Short answer questions: “did you understand?” - based on lectures and readings
- Coding assignments: “can you implement it?” - evaluated on correctness, readability & comprehensibility, cleanliness, comments & documentation, and general software best practices

Students will be assigned the following final letter grades, based on calculations coming from the course assessment section.

Grade	Percentage Interval
A	[90-100%)
B	[80-90%)
C	[70-80%)
D	[60-70%)
R (F)	<60%

Grading Policies

- Quizzes are due prior to weekly recitation at **7:30 EDT Wednesday**.
- Homework is due **midnight EDT Sunday night**.
- Need an extension? Ask instructor by **9 AM EDT Sunday morning**.
- Late-work policy: late work will not be accepted without prior extension from instructor
- Make-up work & regrading policy: make up work and regrading are not available.

Course Policies

- **Academic Integrity & Collaboration**: These are **individual projects**: discuss with your classmates (preferably on Canvas) but do your own work**.**
- **Consent to record**: Live recitations on Zoom will be recorded and distributed to other students in this class section for review purposes; by attending recitation, you agree to such recording. **If you do not wish to be recorded, please contact the professor.**
- **Accommodations for students with disabilities**: If you have a disability and require accommodations, please contact Catherine Getchell, Director of Disability Resources, 412-268-6121, getchell@cmu.edu. If you have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate.
- **Statement on student wellness**: As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may diminish your academic performance and/or reduce your ability to participate in daily activities. CMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at: http://www.cmu.edu/counseling/. Support is always available (24/7) from Counseling and Psychological Services: 412-268-2922.

Learning Resources

General references:

- Mode SQL Tutorial
- DuckDB SQL documentation

Fundamentals of Data Engineering ("*FDE*" in reading list)
by Joe Reis, Matt Housley
Released June 2022
Publisher(s): O'Reilly Media, Inc.
ISBN: 9781098108304

Weekly Readings

Bold items are required; everything else is optional. Core concepts are covered in video lectures; these resources expand and further explain those topics. Supplemental readings may be helpful in completing coding projects (assignments) & explaining technical concepts (short answer questions). **Watch Canvas announcements for updates!**

1. **Data ingest (ETL)**
    a. **What is ELT (Extract, Load, Transform)?**
    b. ETL vs ELT: Differences and Similarities | Snowflake Guides
    c. *FDE*: chapter 7, pp. 237-251, 255-259, 266-267
2. **Transform data using dbt**
    a. **Introduction to dbt (data build tool) from Fishtown Analytics**
    b. **How to build a mature dbt project from scratch (w/ Dave Connors)**
    c. **Excelling at dbt: Jinja & Macros for modular and cleaner SQL Queries**
    d. Data Wrangling with SQL | Advanced SQL - Mode

e. Using SQL String Functions to Clean Data | Advanced SQL - Mode

3. **Maintain data quality**
   a. **Tutorial: Using dbt to Test for Schema Changes**
   b. How to Manage Data Quality: A Comprehensive Guide
   c. Open source interviews #14 - Maayan Salom, founder of Elementary Data
   d. Compounding Quality
4. **Compare data modeling strategies**
   Note: ignore all mentions of Data Vault
   a. **Babies and bathwater: Is Kimball still relevant?**
   b. **Introducing the activity schema data modeling with a single table**
   c. Data Vault vs Star Schema vs Third Normal Form: Which Data Model to Use?
   d. Back to the Future: Where Dimensional Modeling Enters the Modern Data Stack
   e. Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault
   f. FDE Chapter 8, pp. 291-310
5. **Advanced SQL for data analytics**
   a. **SQL Joins | Intermediate SQL - Mode**
   b. **SQL Aggregate Functions | Intermediate SQL - Mode**
   c. **SQL Window Functions | Advanced SQL - Mode**
   d. **Use Common Table Expressions (CTE) to Keep Your SQL Clean | Mode**
   e. Back to the Basics With SQL- Date Functions
6. **Data warehouse architecture**
   a. **Databases Demystified Chapter 3 – Row Store vs. Column Store | Blog | Fivetran**
   b. **Benchmark (computing) - Wikipedia** through “Benchmarking Principles”
   c. Understanding Snowflake Table Structures
   d. MPP: The Transformation on Big Data Analytics | by Maggy Hu | Slalom Technology | Medium.
   e. (optional) Designing Data-Intensive Applications (available on Canvas), chapter 3, pp. 90-103, chapter 6 pp. 199-219
7. **Reporting and BI**
   a. **A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional Data | by Dipanjan (DJ) Sarkar**
   b. Reporting vs. Analytics: What's the Difference? | Indeed.com
   c. Difference Between Business Analytics and Predictive Analytics - GeeksforGeeks
   d. Reporting, Predictive Analytics, and Everything in Between (available on Canvas), pp. TBD