Snowflake Data Science

Course outline

Overview of Data Science with Snowflake

• Data science applications

• Common machine learning vocabulary

• Machine learning workflow and pipeline

Snowflake Architecture

• Snowflake Cloud Data Platform overview

• Three-tiered architecture

• Snowflake UI and core capabilities, including elasticity, workload separation, data security, and simplicity

of performance

• Built-in functions for traversing, flattening, and nesting of semi-structured data

• Seamless connectivity using Snowflake connectors for languages such as Python, Spark, and R

• Notebook-based data science development environments

Acquire Data

• Raw and external data sets in object stores

• External tables and direct queries in data lakes

• Native data formats of such as CSV, JSON, and Parquet

• Ingesting into native semi-structured data types without pre-processing

• Snowflake data ingestion best practices

• Serverless continuous ingestion service Snowpipe

• Diverse data including customer demographic data, time-series data, geospatial data

• Private and Public Data Exchange

• Data Marketplace with ready-to-use and third-party datasets for data augmentation

Prepare Data

• Sampling data

• Cloning data and utilizing Time Travel

• Data cleansing techniques to address duplicates, missing values, and outliers

• Bulk ingestion and scheduling data loads with tasks

• Table stream for capturing change data

Exploratory Data Analysis

• Exploration and visualization using Snowsight

• Descriptive exploratory data analysis using statistical, analytic, and approximation functions

• Visual exploratory data analysis using popular and relevant libraries

Feature Engineering

• Employ common feature selection and feature engineering techniques

• Advanced SQL functions for data transformation at scale

• Transform data and perform feature engineering with Snowpark

Training Models (Machine Learning)

• Supervised learning: linear regression with popular ML libraries

• Supervised learning: classification using techniques such as logistic regression, random forests, gradient

boosts, and more

• Identifying, using, and interpreting metrics to evaluate models and performance

• Unsupervised learning

• Developing models using a variety of popular machine learning libraries including Scikit-Learn and more

• Communicating machine learning results

• Integrating with partner platforms for data science automation and democratization around AutoML

Deploying Models

• Storing machine learning results in Snowflake

• Deploying machine learning models using scalable frameworks

• Creating external functions to support prediction and data augmentation through APIs

• Utilizing partner platforms for deployment and practices with ML Ops

Beyond Deployment: ML Ops

• Operationalizing models with Snowflake's extensive partner ecosystem using automation (AutoML) and

ML Ops practices

• Using Snowflake capabilities including Snowpipe, table stream, and tasks for continuous data pipelines to

update machine learning models