

..

Open Data Lakehouse

Date published: 2022-07-24

Date modified: 2022-07-24

CLOUDERA

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

What is Open Data Lakehouse?.....4
Benefits of Open Data Lakehouse.....4

What is Open Data Lakehouse?

CDP supports a Data Lakehouse architecture by pre-integrating and unifying the capabilities of Data Warehouses and Data Lakes, to support data engineering, business intelligence, and machine learning – all on a single platform. Cloudera's 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
- Compute engines ([Impala](#), [Spark](#), Flink, Nifi) 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](#)



Note: CDP Open Data Lakehouse does not support queries of Iceberg tables from the Hive compute engine in this release.

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 CDP. 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 CDP 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](#)