..

Open Data Lakehouse

Date published: 2022-07-24
Date modified: 2022-07-24

CLOUD≡RA

Legal Notice

Contents

What is Open Data Lakehouse?

Cloudera supports a Data Lakehouse architecture by pre-integrating and unifying the capabilities of Cloudera Data Warehouse and Data Lakes, to support data engineering, business intelligence, and machine learning – all on a single platform. Cloudera support for an open data lakehouse brings high-performance, self-service reporting and analytics to your business – simplifying data management for both for data practitioners and administrators.

Open Data Lakehouse components

- Support for Apache Iceberg 1.3 access and processing in CDP Private Cloud Base 7.1.9 and higher versions
- Compute engines (Hive, Impala, Spark, Flink) integration for accessing and processing Iceberg datasets concurrently
- SDX integration with Iceberg catalog
- Iceberg table maintenance from Spark and replication
- Iceberg Catalog set to HiveCatalog for Metastore management of Iceberg Tables
- Certified HDFS and Ozone storage

Related Information
Using Apache Iceberg

Benefits of Open Data Lakehouse

Key benefits and links to more information helps you understand how Open Data Lakehouse might help solve your business problems.

The following benefits are driving many organizations to adopt Iceberg:

- ACID compliance
- Time travel and rollback
- Flexible SQL
- In-place schema and partition evolution
- Hidden partition
- Snapshot isolation
- Multi-engine concurrent read and write

Today, many experience lower response times as volume of data increases. Open Data Lakehouse improves SQL performance for faster time to value. Iceberg is performant at scale, thanks to its metadata layout architecture and additional capabilities, such as hidden partitioning, in-place partition evolution.

To comply with regulations, your complex data structures maintain huge volumes of historical data that are error prone and hard to manage. Open Data Lakehouse time travel functionality helps you meet audit requirements–with no tedious manual snapshotting. Iceberg automatically creates snapshots and manages them, reduces your storage requirements, and offers solutions to new use cases.

Using Iceberg simplifies your business greatly due to the unification of all data. Iceberg tables can be accessed by engines outside of Cloudera. You maintain just a single copy of Iceberg data with better security and governance. Being engine-agnostic, Iceberg reduces analytic costs. Use Snowflake, Trino/Presto, any third-party engines that support the Iceberg form, or Cloudera engines: Impala, Spark, NiFi (a separate component), or Flink. Use best of breed tools to deliver self-service analytics to a Line of Business without moving, copying or transforming data.

Related Information
Using Apache Iceberg