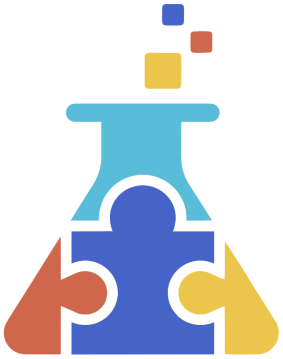
Jigsaw Labs

Data Engineering Career Course 2023 Syllabus



Overview:

This course consists of two modules: backend engineering and data engineering.

In the backend engineering module, students learn coding fundamentals in Python and SQL, and how to build backend APIs with Flask. Students also learn backend engineering skills like writing and reading tests, object oriented design, data modeling, and how to integrate these skills in a large codebase (~1k lines of code).

In the data engineering module, students continue with the curriculum for ⅔ of class time, and spend ⅓ of class time working performing technical work for external organizations through externships (more below). Through the data engineering curriculum, students learn advanced SQL, cloud computing, analytics databases (snowflake, redshift), and building data pipelines (with Airflow, AWS services, and DBT). Students also begin technical interview prep by learning data structures, and completing leetcode style problems in Python and SQL.

The course is designed to teach students the various tools and techniques necessary to become a data engineer, while repeatedly emphasizing fundamental engineering skills: Python, SQL, data modeling and object oriented design.

Schedule

Tues, Weds and Thurs | 6:30 - 9:30 pm EST

Sundays | 12:30 pm - 9:30 pm EST

Next cohort: Jan 24 2023 - July 11 (24 weeks)

Teaching Format

All classes are conducted live online via zoom.

Days consist of 30 minute live lectures followed by interactive readings and labs performed either individually or in pairs.

Quizzes are regularly delivered to assess understanding.

Module 1 - Backend Engineering

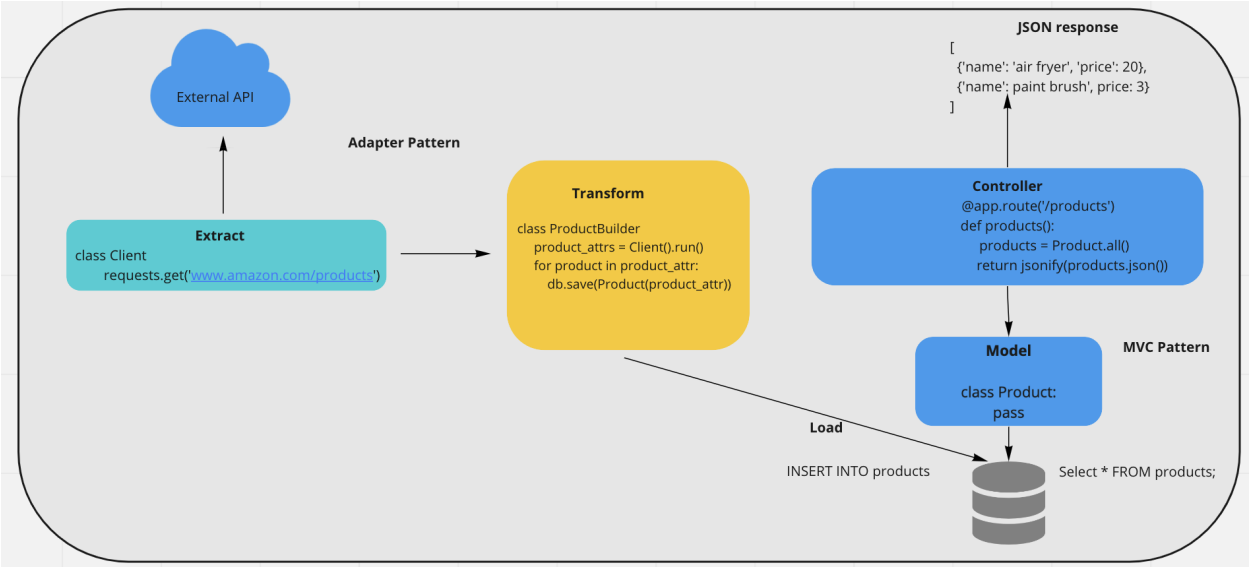
This module prepares students to navigate, write, and contribute to large backend codebases in Python and SQL. As such, we can think of this course as training students on two main subjects: (1) Software Engineering fundamentals and (2) Object Oriented Design Patterns.

Software Engineering Fundamentals

With software engineering fundamentals, students will learn to retrieve and manipulate data in Python, with a focus on writing clean functions to retrieve and coerce the data. We'll also cover SQL, where we will learn both single table queries, and relational queries, and practice data modeling in a normalized database.

Object Oriented Design Patterns

The latter part of the module is focused on object oriented programming in Python, and related design patterns. Here, students learn about the MVC design pattern for building backend Flask applications, the object relational mapping pattern for mapping data from Python to SQL, and the adapter pattern for pulling data from external sources and loading them into the database. Students will learn to read, contribute to, and debug a large code base.



Module 2 - Cloud Computing and Data Pipelines

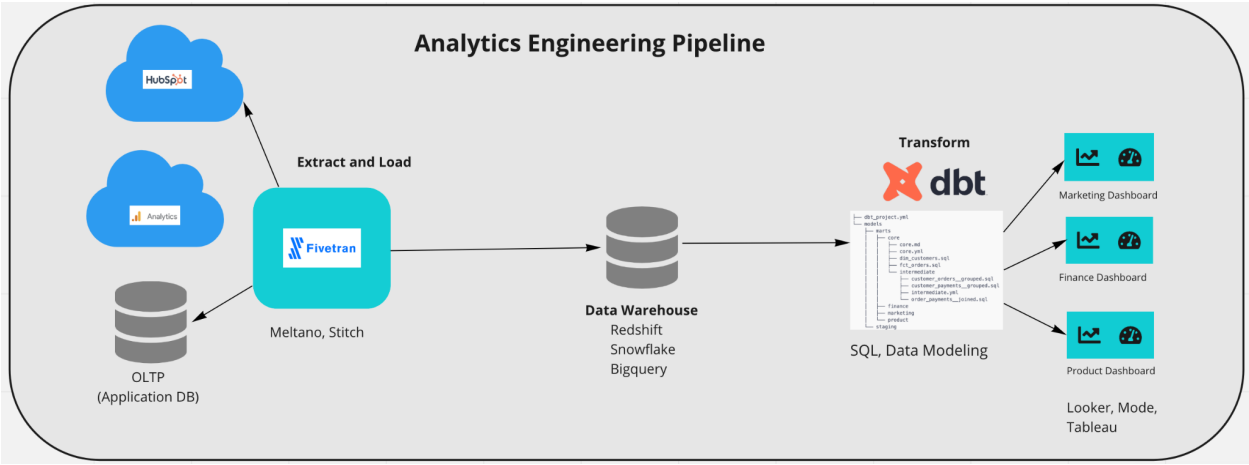
Cloud Computing

Students will deepen their knowledge of working in the terminal, and then deploy their Flask API to AWS instances, connecting the API to an RDS database in AWS. Afterwards we will learn about Docker – using both pre-existing images and building our own.

Data Pipelines

With data pipelines, we'll learn to move data from a transactional to an analytics based database, and the data modeling and data queries that come with it. We'll learn how to build an ETL data pipeline by first writing SQL that coerces and exports our data, then transferring our data from our RDS database to S3, and finally to Redshift. We'll automate this workflow with Airflow.

We'll then learn how to construct an analytics engineering pipeline with an ELT pattern, using Fivetran, Snowflake, and coercing our data using design patterns in DBT.



Externships - Production Projects

Throughout the second half of the course, students are partnered with outside companies or open source projects where they will practice working with a larger engineering team, and contributing to a professional codebase. We allocate six hours of in class time per week – Wednesday and Sunday evenings – and many students devote additional hours outside of class. The externships give students the opportunity to gain technical and non-technical skills involved when working with a project to be used in production.

Supplementary Material

In addition to the material listed in the syllabus, throughout the course we will review Python and SQL using leetcode style questions and covering computer science fundamentals. This includes big O notation and data structures like lists, dictionaries and histograms. We'll similarly review SQL, discussing techniques for tackling leetcode style SQL queries.

Course Schedule

The class is seventeen hours per week with classes running from 12:30 pm to 9:30 pm on Sundays (one hour break), and from 6:30 pm - 9:30 pm on Tuesdays, Wednesdays and Thursdays. All classes are held via Zoom.

Office hours are available to students weekly to review material or receive other assistance.

Module 1 - Backend Engineering (Weeks 1 - 13)			
Week	**Chapter**	**Topic**	**Description**
Week 1	Data Types in Python	Data Coercion and List Comprehension	A deeper look at lists, strings, dictionaries, and coercing to different data types
Week 2		Mapping and Filtering data	Practice coercing and filtering data using list comprehensions. Airbnb lab.
Week 3	Functional Python	Functions and APIs	Learn to write clean, small functions We'll learn how to retrieve and coerce data from APIs
Week 4	Scripting	Command Line, Python Scripts, Tests	Start working in the terminal, and running our code in scripts. We'll learn how to write tests for our functions.
Week 5	Intro to Databases	Single Table and Relational Queries	Learn SQL fundamentals including selecting data, performing aggregates, joins and data modeling
Week 6		Postgres	Learn how to use the postgres from the CLI and the psycopg2 library, practicing our relational queries and database design.

Week 7	Object oriented programming	Intro to Objects	Learn object fundamentals, including self, instance and class methods, and testing objects
Week 8		Object Relations	Relate various classes together with relational queries in Python and practice testing.
Week 9		OOP Projects	Build an OOP project from scratch, isolating models and from service objects
Week 10	Intro to building APIs	Flask and Github	Learn both git and github. We'll introduce Flask and MVC in Python.
Week 11		ORMs from Scratch	Write out and then use ORMs in our Flask application. We'll integrate relational queries in our codebase.
Week 12	ETL with APIs	Adapter Pattern	Use the the adapter pattern to query data from APIs, selecting and coercing the data to load into each corresponding table
Week 13		Adapter with APIs	Build a codebase using the adapter pattern to load data, and MVC to query and serve data

Module 2 - Cloud Computing and Data Pipelines (Weeks 14 - 24) *Externships: Sunday evenings and Wednesdays*			
Week	**Chapter**	**Topic**	**Description**
Week 14	Advanced SQL	Window Functions and Subqueries, and Outer Joins	Learn about aggregates, running totals, and rank based window functions We'll learn about using subqueries that return a scalar, vector, and table
Week 15	AWS Fundamentals Big O	Advanced terminal, EC2, RDS	Learn how to our API to an EC2 machine, and connect the database to an AWS RDS database Learn about Big O, and review Python with the brute force approach.
Week 16	Docker	Docker CLI and containers	Learn the differences between containers and images, and learn to use bind mounts

			and volumes Practice writing and building our own images in Docker
Week 17	ETL Pipelines	OLTP to OLAP	Review querying and modeling data in 3NF. We'll then move to modeling and querying data using the star schema.
Week 18		ETL with Cloud Services	Introduction to Amazon Redshift, and learn how to migrate data from a transactional RDS instance, to S3 to redshift. Learn how to migrate our data from OLTP to OLAP in SQL.
Week 19		Airflow and Python	Work with Airflow 2, working with DAGs, tasks and connections Review Python with scraping, and CS fundamentals like lists, dictionaries and histograms
Week 20	ELT Pipelines	Fivetran and Snowflake	Set up snowflake and learn about snowflake warehouses, computes, and worksheets Learn about analytics databases and column based storage
Week 21		DBT Workflow	Connect DBT to snowflake, and coerce our data using a pattern of sources, staging, integration, and marts
Week 22		Advanced DBT and BI Dashboards	Learn about tests, refs and Jinja in DBT. Integrate DBT with a BI dashboard to produce data visualizations
Week 23	Review ETL and Analytics Engineering	ETL in Python Review Project	Build an end to end ETL project in Python, and review OOP, data modeling, the adapter pattern, and tests
Week 24		Analytics Engineering Review Project	Review SQL material by writing advanced SQL queries, practice comprehending and optimizing SQL codebases, and use SQL to answer analytics questions.