CLOUDERA

DENG-251: Building an Open Data Lakehouse Using Apache Iceberg

Course Overview

Course Type

Instructor-led training

Level

Intermediate

Duration

3 days

Platform

CDP Private Cloud Base 7.1.9

Spark 3.3.x and Iceberg 1.3.0

Topics Covered

- Iceberg Concepts
- Logical Architecture
- Metadata Layer
- Copy-on-Write (COW) and Merge-on-Read (MOR)
- Rollback and Time Travel
- Change Data Capture (CDC)
- Schema Evolution
- Hidden Partitions
- Branches and Tags
- Write-Audit-Publish (WAP)
- Table Maintenance Tasks
- Hive-to-Iceberg
- Table Migration

About This Training

This introduces Apache Iceberg, a high-performance open table format for organizing petabyte-scale analytic datasets on a file system or object store, available on Cloudera Data Warehouse and Cloudera Data Engineering on both Private and Public Cloud. Combined with Cloudera Data Platform, Iceberg can enable users to build an open data lakehouse architecture for multi-function analytics and to deploy large-scale end-to-end pipelines. This course covers various aspects of Apache Iceberg, such as benefits, architecture, internal operation, read and write operations, and advanced functions, all while drawing comparisons to Hive and building on the students' existing knowledge and experience.

What Skills You Will Gain

Participants will:

- Gain a deep understanding of Iceberg's benefits, snapshots, and their functionalities.
- Confidently build external and managed tables, configuring copy-on-write and merge-on-read for optimized data management.
- Perform rollbacks and time travel, navigate schema and partition evolution, and utilize hidden partitions.
- Create and merge table branches, mastering Iceberg's write-audit-publish procedure.
- Efficiently perform table maintenance tasks and tackle data migration challenges.

Who Should Take This Course?

This course is for new and existing customers using Cloudera Data Warehouse or Cloudera Data Engineering on Private or Public Cloud who are interested in benefiting from using Apache Iceberg. The course is designed for Data Engineers, Hive SQL Developers, Kafka Streaming Engineers, Data Scientists, and CDP Admins. A general knowledge of HDFS and experience with Hive and Spark are required.

Other Training That Might Interest You

- Developing Applications with Apache Spark
- Analyzing with Cloudera Data Warehouse

DENG-251: Building an Open Data Lakehouse Using Apache Iceberg

Training Outline (Page 2 of 2)

Introduction
- Apache Hive
- Why Iceberg?
- Data Lakehouses
- What is Iceberg?

Catalogs
- Review Iceberg Catalog Configuration

Iceberg Concepts
- Snapshots
- Metadata Layer: Manifest List, Manifest Files
- Time Travel
- Schema Evolution
- Hidden Partition
- Write-Audit-Publish (WAP)
- Branches, Tags, Zero-Copy-Clone

Iceberg Table Design
- Managed & External Tables
- Table Properties Review
- Copy-On-Write (COW) vs Merge-On-Read (MOR)
- Hidden Partitions
- Compare Hive vs Iceberg Partition Design
- Table Metadata
- Table Maintenance

Data-As-Code
- Iceberg Personas
- Write-Audit-Publish (WAP)
- Branches & Tagging

Hive-to-Iceberg Table Migration
- In-place Migration
- Shallow Migration

VERSION: 1.0.3