UDACITY

NANODEGREE PROGRAM SYLLABUS

Data Engineering

Overview

Learn to design data models, build data warehouses and data lakes, automate data pipelines, and work with massive datasets. At the end of the program, you'll combine your new skills by completing a capstone project.

Students should have intermediate SQL and Python programming skills.

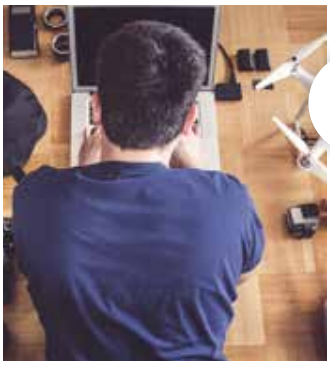
Educational Objectives: Students will learn to

- Create user-friendly relational and NoSQL data models
- Create scalable and efficient data warehouses
- Work efficiently with massive datasets
- Build and interact with a cloud-based data lake
- Automate and monitor data pipelines
- Develop proficiency in Spark, Airflow, and AWS tools

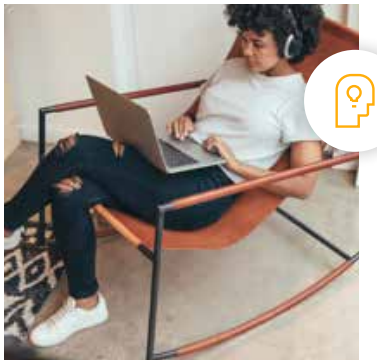
IN COLLABORATION WITH

INSIGHT

Estimated Time:
5 Months at
5 hrs/week



Prerequisites:
Intermediate
Python & SQL



Flexible Learning:
Self-paced, so
you can learn on
the schedule that
works best for you

Need Help?
udacity.com/advisor
Discuss this program
with an enrollment
advisor.

Course 1: Data Modeling

In this course, you'll learn to create relational and NoSQL data models to fit the diverse needs of data consumers. You'll understand the differences between different data models, and how to choose the appropriate data model for a given situation. You'll also build fluency in PostgreSQL and Apache Cassandra.

Course Project
Data Modeling with Postgres

In this project, you'll model user activity data for a music streaming app called Sparkify. You'll create a relational database and ETL pipeline designed to optimize queries for understanding what songs users are listening to. In PostgreSQL you will also define Fact and Dimension tables and insert data into your new tables.

Course Project
Data Modeling with Apache Cassandra

In these projects, you'll model user activity data for a music streaming app called Sparkify. You'll create a database and ETL pipeline, in both Postgres and Apache Cassandra, designed to optimize queries for understanding what songs users are listening to. For PostgreSQL, you will also define Fact and Dimension tables and insert data into your new tables. For Apache Cassandra, you will model your data so you can run specific queries provided by the analytics team at Sparkify.

	LEARNING OUTCOMES	
LESSON ONE	**Introduction to Data Modeling**	• Understand the purpose of data modeling • Identify the strengths and weaknesses of different types of databases and data storage techniques • Create a table in Postgres and Apache Cassandra
LESSON TWO	**Relational Data Models**	• Understand when to use a relational database • Understand the difference between OLAP and OLTP databases • Create normalized data tables • Implement denormalized schemas (e.g. STAR, Snowflake)

LESSON THREE	**NoSQL Data Models**	• Understand when to use NoSQL databases and how they differ from relational databases • Select the appropriate primary key and clustering columns for a given use case • Create a NoSQL database in Apache Cassandran

Course 2: Cloud Data Warehouses

In this course, you'll learn to create cloud-based data warehouses. You'll sharpen your data warehousing skills, deepen your understanding of data infrastructure, and be introduced to data engineering on the cloud using Amazon Web Services (AWS).

Course Project
Build a Cloud Data Warehouse

In this project, you are tasked with building an ELT pipeline that extracts their data from S3, stages them in Redshift, and transforms data into a set of dimensional tables for their analytics team to continue finding insights in what songs their users are listening to.

	LEARNING OUTCOMES	
LESSON ONE	**Introduction to the Data Warehouses**	• Understand Data Warehousing architecture • Run an ETL process to denormalize a database (3NF to Star) • Create an OLAP cube from facts and dimensions • Compare columnar vs. row oriented approaches
LESSON TWO	**Introduction to the Cloud with AWS**	• Understand cloud computing • Create an AWS account and understand their services • Set up Amazon S3, IAM, VPC, EC2, RDS PostgreSQ
LESSON THREE	**Implementing Data Warehouses on AWS**	• Identify components of the Redshift architecture • Run ETL process to extract data from S3 into Redshift • Set up AWS infrastructure using Infrastructure as Code (IaC) • Design an optimized table by selecting the appropriate distribution style and sorting key

UDACITY

Course 3: Spark and Data Lakes

In this course, you will learn more about the big data ecosystem and how to use Spark to work with massive datasets. You'll also learn about how to store big data in a data lake and query it with Spark.

Course Project
Build a Data Lake

In this project, you'll build an ETL pipeline for a data lake. The data resides in S3, in a directory of JSON logs on user activity on the app, as well as a directory with JSON metadata on the songs in the app. You will load data from S3, process the data into analytics tables using Spark, and load them back into S3. You'll deploy this Spark process on a cluster using AWS.

	LEARNING OUTCOMES	
LESSON ONE	**The Power of Spark**	• Understand the big data ecosystem • Understand when to use Spark and when not to use it
LESSON TWO	**Data Wrangling with Spark**	• Manipulate data with SparkSQL and Spark Dataframes • Use Spark for ETL purposes
LESSON THREE	**Debugging and Optimization**	• Troubleshoot common errors and optimize their code using the Spark WebUI
LESSON FOUR	**Introduction to Data Lakes**	• Understand the purpose and evolution of data lakes • Implement data lakes on Amazon S3, EMR, Athena, and Amazon Glue • Use Spark to run ELT processes and analytics on data of diverse sources, structures, and vintages • Understand the components and issues of data lakes

Course 4: Automate Data Pipelines

In this course, you'll learn to schedule, automate, and monitor data pipelines using Apache Airflow. You'll learn to run data quality checks, track data lineage, and work with data pipelines in production.

Course Project
Data Pipelines with Airflow

In this project, you'll continue your work on the music streaming company's data infrastructure by creating and automating a set of data pipelines. You'll configure and schedule data pipelines with Airflow and monitor and debug production pipelines.

		LEARNING OUTCOMES
LESSON ONE	**Data Pipelines**	• Create data pipelines with Apache Airflow • Set up task dependencies • Create data connections using hooks
LESSON TWO	**Data Quality**	• Track data lineage • Set up data pipeline schedules • Partition data to optimize pipelines • Write tests to ensure data quality • Backfill data
LESSON THREE	**Production Data Pipelines**	• Build reusable and maintainable pipelines • Build your own Apache Airflow plugins • Implement subDAGs • Set up task boundaries • Monitor data pipelines

Course 4: Capstone Project

Combine what you've learned throughout the program to build your own data engineering portfolio project.

Course Project
Data Engineering Capstone

The purpose of the data engineering capstone project is to give you a chance to combine what you've learned throughout the program. This project will be an important part of your portfolio that will help you achieve your data engineering-related career goals.

In this project, you'll define the scope of the project and the data you'll be working with. We'll provide guidelines, suggestions, tips, and resources to help you be successful, but your project will be unique to you. You'll gather data from several different data sources; transform, combine, and summarize it; and create a clean database for others to analyze.