

Apache Iceberg Overview

Date published: 2022-03-15

Date modified: 2022-03-15

CLOUDBERA

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

Iceberg overview..... 4

Iceberg overview

Apache Iceberg is a table format for huge analytics datasets on the storage systems such as HDFS. You can efficiently query large Iceberg tables on object stores. Iceberg supports concurrent reads and writes on all storage media.



Note: Cloudera supports Iceberg V1 and V2 on Red Hat OpenShift and Embedded Container Service (ECS) platforms in Cloudera Data Warehouse (CDW) on HDFS and Ozone. This feature is in technical preview and is not recommended for production deployment. Cloudera recommends that you try this feature in test or development environments.

Iceberg tables cannot be accessed from Hive, Impala, Spark, and Flink on the base cluster. You need to be on CDW. Therefore, only selected tables should be migrated to Iceberg according to the available compute resources in the private cloud environment.

You create Iceberg tables and run queries from Hive or Impala in CDP. The Hive metastore stores Iceberg metadata, including the location of the table.

Hive metastore plays a lightweight role in the Catalog operations. Iceberg relieves Hive metastore (HMS) pressure by storing partition information in metadata files on the file system/object store instead of within the HMS. This architecture supports rapid scaling without performance hits.

By default, Hive and Impala use the Iceberg HiveCatalog. Cloudera recommends the default HiveCatalog to create an Iceberg table.

You can use Iceberg when a single table contains tens of petabytes of data, and you can read these tables without compromising performance. From Apache Hive and Apache Impala, you can query Iceberg tables. The following features are included:

- Listing table snapshot and history
- Expiring snapshots from Hive and Impala
- Migrating external tables to iceberg in Hive
- Iceberg table rollback from Hive
- Creating an Iceberg table from Hive with a metadata location
- Expiring and removing old snapshots
- Performance and scalability enhancements

The following table lists Iceberg features you can access from Hive and Impala in CDW Private Cloud 1.5.0:

Feature name	Hive	Impala
Create table	#	#
Read Iceberg V1 tables	#	#
Schema evolution and partition evolution	#	#
Load data	#	#
Create table as select (CTAS)	#	#
Insert into select	#	#
Insert overwrite	#	#
Update, Delete, Merge with Iceberg V2 tables	#	#
Time travel using timestamps and snapshot IDs	#	#
Compaction	#	#
Snapshot expiration	#	#

Apache Iceberg integrates Apache Ranger for security. You can use Ranger integration with Hive and Impala to apply fine-grained access control to sensitive data in Iceberg tables. Iceberg is also integrated with Data Visualization for creating dashboards and other graphics of your Iceberg data.

Related Information

[Apache Software Foundation Iceberg Docs](#)

[Using Apache Iceberg](#)