Delta Lake

BOSS'20 workshop (https://boss-workshop.github.io/boss-2020)

Abstract

A common data engineering pipeline architecture uses tables that correspond to different quality levels, progressively adding structure to the data: data ingestion (“Bronze” tables), transformation/feature engineering (“Silver” tables), and machine learning training or prediction (“Gold” tables). Combined, we refer to these tables as a “multi-hop” architecture. It allows data engineers to build a pipeline that begins with raw data as a “single source of truth” from which everything flows. In this session, we will show how to build a scalable data engineering data pipeline using Delta Lake.

Delta Lake is an open-source storage layer that brings reliability to data lakes. Delta Lake offers ACID transactions, scalable metadata handling, and unifies streaming and batch data processing. It runs on top of your existing data lake and is fully compatible with Apache Spark APIs

In this 1.5 hour session you will learn about:

- The data engineering pipeline architecture
- Data engineering pipeline scenarios
- Data engineering pipeline best practices
- How Delta Lake enhances data engineering pipelines
- The ease of adopting Delta Lake for building your data engineering pipelines

Course assets:

- lectures will be delivered as slide decks
- labs will be run as demos by the instructor using lab docs
- students will be given PDFs of the labs at the end of the class:
    - ☐ Lab 1: How to get up and running on Community Edition
    - ☐ Lab 2: Raw, Bronze, Silver
    - ☐ Lab 3: Silver, Gold

Outline

10 minutes	Setup and introductions // Demo 1: How to get up and running on Community Edition
10 minutes	Lecture 1: The Big Picture, The Lambda Architecture versus the Delta Architecture
30 minutes	Lab 2 demo: Raw, Bronze, Silver
20 minutes	Lab 3 demo: Silver, Gold
10 minutes	Lecture 2: review the Delta architecture, benefits and features
10 minutes	final thoughts, Q&A

Technology

Databricks Community Edition

Google Slides

Delta Lake

Presenters

Kate Sullivan, Databricks

Emma Freeman, Databricks