Data Engineering With Databricks

Bo Zhang
Technical Instructor
Databricks

Tuesday 10 May – Friday 13 May

6:00 AM-11:00 AM IST

8:30 AM-1:30 PM SGT

10:30 AM-3:30 PM AEST

12:30 PM-5:30 PM NZST

Description

Data professionals from all industries will benefit from this comprehensive introduction to the components of the Databricks Lakehouse Platform that directly support putting ETL pipelines into production. You will leverage SQL and Python to define and schedule pipelines that incrementally process new data from a variety of data sources to power analytic applications and dashboards in the lakehouse. This course offers hands-on instructions in Databricks Data Science & Engineering Workspace, Databricks SQL, Delta Live Tables, Databricks Repos, Databricks Task Orchestration, and the Unity Catalog.

Duration

- 4 half days (5 hours per day) from 10 to 13 May

Learning outcomes

- Leverage the Databricks Lakehouse Platform to perform core responsibilities for data pipeline development
- Use SQL and Python to write production data pipelines to extract, transform, and load data into tables and views in the lakehouse
- Simplify data ingestion and incremental change propagation using Databricks-native features and syntax, including Delta Live Tables
- Orchestrate production pipelines to deliver fresh results for ad hoc analytics and dashboarding
- Get certified as a Databricks Certified Data Engineer Associate upon completion of this 4-day course, normally worth US$1.5K

Prerequisites

- Experience using SQL to query data from enterprise data stores
- Familiarity with basic cloud concepts (virtual machines, object storage, identity management)
- Basic familiarity with Python variables, functions and control flow (preferred)

Topic outline

- Delta Lake
- Relational entities on Databricks
- ETL with Spark SQL
- Just enough Python for Spark SQL
- Incremental data processing with Structured Streaming and Auto Loader
- Medallion architecture in the data lakehouse
- Delta Live Tables
- Task orchestration with Databricks Jobs
- Databricks SQL
- Managing permissions in the lakehouse
- Productionizing dashboards and queries on Databricks SQL

Links to Github repo and course materials

- https://github.com/databricks-academy/data-engineering-with-databricks
- Ebooks for the course
    - https://github.com/databricks-academy/data-engineering-with-databricks/releases/download/v2.2.1/Data-Engineering-with-Databricks.pdf
    -

Day 1 Agenda

- Introduction to the Databricks Lakehouse Platform
- Introduction to the Databricks Workspace and Services
    - Using clusters, files, notebooks, and repos
    - Lab 1.3L - Get Started Databricks
- Introduction to Delta Lake
    - Manipulating and optimizing data in Delta tables
    - Lab 2.2L - Manipulating tables
    - Lab 2.4L - Delta Lake Versioning
- Quiz and pickup 3 winners of the day

Day 2 Agenda

- Relational entities on Databricks
    - Databases, Tables
    - Views, Common Table Expressions
    - Lab 3.3L - Databases, Tables, Views
- ETL with Spark SQL
    - Query files from files
    - Provide Options for external data sources
    - Creating Delta tables, writing to tables
    - Lab 4.5L - Extract and Load data
    - Data Cleaning
    - Advanced Data Transformation
    - SQL UDF & Control Flow
    - Lab 4.9L - Reshaping Data

Day 3 Agenda

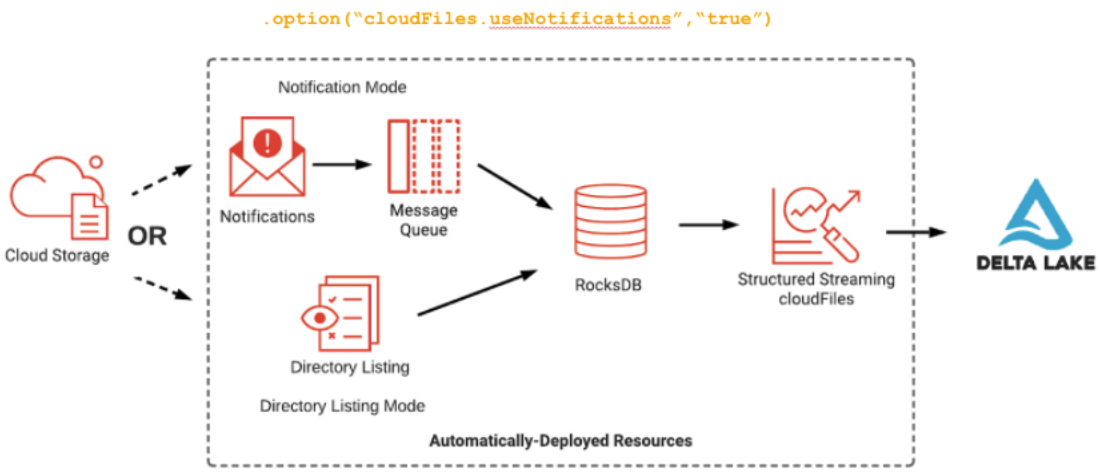
- Incremental Data Ingestion and Processing with Structured Streaming and Auto Loader
    - Lab 6.3L - Using Auto Loader and Structured Streaming with Spark SQL
- Multi-Hop architecture in Lakehouse

    - Lab 7.2L - Propagating Incremental Updates with Structured Streaming and Delta Lake
- Using Delta Live Tables to build reliable data pipeline
    - Lab 8.2L Migrating SQL Notebooks to Delta Live Tables

Day 4 Agenda

- Orchestration with Databricks Workflows
    - Lab - Schedule notebooks and DLT pipelines with dependencies
- Run Your First Databricks SQL Query
- Managing Permissions in the Lakehouse
    - Lab - Configure permissions for databases, tables and views in data lakehouse
- Productionalizing Dashboards and Queries in DBSQL
    - Final Lab - Schedule queries, dashboards and alerts for end-to-end analytic pipelines

Auto Loader Under the Hood



https://community.cloud.databricks.com/login.html

https://docs.databricks.com/data-engineering/delta-live-tables/index.html

https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html

https://docs.databricks.com/_static/notebooks/stream-stream-joins-python.html

work

Workspace Registration

Workspace Registration Link:

https://labs.databricks.com/#/odl/d8805902-1a86-4f4f-9a64-289fafc89501

Workspace Registration Walkthrough Guide

Workspace Registration Steps:

Step 1:

- Complete this short registration form for the hands-on portion of the workshop. Registration link: https://labs.databricks.com/#/odl/d8805902-1a86-4f4f-9a64-289fafc89501

Step 2:

- Click "Launch Lab" and allow 5-10 minutes for the environment to load

Step 3:

- After the environment has been prepared, copy-paste the provided workspace link into an incognito window

  Tip: One of the common issues you may face will be if you are using your company account to log into Azure Directory, which will not work for this workshop. You need to make sure to use the username/password provided by Cloudlabs to access the workspace. If you are having trouble doing that, use an "Incognito Window" to log into the workspace.

Step 4:

- Click the "Sign in with Azure AD" icon
- Select " Use another account"
- Copy-paste the provided username and password from the environment details page

Step 5:

- Arrive in your Databricks Workspace. Find the notebook under the Workspace tab and within your User folder

Notebook

How to import code of the course into your Databricks Workspace

https://www.databricks.training/step-by-step/importing-courseware-from-github/

NEXT: Register for Databricks Certification Preparation – Associate DE May 24/25