DENG-255

Building an Open Data Lakehouse Using Apache Iceberg

Course Overview

Course Type
Instructor-led training

Level
Intermediate

Duration
4 days

Platform
Cloudera on premises

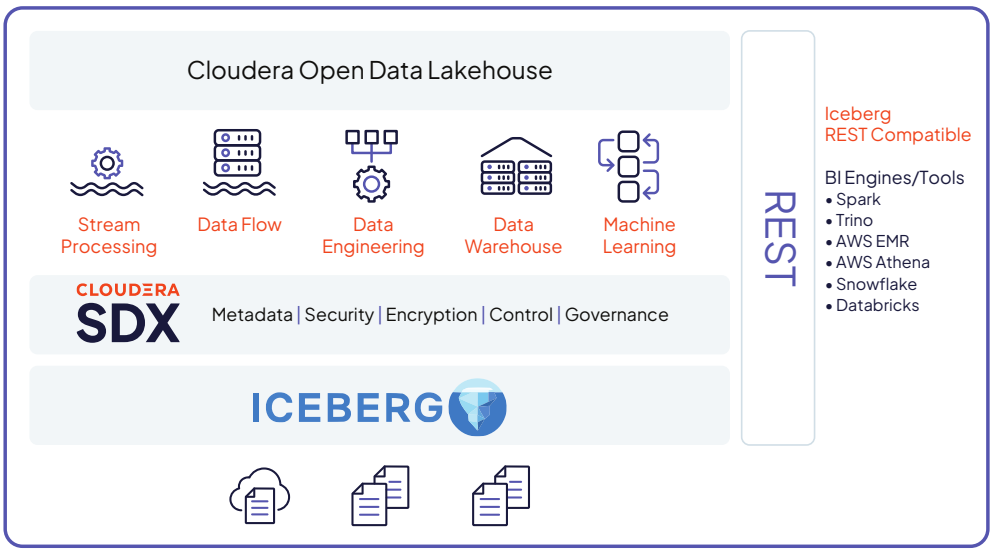
Topics Covered

- Intro to Data Lakehouse
- Apache Ozone Concepts
- Working with Ozone
- Ozone App Integration
- Apache Iceberg Concepts
- Iceberg Table Design
- 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)
- Replication
- Hive-to-Iceberg
- Table Migration
- Table Maintenance Tasks
- GDPR and Hard Deletes
- Late-arriving Data
- Access Control

About This Training

The Open Data Lakehouse is a modern data architecture that enables versatile analytics on streaming and stored data within cloud-native object stores. This architecture can span hybrid and multi-cloud environments.

This course introduces Apache Ozone, a hybrid storage service addressing the limitations of HDFS. You'll also explore Apache Iceberg, an open-table format optimized for petabyte-scale datasets. The course covers Iceberg's benefits, architecture, read/write operations, streaming, and advanced features like time travel, partition evolution, and Data-as-Code. Over 25 hands-on labs and a capstone project will equip you with the skills to build an efficient, performant Open Data Lakehouse within your own environment.



Who Should Take This Course?

This course is designed for data professionals within organizations using Cloudera Data Warehouse or Cloudera Data Engineering solutions. If you're building an Open Data Lakehouse powered by Apache Iceberg, this course will provide the knowledge and skills you need. Ideal roles include Data Engineers, Hive/Impala SQL Developers, Kafka Streaming Engineers, Data Scientists, and Cloudera Admins. A basic understanding of HDFS and experience with Hive and Spark are prerequisites.

DENG-255

Building an Open Data Lakehouse Using Apache Iceberg

Skills You Will Gain

Open Data Lakehouse Fundamentals

- Understand core Open Data Lakehouse concepts and benefits.
- Introduction to Apache Ozone and its integration within the Cloudera Ecosystem.

Apache Ozone Mastery

- Configure Ozone, use CLI commands, and transfer data between HDFS and Ozone.
- Integrate Ozone into applications.

Apache Iceberg Expertise

- Explore Iceberg's integration with Cloudera, architecture, and data lakehouse design principles.
- Master data management, governance, and optimization best practices.
- Understand snapshots and time travel queries.
- Design tables strategically (external/managed, copy-on-write, merge-on-read).
- Employ advanced features: change data capture (CDC), schema/partition evolution, hidden partitions.

Data-as-Code and Compliance

- Implement zero-copy cloning, table branching, and tagging for QA, ML models, and auditing.
- Optimize ETL/ELT data loading and achieve GDPR compliance with Iceberg's write-audit-publish (WAP).

Hive to Iceberg Migration

- Understand catalog differences and migration strategies.
- Manage late-arriving data effectively.

Iceberg Administration

- Perform table maintenance tasks.
- Configure and manage access control settings.

Capstone Project

- Apply all concepts by implementing an Open Data Lakehouse use case in Cloudera.
- Develop a comprehensive Open Data Lakehouse implementation runbook.

Are you ready to transform your data lakehouse and take your organization's data strategy to the next level? Join our immersive four-day course and master the Open Data Lakehouse architecture, focusing on the powerful combination of Apache Iceberg and Apache Ozone.

DENG-255

Building an Open Data Lakehouse Using Apache Iceberg

Class Schedule

Day 1

- Iceberg Introduction
- Data Lake Concepts
- Open Lakehouse
- Hive Architecture and Tables
- Introduction and working with Ozone
- Transfer data between HDFS & Ozone
- Ozone Application Integration
- Iceberg Architecture
- Iceberg Spark, SQL Setup
- Iceberg Catalog Review
- Iceberg Tables: Managed & External
- Table Design and Practice
- Iceberg Table Tune for Read vs Write

Day 2

- Schema Evaluation, Understand various data types issues between Hive and Iceberg during migration
- Hidden Partition: How partition works in the Iceberg table. Compare Hive and Iceberg Partition
- Time Travel. Various ways of Time Travel and How it helps for testing.
- Data-As-Code including WAP - For ETL, branching & Tags - For Zero Copy Clone for Testing QA and ML
- Iceberg Metadata for Maintenance.

Day 3

- Change Data Capture CDC
- Rollback Data
- Migration – Practice various Hive to Iceberg migration
- Shallow Migration
- In-Place Migration
- Hybrid Migration
- Snapshot migration for testing
- Late Late-arriving data migration
- RunBook build
- Table Maintenance
- Streaming

Day 4

The capstone project aims to create a Type 2 table data flow, which is a system for managing historical changes to data in a database table. In a Type 2 table, each record maintains historical information, allowing users to track changes over time. This is crucial for data warehousing and analytics, where historical data is often required for analysis and reporting purposes.