Data Engineer Nanodegree Syllabus

Build Production Ready Data Warehouses at Scale

UDACITY
FOR ENTERPRISE

Data Engineering

Table of Contents

Data Engineering

Before You Start

Prerequisites: 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
- Identify the appropriate use cases for different big data technologies
- 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

Contact Info

While going through the program, if you have questions about anything, you can reach us at enterprise-support@udacity.com. For help from Udacity mentors and peers, please visit the Udacity classroom.

Nanodegree Program Info

Technical Requirements
Hardware Requirements: webcam, microphone
Software and Software Version Requirements: No software required

LENGTH OF PROGRAM*: 6 months
FREQUENCY OF CLASSES: Self-paced
TEXTBOOKS REQUIRED: None

*This is a self-paced program and the length is an estimation of total hours the average student may take to complete all required coursework, including lecture and project time. Actual hours may vary.

Project 1: Data Modeling with Postgres and 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.

Supporting Lesson Content: Data Modeling

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.

Project 2: Data Infrastructure on the Cloud

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.

Supporting Lesson Content: Cloud Data Warehouses

Lesson Title	Learning Outcomes
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.
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 PostgreSQL.
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.

Project 3: Big Data with 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.

Supporting Lesson Content: Data Lakes with Spark

Lesson Title	Learning Outcomes
THE POWER OF SPARK	• Understand the big data ecosystem. • Understand when to use Spark and when not to use it.
DATA WRANGLING WITH SPARK	• Manipulate data with SparkSQL and Spark Dataframes. • Use Spark for ETL purposes.
DEBUGGING AND OPTIMIZATION	• Troubleshoot common errors and optimize their code using the Spark WebUI.
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.

Project 4: 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.

Supporting Lesson Content: Automate Data Pipelines

Lesson Title	Learning Outcomes
DATA PIPELINES	• Create data pipelines with Apache Airflow. • Set up task dependencies. • Create data connections using hooks.

Supporting Lesson Content: Data Lakes with Spark (Continued)

Lesson Title	Learning Outcomes
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.

Project 5: Data Engineering Nanodegree Capstone Project

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.