



UDACITY
FOR ENTERPRISE

THE SCHOOL OF DATA SCIENCE

Data Engineering



NANODEGREE SYLLABUS

Overview

Data Engineering Nanodegree Program

BUILT IN COLLABORATION WITH



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.

Learners in this program will:

- 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.

Program Information



TIME

5 months
Study 10 hours/week



LEVEL

Practitioner



PREREQUISITES

Intermediate Python programming knowledge, including:

- Strings, numbers, and variables; statements, operators, and expressions;
- Lists, tuples, and dictionaries; Conditions, loops;
- Procedures, objects, modules, and libraries;
- Troubleshooting and debugging; Research & documentation;
- Problem solving; Algorithms and data structures.



HARDWARE/SOFTWARE REQUIRED

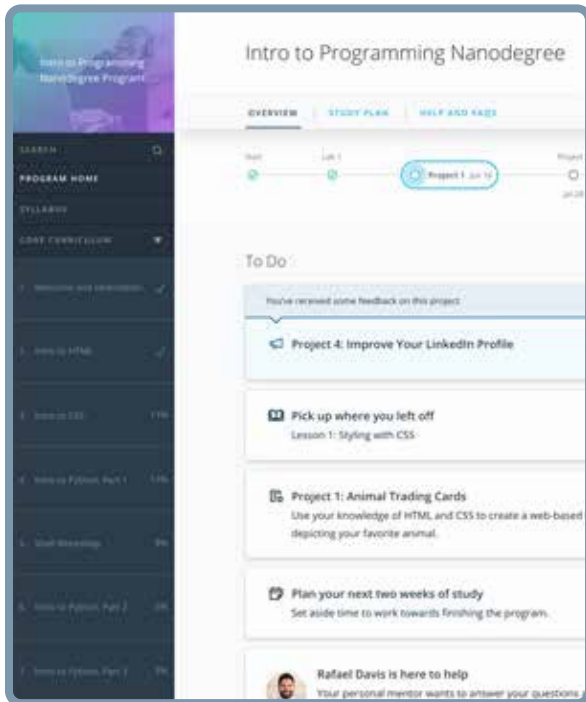
Access to the internet and a 64-bit computer.



LEARN MORE ABOUT THIS NANODEGREE

Contact us at enterpriseNDs@udacity.com.

Our Classroom Experience



REAL-WORLD PROJECTS

Learners build new skills through industry-relevant projects and receive personalized feedback from our network of 900+ project reviewers. Our simple user interface makes it easy to submit projects as often as needed and receive unlimited feedback.

KNOWLEDGE

Answers to most questions can be found with Knowledge, our proprietary wiki. Learners can search questions asked by others and discover in real-time how to solve challenges.

LEARNER HUB

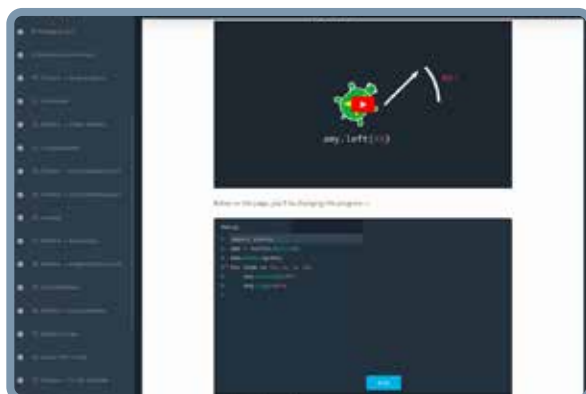
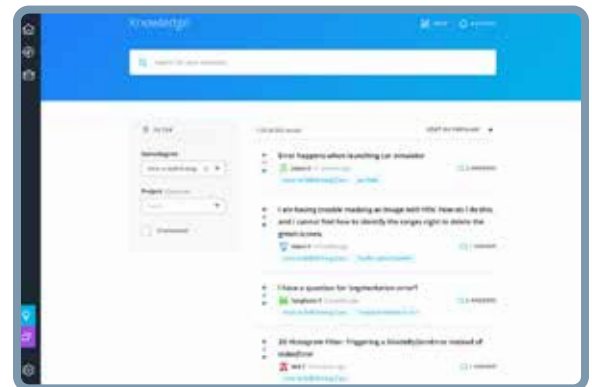
Learners leverage the power of community through a simple, yet powerful chat interface built within the classroom. Learner Hub connects learners with their technical mentor and fellow learners.

WORKSPACES

Learners can check the output and quality of their code by testing it on interactive workspaces that are integrated into the classroom.

QUIZZES

Understanding concepts learned during lessons is made simple with auto-graded quizzes. Learners can easily go back and brush up on concepts at anytime during the course.



CUSTOM STUDY PLANS

Mentors create a custom study plan tailored to learners' needs. This plan keeps track of progress toward learner goals.

PROGRESS TRACKER

Personalized milestone reminders help learners stay on track and focused as they work to complete their Nanodegree program.

Learn with the Best



Amanda Moran

DEVELOPER ADVOCATE AT DATASTAX

Amanda is a developer Advocate for DataStax after spending the last 6 years as a Software Engineer on 4 different distributed databases. Her passion is bridging the gap between customers and engineering. She has degrees from University of Washington and Santa Clara University.



Ben Goldberg

STAFF ENGINEER AT SPOTHERO

In his career as an engineer, Ben Goldberg has worked in fields ranging from Computer Vision to Natural Language Processing. At SpotHero, he founded and built out their Data Engineering team, using Airflow as one of the key technologies.



Sameh El-Ansary

CEO AT NOVELARI & ASSISTANT PROFESSOR AT NILE UNIVERSITY

Sameh is the CEO of Novelari, lecturer at Nile University, and the American University in Cairo (AUC) where he lectured on security, distributed systems, software engineering, blockchain and BigData Engineering.



Olli Iivonen

DATA ENGINEER AT WOLT

Olli works as a Data Engineer at Wolt. He has several years of experience on building and managing data pipelines on various data warehousing environments and has been a fan and active user of Apache Airflow since its first incarnations.



David Drummond

VP OF ENGINEERING AT INSIGHT

David is VP of Engineering at Insight where he enjoys breaking down difficult concepts and helping others learn data engineering. David has a PhD in Physics from UC Riverside.



Judit Lantos

DATA ENGINEER AT SPLIT

Judit was formerly an instructor at Insight Data Science helping software engineers and academic coders transition to DE roles. Currently, she is a Data Engineer at Split where she works on the statistical engine of their full-stack experimentation platform.



Juno Lee

CURRICULUM LEAD AT UDACITY

Juno is the curriculum lead for the School of Data Science. She has been sharing her passion for data and teaching, building several courses at Udacity. As a data scientist, she built recommendation engines, computer vision and NLP models, and tools to analyze user behavior.

Nanodegree Program Overview

Course 1: Data Modelling

In this module, 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.

Project

Data Modelling 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.

Project

Data Modelling with Apache Cassandra

In this project, 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.

LESSON TITLE

LEARNING OUTCOMES

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.

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).

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 Cassandra.



Course 2: Cloud Data Warehouses

In this module, 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).

Project

Build a Cloud Data Warehouse

In this project, you are tasked with building an ETL 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.

LESSON TITLE	LEARNING OUTCOMES
INTRODUCTION TO THE DATA WAREHOUSES	<ul style="list-style-type: none">• 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.
INTRODUCTION TO THE CLOUD WITH AWS	<ul style="list-style-type: none">• Understand cloud computing.• Create an AWS account and understand their services.• Set up Amazon S3, IAM, VPC, EC2, RDS PostgreSQL.
IMPLEMENTING DATA WAREHOUSES ON AWS	<ul style="list-style-type: none">• 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.

Nanodegree Program Overview

Course 3: Spark and Data Lakes

In this module, 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.

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.

LESSON TITLE	LEARNING OUTCOMES
THE POWER OF SPARK	<ul style="list-style-type: none">• Understand the big data ecosystem.• Understand when to use Spark and when not to use it.
DATA WRANGLING WITH SPARK	<ul style="list-style-type: none">• Manipulate data with SparkSQL and Spark Dataframes.• Use Spark for ETL purposes.
DEBUGGING AND OPTIMIZATION	<ul style="list-style-type: none">• Troubleshoot common errors and optimize their code using the Spark WebUI.
INTRODUCTION TO DATA LAKES	<ul style="list-style-type: none">• 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 module, 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.

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.

LESSON TITLE

LEARNING OUTCOMES

DATA PIPELINES

- Create data pipelines with Apache Airflow.
- Set up task dependencies.
- Create data connections using hooks.

DATA QUALITY

- Track data lineage.
- Set up data pipeline schedules.
- Partition data to optimize pipelines.
- Write tests to ensure data quality.
- Backfill data.

PRODUCTION DATA PIPELINES

- Build reusable and maintainable pipelines.
- Build your own Apache Airflow plugins.
- Implement subDAGs.
- Set up task boundaries.
- Monitor data pipelines.

Nanodegree Program Overview

Capstone Project

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

Project

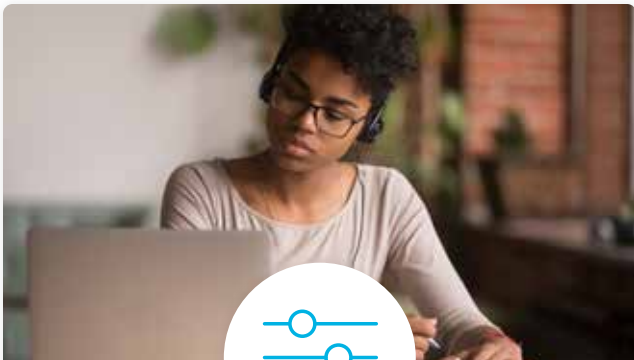
Data Engineering Capstone

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.



Our Nanodegree Programs Include:



Pre-Assessments

Our in-depth workforce assessments identify your team's current level of knowledge in key areas. Results are used to generate custom learning paths designed to equip your workforce with the most applicable skill sets.



Dashboard & Progress Reports

Our interactive dashboard (enterprise management console) allows administrators to manage employee onboarding, track course progress, perform bulk enrollments and more.



Industry Validation & Reviews

Learners' progress and subject knowledge is tested and validated by industry experts and leaders from our advisory board. These in-depth reviews ensure your teams have achieved competency.



Real World Hands-on Projects


Through a series of rigorous, real-world projects, your employees learn and apply new techniques, analyze results, and produce actionable insights. Project portfolios demonstrate learners' growing proficiency and subject mastery.

Our Review Process

Real-life Reviewers for Real-life Projects

Real-world projects are at the core of our Nanodegree programs because hands-on learning is the best way to master a new skill. Receiving relevant feedback from an industry expert is a critical part of that learning process, and infinitely more useful than that from peers or automated grading systems. Udacity has a network of over 900 experienced project reviewers who provide personalized and timely feedback to help all learners succeed.


All Learners Benefit From:




Line-by-line feedback for coding projects



Industry tips and best practices



Advice on additional resources to research



Unlimited submissions and feedback loops


How it Works

Real-world projects are integrated within the classroom experience, making for a seamless review process flow.


- Go through the lessons and work on the projects that follow
- Get help from your technical mentor, if needed
- Submit your project work
- Receive personalized feedback from the reviewer
- If the submission is not satisfactory, resubmit your project
- Continue submitting and receiving feedback from the reviewer until you successfully complete your project

About our Project Reviewers


Our expert project reviewers are evaluated against the highest standards and graded based on learners' progress. Here's how they measure up to ensure your success.




Expert Project Reviewers
Are hand-picked to provide detailed feedback on your project submissions.



Projects Reviewed
Our reviewers have extensive experience in guiding learners through their course projects.



Hours Average Turnaround
You can resubmit your project on the same day for additional feedback.



Average Reviewer Rating
Our learners love the quality of the feedback they receive from our experienced reviewers.



Vaibhav
UDACITY LEARNER

"I never felt overwhelmed while pursuing the Nanodegree program due to the valuable support of the reviewers, and now I am more confident in converting my ideas to reality."

now at
CODING VISIONS INFOTECH



UDACITY
FOR ENTERPRISE

Udacity © 2020

2440 W El Camino Real, #101
Mountain View, CA 94040, USA - HQ

For more information visit: www.udacity.com/enterprise