Syllabus for I320: Data Engineering

Welcome to I320: Data Engineering in the School of Information at the University of Texas at Austin.

Semester	Fall 2023
Classroom	CBA 4.330
Class times	12:30-2:00 Tuesdays and Thursdays
Unique Number	28370

Jump to Course Schedule.

Lab link

Instructions for how to access your VM environment are here: AWS VM Environment Access
Superset access: https://superset.dei320.net
Jupyter access: https://notebook.dei320.net

Course Description

This class will be a foundational course in Data Engineering principles and practices. This course will enable you to:

- Understand the professional role of data engineers in organizations and career paths for data professionals.
- Understand the data engineering lifecycle.
- Use SQL to transform and query data.
- Understand data modeling techniques for organizing and managing data.
- Build data pipelines to collect, transform, analyze, and visualize data from operational source systems.

The class will balance general principles with hands-on experience with some of the tools, languages, and techniques of the modern data stack. Emphasis will be placed on SQL as the primary language of data engineering along with low- or no-code tools that leverage SQL. We'll walk through building data pipelines end-to-end, from ingesting source data to creating analytical data products that deliver value to organizations. We'll use business intelligence tools to build visualizations using those data products. We will look at both batch processing and streaming systems to understand their pros and cons. We'll talk about data lakes, data warehouses, ETL/ELT, and batch and streaming systems to understand the pros and cons of each. Time permitting, we will look at issues around data quality, understand the uses of data catalogs, examine data

lineage and data profiling tools, and discuss data governance in organizations. Finally, we'll discuss trends and future directions in data engineering.

Some python or other programming languages are helpful. INF 320D Database Design and INF 320D Data Visualization, are also helpful.

Professors and Office Hours

This course is taught by Chip Young, a working data professional.

Office hours are by email request. Also happy to talk before class, please email to set that up. Feel free to reach out, happy to chat about anything, including career paths, other courses, life in organizations, and getting the best out of your UTexas experience.

Canvas link

The course space on Canvas is available at: https://utexas.instructure.com/courses/1341797

Land Acknowledgement

We would like to acknowledge that when we are meeting on Indigenous land. Moreover, we would like to acknowledge and pay our respects to the Carrizo & Comecrudo, Coahuiltecan, Caddo, Tonkawa, Comanche, Lipan Apache, Alabama-Coushatta, Kickapoo, Tigua Pueblo, and all the American Indian and Indigenous Peoples and communities who have been or have become a part of these lands and territories in Texas, here on Turtle Island.

Course Objectives

Learn fundamentals of data engineering.
Be able to apply the principles used in class to build a simple data pipeline and visualize the data.
Prepare students for careers as data professionals.

Computing Resources

You need a laptop with a browser to access the data visualization tool we will use. You will be using a virtual machine (VM) on a cloud service to do most of your work. The software used in this class will be installed on your VM or as a cloud service. However, you will need your own laptop for class, able to access the utexas Wi-Fi network. If you do not have a laptop, or yours stops working, the school and university has resources available. Please check these university resources. Check the "Before your classes" section; I believe that you

reach out to the Texas One Stop. We will work to have one or two loaner laptops available since we know things sometimes break just before class.

Class Recordings

The course is an in person course; you should plan to attend each and every class. However, we know that recordings can be very useful for unavoidable missed classes and for reviewing in-class material when working on homework or studying.

Therefore, this class is using the Lectures Online recording system. This system records the audio and video material presented in class for you to review after class. Links for the recordings will appear in the Lectures Online tab on the Canvas page for this class. You will find this tab along the left side navigation in Canvas.

To review a recording, simply click on the Lectures Online navigation tab (in Canvas) and follow the instructions presented to you on the page. You can learn more about how to use the Lectures Online system at http://sites.la.utexas.edu/lecturesonline/students/how-to-access-recordings/.

You can find additional information about Lectures Online at: https://sites.la.utexas.edu/lecturesonline/.

Course Texts

There are no required texts for the course, but you will find these resources to be useful.

An intro book for MySQL that's available online at UT is: Learning MySQL.

As a member of this class you will have free access to the DataCamp site through their support for education (more at datacamp.com/groups/education, I believe that access extends for 6 months. I will establish the access a few weeks into the semester, causing an invitation email to come to the email address registered with the University. The most relevant courses are:

- Introduction to Python
- Intro to SQL for Data Science
- Joining Data in SQL

Course Schedule

[subject to change as course materials are developed]

Week 1: Introduction to Data Engineering (Aug 22/24)

Introductions
Syllabus review
Definition and Overview of Data Engineering

What is Data Engineering presentation
Overview of example end-to-end project
Overview of semester project

- Data Engineering Introduction Slides
- Data Transformation Exercise

Week 2: Introduction to Data Pipelines/End-to-End Presentation (Aug 29/Aug 31)

Presentation and distribution of sample end-to-end project

- Data Engineering Pipeline Overview

Discussion of semester projects

Weeks 3-5.1: SQL Review (Sep 5/7 12/14 19)

Link to SQL material

Into to Postgres and psql
SQL Basics Review
Aggregate Functions (COUNT, SUM, AVG) Different types of Joins especially Outer Joins
SQL mini-quiz

Week 5.2: Introduction to Semester Project (Sep 21)

Semester Project Instructions:

Semester Project

Requirements Documents:

Austin Animal Center Outcomes
COVID 19 Cases by US County
Dancing with the Stars Internet Movie Database
Olympic Events for all Olympics
Saturday Night Live

Week 6: Source Systems and Data Ingestion (Sep 26/28)

Link to Ingestion Presentation
Link to Bulk Load

What is a Data Lake?
What is a Data Warehouse?
Data Lakehouses

Source Systems
Replication of source data
Batch Processing
Streaming
Bulk ingestion using the Copy command
Workshop on ingesting data for semester project

Weeks 7-8: Data Modeling (Oct 3/5 and 10/12)

ER diagrams and modeling transactional systems

- one-to-many
- many-to-many
- many-to-many with attributes Normalization
  Dimensional Modeling (Star Schema)
- Dimensional Modeling, Part 1
- Dimensional Modeling, Part 2 Assignment: Queries from Star Schema

Week 9: Data Transformation (Oct 17/19)

Data Products
Data Products
Introduction to dbt
dbt Intro

Week 10: Data Presentation and Visualization (Oct 24/26)

Business Intelligence Tools - Superset
Introduction to Superset
More Advanced visualizations
Creating visualizations with Superset

Week 11: Workshop on Semester Projects (Oct 31/Nov 2)

Review Semester Project Instructions:

Semester Project

Project Report Template:

Project Report Template

Week 12: Workshop on Semester Projects (Nov 7/9)

Week 13: Workshop on Semester Projects (Nov 14/Nov 16)

Thanksgiving break (Nov 20-25))

Week 15: Trends and New Directions in Data Engineering (Nov 28/30)

Nov 28 - Project Presentations
Nov 30 - Wrap-Up: Data Engineering concepts, careers, and interviewing

Assessments

Course grades will be assiged based on performance in the course assessments (see below for details):

- Weekly assignments (~10): 40%
- Semester Project: 60%
  - Group Pipeline code and dashboard: 40%, Due Nov Friday 16 (before Thanksgiving break)
  - Group presentation: 5%, Tuesday Nov 28 (just after Thanksgiving break)
  - Individual interview about group project: 15% (scheduled during last week of classes (Mon Nov 27-Mon Dec 4).

The assignments and grading scheme (A, A-, B, ..., F) are shown in Canvas. Assignments will be submitted through Canvas.

Weekly Assignments

40% of your courser grade will come from Weekly Assignments. These are assignments each week for this course, covering the material addressed that week, with the assignment released and introduced during class on Thursday. The weekly assignments are due 11:59 pm on Sunday (this is to ensure that we can grade them before Tuesday class). Late assignments *will receive a grade of zero* but you can drop your 2 lowest grades. It's always worth turning in the assignment, even if late, because the assignments test and drive your learning and your performance helps guide me on material. So not completing an assignment is a sure way to fall behind. Students have used their drops in the past and then been very sad when an actual emergency meant they could not complete their homework.

If technology fails you (broken laptops, server issues) and this means that you have difficulty with your homework you should complete and submit as much as can be done without the computer (e.g., hand drawn diagrams, writing out parts of queries, describing pipeline elements). Describe the issues that you have faced and the professors will consider excusing the remainder of the assignment, or may require you to use one of your drops.

If you've uploaded a PDF as part of the assignhment there will be comments left on the PDF, in addition to any text comments in Canvas. You can see the comments on the PDF via by using the "viewing feedback" button.

Weekly assignments should take about 1-2 hours. If you are spending 3 or more hours on the homework you are spending more time than expected; reach out for a meeting with a course professor.

Project

60% of your course grade will come from a group project to a data engineering workflow (sometimes also called a "data pipeline").

The project will be done in groups (to be determined but likely 4 students).

The project will consist of three elements, two graded as a group and one graded individually.

1. A working data pipeline, using the technologies taught in this class, including vizualization dashboards, together with a written commentary (README.md) describing the pipeline and the challenges overcome will be submitted (3-4 pages including illustrations). Worth 40% of course grade.
2. A group presentation built using Markdown and presented in class on Tuesday Nov 28 (presentation 5-8 minutes, around 6 slides). Worth 5% of course grade.
3. An individual interview with course professor to explain the project and answers questions about the techniques and results. Worth 15% of course grade.

Working in groups for this course does not mean dividing up the work; We require each group member to understand everything about their project. If there is a part of your project that you don't understand or couldn't work with then you are missing a crucial learning opportunity. We work in groups to work together (supporting each other's learning) not to reduce the amount or diversity of the work that we do. The individual interview helps us assess how each group member understands the project as whole.

We will provide each group with data and requirements for their project. The datasets will be randomly allocated to groups. Only a single group will be working with each dataset.

The group project should take 15-20 hours over a number of weeks. We will cover each element expected in the project during the course materials and have in-class workshops for 3 weeks in October and November.

Policies

Academic Integrity

Each student in the course is expected to abide by the University of Texas Honor Code: "As a student of The University of Texas at Austin, I shall abide by the core values of the University and uphold academic integrity." Plagiarism is taken very seriously at UT. Therefore, if you use words or ideas that are not your own (or that you have used in previous class), you must cite your sources and use quote marks appropriately. Otherwise, you will be guilty of plagiarism and subject to academic disciplinary action, including failure of the course. In particular, students are reminded that proper citation requires mentioning sources when you use them, not just in a general list of references at the end of a document. You are responsible for understanding UT's Academic Honesty and the University Honor Code. If this is at all confusing, please take this Plagiarism Tutorial.

In particular, **any time you use the copy function** from someone else's writing (e.g., an article, blog post) you must have a plan about how you will use those words, how you will use quote marks ("), and how you will cite the work.

Collaboration policy

The weekly assignments are individual work. However, as long as you meet the condition below, I give you explicit permission to work together with other classmates on the assignments or on your projects. With the same condition, you are also welcome to seek input from people outside the class, such as friends and family.

The one condition is that you add a note to your homework (ideally through a comment in the Canvas submission) indicating how the work was done and identifying with whom you worked and how (thus ensuring that we are following the Academic Integrity policy above). For example, you might say "Daria and I worked on this in the lab together, when we started out we were confused about X but I figured it out and shared that with Daria. Our code is very similar because we worked together". Or perhaps "I was confused about how to pad a string with spaces, and after working at it for 30 minutes I chatted about it with my partner who suggested the xyz method. I was pleased when I got that working myself." When you have worked together your code will have similarities, but you must not turn in identical code; rather you should take code you've worked on together and personalize it through comments that explain what is happening in the code. The comments must be your own, individual, work.

Neither "working together" nor "seeking input" means having others do the work for you; you should always be certain that you are learning and that you understand the code that you have submitted.

If you have questions on this policy please ask in the Assignment Discussion forum on Canvas and I will answer there. I have this policy because learning to program is both individual hard work and learning how to get help from others. Sometimes chatting through with another class member is just what is needed.

Sharing of Course Assignment Materials is Prohibited

No assignment materials used in this class, including, but not limited to quizzes, exams, papers, projects, homework assignments, review sheets, and additional problem sets, may be shared online or with anyone outside of the class unless you have my explicit, written permission. Unauthorized sharing of materials promotes cheating. It is a violation of the University's Student Honor Code and an act of academic dishonesty. I am well aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course.

Class Recordings

Class recordings are reserved only for students in this class for educational purposes and are protected under FERPA. The recordings should not be shared outside the class in any form. Violation of this restriction by a student could lead to Student Misconduct proceedings.

COVID Caveats

To help keep everyone at UT and in our community safe, it is critical that students report COVID-19 symptoms and testing, regardless of test results, to University Health Services, and faculty and staff report to the

HealthPoint Occupational Health Program (OHP) as soon as possible. Please see this link to understand what needs to be reported.
In addition, to help understand what to do if a fellow student in the class (or the instructor or TA) tests positive for COVID, see this University Health Services link.

Student rights and responsibilities

- You have a right to a learning environment that supports mental and physical wellness.
- You have a right to respect.
- You have a right to be assessed and graded fairly.
- You have a right to freedom of opinion and expression.
- You have a right to privacy and confidentiality.
- You have a right to meaningful and equal participation, and to self-organize groups to improve your learning environment.
- You have a right to learn in an environment that is welcoming to all people. No student shall be isolated, excluded or diminished in any way.

With these rights come responsibilities:

- You are responsible for taking care of yourself, managing your time, and communicating with the teaching team and with others if things start to feel out of control or overwhelming.
- You are responsible for acting in a way that is worthy of respect and always respectful of others.
- Your experience with this course is directly related to the quality of the energy that you bring to it, and your energy shapes the quality of your peers' experiences.
- You are responsible for creating an inclusive environment and for speaking up when someone is excluded. In particular, you are responsible for ensuring that your participation does not exclude the participation of others. Office hours are available for in-depth further discussion of advanced topics or other interests that pursuing in depth during class would exclude others.
- You are responsible for holding yourself accountable to these standards, holding each other to these standards, and holding the teaching team accountable as well.

Personal Pronoun Preference and Pronounciation

Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, culture, religion, politics, sexual orientation, gender identity & expression, and nationalities. Class rosters are provided to the instructor with the student's legal name, unless they have added a "chosen name" with the registrar's office, which you can do so here: https://utdirect.utexas.edu/apps/ais/chosen_name/.

We will gladly honor your request to address you by a name that is different from what appears on the official roster, and by the pronouns you use (she/he/they/ze, etc.). Please advise us of any changes early in the semester so that I may make appropriate updates to my records. For instructions on how to add your pronouns to Canvas, visit https://utexas.instructure.com/courses/633028/pages/profile-pronouns. More resources available on the Gender and Sexuality Center's website, www.utgsc.org.

While I wish that we were all able to pronouce each others names from the way they are written and replicate the correct way each of us says our names, this is a challenge we all face. To help us learn to match correct pronounciations I have two suggestions:

The canvas site has NameCoach enabled to help us all listen to learn to pronounce each other's names. Please ensure you have a recording in NameCoach.

Please also consider creating a respelling pronounciation guide and include it in your emails. For example, James can be rendered as "Jaymz" or the US state of Arkansas as "Ar-kuhn-saw", or Beyonce as "Be-yon-say". James finds the Wikipedia respelling key the most useful starting point.

Basic Needs Security

Any student who faces challenges securing their food or housing and believes this may affect their performance in the course is urged to contact the Dean of Students for support. UT maintains the UT Outpost (https://deanofstudents.utexas.edu/emergency/utoutpost.php) which is a free on-campus food pantry and career closet. Furthermore, please notify the professor if you are comfortable in doing so. This will enable him to provide any resources that he may possess.

Mental Health Resources

I urge students who are struggling for any reason and who believe that it might impact their performance in the course to reach out to me if they feel comfortable. This will allow me to provide any resources or accommodations that I can. If immediate mental health assistance is needed, call the Counseling and Mental Health Center (CMHC) at 512-471-3515 or you may also contact Bryce Moffett, LCSW (iSchool CARE counselor) at 512-232-2983. Outside CMHC business hours (8a.m.-5p.m., Monday-Friday), contact the CMHC 24/7 Crisis Line at 512-471-2255."

Drop Policy

If you want to drop a class after the 12th class day, you'll need to execute a Q drop before the Q- drop deadline, which typically occurs near the middle of the semester. Under Texas law, you are only allowed six Q drops while you are in college at any public Texas institution. For more information, see: http://www.utexas.edu/ugs/csacc/academic/adddrop/qdrop

International students *must* meet with the international office before dropping a class that would put them below full-time status.

University Resources for Students

Your success in this class is important to me. We will all need accommodations because we all learn differently. If there are aspects of this course that prevent you from learning or exclude you, please let me know as soon as possible. Together we'll develop strategies to meet both your needs and the requirements of the course. There are also a range of resources on campus, detailed below.

Services for Students with Disabilities

This class respects and welcomes students of all backgrounds, identities, and abilities. If there are circumstances that make our learning environment and activities difficult, if you have medical information that you need to share with me, or if you need specific arrangements in case the building needs to be evacuated, please let me know.

I am committed to creating an effective learning environment for all students, but I can only do so if you discuss your needs with me as early as possible. Requests for accommodations are quite normal and quite frequent and I promise to maintain the confidentiality of these discussions. If appropriate, also contact Services for Students with Disabilities.

Counseling and Mental Health Center

All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. http://www.cmhc.utexas.edu/individualcounseling.html

The Sanger Learning Center

All students, including graduate students, are welcome to take advantage of Sanger Center's classes and workshops, private learning specialist appointments, peer academic coaching, and tutoring for more than 70 courses in 15 different subject areas. For more information, please visit https://ugs.utexas.edu/slc/grad or call 512-471-3614 (JES A332).

University Writing Center free programs for grad students

Libraries

IT services

Student Emergency Services

Important Safety Information

If you have concerns about the safety or behavior of fellow students, TAs or Professors, call BCAL (the Behavior Concerns Advice Line): 512-232-5050. Your call can be anonymous. If something doesn't feel right – it probably isn't. Trust your instincts and share your concerns.