CTR-601 Databricks Data and ML Engineering

About This Course

In this course, you'll learn how to use Delta Live Tables with Spark SQL and Python to define and schedule pipelines that incrementally process new data from a variety of data sources into the Lakehouse. You'll also learn how to orchestrate tasks with Databricks Workflows and promote code with Databricks Repos. This course will equip you with the skills you need to navigate the intricacies of building, deploying, and operating machine learning models using Databricks.

Target Population

Developers and Devops engineers who wish to quickly get familiar and practice the Databricks platform

Pre-requisites

Basic data and ML pipelines

Course Outline

Module 1: Get Started with Data Engineering on Databricks

- Core components that make up the Databricks Lakehouse platform
- Navigating the Databricks Workspace UI
- Developing and running code in multi-cell Databricks notebooks
- Integrating git support using Databricks Repos

Module 2: Transform Data with Spark

- Extracting data from variety of file formats and data sources
- Applying a number of common transformations to clean data
- Reshaping and manipulating complex data using advanced built-in functions
- Leveraging UDFs for reusable code and applying best practices for performance in Spark

Module 3: Managing data with Delta Lake

- Delta lake as the foundations of the data lakehouse architecture
- Using SQL to perform complete and incremental updates to existing table

Module 4: Build Data Pipelines with Delta Live Tables

- Configure and run data pipelines using the Delta Live Tables
- Use the Auto Loader and Delta Live Tables to define pipelines that ingest and process data through multiple tables
- Review event logs and data artifacts created by pipelines

Module 5: How to use Databricks Workflows to orchestrate tasks

- How to orchestrate tasks with Databricks Workflow jobs
- How to configure and scheduled dashboards and alerts to reflect updates to production data pipelines

Module 6: Key aspects of Unity Catalog

- Unity Catalog key concepts, and how the catalog integrates with the Databricks platform
- How to access the Unity Catalog through clusters and SQL warehouses
- How to create and govern data assets in Unity Catalog
- How to adopt Databricks recommendations into your organization's Unity Catalog based solutions

Module 7: Introduction to Feature Store

- Creating and interacting with feature stores
- Using the Feature Engineering Client in Databricks

Module 8: Model Development Workflow

- The benefits of MLflow tracking in enhancing the model development process
- Understand the crucial relationship between experiments and runs with MLflow
- Explore the automated logging capabilities of Databricks Autologging, streamlining your workflow

Module 9: AutoML

- The benefits of using AutoML for generating machine learning models
- The glass-box nature of Databricks AutoML and its advantages
- Initiate an AutoML experiment using both the user interface and the programming API
- Open and edit notebooks generated by AutoML for enhanced customization
- Identify the best model generated by AutoML based on specific metrics
- Modify the best model generated by AutoML to suit your requirements

Module 10: Model Deployment Fundamentals

- Understanding of what model deployment entails
- Understand different deployment types — batch, pipeline, and real-time
- Compare batch, pipeline, and real-time deployments specifically within the Databricks environment
- Key deployment features offered by MLflow to streamline the deployment process

Module 11: Batch Deployment

- Describe batch deployment
- Analyze the pros and cons of deploying a model via batch processing
- Load a logged Model Registry model
- Conduct batch inference using Feature Store's score_batch functionality

Module 12: Pipeline Deployment

- The intricacies of pipeline deployment
- How to develop a simple Delta Live Tables pipeline for performing streaming-based inference

Module 13: Real-Time Deployment and Online Stores

- Define real-time deployment
- The challenges associated with real-time serving
- Databricks Model Serving
- A/B testing in the context of model deployment