Under development

I 320D
Topics in Human-Centered Data Science : Data Engineering
Course Syllabus

Course Description

This class will be a foundational course in Data Engineering principles and practices. We will focus on the data engineering lifecycle and how to build data pipelines to collect, transform, analyze, and visualize data from multiple source systems. We will discuss data modeling techniques for organizing and managing data. We will look at data as an organizational asset and as a product. We will examine the various roles data engineers can have in an organization and career paths for data professionals.

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, plus a little python. 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. 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. Time permitting, we'll also discuss trends and future directions in data engineering.

Prerequisites

You need to know basic fundamentals of SQL. Some python or other programming language is helpful. INF 385M (Database Management), INF 385T.9 (Data Wrangling), or INF 380P (Introduction to Programming).

Class Meetings

Tues/Thurs 11-12:30.

Office Hours

As requested. We will meet with students at any time that's convenient with at least 24 hours notice.

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 at least 8 GB of memory. The software used in this class will run in the cloud (via an internet connection). 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).

Course Schedule

1) Introduction to Data Engineering

- Definition
- A Brief History of Data Engineering
- “Data Is The New Oil” presentation
- Mini-quiz

2) Introduction to Data Pipelines/End-to-End Presentation
    - Presentation and distribution of sample end-to-end project
    - Installation
    - Processing steps in the data pipeline w/examples
        - Source systems
        - Ingestion
        - Data cleansing and validation
        - Data transformation
        - Presentation and Visualization
    - ELT vs ETL

    Semester Project

    - Assignments of data sources
    - Project Definition
    - Assignment: Make enhancements to sample project

3) SQL Review
    - Into to psql/Snowflake (depending on whether we use Postgres or Snowflake)
    - SQL Basics Review
    - Different types of Joins especially Outer Joins
    - Advanced SQL Features - subqueries, CTE’s, and Window functions
    - SQL mini-quiz

4) Source Systems and Data Ingestion
    - Source Systems
        - Replication of source data
        - Batch Processing
        - Streaming
    - Bulk ingestion using the Copy command
    - Workshop on ingesting data for semester project

5) Data Lakes and Data Warehouses
    - What is a Data Lake?
    - What is a Data Warehouse?
    - Data Lakehouses

    Data Cleansing and Validation

- Data Quality of Source Systems
- Statistical validation
- Rule-based validation

6) Data Modeling
    - Normalization
    - Dimensional Modeling
    - Creating Tables
    - Schema Migration
    - Assignment: Create dimensional model for semester project

7) Data Transformation
    - Building the data warehouse
    - Transforming source data into dimensional models
    - Building data products (data marts)
    - Assignment: Populate dimensional mode for semester project

8) Data Presentation and Visualization
    - Business Intelligence Tools
    - Introduction to Superset
    - Creating visualizations
    - Assignment: Create visualization for semester project

9) Workshop on Semester Projects

10) Data and Metadata Management and Governance
    - Data Quality
    - Data Catalogs, Data Lineage, and Data Governance
    - DataOps and Data Observability
    - Mini-quiz

11) Trends and New Directions in Data Engineering
    - Data Mesh
    - CDC/Streaming/Event Processing for near real-time analytics
    - Reverse ETL

12) Careers in Data Engineering
    - Data Engineer as an Analytics Engineer
    - Data Engineer as a Software Engineer
    - Data Engineer as a Data Scientist
    - Data Engineer as an Infrastructure Engineer

Assignments

Assessment will be based on weekly assignments and a group project.

Weekly assignments will take between 1-3 hours to complete, and make up 60% of the course grade.

The group project will ask students to work with their choice from 5 projects curated by the professor. Each project will include around 5 source datasets which the project will require you to integrate, transform, validate, analyze, and visualize. Each group will be asked to locate additional datasets online to give their particular spin to the underlying project. Each project will include changes released by the professor during the project (e.g., shifts in schemas, shifts in scale, new data sources to integrate).

Projects to choose from will include cases in Health, Business, Social Media Analysis, Cultural Heritage and other topics. Each will include either public domain or simulated examples of relevant datasets (e.g., ERP, CRM, ticketing systems for business, EHR and demographic data for Health, Library/Archive catalog and patron system datasets for Cultural Heritage). Additional data cases might include topics in Sports (e.g., Title 9 analysis of university sports equity), or Culture (e.g., analysis of reality TV shows).

The group project will include individual assignments early in the semester where each member of the group works alone through the material, ensuring each member has familiarity with the material before beginning group work.