..

Apache Iceberg Overview

Date published: 2022-03-15
Date modified: 2022-03-15

Legal Notice

Contents

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 on Red Hat OpenShift and Embedded Container Service (ECS) platforms in Cloudera Data Warehouse (CDW). This feature is in technical preview in the CDP Private Cloud 1.5.0 release 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