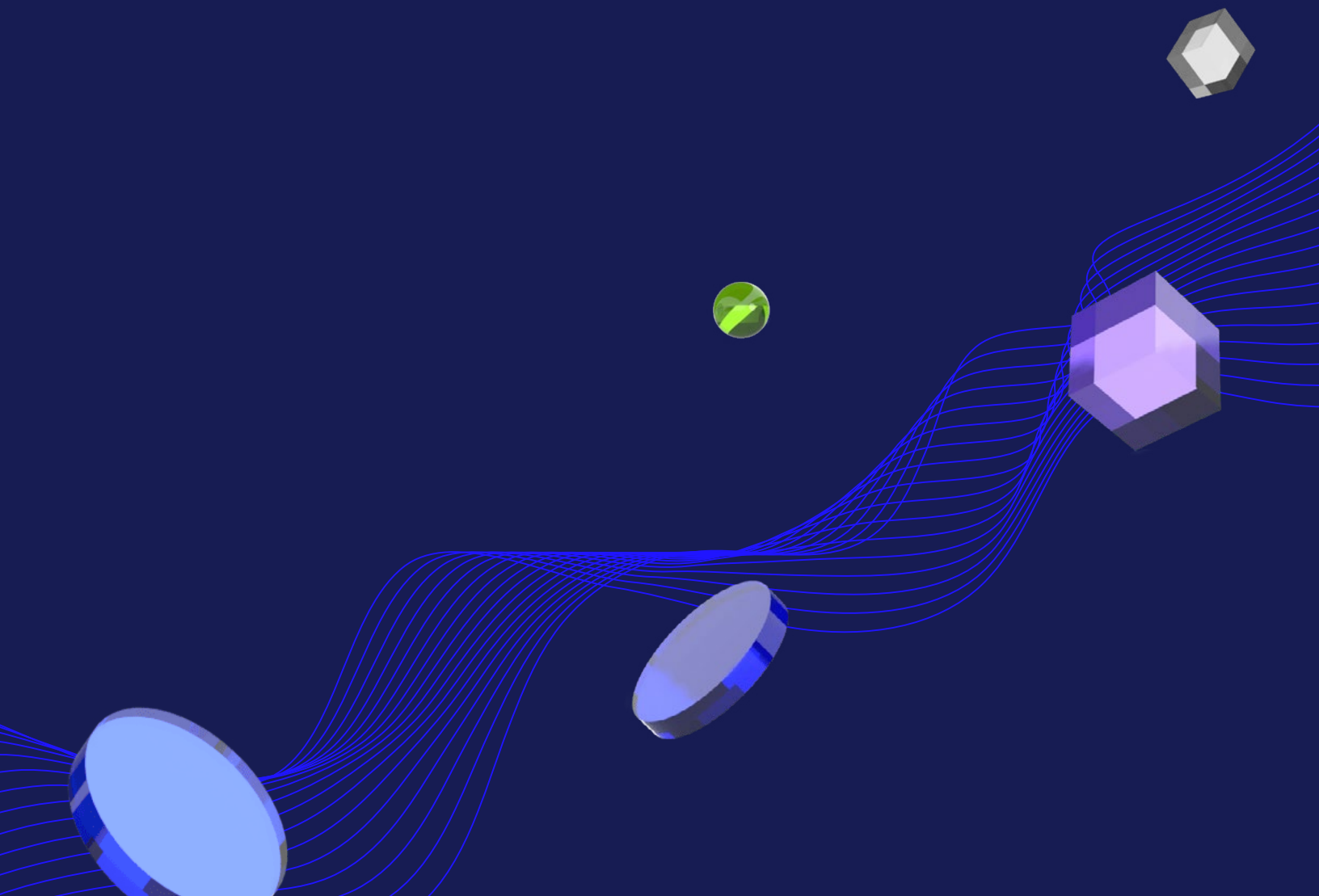
ENTERPRISE

UDACITY

SCHOOL OF DATA SCIENCE

Data Engineering with Microsoft Azure

Nanodegree Program Syllabus



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.

Learning 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.

Program information

Estimated Time

4 months at 5-10hrs/week*

Skill Level

Intermediate

Prerequisites

A well-prepared learner should have:

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

Required Hardware/Software

None

*The length of this program is an estimation of total hours the average student may take to complete all required coursework, including lecture and project time. If you spend about 5-10 hours per week working through the program, you should finish within the time provided. Actual hours may vary.

Course 1

Data Modeling

Learn to create relational and NoSQL data models to fit the diverse needs of data consumers. Understand the differences between different data models and how to choose the appropriate data model for a given situation. Additionally, build fluency in PostgreSQL and Apache Cassandra.

Course Project

Data Modeling with Postgres

In this project, model user activity data for a music streaming app called Sparkify. Create a relational database and ETL pipeline designed to optimize queries for understanding what songs users are listening to. In PostgreSQL, define fact and dimension tables and insert data into the new tables.

Course Project

Data Modeling with Apache Cassandra

In this project, model user activity data for the same music streaming service—this time using Apache Cassandra. Create a database and ETL pipeline designed to optimize queries for understanding what songs users are listening to. Model the data to run specific queries provided by the analytics team at Sparkify.

Lesson 1

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 2

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 3

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

Learn how to create cloud-based data warehouses and sharpen data warehousing skills, deepen knowledge of data infrastructure, and be introduces to data engineering on the cloud using Azure. Start with an introduction to data warehouses and ETL, followed by an introduction to ELT and data warehouse technology in the cloud. Lastly, learn about cloud data warehouse technology in Azure, including Azure Synaps Analytics.

Course Project

Building an Azure Data Warehouse for Bikeshare Data Analytics

Create a data warehouse solution using Azure Synaps Analytics to better understand Divvy, a bike-sharing program. 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.

Lesson 1

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, rollup, and drill down.
- Compare columnar vs. row-oriented approaches.
- Implement columnar vs. row-oriented approaches.

Lesson 2

ELT & 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.

Lesson 3

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 4

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 & Lakehouse with Spark & Azure Databricks

Learn about the big data ecosystem and how to use Spark to work with massive datasets. Additionally, 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

Build a data lake solution for Divvy bikeshare with Azure Databricks using a lakehouse architecture. Design a star schema based on business outcomes and create a Bronze data store. Then create a Gold data store in Delta Lake tables and transform the data into the star schema for a Gold data store.

Lesson 1

Big Data Ecosystem, Data Lakes & 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 2

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 3

Spark Debugging & 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.

Lesson 4

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 5

Data Lakes on Azure with Asure Databricks

- Implement key features of data lakes on Azure.
- Use Spark and Databricks to run ELT processes and analytics on data of diverse sources, structures, and vintages.

Course 4

Data Pipelines with Azure

Learn to build, orchestrate, automate, and monitor data pipelines in Azure using Azure Data Factory and pipelines in Azure Synapse Analytics. 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.

Learners will act as data engineers by creating 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 learners will build pipelines using Azure Data Factory for historical and new data to be processed in a NYC data warehouse in Azure Synapse Analytics.

Lesson 1

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 2

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 3

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 4

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.

Meet your instructors.

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

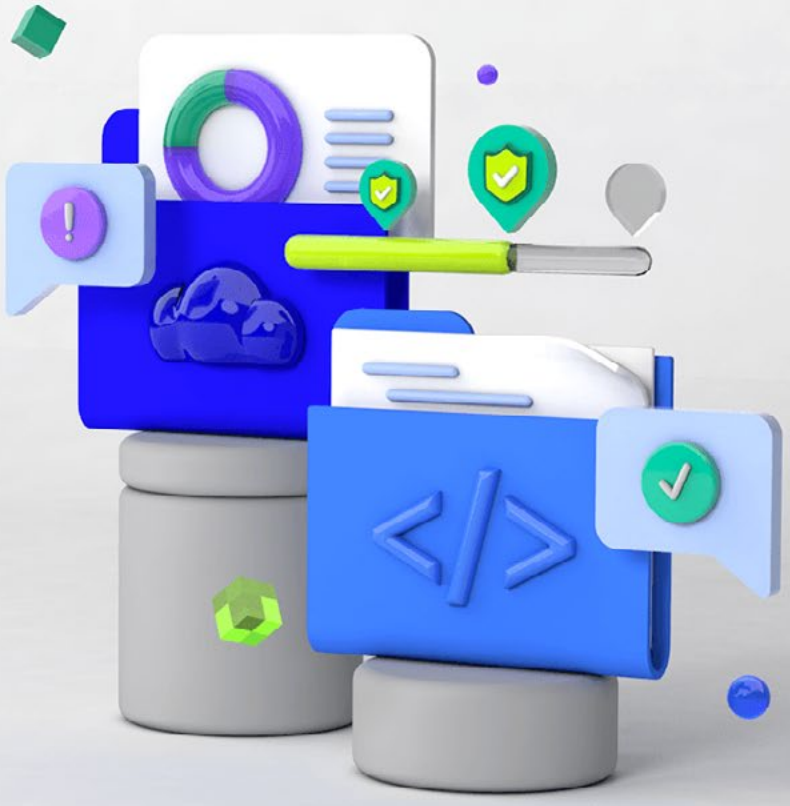
Amanda is a developer advocate for DataStax after spending the last 6 years as a software engineer on 4 different distributed databases. Her passion is bridging the gap between customers and engineering. She has degrees from the University of Washington and Santa Clara University.

Vishnu (Lucky) Pamula

Sr. Cloud Solution Architect at Microsoft

Lucky is a data and 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.

Udacity's learning experience



Hands-on Projects

Open-ended, experiential projects are designed to reflect actual workplace challenges. They aren't just multiple choice questions or step-by-step guides, but instead require critical thinking.

Quizzes

Auto-graded quizzes strengthen comprehension. Learners can return to lessons at any time during the course to refresh concepts.

Knowledge

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

Custom Study Plans

Create a personalized study plan that fits your individual needs. Utilize this plan to keep track of movement toward your overall goal.

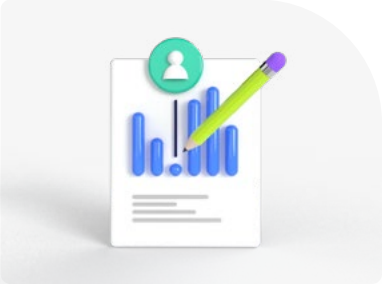
Workspaces

See your code in action. Check the output and quality of your code by running it on interactive workspaces that are integrated into the platform.

Progress Tracker

Take advantage of milestone reminders to stay on schedule and complete your program.

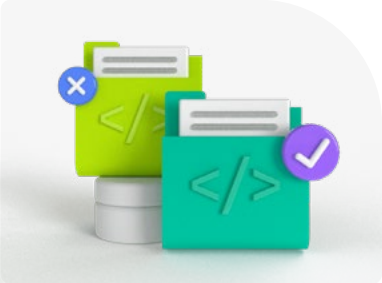
Our proven approach for building job-ready digital skills.



Pre-Assessments

Identify skills gaps.

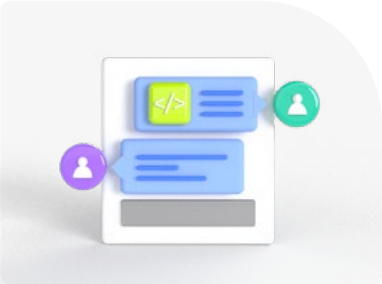
- In-depth assessments benchmark your team's current level of knowledge in key areas.
- Results are used to generate custom learning paths.



Experienced Project Reviewers

Verify skills mastery.

- Personalized project feedback and critique includes line-by-line code review from skilled practitioners with an average turnaround time of 1.1 hours.
- Project review cycle creates a feedback loop with multiple opportunities for improvement—until the concept is mastered.
- Project reviewers leverage industry best practices and provide pro tips.



Technical Mentor Support

24/7 support unblocks learning.

- Learning accelerates as skilled mentors identify areas of achievement and potential for growth.
- Unlimited access to mentors means help arrives when it's needed most.
- 2 hr or less average question response time assures that skills development stays on track.

Mentor Network

Highly vetted for effectiveness.

- Mentors must complete a 5-step hiring process to join Udacity's selective network.
- After passing an objective and situational assessment, mentors must demonstrate communication and behavioral fit for a mentorship role.
- Mentors work across more than 30 different industries and often complete a Nanodegree program themselves.

Dashboard & Reporting

Track course progress.

- Udacity's enterprise management console simplifies management of bulk enrollments and employee onboarding.
- Interactive views help achieve targeted results to increase retention and productivity.
- Maximize ROI while optimizing job readiness.

UDACITY

Learn more at

udacity.com/enterprise →

12.16.22 | V1.0