UDACITY

NANODEGREE PROGRAM SYLLABUS

Data Engineering with Microsoft Azure

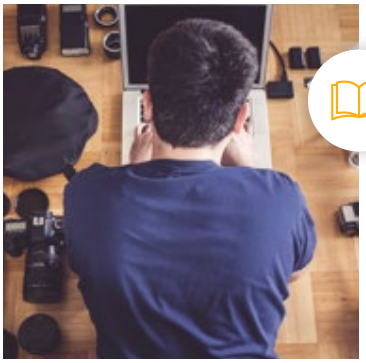
Overview

Learn to design data models, build data warehouses, build data lakes and lakehouse architecture, create data pipelines, and work with large datasets on the Azure platform using Azure Synapse Analytics, Azure Databricks, and Azure Data Factory.

Educational objectives:
A graduate of this program will be able to:

- Create relational and NoSQL data models
- Create data warehouses on the Azure cloud platform
- Work with large datasets using Spark and Azure Databricks
- Build and interact with Azure data lakes and lakehouse architecture
- Create data pipelines using Azure Data Factory and Synapse Analytics
- Develop proficiency in Spark, Azure Databricks, and Azure Databases

Estimated Time:
4 months at
5-10 hrs/week



Prerequisites:

- Intermediate SQL programming skills
- Intermediate Python programming skills
- Familiarity with the Azure cloud platform
- Experience with Github

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

Technical Mentor Support:
Our knowledgeable mentors guide your learning and are focused on answering your questions, motivating you and keeping you on track

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 1
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 2
Data Modeling 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.

		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 Cassandra

Course 2: Cloud Data Warehouses with Azure

In this course, you will learn how to create cloud-based data warehouses and sharpen your data warehousing skills, deepen your knowledge of data infrastructure, and be introduced to data engineering on the cloud using Azure. You will start with an introduction to data warehouses and ETL, followed by an introduction to ELT and data warehouse technology in the cloud. After this you will learn about cloud data warehouse technology in Azure, including Azure Synaps Analytics.

Course Project
Building an Azure Data Warehouse for Bikeshare Data Analytics

In this project, you'll create a data warehouse solution using Azure Synaps Analytics to better understand Divvy, a bike-sharing program. You'll start by importing data into Synapse Analytics, then transform the data into a star schema and view reports from Analytics to identify how much time and money is spent per ride.

		LEARNING OUTCOMES
LESSON ONE	**Introduction to Data Warehouses**	• Explain how OLAP may support certain business users better than OLTP • Implement ETL for OLAP Transformations with SQL • Describe Data Warehouse Architecture • Describe OLAP cube from facts and dimensions to slice, dice, roll-up, and drill down operations • Implement OLAP cubes from facts and dimensions to slice, dice, roll-up, and drill down • Compare columnar vs. row-oriented approaches • Implement columnar vs. row-oriented approaches
LESSON TWO	**ELT and Data Warehouse Technology in the Cloud**	• Explain the differences between ETL and ELT • Differentiate scenarios where ELT is preferred over ETL • Implement ETL for OLAP Transformations with SQL • Select appropriate cloud data storage solutions • Select appropriate cloud pipeline solutions • Select appropriate cloud data warehouse solutions

	LEARNING OUTCOMES	
LESSON THREE	**Azure Data Warehouse Technologies**	• Explain the benefits of Azure cloud computing services in data engineering • Describe Azure data engineering services • Set up key Azure features • Implementing Data Warehouse on Azure with Synapse Analytics
LESSON FOUR	**Implementing Data Warehouses in the Cloud**	• Identify components of Azure Data Warehouse Architecture • Set up Azure infrastructure using Infrastructure as Code (IaC) • Run ELT process to extract data from Azure data storage into Synapse Analytics

Course 3: Data Lakes and Lakehouse with Spark and Azure Databricks

In this course, you'll learn about the big data ecosystem and how to use Spark to work with massive datasets. You will also store big data in a data lake and develop lakehouse architecture on the Azure Databricks platform.

Course Project
Building an Azure Data Lake for Bikeshare Data Analytics

In this project, you'll build a data lake solution for Divvy bikeshare with Azure Databricks using a lakehouse architecture. You will design a star schema based on business outcomes and create a Bronze data store. Then you'll create a gold data store in Delta Lake tables and transform the data into the star schema for a Gold data store.

	LEARNING OUTCOMES	
LESSON ONE	**Big Data Ecosystem, Data Lakes, and Spark**	• Identify what constitutes the big data ecosystem for data engineering • Explain the purpose and evolution of data lakes in the big data ecosystem • Compare the Spark framework with Hadoop framework • Identify when to use Spark and when not to use it • Describe the features of lakehouse architecture
LESSON TWO	**Data Wrangling with Spark**	• Identify what constitutes the big data ecosystem for data engineering • Explain the purpose and evolution of data lakes in the big data ecosystem • Compare the Spark framework with Hadoop framework • Identify when to use Spark and when not to use it
LESSON THREE	**Spark Debugging and Optimization**	• Troubleshoot common errors and optimize their code using Spark WebUI • Identify common Spark bugs including errors in code syntax and issues with data • Diagnose errors in a distributed cluster to correct for them.

	LEARNING OUTCOMES	
LESSON FOUR	**Azure Databricks**	• Set up Spark Clusters in Azure Databricks • Produce Spark code in Databricks using Jupyter Notebooks and Python scripts • Implement distributed data storage using Azure Data Storage options
LESSON FIVE	**Data Lakes on Azure with Azure Databricks**	• Implement key features of data lakes on Azure • Use Spark and Databricks to run ELT processes and Storage options analytics on data of diverse sources, and vintages

UDACITY

Course 4: Data Pipelines with Azure

In this course, you'll learn to build, orchestrate, automate and monitor data pipelines in Azure using Azure Data Factory and pipelines in Azure Synapse Analytics. You'll build, trigger, and monitor data pipelines on the Azure platform for analytical workloads and run data transformations, optimize data flows, and work with data pipelines in production.

Course Project
Data Integration Pipelines for NYC Payroll Data Analytics

The City of New York would like to develop a data analytics platform using Azure Synapse Analytics to analyze how the city's financial resources are allocated and how much of the city's budget is being devoted to overtime.

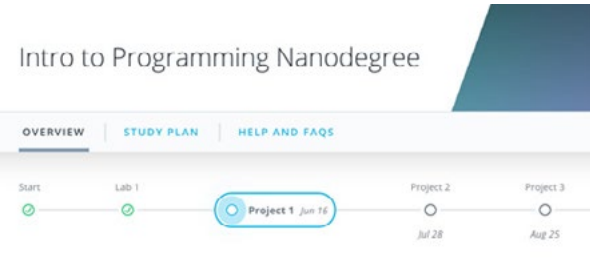
You have been hired as a data engineer to create high-quality data pipelines that are dynamic, can be automated, and can be monitored for efficient operation.

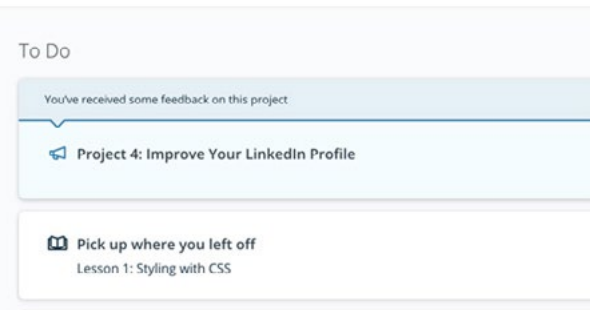
The source data resides in Azure Data Lake and you will build pipelines using Azure Data Factory for historical and new data to be processed in a NYC data warehouse in Azure Synapse Analytics.

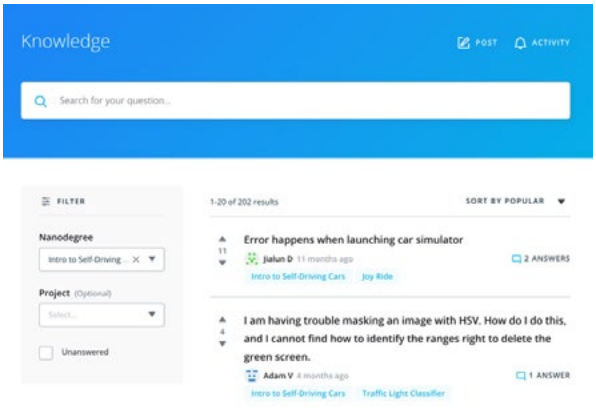
	LEARNING OUTCOMES	
LESSON ONE	**Azure Data Pipeline Components**	• Create and configure Azure data pipeline components • Create pipelines and associated components in Azure Data Factory or Azure Synapse • Configure linked service and dataset pipeline components • Choose integration runtimes for data pipelines
LESSON TWO	**Transforming Data in Azure Data Pipelines**	• Create and trigger Mapping data flows and Azure pipeline activities to transform and move data • Transform data in Azure Data factory and synapse pipelines with data flows • Debug, trigger, and monitor pipeline activities containing data flows • Develop pipelines in multiple ways in Azure Data Factory and Synapse Pipelines • Integrate Power Query in Azure Pipelines

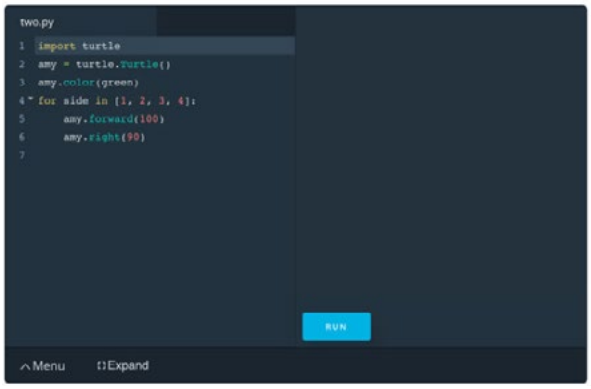
LESSON THREE	**Azure Pipeline Data Quality**	• Use common techniques optimize Azure data pipelines for data quality and flow • Manage data changing over time in pipeline data flows • Explain strategies and tools for data governance in Azure data pipelines
LESSON FOUR	**Azure Data Pipelines in Production**	• Implement production aspects of Azure data pipelines • Add Parameters to data pipelines in Azure Data Factory or Synapse Pipelines • Create pipeline objects programmatically • Automate data pipeline deployment with Azure DevOps or Github

Our Classroom Experience









REAL-WORLD PROJECTS

Build your skills through industry-relevant projects. Get personalized feedback from our network of 900+ project reviewers. Our simple interface makes it easy to submit your projects as often as you need and receive unlimited feedback on your work.

KNOWLEDGE

Find answers to your questions with Knowledge, our proprietary wiki. Search questions asked by other students, connect with technical mentors, and discover in real-time how to solve the challenges that you encounter.

WORKSPACES

See your code in action. Check the output and quality of your code by running them on workspaces that are a part of our classroom.

QUIZZES

Check your understanding of concepts learned in the program by answering simple and auto-graded quizzes. Easily go back to the lessons to brush up on concepts anytime you get an answer wrong.

CUSTOM STUDY PLANS

Create a custom study plan to suit your personal needs and use this plan to keep track of your progress toward your goal.

PROGRESS TRACKER

Stay on track to complete your Nanodegree program with useful milestone reminders.

Learn with the Best

Matt Swaffer, PhD

DATA SCIENCE PRACTICE LEAD AT COGNITELL

Matt is a data science professional whose career has spanned software development, user experience design, and data visualization. He earned his PhD in the research area of cognitive psychology in human learning and is an adjunct professor teaching software design courses.

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 the University of Washington and Santa Clara University.

Vishnu (Lucky) Pamula

SR. CLOUD SOLUTION ARCHITECT AT MICROSOFT

Lucky is a data & AI evangelist with a track record of successfully helping organizations build analytics solutions. Besides his day job, he teaches as an adjunct professor, delivers lunch & learns, mentors students, and evangelizes Azure Quantum as an ambassador.

All Our Nanodegree Programs Include:

EXPERIENCED PROJECT REVIEWERS

REVIEWER SERVICES

- Personalized feedback & line by line code reviews
- 1,600+ Reviewers with a 4.85/5 average rating
- 3 hour average project review turnaround time
- Unlimited submissions and feedback loops
- Practical tips and industry best practices
- Additional suggested resources to improve

TECHNICAL MENTOR SUPPORT

MENTORSHIP SERVICES

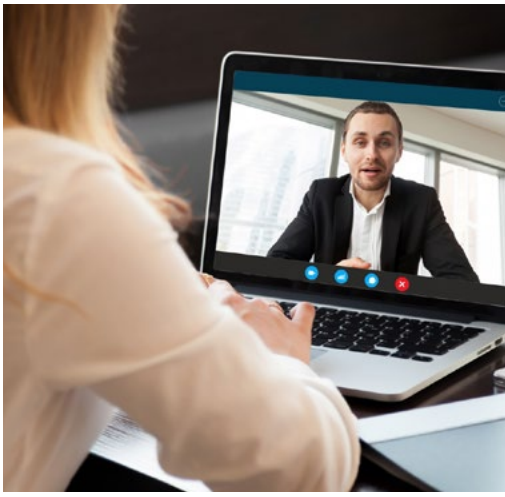
- Questions answered quickly by our team of technical mentors
- 1,000+ mentors with a 4.7/5 average rating
- Support for all your technical questions

PERSONAL CAREER SERVICES

CAREER SUPPORT

- Github portfolio review
- LinkedIn profile optimization

Frequently Asked Questions



PROGRAM OVERVIEW

WHY SHOULD I ENROLL?

This program is designed to help you take advantage of the growing need for skilled Microsoft Azure data engineers. Prepare to meet the demand for qualified Microsoft Azure data engineers who can respond to real-life, high-stakes workplace challenges.

WHAT JOBS WILL THIS PROGRAM PREPARE ME FOR?

The skills you will gain from this Nanodegree program will qualify you for jobs in several industries, as countless companies are trying to design data models and perform other tasks using Microsoft Azure.

HOW DO I KNOW IF THIS PROGRAM IS RIGHT FOR ME?

The program is for individuals who are looking to advance their Microsoft Azure data engineering careers with skills in a burgeoning field.

ENROLLMENT AND ADMISSION

DO I NEED TO APPLY? WHAT ARE THE ADMISSION CRITERIA?

No. This Nanodegree program accepts all applicants regardless of experience and specific background.

WHAT ARE THE PREREQUISITES FOR ENROLLMENT?

A well-prepared student currently has:

- Intermediate SQL programming skills
- Intermediate Python programming skills
- Familiarity with the Azure cloud platform
- Experience with Github

IF I DO NOT MEET THE REQUIREMENTS TO ENROLL, WHAT SHOULD I DO?

We have a number of Nanodegree programs and free courses that can help you prepare, including:

- **Learn SQL Nanodegree program**
- **Programming for Data Science with Python Nanodegree program**

TUITION AND TERM OF PROGRAM

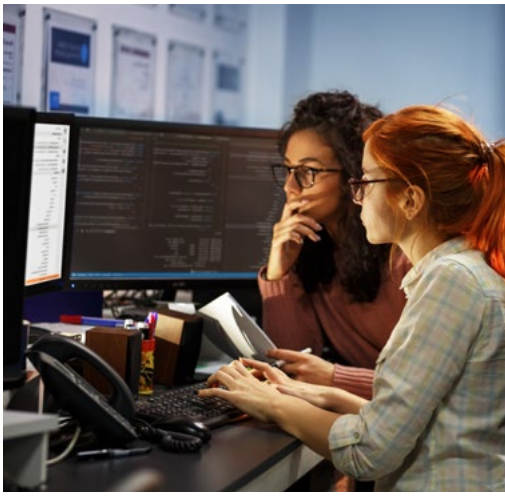
HOW IS THIS NANODEGREE PROGRAM STRUCTURED?

The Data Engineering with Microsoft Azure Nanodegree program is comprised of content and curriculum to support five (5) projects. We estimate that students can complete the program in four (4) months, working 5-10 hours per week.

Each project will be reviewed by the Udacity reviewer network. Feedback will

FAQs Continued

be provided and if you do not pass the project, you will be asked to resubmit the project until it passes.



HOW LONG IS THIS NANODEGREE PROGRAM?

Access to this Nanodegree program runs for the length of time specified above. If you do not graduate within that time period, you will continue learning with month-to-month payments. See the **Terms of Use** and **FAQs** for other policies regarding the terms of access to our Nanodegree programs.

CAN I SWITCH MY START DATE? CAN I GET A REFUND?

Please see the Udacity Nanodegree program **FAQs** for policies on enrollment in our programs.

SOFTWARE AND HARDWARE

WHAT SOFTWARE AND VERSIONS WILL I NEED IN THIS PROGRAM?

There are no software and version requirements to complete this Nanodegree program. All coursework and projects can be completed via Student Workspaces and the Azure portal in the Udacity online classroom. Udacity's basic tech requirements can be found at https://www.udacity.com/tech/requirements.