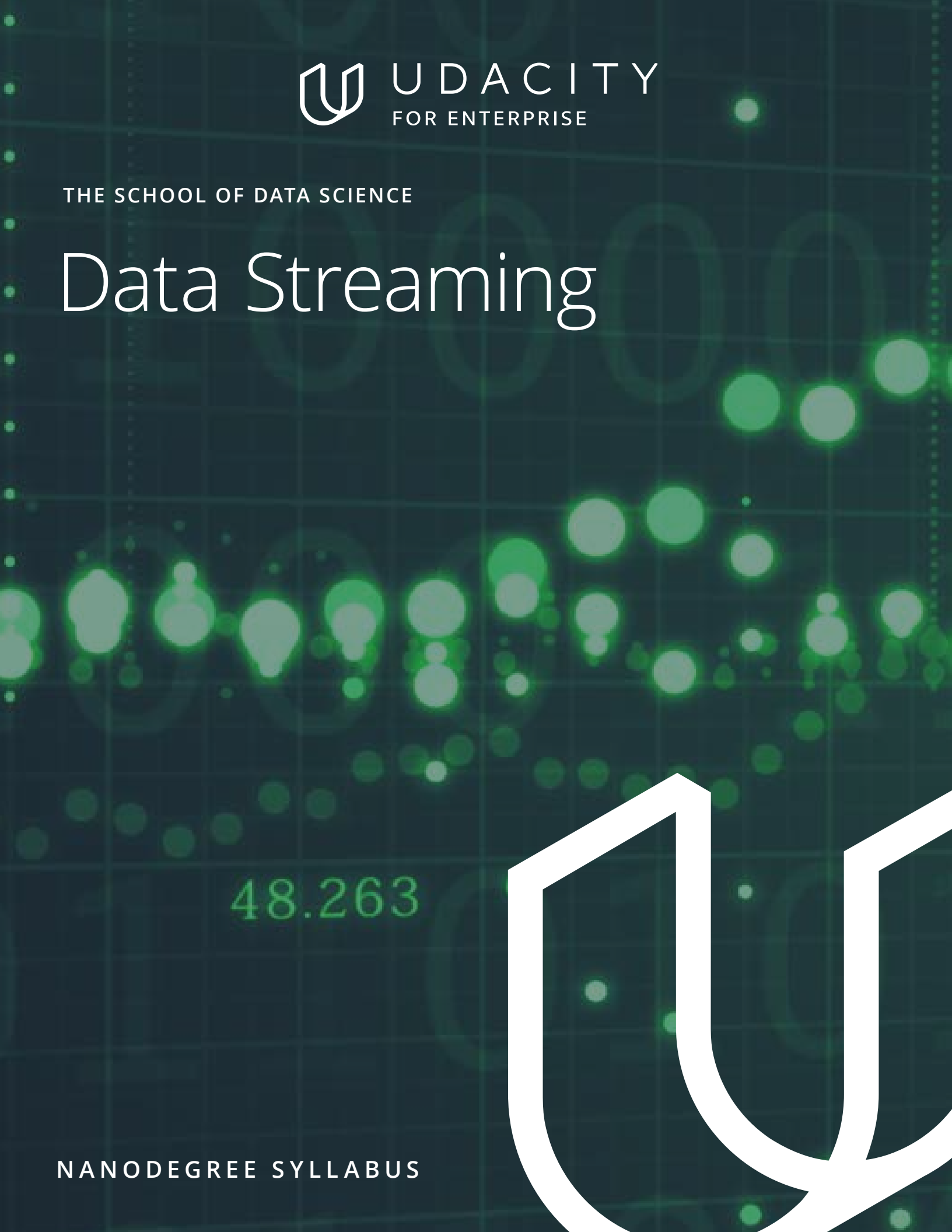
UDACITY
FOR ENTERPRISE

THE SCHOOL OF DATA SCIENCE

Data Streaming



NANODEGREE SYLLABUS

Overview

Data Streaming Nanodegree Degree Program

The Data Streaming Nanodegree program provides learners with the latest skills to process data in real-time by building fluency in modern data engineering tools, such as Apache Spark, Kafka, Spark Streaming, and Kafka Streaming. A graduate of this program will be able to:

- Understand the components of data streaming systems. Ingest data in real-time using Apache Kafka and Spark and run analyses.
- Use the Faust Stream Processing Python library to build a real-time stream-based application. Compile real-time data and run live analytics, as well as draw insights from reports generated by the streaming console.
- Learn about the Kafka ecosystem, and the types of problems each solution is designed to solve. Use the Confluent Kafka Python library for simple topic management, production, and consumption.
- Explain the components of Spark Streaming (architecture and API), integrate Apache Spark Structured Streaming and Apache Kafka, manipulate data using Spark, and understand the statistical report generated by the Structured Streaming console.

Program Information

TIME
2 months
Study 10 hours/week

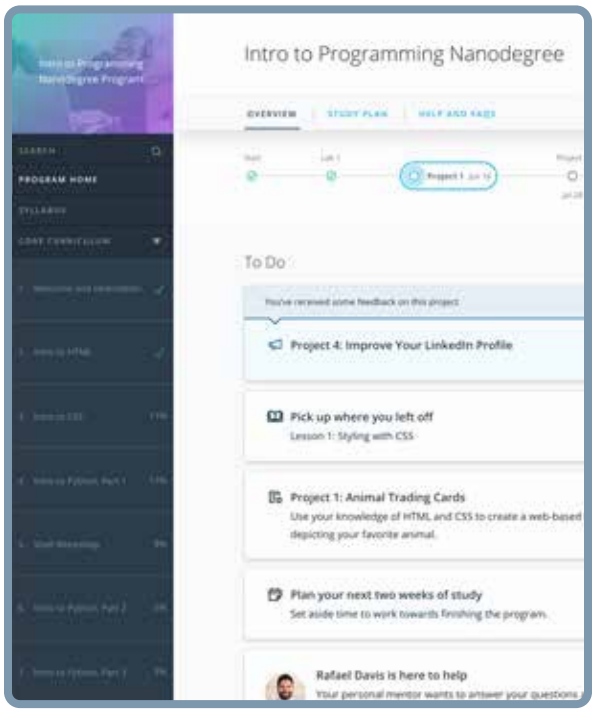
LEVEL
Specialist

PREREQUISITES
- Intermediate SQL, and Python. And experience with ETL.
- Basic familiarity with traditional batch processing and traditional service architectures is desired, but not required.

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

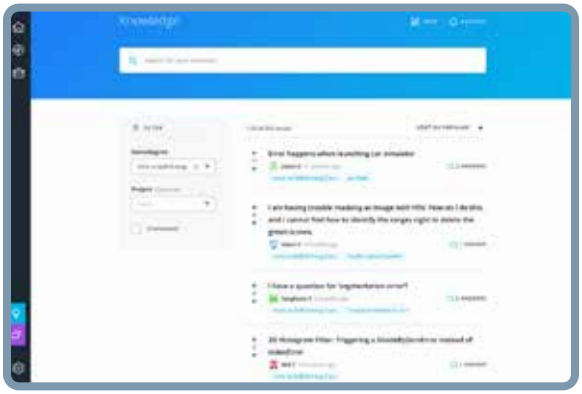
Learners leverage the power of community through a simple, yet powerful chat interface built within the classroom. Learner Hub connects learners with their technical mentor and fellow learners.

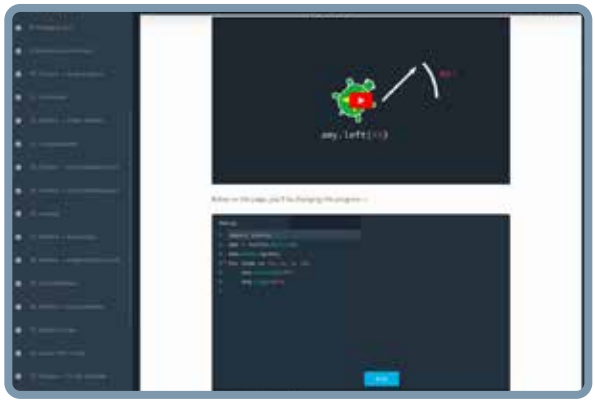
WORKSPACES

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

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.

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 holds a PhD in Physics from UC Riverside.

Judit Lantos

SENIOR DATA ENGINEER AT NETFLIX

Currently, Judit is a Senior Data Engineer at Netflix. Formerly a Data Engineer at Split, where she worked on the statistical engine of their full-stack experimentation platform, she has also been an instructor at Insight Data Science, helping software engineers and academic coders transition to DE roles.

Sean Murdock

FACULTY, BYU - IDAHO

Sean has worked as an Architect or Software Engineer for Columbia Ultimate, Firstsource Global, Intermountain Healthcare, General Motors, The Church of Jesus Christ, Northrup Grumman, Zions Bank, and Ancestry. He currently teaches DevOps and Cybersecurity at Brigham Young University - Idaho.

Course 1: Foundations of Data Streaming, and SQL & Data Modeling for the Web

The goal of this course is to demonstrate knowledge of the tools taught throughout, including Kafka Consumers, Producers, & Topics; Kafka Connect Sources and Sinks, Kafka REST Proxy for producing data over REST, Data Schemas with JSON and Apache Avro/Schema Registry, Stream Processing with the Faust Python Library, and Stream Processing with KSQL.

Project

Optimize Chicago Bus and Train Availability Using Kafka

For your first project, you'll be streaming public transit status using Kafka and the Kafka ecosystem to build a stream processing application that shows the status of trains in real-time. Based on the skills you learn, you will be able to optimize the availability of buses and trains in Chicago based on streaming data. You will learn how to have your own Python code produce events, use REST Proxy to send events over HTTP, and use Kafka Connect to collect data from a Postgres database to produce streaming data from a number of sources into Kafka. Then, you will use KSQL to combine related data models into a single topic ready for consumption by the downstream Python applications, and complete a simple Python application that ingests data from the Kafka topics for analysis. Finally, you will use the Faust Python Stream Processing library to further transform train station data into a more streamlined representation: using stateful processing, this library will show whether passenger volume is increasing, decreasing, or staying steady.

LESSON TITLE	LESSON OUTCOMES
INTRODUCTION TO STREAM PROCESSING	• Describe and explain streaming data stores and stream processing. • Describe and explain real-world usages of stream processing. • Describe and explain append-only logs, events, and how stream processing differs from batch processing. • Utilize Kafka CLI tools and the Confluent Kafka Python library for topic management, production, and consumption.

Course 1: Foundations of Data Streaming, and SQL & Data Modeling for the Web, cont.

LESSON TITLE	LESSON OUTCOMES
APACHE KAFKA	• Understand Kafka architecture, topics, and configuration. • Utilize Confluent Kafka Python to create topics and configuration. • Understand Kafka producers, consumers, and configuration. • Utilize Confluent Kafka Python to create producers and configuration. • Utilize Confluent Kafka Python to create topics, configuration, and manage offsets. • Describe and explain user privacy considerations. • Describe and explain performance monitoring for consumers, producers, and the cluster itself.
DATA SCHEMAS AND APACHE AVRO	• Understand what a data schema is and the value it provides. • Understand what Apache Avro is and what value it provides. • Utilize AvroProducer and AvroConsumer in Confluent Kafka Python. • Describe and explain schema evolution and data compatibility types. • Utilize Schema Registry components in Confluent Kafka Python to manage compatibility.
KAFKA CONNECT AND REST PROXY	• Describe and explain what problem Kafka Connect solves for and where it would be more appropriate than a traditional consumer. • Describe and explain common connectors and how they work. • Utilize Kafka Connect FIleStream & JDBC Source and Sink. • Describe and explain what problem Kafka REST Proxy solves for and where it would be more appropriate than alternatives. • Describe, explain, and utilize the REST Proxy metadata and administrative APIs. • Describe and explain the REST Proxy consumer APIs. • Utilize the REST Proxy consumer, subscription, and offset APIs. • Describe, explain, and utilize the REST Proxy producer APIs.

Nanodegree Program Overview

LESSON TITLE	LESSON OUTCOMES
STREAM PROCESSING FUNDAMENTALS	• Describe and explain common scenarios for stream processing, and where you would use stream versus batch. • Describe and explain common stream processing strategies. • Describe and explain how time and windowing works in stream processing. • Describe and explain what a stream versus a table is in stream processing, and where you would use on over the other. • Describe and explain how data storage works in stream processing applications and why it is needed.
STREAM PROCESSING WITH FAUST	• Describe and explain the Faust Stream Processing Python library, and how it fits into the ecosystem relative to solutions like Kafka Streams. • Describe and explain Faust stream-based processing. • Utilize Faust to create a stream-based application. • Describe and explain how Faust table-based processing works. • Utilize Faust to create a table-based application. • Describe and explain Faust processors and function usage. • Utilize Faust processor and function. • Describe and explain Faust serialization and deserialization. • Utilize Faust serialization and deserialization.
KSQL	• Describe and explain how KSQL fits into the Kafka ecosystem, and why you would choose it over a stream processing application built from scratch. • Describe and explain KSQL architecture. • Describe and explain how to create KSQL streams and tables from topics. • Understand the importance of KEY and schema transformations. • Utilize KSQL to create tables and streams. • Describe and explain KSQL selection syntax. • Utilize KSQL syntax to query tables and streams. • Describe and explain KSQL windowing. • Utilize KSQL windowing within the context of table analysis. • Describe and explain KSQL grouping and aggregates. • Utilize KSQL grouping and aggregates within queries.

Course 2: Streaming API Development and Documentation

The goal of this course is to grow your expertise in the components of streaming data systems, and build a real time analytics application. Specifically, you will be able to identify components of Spark Streaming (architecture and API), build a continuous application with Structured Streaming, consume and process data from Apache Kafka with Spark Structured Streaming (including setting up and running a Spark Cluster), create a DataFrame as an aggregation of source DataFrames, sink a composite DataFrame to Kafka, and visually inspect a data sink for accuracy.

Project: Evaluate Human Balance with Spark Streaming

In this project, you will work with a real-life application called the Step Trending Electronic Data Interface (STEDI). It is a working application used to assess fall risk for seniors. When a senior takes a test, they are scored using an index which reflects the likelihood of falling, and potentially sustaining an injury in the course of walking. STEDI uses a Redis datastore for risk score and other data. The Data Science team has completed a working graph for population risk at a STEDI clinic. The problem is the data is not populated yet. You will work with Kafka Connect Redis Source events and Business Events to create a Kafka topic containing anonymized risk scores of seniors in the clinic.

LESSON TITLE	LESSON OUTCOMES
STREAMING DATAFRAMES	• Start a Spark Cluster and Deploy a Spark Application • Create a Spark Streaming DataFrame with a Kafka Source • Create a Spark View • Query a Spark View
JOINS AND JSON	• Parse a JSON Payload Into Separate Fields for Analysis • Join Two Streaming DataFrames from Different Data Sources • Write a Streaming DataFrame to Kafka with Aggregated Data
REDIS, BASE64 AND JSON	• Manually Save to Redis and Read the Same Data from a Kafka Topic • Parse Base64 Encoded Information • Sink a Subset of JSON Fields

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.

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

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

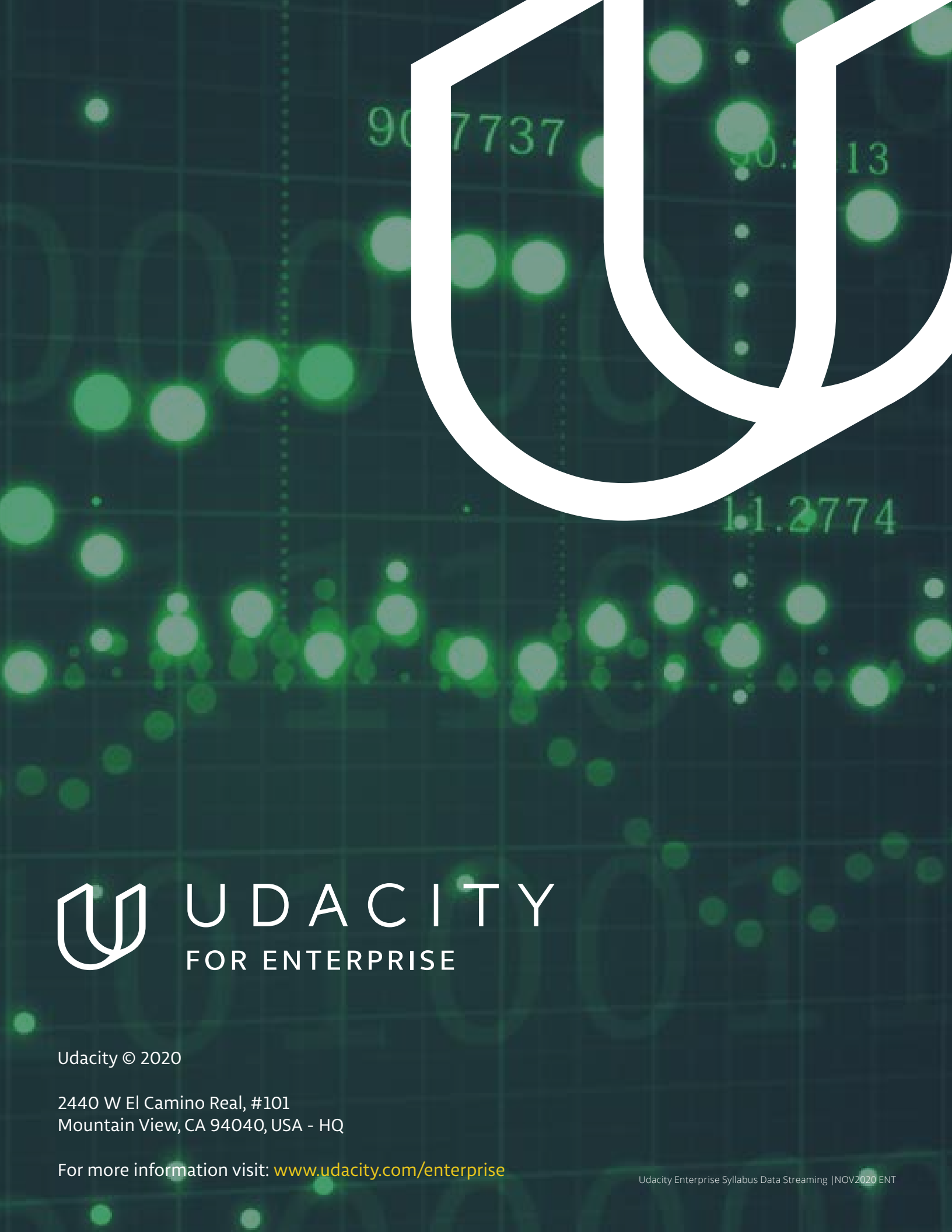
Our expert project reviewers are evaluated against the highest standards and graded based on learners' progress. **Here's how they measure up to ensure your success.**

900+
Expert Project Reviewers
Are hand-picked to provide detailed feedback on your project submissions.

1.8M
Projects Reviewed
Our reviewers have extensive experience in guiding learners through their course projects.

3
Hours Average Turnaround
You can resubmit your project on the same day for additional feedback.

4.85 /5
Average Reviewer Rating
Our learners love the quality of the feedback they receive from our experienced reviewers.



UDACITY
FOR ENTERPRISE

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

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