Project Portfolio

Real-world projects are integral to every Udacity Nanodegree program. They become the foundation for a job-ready portfolio to help learners advance their careers in their chosen field. The projects in the Data Engineer Nanodegree program were designed in collaboration with a group of highly talented industry professionals to ensure you develop the most in-demand skills. Every project in a Nanodegree program is human-graded by a member of Udacity's mentor and reviewer network. These project reviews include detailed, personalized feedback on how you can improve their work. Udacity graduates consistently rate projects and project reviews as one of the best parts of their experience with Udacity.

The Project Journey

The projects will take you on a journey where you'll assume the role of a Data Engineer at a fabricated data streaming company called "Sparkify" as it scales its data engineering in both size and sophistication. You'll work with simulated data of listening behavior, as well as a wealth of metadata related to songs and artists. You'll start working with a small amount of data, with low complexity, processed and stored on a single machine. By the end, you'll develop a sophisticated set of data pipelines to work with massive amounts of data processed and stored on the cloud. There are five projects in the program. Below is a description of each.

Project 1 - Data Modeling

In this project, you'll model user activity data for a music streaming app called Sparkify. The project is done in two parts. You'll create a database and import data stored in CSV and JSON files, and model the data. You'll do this first with a relational model in Postgres, then with a NoSQL data model with Apache Cassandra. You'll design the data models 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 to help the data team at Sparkify answer queries about app usage. You will set up your Apache Cassandra database tables in ways to optimize writes of transactional data on user sessions.

Project 2 - Cloud Data Warehousing

In this project, you'll move to the cloud as you work with larger amounts of data. You are tasked with building an ELT pipeline that extracts Sparkify's data from S3, Amazon's popular storage system. From there, you'll stage the data in Amazon Redshift and transform it into a set of fact and dimensional tables for the Sparkify analytics team to continue finding insights in what songs their users are listening to.

Project 3 - Data Lakes with Apache Spark

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.

Project 4 - Data Pipelines with Apache Airflow

In this project, you'll continue your work on Sparkify's data infrastructure by creating and automating a set of data pipelines. You'll use the up-and-coming tool Apache Airflow, developed and open-sourced by Airbnb and the Apache Foundation. You'll configure and schedule data pipelines with Airflow, setting dependencies, triggers, and quality checks as you would in a production setting.

Project 5 - Data Engineering Capstone

The capstone project is an opportunity for you to combine what you've learned throughout the program into a more self-driven project. 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.

We're excited to see what you build!