databricks

Welcome to Data Engineering with Databricks

Course Objectives

1. Use the Databricks Data Science and Engineering Workspace to perform common code development tasks in a data engineering workflow.
2. Use Spark to extract data from a variety of sources, apply common cleaning transformations, and manipulate complex data with advanced functions.
3. Define and schedule data pipelines that incrementally ingest and process data through multiple tables in the lakehouse using Delta Live Tables.
4. Orchestrate data pipelines with Databricks Workflow Jobs and schedule dashboard updates to keep analytics up-to-date.
5. Configure permissions in Unity Catalog to ensure that users have proper access to databases for analytics and dashboarding.

Course Overview

Module 0: Get Started with PySpark Programming (OPTIONAL)

Module 1: Get Started with Databricks Data Science and Engineering Workspace

Module 2: Transform Data with Spark (SQL or PySpark)

Module 3: Manage Data with Delta Lake

Module 4: Build Data Pipelines with Delta Live Tables (SQL or PySpark)

Module 5: Deploy Workloads with Databricks Workflows

Module 6: Manage Data Access for Analytics with Unity Catalog

Module Agendas

Module Agenda

Get Started with PySpark Programming

Spark SQL Overview

DE 0.1 - Spark SQL

DE 0.2L - Spark SQL Lab

DE 0.3 - DataFrame & Column

DE 0.4L - Purchase Revenues Lab

DE 0.5 - Aggregation

DE 0.6L - Revenue by Traffic Lab

Module Agenda

Get Started with Databricks Data Science and Engineering Workspace

Introduction to the Databricks Lakehouse Platform

Databricks Architecture and Services

Demo – Navigating the Workspace

DE 1.1 – Create and Manage Clusters Interactively

DE 1.2 – Notebook Basics

Git Versioning with Databricks Repos

Demo – Using Databricks Repos

DE 1.3L – Getting Started with the Databricks Lakehouse Platform Lab

Module Agenda

Transform Data with Spark SQL

DE 2.1 – Querying Files Directly

DE 2.2 – Options for External Sources

DE 2.3L – Extract Data Lab

DE 2.4 – Cleaning Data

DE 2.5 – Complex Transformations

DE 2.6 – UDFs and Control Flow

DE 2.7L – Reshape Data Lab

Transform Data with PySpark

DE 3.1 – Querying Files Directly

DE 3.2 – Reader & Writer

DE 3.3L – Extract Data Lab

DE 3.4 – Cleaning Data

DE 3.5 – Complex Transformations

DE 3.6 – UDFs

DE 3.7L – Reshape Data Lab

Module Agenda

Manage Data with Delta Lake

What is Delta Lake

DE 4.1 – Schemas and Tables

DE 4.2 – Version and Optimize Delta Tables

DE 4.3L – Manipulate Delta Tables Lab

DE 4.4 – Set Up Delta Tables

DE 4.5 – Load Data into Delta Lake

DE 4.6 – Load Data Lab

Module Agenda

Build Data Pipelines with Delta Live Tables

Introduction to Delta Live Tables

DE 5.1 – DLT UI Walkthrough

DE 5.1A – SQL Pipelines

DE 5.1B – Python Pipelines

DE 5.2 – Python vs SQL

DE 5.3 – Pipeline Results

DE 5.4 – Pipeline Event Logs

Module Agenda

Deploy Workloads with Databricks Workflows

Introduction to Workflows

Building and Monitoring Workflow Jobs

DE 6.1 - Scheduling Tasks with the Jobs UI

DE 6.2L - Jobs Lab

DE 6.3 - Navigating Databricks SQL and Attaching to Endpoints

DE 6.4 - Last Mile ETL with DBSQL

Module Agenda

Manage Data Access for Analytics with Unity Catalog

Introduction to Unity Catalog

DE 7.1 – Managing principals in Unity Catalog

DE 7.2 – Managing Unity Catalog metastores

DE 7.3 – Creating compute resources for Unity Catalog access

DE 7.4 – Creating and governing data objects with Unity Catalog

DE 7.5 – Create and Share Tables in Unity Catalog

DE 7.6 – Create external tables in Unity Catalog

DE 7.7 – Upgrade a table to Unity Catalog

DE 7.8 – Create views and limit table access