..

Using Apache Iceberg

Date published: 2022-03-15
Date modified: 2026-08-28



https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2026. 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

Apache Iceberg features **5**
Alter table feature 7
Create table feature 8
Create table as select feature 11
Create partitioned table as select feature 11
Create table … like feature 12
Data compaction 12
Delete data feature 14
Describe table metadata feature 15
Drop table feature 16
Drop partition feature 17
Expire snapshots feature 17
Insert table data feature 19
Load data inpath feature 20
Load or replace partition data feature 20
Materialized view feature 21
Materialized view rebuild feature 22
Merge feature 24
Migrate Hive table to Iceberg feature 25
Securing Iceberg table data 26
Flexible partitioning 26
Partition evolution feature 27
Partition transform feature 27
Insert into/overwrite partition feature 29
Truncate partition feature 30
Query metadata tables feature 30
Rollback table feature 32
Row-level operations 32
Row lineage feature [Technical Preview] 34
Puffin statistics feature 34
Select Iceberg data feature 34
Schema evolution feature 35
Schema inference feature 36
Snapshot management 37
Branching feature 37
Tagging feature 38
Table repair feature 39
LIKE predicate pushdown feature 39
Time travel feature 40
Truncate table feature 41
Update data feature 41
Best practices for Iceberg in Cloudera **42**
Making row-level changes on V2 tables only 43
Performance tuning **43**
Caching manifest files 44
Configuring manifest caching in Cloudera Manager 44
Unsupported features and limitations **45**
Accessing Iceberg tables **46**
Opening Ranger in Cloudera Data Hub 47
Editing a storage handler policy to access Iceberg files on the file system 48

Creating a SQL policy to query an Iceberg table....51
Accessing Iceberg files in Ozone....52
Creating an Iceberg table....54
Creating an Iceberg partitioned table....55
Expiring snapshots....56
Inserting data into a table....56
Migrating a Hive table to Iceberg....56
Selecting an Iceberg table....57
Running time travel queries....58
Updating an Iceberg partition....58
Test driving Iceberg from Impala....59
Hive demo data....61
Test driving Iceberg from Hive....64
Iceberg data types....65
Iceberg table properties....66
Cloudera Lakehouse Optimizer for Iceberg table maintenance....67
Cloudera Lakehouse Optimizer for Iceberg table maintenance....67
Cloudera Lakehouse Optimizer features....68
Best practices to use Cloudera Lakehouse Optimizer....72
Use cases for Cloudera Lakehouse Optimizer....73
Understand and prepare to use Cloudera Lakehouse Optimizer....73
How Cloudera Lakehouse Optimizer performs table maintenance....73
What are Cloudera Lakehouse Optimizer policy resources....74
High availability for Cloudera Lakehouse Optimizer....85
State persistence for Cloudera Lakehouse Optimizer....88
Auditing for Cloudera Lakehouse Optimizer....90
Lakehouse Optimizer UI....93
Deploying and configuring Cloudera Lakehouse Optimizer UI....93
Access Lakehouse Optimizer UI....94
Preparing to create Cloudera Lakehouse Optimizer policies in UI....104
Creating a Cloudera Lakehouse Optimizer policy in Lakehouse Optimizer UI....105
Performing dry run on Cloudera Lakehouse policies in Lakehouse Optimizer UI....114
Cloudera Lakehouse Optimizer REST APIs....114
Preparing and defining Cloudera Lakehouse Optimizer policies using REST APIs....115
Performing manual Iceberg table maintenance using Cloudera Lakehouse Optimizer REST APIs....123
Puffin statistics for Cloudera Lakehouse Optimizer....124
Data file compression for Cloudera Lakehouse Optimizer....124
Cloudera Lakehouse Optimizer REST APIs....126
Manage and monitor Cloudera Lakehouse Optimizer table maintenance tasks....131
Associating policies and fetching details using Cloudera Lakehouse Optimizer REST APIs....131
Pausing and resuming table maintenance....132
Viewing table maintenance status....133
Viewing logs for Cloudera Lakehouse Optimizer....134
Troubleshooting: Configuring the Spark engine to optimize the rewrite compaction job....135

Apache Iceberg features

You can quickly build on your past experience with SQL to analyze Iceberg tables.

From Hive or Impala, you run SQL queries to create and query Iceberg tables. Impala queries are table-format agnostic. For example, Impala options are supported in queries of Iceberg tables. You can run nested, correlated, or analytic queries on all supported table types. Most Hive queries are also table-format agnostic.

You can use Impala for manipulating Iceberg tables. The provided examples show how to run queries on Iceberg tables from Impala but do not list every supported Impala query.

If your environment uses HDFS HA, you must enable it before creating Iceberg tables. For instructions, see Enable HDFS HA before you create Iceberg tables.

Support for Apache Iceberg library version 1.10.0

Starting with Cloudera Runtime 7.3.2 SP1, Apache Iceberg library version 1.10.0 is supported for Hive and Impala. Existing Iceberg V1 and V2 table operations are unchanged; you do not need new SQL syntax for current workflows.

Variant data type

Hive supports the Iceberg variant data type on Iceberg V3 tables. In Hive, use PARSE_JSON to convert a JSON string into a variant value and VARIANT_GET to extract a nested field using a JSON path and return type.

For more information, see Iceberg data types.

Deletion vector support for Hive

Hive supports deletion vectors on Iceberg V3 tables. Deletion vectors are the Iceberg V3 encoding for position deletes. When you delete rows from a V3 table, Hive records which rows are deleted by position in the referenced data file.

Deletion vectors are a a bitmap of deleted row positions within a single data file. On V3 tables, Hive writes deletion vectors instead of V2 position delete files for new delete operations. Read operations merge active rows with the deletion vectors to return query results. You need not run a new SQL syntax to write deletion vectors. Create an Iceberg table with 'format-version'='3' and run a standard DELETE statement. Hive encodes the delete operation using deletion vectors automatically.

For more information, see Delete data feature and Row-level operations.

Deletion vector support for Impala [Technical Preview]



Note: This feature is in Technical Preview and is not ready for production deployments. Cloudera recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

Impala supports deletion vectors on Iceberg V3 tables. Deletion vectors are the Iceberg V3 encoding for position deletes. When you delete rows from a V3 table from Impala, Impala records which rows are deleted by position in the referenced data file.

On V3 tables, Impala writes deletion vectors instead of V2 position delete files for new delete operations. Read operations merge active rows with the deletion vectors to return query results. You need not run a new SQL syntax to write deletion vectors. Create an Iceberg table with 'format-version'='3' and run a standard DELETE statement. Impala encodes the delete operation using deletion vectors automatically.

For more information, see Delete data feature and Row-level operations.

Row lineage support for Impala [Technical Preview]

Note: This feature is in Technical Preview and is not ready for production deployments. Cloudera recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

Impala supports row lineage tracking on Iceberg format version 3 (V3) tables. Row lineage is an Iceberg V3 capability that assigns a unique row identifier to each row and records which snapshot last updated a row. This metadata supports change data capture (CDC) use cases. To use row lineage from Impala, create an Iceberg V3 table by setting 'format-version'='3'. When Impala writes to the table, Impala maintains the row lineage metadata required; you need not run a new SQL syntax to enable row lineage. Row lineage applies to newly written rows on V3 tables. It does not apply to rows updated using equality deletes.

For more information, see Row lineage feature.

Puffin statistics support for Impala

Impala supports writing Puffin statistics files on Iceberg V3 tables. Puffin is a lightweight container format that stores binary metadata such as deletion vector sidecar files required by the Iceberg V3 specification. To use Puffin statistics from Impala, create an Iceberg V3 table by setting 'format-version'='3'. When Impala performs supported V3 write operations, Impala writes Puffin files as required by the Iceberg V3 specification; you need not run a new SQL syntax to enable Puffin statistics writes.

For more information, see Puffin statistics feature.

Table repair support for Impala

Impala supports the repair_metadata() function for Iceberg tables. Use this function to recover tables that are inaccessible due to missing data files after manual file deletions in the underlying storage. The function synchronizes Iceberg metadata by removing references to missing data files; it does not restore deleted content.

For more information, see Table repair feature.

LIKE predicate pushdown support for Impala

Impala pushes eligible SQL LIKE predicates on Iceberg tables to Iceberg startsWith() and equal() expressions during query planning. This optimization uses Iceberg manifest statistics for file and partition pruning.

For more information, see LIKE predicate pushdown feature.

Supported ACID transaction properties

Iceberg supports atomic and isolated database transaction properties. Writers work in isolation, not affecting the live table, and perform a metadata swap only when the write is complete, making the changes in one atomic commit.

Iceberg uses snapshots to guarantee isolated reads and writes. You see a consistent version of table data without locking the table. Readers always see a consistent version of the data without the need to lock the table. Writers work in isolation, not affecting the live table, and perform a metadata swap only when the write is complete, making the changes in one atomic commit.

Iceberg partitioning

The Iceberg partitioning technique has performance advantages over conventional partitioning, such as Apache Hive partitioning. Iceberg hidden partitioning is easier to use. Iceberg supports in-place partition evolution; to change a partition, you do not rewrite the entire table to add a new partition column, and queries do not need to be rewritten for the updated table. Iceberg continuously gathers data statistics, which supports additional optimizations, such as partition pruning.

Iceberg uses multiple layers of metadata files to find and prune data. Hive and Impala keep track of data at the folder level and not at the file level, performing file list operations when working with data in a table. Performance problems occur during the execution of multiple list operations. Iceberg keeps track of a complete list of files within a table

using a persistent tree structure. Changes to an Iceberg table use an atomic object/file level commit to update the path to a new snapshot file. The snapshot points to the individual data files through manifest files.

The manifest files track several data files across many partitions. These files store partition information and column metrics for each data file. A manifest list is an additional index for pruning entire manifests. File pruning increases efficiency.

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.

Alter table feature

In Hive or Impala, you can use ALTER TABLE to set table properties. From Impala, you can use ALTER TABLE to rename a table, to change the table owner, or to change the role of the table owner. From Hive, you can alter the metadata location of the table if the new metadata does not belong to another table; otherwise, an exception occurs.

You can convert an Iceberg v1 table to v2 by setting a table property as follows:'format-version' = '2'.

Hive or Impala syntax

```
ALTER TABLE table_name SET TBLPROPERTIES table_properties;
```

- table_properties

  A list of properties and values using the following syntax:

  ```
  ('key' = 'value', 'key' = 'value', ... )
  ```

Impala syntax

```
ALTER TABLE table_name RENAME TO new_table_name;

ALTER TABLE table_name SET OWNER USER user_name;

ALTER TABLE table_name SET OWNER ROLE role_name;
```

Hive example

```
ALTER TABLE test_table SET TBLPROPERTIES('metadata_location'='hdfs://ice_tab
le/metadata/v1.metadata.json');
ALTER TABLE test_table2 SET TBLPROPERTIES('format-version' = '2');
```

Impala examples

```
ALTER TABLE t1 RENAME TO t2;

ALTER TABLE ice_table1 set OWNER USER john_doe;

ALTER TABLE ice_table2 set OWNER ROLE some_role;
ALTER TABLE ice_8 SET TBLPROPERTIES ('read.split.target-size'='268435456');

ALTER TABLE ice_table3 SET TBLPROPERTIES('format-version' = '2');
```

Create table feature

You use CREATE TABLE from Impala or CREATE EXTERNAL TABLE from Hive to create an external table in Iceberg. You learn the subtle differences in these features for creating Iceberg tables from Hive and Impala. You also learn about partitioning.

Hive and Impala handle external table creation a little differently, and that extends to creating tables in Iceberg. By default, Iceberg tables you create are v1. To create an Iceberg v2 table from Hive or Impala, you need to set a table property as follows:'format-version' = '2'.

The VARIANT data type is an Iceberg format version 3 feature. To create a Hive Iceberg table that supports the VARIANT data type, set the table property 'format-version'='3', similar to how ice_6 sets 'format-version'='2' in the Hive examples below.

Enable HDFS HA before you create Iceberg tables

Learn about how Apache Iceberg stores absolute paths in its metadata manifest files, and how HDFS HA ensures that the table is accessible regardless of which NameNode is active.

Background

Apache Iceberg stores absolute paths in its metadata manifest files for Iceberg tables. Opposed to this, Apache Hive writes the relative path to the table into the metadata, while building the absolute path based on the HDFS configuration.

The format of these paths depends on your HDFS configuration at the time of table creation. Without HDFS HA, Iceberg writes the paths using the specific hostname and port of the active NameNode. With HDFS HA, Iceberg writes the paths using the logical nameservice. Using the HDFS nameservice ensures that the paths remain valid regardless of which NameNode is active.

To avoid possible data loss when switching from one HDFS configuration to another, Cloudera highly recommends enabling HDFS HA in the environment.

Explanation

Writing a table in Hadoop in either Hive or Iceberg format includes the data itself and the metadata of the table.

While Cloudera stores the metadata in the Hive Metastore (HMS) for both Hive and Iceberg tables, the method that table location is written in the metadata varies:

- Hive uses a relative path, where the table location is concatenated with the relative path in the metadata and the HDFS definition for the NameNode or NameService that is active.
- Iceberg uses an absolute path in the table metadata to describe the location of the table at the time of its creation. If HDFS runs in non-HA mode, the cluster uses either the Active NameNode and Standby NameNode, and the used one is written to the table information in the hdfs://node4.cloudera.com:8020/warehouse/tablespace/external/hive/test.db/ice_table/ format.

  That path is accessible as long as that NameNode (node4.cloudera.com) is active. However, if the node becomes unavailable, and the secondary (standby) node takes over, the table becomes inaccessible because the original node that was used to write the table is not active. This applies to both the table information itself and the manifest JSON file associated with it, containing the absolute path to the table location.

When you enable HDFS HA, the service removes the NameNode references, and uses nameservice as the only reference. The nameservice internally manages its NameNodes, such as Load Balancer, exposing only the nameservice as the access point.

Tables created after HDFS HA is enabled use the following format (assuming the nameservice is named nameservice01): hdfs://nameservice01/warehouse/tablespace/external/hive/test.db/ice_table/

This format ensures that the table is accessible regardless of which NameNode is active, as the NameService takes care of the path, thereby supporting the failover cases.

Note: If the solution includes another cluster to which the primary cluster synchronizes data, such as in the case of Disaster Recovery, the same NameService name must be maintained on the destination cluster.

Iceberg table creation from Hive

From Hive, CREATE EXTERNAL TABLE is recommended to create an Iceberg table in Cloudera.

When you use the EXTERNAL keyword to create the Iceberg table, by default only the schema is dropped when you drop the table. The actual data is not purged. Conversely, if you do not use EXTERNAL, by default the schema and actual data is purged. You can override the default behavior. For more information, see the Drop table feature.

From Hive, you can create a table that reuses existing metadata by setting the metadata_location table property to the object store path of the metadata. The operation skips generation of new metadata and re-registers the existing metadata. Use the following syntax:

```
CREATE EXTERNAL TABLE ice_fm_hive (i int) STORED BY ICEBERG TBLPROPERTIES ('
metadata_location'='<object store or file system path>')
```

See examples below.

Iceberg table creation from Impala

From Impala, CREATE TABLE is recommended to create an Iceberg table in Cloudera. Impala creates the Iceberg table metadata in the metastore and also initializes the actual Iceberg table data in the object store.

The difference between Hive and Impala with regard to creating an Iceberg table is related to Impala compatibility with Kudu, HBase, and other tables. For more information, see the Apache documentation, "Using Impala with Iceberg Tables".

Iceberg table creation from a non-Iceberg table (Impala)

From Impala, you can use `CREATE TABLE LIKE` with STORED BY ICEBERG to create an Iceberg table that uses the column schema of an existing non-Iceberg table in the metastore. Supported source tables include HDFS-backed formats such as Parquet, ORC, Avro, and Text.

The operation creates Iceberg table metadata only; it does not copy table data. Use `INSERT INTO` or `INSERT OVERWRITE` to load data into the new Iceberg table after you create it.

```
CREATE TABLE my_ice_tbl LIKE functional_parquet.complextypestbl STORED BY IC
EBERG;
```

Metadata storage of Iceberg tables

When you create an Iceberg table using CREATE EXTERNAL TABLE in Hive or using CREATE TABLE in Impala, HiveCatalog creates an HMS table and also stores some metadata about the table on your object store, such as S3. Creating an Iceberg table generates a metadata.json file, but not a snapshot. In the metadata.json, the snapshot-id of a new table is -1. Inserting, deleting, or updating table data generates a snapshot. The Iceberg metadata files and data files are stored in the table directory under the warehouse folder. Any optional partition data is converted into Iceberg partitions instead of creating partitions in the Hive Metastore, thereby removing the bottleneck.

To create an Iceberg table from Hive or from Impala, you associate the Iceberg storage handler with the table using one of the following clauses, respectively:

- Hive: STORED BY ICEBERG
- Impala: STORED AS ICEBERG or STORED BY ICEBERG

Supported file formats

You can write Iceberg tables in the following formats:

- From Hive: Parquet (default), Avro, ORC
- From Impala: Parquet

Impala supports writing Iceberg tables in only Parquet format. Impala does not support defining both file format and storage engine. For example, CREATE TABLE tbl ... STORED AS PARQUET STORED BY ICEBERG works from Hive, but not from Impala.

You can read Iceberg tables in the following formats:

- From Hive: Parquet, Avro, ORC
- From Impala: Parquet, Avro, ORC



Note: Reading Iceberg tables in Avro format from Impala is available as a technical preview. Cloudera recommends that you use this feature in test and development environments. It is not recommended for production deployments.

Hive syntax

```
CREATE [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name
  [(col_name data_type, ... )]
  [PARTITIONED BY [SPEC]([col_name][, spec(value)][, spec(value)]...)]]
   STORED BY ICEBERG
  [STORED AS file_format]
  [TBLPROPERTIES ('key'='value', 'key'='value', ...)]
```

Impala syntax

```
CREATE TABLE [IF NOT EXISTS] [db_name.]table_name
  [(col_name data_type, ... )]
  [PARTITIONED BY [SPEC]([col_name][, spec(value)][, spec(value)]...)]]
   STORED {AS | BY} ICEBERG
  [TBLPROPERTIES (property_name=property_value, ...)]
```

Hive examples

```
CREATE EXTERNAL TABLE ice_1 (i INT, t TIMESTAMP, j BIGINT) STORED BY ICEBERG
;
      CREATE EXTERNAL TABLE ice_2 (i INT, t TIMESTAMP) PARTITIONED BY (j B
IGINT) STORED BY ICEBERG;
      CREATE EXTERNAL TABLE ice_4 (i int) STORED BY ICEBERG STORED AS ORC;
      CREATE EXTERNAL TABLE ice_5 (i int) STORED BY ICEBERG TBLPROPERTIES (
'metadata_location'='hdfs://ice_table/metadata/v1.metadata.json')
      CREATE EXTERNAL TABLE ice_6 (i int) STORED BY ICEBERG STORED AS ORC
TBLPROPERTIES ('format-version' = '2');
```

```
CREATE EXTERNAL TABLE ice_3 (v VARIANT) STORED BY ICEBERG TBLPROPERTIES ('fo
rmat-version'='3');
```

Impala examples

```
CREATE TABLE ice_7 (i INT, t TIMESTAMP, j BIGINT) STORED BY ICEBERG; //creat
es only the schema
CREATE TABLE ice_8 (i INT, t TIMESTAMP) PARTITIONED BY (j BIGINT) STORED BY
 ICEBERG; //creates schema and initializes data
CREATE TABLE ice_v2 (i INT, t TIMESTAMP) PARTITIONED BY (j BIGINT) STORED BY
 ICEBERG TBLPROPERTIES ('format-version' = '2'); //creates a v2 table
```

Related Information

Drop table feature
Partition transform feature

Create table as select feature

You can create an Iceberg table based on an existing Hive or Impala table.

The create table as select (CTAS) query can optionally include a partitioning spec for the table being created.

Hive examples

```
CREATE EXTERNAL TABLE ctas STORED BY ICEBERG AS SELECT i, t, j FROM ice_1;
```

Impala examples

```
CREATE TABLE ctas STORED BY ICEBERG AS SELECT i, b FROM ice_11;
```

Create partitioned table as select feature

You can create a partitioned Iceberg table by selecting another table. You see an example of how to use PARTITIONED BY and TBLPROPERTIES to declare the partition spec and table properties for the new table.

You see an example of using a partition transform with the PARTITIONED BY SPEC clause.

The newly created table does not inherit the partition spec and table properties from the source table in SELECT. The Iceberg table and the corresponding Hive table is created at the beginning of the query execution. The data is inserted / committed when the query finishes. So for a transient period the table exists but contains no data.

Hive syntax

```
CREATE [EXTERNAL] TABLE prod.db.sample
    USING iceberg
    PARTITIONED BY (part)
    TBLPROPERTIES ('key'='value')
    AS SELECT ...
```

Hive examples

```
CREATE EXTERNAL TABLE ctas STORED BY ICEBERG AS SELECT i, t, j FROM ice_1;

CREATE EXTERNAL TABLE ctas_part PARTITIONED BY(z) STORED BY ICEBERG TBLPROPE
RTIES ('format-version'='2')
AS SELECT x, ts, z FROM t;
CREATE EXTERNAL TABLE ctas_part_spec PARTITIONED BY SPEC (month(d)) STORED
BY ICEBERG TBLPROPERTIES ('format-version'='2')
AS SELECT x, ts, d FROM source_t;
```

Impala examples

```
CREATE TABLE ctas STORED BY ICEBERG AS SELECT i, b FROM ice_11;

CREATE TABLE ctas_part PARTITIONED BY(b) STORED BY ICEBERG AS SELECT i, s, b
 FROM ice_11;
```

```
CREATE TABLE ctas_part_spec PARTITIONED BY SPEC (month(d)) STORED BY ICEBERG
 TBLPROPERTIES ('format-version'='2')
AS SELECT x, ts, d FROM source_t;
```

Create table … like feature

You learn by example how to create an empty table based on another table.

From Hive or Impala, you can create an Iceberg table schema based on another table. The table contains no data. The table properties of the original table are carried over to the new table definition. The following examples show how to use this feature:

Hive example

```
CREATE EXTERNAL TABLE target LIKE source STORED BY ICEBERG;
```

Impala example

```
CREATE TABLE target LIKE source STORED BY ICEBERG;
```

Data compaction

From Hive and Impala, you can compact Iceberg tables and optimize them for read operations. Compaction is an essential table maintenance activity that creates a new snapshot, which contains the table content in a compact form.

Hive Iceberg compaction supports WHERE clause predicates only on partition columns.

Frequent updates and row-level modifications on Iceberg tables can result in many small data files and delete files, which have to be merged-on-read. This degrades the query performance over time. You can use the following Hive and Impala SQL statements to compact Iceberg tables and optimize the table for reading.

Impala syntax

```
OPTIMIZE TABLE [db_name.]table_name (FILE_SIZE_THRESHOLD_MB=<value>);
```

The FILE_SIZE_THRESHOLD_MB enables you to specify the maximum size of files (in MB) that should be considered for compaction. Data files larger than the specified limit are rewritten only if they are referenced from delete files.



Note:

- If FILE_SIZE_THRESHOLD_MB is specified, only the files meeting the file size criteria are rewritten according to the latest schema and partition spec and the remaining data files might still have an older schema or partition layout. Therefore, run the OPTIMIZE TABLE statement without the file size threshold option to rewrite the entire table according to the latest schema and partititon layout.
- OPTIMIZE TABLE statement without a specified FILE_SIZE_THRESHOLD_MB rewrites the entire table and the operation can take longer to complete depending on the table size. Therefore, it is recommended that you specify a file size threshold for recurring table maintenance jobs to save compute resources.

Impala example

```
OPTIMIZE TABLE ice_table (FILE_SIZE_THRESHOLD_MB=128);
```

Hive syntax

```
ALTER TABLE [database_name.]table_name COMPACT 'compaction_type' [AND WAIT];
```

```
OPTIMIZE TABLE [database_name.]table_name REWRITE DATA;
```

Hive example

```
ALTER TABLE ice_table COMPACT 'MAJOR';
```

```
OPTIMIZE TABLE ice_table REWRITE DATA;
```



Important: When Cloudera Data Warehouse and Cloudera Data Hub are deployed in the same environment and use the same Hive Metastore (HMS) instance, the Cloudera Data Hub compaction workers can inadvertently pick up Iceberg compaction tasks. Since Iceberg compaction is not yet supported in the latest Cloudera Data Hub version, the compaction tasks will fail when they are processed by the Cloudera Data Hub compaction workers.

In such a scenario where both Cloudera Data Warehouse and Cloudera Data Hub share the same HMS instance and there is requirement to run both Hive ACID and Iceberg compaction jobs, it is recommended that you use the Cloudera Data Warehouse environment for these jobs. If you want to run only Hive ACID compaction tasks, you can choose to use either the Cloudera Data Warehouse or Cloudera Data Hub environments.

If you want to run the compaction jobs without changing the environment, it is recommended that you use Cloudera Data Warehouse. To avoid interference from Cloudera Data Hub, change the value of the hive .compactor.worker.threads Hive Server (HS2) property in Cloudera Data Hub to '0'. This ensures that the compaction jobs are not processed by Cloudera Data Hub. You can find this property in Cloudera Manager Hive Configuration .

This behavior is applicable only to the Hive Iceberg manual compaction and does not affect Cloudera Lakehouse Optimizer.

To perform table optimization, ensure that the following prerequisites are met:

- The user performing compaction must have the 'ALL' permissions on the table, which can be set through Ranger.
- Impala can only write Parquet files, therefore the write.format.default table property must be set to parquet. Hive can write both Parquet and ORC file formats.
- Impala cannot compact tables with complex data types.
- Impala cannot compact views.

The OPTIMIZE TABLE statement rewrites the entire table, performing the following tasks:

- Compact small files into larger files
- Merge delete files created due to previously run DELETE and UPDATE operations
- Rewrite all files, converting them to the latest table schema
- Rewrite all partitions according to the latest partition specification
- Compact tables with partition evolution

When an Iceberg table is optimized, a new snapshot is created where all the old files of the table are replaced with newly written files. Note that if the FILE_SIZE_THRESHOLD_MB option is specified, only the files meeting the file size threshold limit are replaced. The old table state and old files can still be queried using time travel, because the rewritten data and delete files are not removed from the system. This can lead to the accumulation of unused files that belong to old snapshots. Use the *Expire Snapshots* feature to permanently remove the old files from the file system.



Note: For full compaction, since the compaction process rewrites the entire table, the operation can take a long time to complete depending on the table size.

Delete data feature

Cloudera supports row-level deletes in Iceberg V2 tables.

From Hive and Impala, you can use the following formatting types defined by the Iceberg Spec:

- position deletes
- equality deletes



Note: The equality deletes feature is in technical preview and not recommended for use in production deployments. Cloudera recommends that you try this feature in test and development environments.

- deletion vectors (Iceberg V3 table format only)

For equality deletes, you must be aware of the following considerations:

- If you are using Apache Flink or Apache NiFi to write equality deletes, then ensure that you provide a PRIMARY KEY for the table. This is required for engines to know which columns to write into the equality delete files.
- If the table is partitioned then the partition columns have to be part of the PRIMARY KEY
- For Apache Flink, the table should be in 'upsert-mode' to write equality deletes
- Partition evolution is not allowed for Iceberg tables that have PRIMARY KEYs

Position delete files contain the following information:

- file_path, which is a full URI
- pos, the file position of the row

Delete files are sorted by file_path and pos. The following table shows an example of delete files in a partitioned table:

Partition	Partition information	Delete file
YEAR(ts)=2022	ts_year=2022/data-abcd.parquet ts_year=2022/data-bcde.parquet	ts_year=2022/delete-wxyz.parquet
YEAR(ts)=2023	ts_year=2023/data-efgh.parquet	ts_year=2023/delete-hxkw.parquet
MONTH(ts)=2023-06	ts_month=2023-06/data-ijkl.parquet	ts_month=2023-06/delete-uzwd.parquet
MONTH(ts)=2023-07	ts_month=2023-07/data-mnop.parquet	ts_month=2023-07/delete-udqx.parquet

Inserting, deleting, or updating table data generates a snapshot.

You use a WHERE clause in your DELETE statement. For example:

```
delete from tbl_ice where a <= 2,1;
```

Hive and Impala evaluate rows from one table against a WHERE clause, and delete all the rows that match WHERE conditions. If you want delete all rows, use the Truncate feature. The WHERE expression is similar to the WHERE expression used in SELECT. The conditions in the WHERE clause can refer to any columns.

Concurrent operations that include DELETE do not introduce inconsistent table states. Iceberg runs validation checks to check for concurrent modifications, such as DELETE+INSERT. Only one will succeed. On the other hand, DELETE+DELETE, and INSERT+INSERT can both succeed, but in the case of a concurrent DELETE+UPDATE, UPDATE+UPDATE, DELETE+INSERT, UPDATE+INSERT from Hive, only the first operation will succeed.

From joined tables, you can delete all matching rows from one of the tables. You can join tables of any kind, but the table from which the rows are deleted must be an Iceberg table. The FROM keyword is required in this case, to separate the name of the table whose rows are being deleted from the table names of the join clauses.

For deletion vectors, you need not run a new SQL syntax to write deletion vectors. Run a standard DELETE statement on a V3 Iceberg table. Set 'format-version'='3' when you create the table. Hive encodes position deletes as deletion vectors.

Hive or Impala examples

For Impala:

Note: This feature is in Technical Preview and is not ready for production deployments. Cloudera recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

```
CREATE EXTERNAL TABLE ice_3 (i INT, s STRING) STORED BY ICEBERG TBLPROPERTIE
S ('format-version'='3');

INSERT INTO ice_3 VALUES (1, 'one'), (2, 'two'), (3, 'three');

DELETE FROM ice_3 WHERE i = 2;
```

Hive or Impala syntax

```
delete from tablename [where expression]
```

```
delete joined_tablename from [joined_tablename, joined_tablename2, ...] [ wh
ere expression ]
```

Hive or Impala examples

```
create external table tbl_ice(a int, b string, c int) stored by iceberg tblp
roperties ('format-version'='2');

insert into tbl_ice values (1, 'one', 50), (2, 'two', 51), (3, 'three', 5
2), (4, 'four', 53), (5, 'five', 54), (111, 'one', 55), (333, 'two', 56);

delete from tbl_ice where a <= 2,1;
```

The following example deletes 0, 1, or more rows of the table. If col1 is a primary key, 0 or 1 rows are deleted:

```
delete from ice_table where col1 = 100;
```

Describe table metadata feature

You can use certain Hive and Impala show and describe commands to get information about table metadata. You can also query metadata tables.

Note: Impala does not support COMPUTE INCREMENTAL STATS for Iceberg tables. If you run this command, Impala defaults to performing a complete COMPUTE STATS operation on the entire table.

For large Iceberg tables where a complete scan is too resource-intensive, use the TABLESAMPLE clause to compute statistics on a subset of the data.

The following table lists SHOW and DESCRIBE commands supported by Hive and Impala:

Command Syntax	Description	SQL engine support
SHOW CREATE TABLE table_name	Displays the schema that created the table.	Hive and Impala
SHOW FILES IN table_name	Lists the files related to the table.	Impala
SHOW PARTITIONS table_name	Returns the Iceberg partition spec, just the column information, not actual partitions or files.	Impala

Command Syntax	Description	SQL engine support
DESCRIBE [EXTENDED] table_name	EXTENDED: Displays all the metadata for the table in Thrift serialized form, which is useful for debugging.	Hive and Impala
DESCRIBE [FORMATTED] table_name	FORMATTED: Displays the metadata in tabular format.	Hive
DESCRIBE HISTORY table_name [BET WEEN timestamp1 AND timestamp2]	Limits the output history to a period of time, optionally.	Impala
SHOW FILES IN [db_name.]table_name [PARTITION (key_col_expression [, ...])]	Lists the data files related to the table or a specific partition.	Impala

Hive example

```
DESCRIBE t;
```

Hive output includes the following information:

col_name	data_type	comment
x	int	
y	int	
	NULL	NULL
# Partition Transform Information	NULL	NULL
# col_name	transform_type	NULL
y	IDENTITY	NULL

The output of DESCRIBE HISTORY includes the following columns about the snapshot. The first three are self-explanatory. The is_current_ancestor column value is TRUE if the snapshot is the ancestor of the table:

- creation_time
- snapshot_id
- parent_id
- is_current_ancestor

Impala examples

```
DESCRIBE HISTORY ice_t FROM '2022-01-04 10:00:00';
DESCRIBE HISTORY ice_t FROM now() - interval 5 days;
DESCRIBE HISTORY ice_t BETWEEN '2022-01-04 10:00:00' AND '2022-01-05 10:00:0
0';
```

Drop table feature

The syntax you use to create the table determines the default behavior when you drop the Iceberg table from Hive or Impala.

If you use CREATE TABLE, the external.table.purge flag is set to true. When the table is dropped, the contents of the table directory (actual data) are removed. If you use CREATE EXTERNAL TABLE from Hive, the external.table.purge flag is set to false. Dropping a table purges the schema only. The actual data is not removed. You can explicitly set the external.table.purge property to true to drop the data as well as the schema.

To prevent data loss during migration of a table to Iceberg, do not drop or move the table during migration. Exception: If you set the table property 'external.table.purge'='FALSE', no data loss occurs if you drop the table.

Hive or Impala syntax

```
DROP TABLE [IF EXISTS] table_name
```

Hive or Impala example

```
ALTER TABLE t SET TBLPROPERTIES('external.table.purge'='true');
DROP TABLE t;
```

Related Information

Create table feature

Drop partition feature

You can easily remove a partition from an Iceberg partition using an alter table statement from Hive or Impala.

Removing a partition does not affect the table schema. The column is not removed from the schema.

Prerequisites

- The filter expression in the drop partition syntax below is a combination of, or at least one of, the following predicates:
  - Minimum of one binary predicate
  - IN predicate
  - IS NULL predicate
- The argument of the predicate in the filter expression must be a partition transform, such as 'YEARS(col-name)' or 'column'.
- Any non-identity transform must be included in the select statement. For example, if you partition a column by days, the filter must be days.

Hive or Impala syntax

```
alter table table-name drop partition (<filter expression>)
```

The following operators are supported in the predicate: =, !=,<, >, <=, >=

Hive or Impala example

```
CREATE TABLE ice_table PARTITIONED BY SPEC (day(d)) STORED BY ICEBERG TBLPRO
PERTIES ('format-version'='2') AS SELECT x, ts, d FROM source_t;

INSERT INTO ice_table(d) VALUES ('2024-04-30');

ALTER TABLE ice_table DROP PARTITION (day(d) = '2024-04-30');
```

Expire snapshots feature

You can expire snapshots that Iceberg generates when you create or modify a table. During the lifetime of a table the number of snapshots of the table accumulate. You learn how to remove snapshots you no longer need.

Cloudera recommends periodically expiring snapshots to delete data files that are no longer needed and to reduce the size of table metadata. Each write to an Iceberg table from Hive creates a new snapshot, or version, of a table. Snapshots accumulate until expired.

You can expire snapshots based on the following conditions:

- (For Hive and Impala) All snapshots older than a timestamp or timestamp expression
- (For Hive) A snapshot having a given ID
- (For Hive) Snapshots having IDs matching a given list of IDs
- (For Hive) Snapshots within the range of two timestamps

You can keep snapshots you are likely to need, for example recent snapshots, and expire old snapshots. For example, you can keep daily snapshots for the last 30 days, then weekly snapshots for the past year, then monthly snapshots for the last 10 years. You can remove specific snapshots to meet GDPR “right to be forgotten” requirements.

Hive syntax

```
ALTER TABLE <table Name> EXECUTE EXPIRE_SNAPSHOTS(<timestamp expression>)

ALTER TABLE <table Name> EXECUTE EXPIRE_SNAPSHOTS('<Snapshot Id>')

ALTER TABLE <table Name> EXECUTE EXPIRE_SNAPSHOTS('<Snapshot Id1>,<Snapshot
 Id2>... ')

ALTER TABLE <table Name> EXECUTE EXPIRE_SNAPSHOTS BETWEEN (<timestamp exp
ression>) AND (<timestamp expression>)
```

Hive example

The first example removes snapshots having a timestamp older than August 15, 2022 1:50 pm. The second example removes snapshots from 10 days ago and before.

```
ALTER TABLE ice_11 EXECUTE EXPIRE_SNAPSHOTS('2022-08-15 13:50:00');
ALTER TABLE ice_t EXECUTE EXPIRE_SNAPSHOTS(CURRENT_TIMESTAMP - interval 10
days);
```

Impala syntax

```
ALTER TABLE <table Name> EXECUTE EXPIRE_SNAPSHOTS(<timestamp expression>)
```

Impala example

The first example removes snapshots having a timestamp older than August 15, 2022 1:50 pm. The second example removes snapshots from 10 days ago and before.

```
ALTER TABLE ice_11 EXECUTE EXPIRE_SNAPSHOTS('2022-08-15 13:50:00');
ALTER TABLE ice_t EXECUTE EXPIRE_SNAPSHOTS(now() - interval 10 days);
```

Preventing snapshot expiration

You can prevent expiration of recent snapshots by configuring the history.expire.min-snapshots-to-keep table property. You can use the alter table feature to set a property. The history.expire.min-snapshots-to-keep property refers to a number of snapshots, not a time delta. For example, assume you always want to keep all snapshots of your table for the last 24 hours. You configure history.expire.min-snapshots-to-keep as a safety mechanism to enforce this. If your table receives only one modification (insert / update / merge) per hour, then setting history.expire.min-snapshots-to-keep = 24 is sufficient to meet your requirement. However, if your table was consistently receiving updates every minute, then the last 24 hour period would entail 1440 snapshots, and the history.expire.min-snapshots-to-keep setting would need to be configured appropriately.

Table data and orphan maintenance

The contents of the table directory (actual data) might, or might not, be removed when you drop the table. An orphan data file can remain when you drop an Iceberg table, depending on the external.table.purge flag table property. An orphaned data file is one that has contents in the table directory, but no snapshot.

Expiring a snapshot does not remove old metadata files by default. You must clean up metadata files using writ e.metadata.delete-after-commit.enabled=true and write.metadata.previous-versions-max table properties. For more information, see "Iceberg table properties" below. Setting this property controls automatic metadata file removal after metadata operations, such as expiring snapshots or inserting data.

Related Information

Expiring snapshots

Iceberg table properties

Insert table data feature

From Hive and Impala, you can insert data into Iceberg tables using the standard INSERT INTO a single table. INSERT statements work for V1 and V2 tables.

You can replace data in the table with the result of a query. To replace data, Hive and Impala dynamically overwrite partitions that have rows returned by the SELECT query. Partitions that do not have rows returned by the SELECT query, are not replaced. Using INSERT OVERWRITE on tables that use the BUCKET partition transform is not recommended. Results are unpredictable because dynamic overwrite behavior would be too random in this case.

From Hive, Cloudera also supports inserting into multiple tables as a technical preview; however, this operation is not atomic, so data consistency of Iceberg tables is equivalent to that of Hive external tables. Changes within a single table will remain atomic.

Inserting, deleting, or updating table data generates a snapshot. A new snapshot corresponds to a new manifest list. Manifest lists are named snap-*.avro.

Iceberg specification defines sort orders. At this point, Hive doesn't support defining sort orders. But if there are sort orders defined by using other engines Hive can utilize them on write operations. For more information about sorting, see sort orders specification.

Hive or Impala syntax

```
INSERT INTO TABLE tablename VALUES values_row [, values_row ...]

INSERT INTO TABLE tablename1 select_statement1 FROM tablename2

INSERT OVERWRITE TABLE tablename1 select_statement1 FROM tablename2
```

Hive or Impala examples

```
CREATE TABLE ice_10 (i INT, s STRING, b BOOLEAN) STORED BY ICEBERG;
INSERT INTO ice_10 VALUES (1, 'asf', true);
CREATE TABLE ice_11 (i INT, s STRING, b BOOLEAN) STORED BY ICEBERG;
INSERT INTO ice_11 VALUES (2, 'apache', false);
INSERT INTO ice_11 SELECT * FROM ice_10;
SELECT * FROM ice_11;
INSERT OVERWRITE ice_11 SELECT * FROM ice_10;
```

Hive example

```
FROM customers
    INSERT INTO target1 SELECT customer_id, first_name;
```

```
    INSERT INTO target2 SELECT last_name, customer_id;
```

Load data inpath feature

From Hive or Impala, you can load Parquet or ORC data from a file in a directory on your file system or object store into an Iceberg table. For Impala, you might need to set the mem_limit or pool configuration (max-query-mem-limit, min-query-mem-limit) to accommodate the load.

Hive syntax

```
LOAD DATA [LOCAL] INPATH '<path to file>' [OVERWRITE] INTO TABLE tablename;
```

Hive example

```
LOAD DATA LOCAL INPATH '/tmp/some_db/files/part.orc' INTO TABLE ice_orc;

LOAD DATA LOCAL INPATH '/tmp/some_db/files/part.orc' OVERWRITE INTO TABLE
ice_orc;
```



Note:

- Specifying the LOCAL keyword looks for a file path in the local file system.
- The OVERWRITE keyword deletes the records in the target table and replaces it with the files specified in the filepath, else, the command adds to the existing content in the target table.

Impala syntax

```
LOAD DATA INPATH '<path to file>' INTO TABLE tablename;
```

Impala example

In this example, you create a table using the LIKE clause to point to a table stored as Parquet. This is required for Iceberg to infer the schema. You also load data stored as ORC.

```
CREATE TABLE test_iceberg LIKE my_parquet_table STORED AS ICEBERG;
SET MEM_LIMIT=1GB;

LOAD DATA INPATH '/tmp/some_db/parquet_files/' INTO TABLE iceberg_tbl;

LOAD DATA INPATH '/tmp/some_db/orc_files/' INTO TABLE iceberg2_tbl;
```

Load or replace partition data feature

There is no difference in the way you insert data into a partitioned or unpartitioned Iceberg table.

Working with partitions is easy because you write the query in the same way for the following operations:

- Insert into, or replace, an unpartitioned table
- Insert into, or replace, an identity partitioned table
- Insert into, or replace, a transform-partitioned table

Do not use INSERT OVERWRITE on tables that went through partition evolution. Truncate such tables first, and then INSERT the tables.

Impala example

```
CREATE TABLE ice_12 (i int, s string, t timestamp, t2 timestamp) STORED BY I
CEBERG;
```

Hive or Impala examples

```
INSERT INTO ice_12 VALUES (42, 'impala', now(), to_date(now()));
INSERT OVERWRITE ice_t VALUES (42, 'impala', now(), to_date(now()));
```

Materialized view feature

Using a materialized view can accelerate query execution. Creating a materialized view on top of Iceberg tables in Cloudera can further accelerate the performance. You can create a materialized view of an Iceberg V1 or V2 table based on an existing Hive or Iceberg table.

The materialized view is stored in Hive ACID or Iceberg format. Materialized view source tables either must be native ACID tables or must support table snapshots. Automatic rewriting of a materialized view occurs under the following conditions:

- The view definition contains the following operators only:
  - TableScan
  - Project
  - Filter
  - Join(inner)
  - Aggregate
- Source tables are native ACID or Iceberg v1 or v2
- The view is not based on time travel queries because those views do not have up-to-date data by definition.

Hive example

The following example creates a materialized view of an Iceberg table from Hive.

```
drop table if exists tbl_ice;
create external table tbl_ice(a int, b string, c int) stored by iceberg st
ored as orc tblproperties ('format-version'='2');

create materialized view mat1 as
select b, c from tbl_ice for system_version as of 5422037307753150798;
```

The following example creates a materialized view of two Iceberg tables. Joined tables must be in the same table format, either Iceberg or Hive ACID.

```
drop table if exists tbl_ice2;
create table tbl_ice2(a int, b string, c int) stored as orc TBLPROPERTIES ('
transactional'='true');
INSERT INTO tbl_ice2 VALUES (2, 'apache', 3);

drop table if exists tbl_ice;
create external table tbl_ice(a int, b string, c int) stored by iceberg sto
red as orc tblproperties ('format-version'='2');
INSERT INTO tbl_acid VALUES (4, 'iceberg', 5);
create materialized view mat1 as
select tbl_ice2.b, tbl_ice2.c from tbl_ice join tbl_ice2 on tbl_ice.a = tb
l_ice2.a;
```

The following example uses explain to examine a materialized view and then creates a materialized view of an Iceberg V1 table from Hive.

```
drop materialized view if exists mat1;
drop table if exists tbl_ice;

create table tbl_ice(a int, b string, c int) stored by iceberg stored as orc
 tblproperties ('format-version'='1');
insert into tbl_ice values (1, 'one', 50), (2, 'two', 51), (3, 'three', 52),
 (4, 'four', 53), (5, 'five', 54);
explain create materialized view mat1 stored by iceberg stored as orc tblpr
operties ('format-version'='1') as
select tbl_ice.b, tbl_ice.c from tbl_ice where tbl_ice.c > 52;

create materialized view mat1 stored by iceberg stored as orc tblproperties
 ('format-version'='1') as
select tbl_ice.b, tbl_ice.c from tbl_ice where tbl_ice.c > 52;
```

Related Information

Materialized view rebuild feature

Materialized view rebuild feature

Updates to materialized view contents when new data is added to the underlying table are critical; otherwise, queries can return stale data.

An update can occur under the following conditions:

- As a row-level incremental rebuild of the view after inserting data into a table

  Source tables can be Iceberg V2 or Hive full ACID.
- As a full rebuild of the view

A full rebuild can be expensive. An incremental rebuild updates only the affected parts of the materialized view, decreasing rebuild step execution time.

An incremental rebuild occurs automatically when you insert (append) data into a source table; otherwise, after you make some other type of change, for example a delete, you must manually start a full rebuild.

You use the ALTER command to manually start a full rebuild of the materialized view from Hive as follows:

Hive syntax

```
ALTER MATERIALIZED VIEW <name of view> REBUILD;
```

Full rebuild example

In this example, first you set required properties. Next, you create Iceberg tables, a V1 table and a V2 table, from Hive. You insert data into the tables and create a materialized view of the joined tables. You insert some new values into one of the source tables, rendering the materialized view stale. Finally, you rebuild the materialized view using explain cbo to show the rebuild plan The rebuild plan indicates a full rebuild will occur, which means the definition query will be executed.

```

drop table if exists tbl_ice;
drop table if exists tbl_ice_v2;
create external table tbl_ice(a int, b string, c int) stored by iceberg sto
red as orc tblproperties ('format-version'='1');
create external table tbl_ice_v2(d int, e string, f int) stored by iceberg
stored as orc tblproperties ('format-version'='2');
```

```
insert into tbl_ice_v2 values (1, 'one v2', 50), (4, 'four v2', 53), (5, 'f
ive v2', 54);

create materialized view mat1 as
select tbl_ice.b, tbl_ice.c, tbl_ice_v2.e from tbl_ice
join tbl_ice_v2 on tbl_ice.a=tbl_ice_v2.d where tbl_ice.c > 52;
group by tbl_ice.b tbl_ice.c;

-- view should be empty
select * from mat1;

-- view is up-to-date, use it
explain cbo
select tbl_ice.b, tbl_ice.c, tbl_ice_v2.e from tbl_ice join tbl_ice_v2 on
tbl_ice.a=tbl_ice_v2.d where tbl_ice.c > 52;

-- insert some new values to one of the source tables
insert into tbl_ice values (1, 'one', 50), (2, 'two', 51), (3, 'three', 52),
 (4, 'four', 53), (5, 'five', 54);

-- view is outdated, cannot be used
explain cbo
select tbl_ice.b, tbl_ice.c, tbl_ice_v2.e from tbl_ice join tbl_ice_v2 on tb
l_ice.a=tbl_ice_v2.d where tbl_ice.c > 52;

explain cbo
alter materialized view mat1 rebuild;

-- view should contain data
select * from mat1;

-- view is up-to-date again, use it
explain cbo
select tbl_ice.b, tbl_ice.c, tbl_ice_v2.e from tbl_ice join tbl_ice_v2 on tb
l_ice.a=tbl_ice_v2.d where tbl_ice.c > 52;
group by tbl_ice.b tbl_ice.c;
```

Automatic query rewrite example

In this example, you create an Iceberg table, insert some values, and create the materialized view. The view is partitioned using a partition specification and stored in the Iceberg ORC format. The v1 format version is specified in this example (the v2 format is also supported). You then look at a description of the view and see that the query rewrite option is enabled by default. An automatic incremental rebuild is possible when this option is enabled.

```
   drop table if exists tbl_ice;
   create external table tbl_ice(a int, b string, c int) stored by iceberg
 stored as orc tblproperties ('format-version'='1');

   -- insert some new values into one of the source tables

   insert into tbl_ice values (1, 'one', 50), (2, 'two', 51), (3, 'three', 5
2), (4, 'four', 53), (5, 'five', 54);

   explain
   create materialized view mat1 partitioned on spec (bucket(16, b), trunc
ate(3, c)) stored by Iceberg stored as orc tblproperties(‘format-version’=’1
’)as
   select tbl_ice.b, tbl)ice.c from tbl_ice where tbl_ice.c > 52;

   -- the output query plan query indicates a rewrite is enabled
```

```
   POSTHOOK: query: explain
   create materialized view mat1 …
   …
   STAGE PLANS
  …
-- In stage one, the materialized view is created by calling the Iceberg
API to create the table object.
rewrite enabled
…
```

You use the DESCRIBE command to see the output query plan, which shows details about the view, including if it can be used in automatic query rewrites.

```
-- check the materialized view details
    describe formatted mat1;
    …
    #col_name     data_type      comment
    b             string
    c             int

    #Partition Transform Information
    #col_name     transform_type
    b             bucket(16)
    c             TRUNCATE[3]


    #detailed table information
    …
    Table_type:    MATERIALIZED_VIEW
    Table Parameters:
    …
    current-snapshot-id   563939E424367334713
    …
    metadata_location     …
    …
    Table_type            Iceberg
    …
    #Materialized View Information
    Original Query:  …
    Expanded Query:  …
    Rewrite Enabled: Yes
```

With the query rewrite option enabled, you insert data into the source table, and incremental rebuild occurs automatically. You do not need to rebuild the view manually before running queries.

Related Information

Materialized view feature

Merge feature

From Hive and Impala, you can perform actions on an Iceberg table based on the results of a join between a target and source v2 Iceberg table.

The MERGE statement supports multiple WHEN clauses, where each clause can specify actions like UPDATE, DELETE, or INSERT. Actions are applied based on the join conditions defined between the source and target tables.

Hive and Impala syntax

```
MERGE INTO <target table> AS T USING <source expression/table> AS S
ON <boolean expression1>
```

```
WHEN MATCHED [AND <boolean expression2>] THEN UPDATE SET <set clause list>
WHEN MATCHED [AND <boolean expression3>] THEN DELETE
WHEN NOT MATCHED [AND <boolean expression4>] THEN INSERT VALUES <value list>
```

Hive and Impala example

Use the MERGE INTO statement to update an Iceberg table based on a staging table:

```
MERGE INTO customer USING new_customer_stage source ON source.id = customer.
id
    WHEN MATCHED THEN UPDATE SET name = source.name, state = source.new_stat
e
    WHEN NOT MATCHED THEN INSERT VALUES (source.id, source.name, source.s
tate);
```

Create an Iceberg table and merge it with a non-Iceberg table.

```
create external table target_ice(a int, b string, c int) partitioned by spec
 (bucket(16, a), truncate(3, b)) stored by iceberg stored as orc tblproperti
es ('format-version'='2');
create table source(a int, b string, c int);
...

merge into target_ice as t using source src ON t.a = src.a
when matched and t.a > 100 THEN DELETE
when matched then update set b = 'Merged', c = t.c + 10
when not matched then insert values (src.a, src.b, src.c);
```

The Impala MERGE INTO statement supports the following capabilities:

- WHEN NOT MATCHED BY SOURCE merge clause - Useful in situations when a source table's rows do not match the target table rows. For example:

  ```
  MERGE INTO target_ice t using source s on t.id = s.id
  WHEN NOT MATCHED BY SOURCE THEN UPDATE set t.column = "a";
  ```

- Supports INSERT * syntax for the WHEN NOT MATCHED clause and UPDATE SET * syntax for the WHEN MATCHED clause.

  INSERT * enumerates all expressions from the source table or subquery to simplify inserting for target tables with large number of columns. The semantics is the same as the regular WHEN NOT MATCHED THEN INSERT clause.

  UPDATE SET * creates assignments for each target table column by enumerating the table columns and assigning source expressions by index.

Migrate Hive table to Iceberg feature

Cloudera supports Hive table migration from Hive to Iceberg tables using ALTER TABLE to set the table properties.



Note: Do not drop or move the old table during a migration operation. Doing so will delete the data files of the old and new tables. Exception: If you set the table property 'external.table.purge'='FALSE', no data loss occurs when you drop the table.

In-place table migration process

In-place table migration saves time generating Iceberg tables. There is no need to regenerate data files. Only metadata, which points to source data files, is regenerated.

To convert a Hive table to an Iceberg V1 table, use the following syntax:

```
ALTER TABLE table_name CONVERT TO ICEBERG;
```

To convert a Hive table to an Iceberg V2 table, you must run two queries. Use the following syntax:

```
ALTER TABLE table_name CONVERT TO ICEBERG

ALTER TABLE table_name SET TBLPROPERTIES ('format-version' = '2'
    ...)
```

In-place table migration saves time generating Iceberg tables. There is no need to regenerate data files. Only metadata, which points to source data files, is regenerated.

Securing Iceberg table data

Learn how to prevent Hive and Impala from reading data files from an Iceberg table that are outside of the table location.

An unauthorized user who knows the underlying file layout of a table can gain access to the data in an Iceberg table in the following way:

User1 owns a table, table1, which User2 does not have permission to read. However, User2 could execute an attack as follows: User2 creates a new table, table2, to which they have access rights. User2 then modifies the metadata files of their own table (table2) to reference data files from User1's table (table1), effectively including these files as part of table2. By accessing their own table (table2), User2 can then read the data files that belong to User1's table (table1).

You can prevent this by configuring Hive and Impala to prevent reading data files that are outside of the Iceberg table location.

Hive

Add the following property to HiveServer2 Advanced Configuration Snippet (Safety Valve) for hive-site.xml and set the value to "true":

```
hive.iceberg.allow.datafiles.in.table.location.only
```

Impala

Ensure that the following Impala startup flag is set to "true". The value of this flag is set to true by default.

```
iceberg_allow_datafiles_in_table_location_only
```

Flexible partitioning

Iceberg partition evolution, which is a unique Iceberg feature, and the partition transform feature, greatly simplify partitioning tables and changing partitions.

Partitions based on transforms are stored in the Iceberg metadata layer, not in the directory structure. You can change the partitioning completely, or just refine existing partitioning, and write new data based on the new partition layout--no need to rewrite existing data files. For example, change a partition by month to a partition by day.

Use partition transforms, such as IDENTITY, TRUNCATE, BUCKET, YEAR, MONTH, DAY, HOUR. Iceberg solves scalability problems caused by having too many partitions. Partitioning can also include columns with a large number of distinct values. Partitioning is hidden in the Iceberg metadata layer, eliminating the need to explicitly write partition columns (YEAR, MONTH for example) or to add extra predicates to queries for partition pruning.

```
SELECT * FROM tbl WHERE ts = '2023-04-21 20:56:08'
AND YEAR = 2023 AND MONTH = 4 AND DAY = 21
```

Year, month, and day can be automatically extracted from '2023-04-21 20:56:08' if the table is partitioned by DAY(ts)

Partition evolution feature

Partition evolution means you can change the partition layout of the table without rewriting existing data files. Old data files can remain partitioned by the old partition layout, while newly added data files are partitioned based on the new layout.

You can use the ALTER TABLE SET PARTITION SPEC statement to change the partition layout of an Iceberg table. A change to the partition spec results in a new metadata.json and a commit, but does not create a new snapshot.

The VOID transform might be used to replace the transform in an existing partition field so that the field is effectively dropped in v1 tables.

Hive or Impala syntax

```
ALTER TABLE table_name SET PARTITION SPEC ([col_name][, spec(value)][, spec(
value)]...)]
```

spec: The specification for a transform listed in the next topic, "Partition transform feature".

Hive or Impala examples

```
ALTER TABLE t
SET PARTITION SPEC ( TRUNCATE(5, level), HOUR(event_time),
BUCKET(15, message), price);
ALTER TABLE ice_p
SET PARTITION SPEC (VOID(i), VOID(d), TRUNCATE(3, s), HOUR(t), i);
```

Related Information

Partition transform feature

Partition transform feature

From Hive or Impala, you can use one or more partition transforms to partition your data. Each transform is applied to a single column. Identity-transform means no transformation; the column values are used for partitioning. The other transforms apply a function to the column values and the data is partitioned by the transformed values.

Using CREATE TABLE ... PARTITIONED BY you create identity-partitioned Iceberg tables. Identity-partitioned Iceberg tables are similar to the Hive or Impala partitioned tables, which are stored in the same directory structure as the data files. Iceberg stores the partitioning columns of identity-partitioned Iceberg tables in a different directory structure from the data files if the tables are migrated to Iceberg from Hive external tables. Iceberg handles the tables and files regardless of the location.

Hive and Impala support Iceberg advanced partitioning through the PARTITION BY SPEC clause. Using this clause, you can define the Iceberg partition fields and partition transforms.

The following table shows the available partition transforms and their corresponding transform specifications:

Partition transform	Transform specification	Supported SQL engine
Partition by year	years(time_stamp) | year(time_stamp)	Hive and Impala
Partition by month	months(time_stamp) | month(time_stamp)	Hive and Impala
Partition by a date value stored as int (dateint)	days(time_stamp) | date(time_stamp)	Hive
Partition by hours	hours(time_stamp)	Hive
Partition by a dateint in hours	date_hour(time_stamp)	Hive
Partition by hashed value mod N buckets	bucket(N, col)	Hive and Impala

Partition transform	Transform specification	Supported SQL engine
Partition by value truncated to L, which is a number of characters	truncate(L, col)	Hive and Impala
Partition by hashed value mod N buckets	bucket(N, col)	Hive and Impala
Partition by value truncated to L, which is a number of characters	truncate(L, col)	Hive and Impala
Partition by value	identity(col)	Impala
Partition by null values	void(col)	Impala

The BUCKET, TRUNCATE, IDENTITY, and VOID transforms now support the BINARY type.

The VOID transform might be used to replace the transform in an existing partition field so that the field is effectively dropped in v1 tables.

Strings are truncated to length L. Integers and longs are truncated to bins. For example, truncate(10, i) yields partitions 0, 10, 20, 30 …

The idea behind transformation partition by hashed value mod N buckets is the same as hash bucketing for Hive tables. A hashing algorithm calculates the bucketed column value (modulus). For example, for 10 buckets, data is stored in column value % 10, ranging from 0-9 (0 to n-1) buckets.

You use the PARTITIONED BY SPEC clause to partition a table by an identity transform.

Hive syntax

```
CREATE [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name
  [(col_name data_type][, time_stamp TIMESTAMP] )]
  [PARTITIONED BY SPEC([col_name][, spec(value)][, spec(value)]...)]
  [STORED AS file_format]
   STORED BY ICEBERG
  [TBLPROPERTIES (property_name=property_value, ...)]
```

Where spec (value) represents one or more of the following transforms:

- YEARS(col_name)
- MONTHS(col_name)
- DAYS(col_name)
- BUCKET(bucket_num,col_name)
- TRUNCATE(length, col_name)

Impala syntax

```
CREATE TABLE [IF NOT EXISTS] [db_name.]table_name
  [(col_name data_type, ... )]
  [PARTITIONED BY SPEC([col_name][, spec(value)][, spec(value)]...)]
   STORED (AS | BY) ICEBERG
  [TBLPROPERTIES (property_name=property_value, ...)]
```

Where spec(value) represents one or more of the following transforms:

- YEARS(col_name)
- MONTHS(col_name)
- DAYS(col_name)
- BUCKET(bucket_num,col_name)
- TRUNCATE(length, col_name)

Hive example

The following example creates a top level partition based on column i, a second level partition based on the hour part of the timestamp, and a third level partition based on the first 1000 characters in column j.

```
CREATE EXTERNAL TABLE ice_3 (i INT, t TIMESTAMP, j BIGINT) PARTITIONED BY SP
EC (i, HOUR(t), TRUNCATE(1000, j)) STORED BY ICEBERG;
```

Impala examples

```
CREATE TABLE ice_13 (i INT, t TIMESTAMP, j BIGINT) PARTITIONED BY SPEC (i, H
OUR(t), TRUNCATE(1000, j)) STORED BY ICEBERG;
```

The following examples show how to use the PARTITION BY SPEC clause in a CREATE TABLE query from Impala.The same transforms are available in a CREATE EXTERNAL TABLE query from Hive.

```
CREATE TABLE ice_t(id INT, name STRING, dept STRING)
PARTITIONED BY SPEC (bucket(19, id), dept)
STORED BY ICEBERG
TBLPROPERTIES ('format-version'='2');
```

```
CREATE TABLE ice_ctas
PARTITIONED BY SPEC (truncate(1000, id))
STORED BY ICEBERG
TBLPROPERTIES ('format-version'='2')
AS SELECT id, int_col, string_col FROM source_table;
```

Related Information

Creating an Iceberg partitioned table

Create table feature

Partition evolution feature

Insert into/overwrite partition feature

From Hive you can insert into, or overwrite data in, Iceberg tables that are statically or dynamically partitioned.

You can also insert data into a table using static and dynamic partitions. INSERT OVERWRITE queries do not work on tables that have undergone partition evolution.

Static partition syntax

```
INSERT INTO|OVERWRITE TABLE name PARTITION(col = val) VALUES (...)
INSERT INTO|OVERWRITE TABLE name PARTITION(col = val) SELECT <expression>
```

Dynamic partition syntax

```
INSERT INTO|OVERWRITE TABLE name PARTITION(col) VALUES (...)
INSERT INTO|OVERWRITE TABLE name PARTITION(col) SELECT <expression>
```

Static and dynamic partition syntax

Insertion into the static partition must precede the insertion into the dynamic partition.

```
INSERT INTO|OVERWRITE TABLE name PARTITION(col1 = val, col2) VALUES (...)
INSERT INTO|OVERWRITE TABLE name PARTITION(col1 = val, col2) SELECT <expre
ssion>
```

Truncate partition feature

You can truncate Iceberg partitioning from Hive.

You specify columns in the TRUNCATE TABLE ... PARTITION command as shown in the following syntax. This command does not support partitioning on columns that have partition transforms other than identity-transform.

Hive syntax

```
TRUNCATE TABLE t PARTITION (parCol1 = Val1, parCol2 = Val2 ... );
```

file:/Users/nisha.sridhar/Sandbox/newco-docs/topics/iceberg-truncate-partition.xml

Query metadata tables feature

Apache Iceberg stores extensive metadata for its tables. From Hive and Impala, you can query the metadata tables as you would query a regular table. For example, you can use projections, joins, filters, and so on.

Iceberg metadata tables include information that is useful for efficient table maintenance (about snapshots, manifests, data, delete files, etc.) as well as statistics that help query engines plan and execute queries more efficiently (value count, min-max values, number of NULLs, etc.).

Note: Iceberg metadata tables are read-only. You cannot add, remove, or modify records in the tables. Also, you cannot drop or create new metadata tables.

For more information about the Apache Iceberg Iceberg metadata table types, see the *Apache Iceberg MetadataTableType enumeration*.

For more information about querying Iceberg metadata, see the *Apache Iceberg Spark documentation*.

The following sections describe how you can interact with and query Iceberg metadata tables:

List metadata tables

The SHOW METADATA TABLES statement lists all metadata tables belonging to an Iceberg table. You can also filter the tables according to a specific pattern.

Note: The SHOW METADATA TABLES statement is only available in Impala and is not supported in Hive.

Impala Syntax:

```
SHOW METADATA TABLES IN [database_name.]table_name [[LIKE] "pattern"];
```

Impala Example:

```
SHOW METADATA TABLES IN default.ice_table;

Output:
+----------------------+
| name                 |
+----------------------+
| all_data_files       |
| all_delete_files     |
| all_entries          |
| all_files            |
| all_manifests        |
| data_files           |
| delete_files         |
| entries              |
| files                |
```

```
| history              |
| manifests            |
| metadata_log_entries |
| partitions           |
| position_deletes     |
| refs                 |
| snapshots            |
+----------------------+
```

Query metadata tables

You can use the regular SELECT statement from Hive or Impala to query Iceberg metadata tables. To reference a metadata table, use the fully qualified path as shown in the syntax.

Hive or Impala Syntax:

```
SELECT … FROM database_name.table_name.metadata_table_name;
```

Hive or Impala Example:

```
SELECT * FROM default.ice_table.files;
```

You can select any subset of the columns or all of them using '*'. In comparison to regular tables, running a SELE CT * from Impala on metadata tables always includes complex-typed columns in the result. The Impala query option, EXPAND_COMPLEX_TYPES only applies to regular tables. However, Hive always includes complex columns irresepctive of whether SELECT queries are run on regular tables or metadata tables.

For Impala queries that have a mix of regular tables and metadata tables, a SELECT * expression where the sources are metadata tables always includes complex types, whereas for SELECT * expressions where the sources are regular tables, complex types are included only if the EXPAND_COMPLEX_TYPES query option is set to 'true'.

In the case of Hive, columns with complex types are always included.

You can also filter the result set using a WHERE clause, use aggregate functions such as MAX or SUM, JOIN metadata tables with other metadata tables or regular tables, and so on.

Example:

```
SELECT
    s.operation,
    h.is_current_ancestor,
    s.summary
FROM default.ice_table.history h
JOIN default.ice_table.snapshots s
  ON h.snapshot_id = s.snapshot_id
WHERE s.operation = 'append'
ORDER BY made_current_at;
```

Limitations

- Impala does not support the DATE and BINARY data types. NULL is returned instead of their actual values.
- Impala does not support unnesting collections from metadata tables.

DESCRIBE metadata tables

Like regular tables, Iceberg metadata tables have schemas that can be explored using the DESCRIBE statement. The DESCRIBE statement displays metadata about a table, such as the column names and their data types.

To reference the metadata table, use the fully qualified path as shown in the syntax.

Note: DESCRIBE FORMATTED|EXTENDED is not available for metadata tables. In Impala, using this statement results in an error whereas Hive displays the same output as the regular DESCRIBE statement.

Hive or Impala Syntax:

```
DESCRIBE database_name.table_name.metadata_table_name;
```

Hive or Impala Example:

```
DESCRIBE default.ice_table.history;
Output:
+---------------------+-----------+---------+----------+
| name                | type      | comment | nullable |
+---------------------+-----------+---------+----------+
| made_current_at     | timestamp |         | true     |
| snapshot_id         | bigint    |         | true     |
| parent_id           | bigint    |         | true     |
| is_current_ancestor | boolean   |         | true     |
+---------------------+-----------+---------+----------+
```

Related Information

Apache Iceberg MetadataTableType

Apache Spark documentation

Rollback table feature

In the event of a problem with your table, you can reset a table to a good state as long as the snapshot of the good table is available. You can roll back the table data based on a snapshot id or a timestamp.

When you modify an Iceberg table, a new snapshot of the earlier version of the table is created. When you roll back a table to a snapshot, a new snapshot is created. The creation date of the new snapshot is based on the Timezone of your session. The snapshot id does not change.

Hive and Impala syntax

```
ALTER TABLE test_table EXECUTE rollback(snapshotID);
ALTER TABLE test_table EXECUTE rollback('timestamp');
```

Hive and Impala examples

The following example rolls back to an earlier table, creating a new snapshot having a new creation date timestamp, but keeping the same snapshot id 3088747670581784990.

```
ALTER TABLE ice_t EXECUTE ROLLBACK(3088747670581784990);
```

The following example rolls the table back to the latest snapshot having a creation timestamp earlier than '2022-08-08 00:00:00'.

```
ALTER TABLE ice_7 EXECUTE ROLLBACK('2022-08-08 00:00:00')
```

Row-level operations

Hive supports the copy-on-write (COW) as well as merge-on-read (MOR) modes for handling Iceberg row-level updates and deletes. Impala supports only the MOR mode and will fail if configured for copy-on-write. Impala does support reading copy-on-write tables.

Depending on the COW or MOR setting, Hive performs updates and deletes to Iceberg tables as follows:

- COW: Hive creates a new version of files for each update/delete. Use COW when updating/deleting a large number of rows, or when reading data frequently.
- MOR (default): Updates/deletes are logged to delta files, which tends to be faster than creating new versions of the files. Later a compaction can eliminate the delete files and rewrite the affected data files.

Impala uses only the MOR method. Impala does not support copy-on-write and will fail if configured for copy-on-write. Impala does support reading copy on write tables.

Configuring Hive tables to support COW or MOR

You use the CREATE TABLE or ALTER commands to set either COW or MOR mode in table properties as shown in the following examples:

```
ALTER TABLE ice01 SET TBLPROPERTIES ('write.delete.mode'='copy-on-write');
ALTER TABLE ice01 SET TBLPROPERTIES ('write.update.mode'='copy-on-write');
```

When to use COW or MOR

Set either COW or MOR based on your use case and rate of data change. Consider the following advantages and disadvantages of the modes:

- MOR
  - Writes are efficient.
  - Reads are inefficient due to read amplification, but regularly scheduled compaction can reduce inefficiency.
  - A good choice when streaming.
  - A good choice when frequently writing or updating, such as running hourly batch jobs.
  - A good choice when the percentage of data change is low.

  COW
  - Reads are efficient.
  - A good choice for bulk updates and deletes, such as running a daily batch job.
  - Writes are inefficient due to write amplification, but the need for compaction is reduced.
  - A good choice when the percentage of data change is high.

Position or equality delete files

By default, Hive and Impala delete Iceberg V2 table data using position delete files. Hive and Impala can read equality deletes, which you might encounter in a table created and written to by Flink or another engine that supports writing equality deletes. Hive and Impala cannot write equality deletes.

Starting with Cloudera Runtime 7.3.2 SP1, Hive and Impala [Technical Preview] writes position deletes on Iceberg V3 tables as deletion vectors instead of position delete files. For more information, see Delete data feature.

Choose the MOR or COW mode based on your use case. The MOR mode keeps track of deleted rows, so it's expensive when you're making many deletions.

The position and equality deletes are summarized below:

Position delete: DELETE operations find the data records that need to be deleted, then write out their file path and position into delete files. Postponing the rewrite of the files speeds writes of updates/deletes. Read operations need to merge the data and delete files to retrieve the active rows.

Equality delete: DELETE operations write out column values (for example, primary keys) that identify the records to be deleted. If you already know the column values of the records you want to delete, then there is no need to read the whole table, which means equality deletes can be written faster than position deletes. Read operations need to merge the data and delete files to retrieve the active rows. Processing equality deletes is typically slower than processing position deletes.

Row lineage feature [Technical Preview]

Impala supports row lineage tracking on Iceberg V3 tables. Row lineage is an Iceberg V3 capability. Row lineage assigns a unique row identifier to each row and records which snapshot last updated a row. This metadata supports change data capture (CDC) and row provenance use cases.

Note: This feature is in Technical Preview and is not ready for production deployments. Cloudera recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

To use row lineage from Impala, create an Iceberg V3 table by setting 'format-version'='3'. When Impala writes to the table, Impala maintains the row lineage metadata that the specification requires; you need not use new SQL syntax to enable row lineage.

Row lineage applies to newly written rows on V3 tables. It does not apply to rows updated using equality deletes.

Impala example

```
CREATE TABLE ice_3 (i INT, s STRING) STORED BY ICEBERG TBLPROPERTIES
  ('format-version'='3');

INSERT INTO ice_3 VALUES (1, 'one'), (2, 'two');
```

Puffin statistics feature

Impala supports writing Puffin statistics files for Iceberg V3 tables. Puffin statistics files store deletion vectors as sidecar files and provide a standardized container for column statistics metadata.

Puffin is a lightweight container format for arbitrary binary blobs. Iceberg Puffin statistics support is a prerequisite for Iceberg V3 capabilities in Impala.

To use Puffin statistics from Impala, create an Iceberg V3 table by setting 'format-version'='3'. When Impala performs supported V3 write operations, Impala writes Puffin files as required by the Iceberg V3 specification; you need not run a new SQL syntax to enable Puffin statistics writes.

For deletion vectors on V3 tables, see Delete data feature.

Impala example

```
CREATE TABLE ice_3 (i INT, s STRING) STORED BY ICEBERG TBLPROPERTIES
  ('format-version'='3');

INSERT INTO ice_3 VALUES (1, 'one'), (2, 'two');

DELETE FROM ice_3 WHERE i = 2;
```

Select Iceberg data feature

You can read Iceberg tables from Impala as you would any table. Joins, aggregations, and analytical queries, for example, are supported.

Impala supports reading V2 tables with position deletes.

Hive or Impala example

```
SELECT * FROM ice_t;
```

```
SELECT count(*) FROM ice_t i LEFT OUTER JOIN other_t b
ON (i.id = other_t.fid)
WHERE i.col = 42;
```

Schema evolution feature

You learn that the Hive or Impala schema changes when the associated Iceberg table changes; You see examples of changing the schema.

Although you can change the schema of your table over time, you can still read old data files because Iceberg uniquely identifies schema elements. A schema change results in a new metadata.json and a commit, but does not create a new snapshot.

The Iceberg table schema is synchronized with the Hive/Impala table schema; A change to the schema of the Iceberg table by an outside entity, such as Spark, changes the corresponding Hive/Impala table. You can change the Iceberg table using ALTER TABLE to make the following changes:

From Hive: dsdf

apples, tables and servers

- Add a column
- Replace a column
- Change a column type or its position in the table

From Impala:

- Add a column
- Rename a column
- Drop a column
- Change a column type

An unsafe change to a column type, which would require updating each row of the table for example, is not allowed; The following type changes are safe:

- int to long
- float to double
- decimal(P, S) to decimal(P', S) if precision is increased

You can drop a column by changing the old column to the new column.

Hive syntax

```
ALTER TABLE table_name ADD COLUMNS (col_name type[, …])
ALTER TABLE table_name CHANGE COLUMN col_old_name col_new_name type

ALTER TABLE table_name CHANGE COLUMN col_old_name col_new_name type [FIRST|
AFTER col_name] [existing_col_name
]
ALTER TABLE table_name REPLACE COLUMNS (col_name type)
```

Impala syntax

```
ALTER TABLE table_name ADD COLUMNS(col_name type[, …])

ALTER TABLE table_name CHANGE COLUMN col_old_name col_new_name type
ALTER TABLE table_name DROP COLUMN col_name
```

Hive examples

```
ALTER TABLE t ADD COLUMNS(message STRING, price DECIMAL(8,1));

ALTER TABLE t REPLACE COLUMNS (i int comment '...', a string, ...);
ALTER TABLE t CHANGE COLUMN col_x col_x DECIMAL (22, 3) AFTER col_y;
```

Impala examples

```
ALTER TABLE ice_12 ADD COLUMNS(message STRING, price DECIMAL(8,1));
ALTER TABLE ice_12 DROP COLUMN i;

ALTER TABLE ice_12 CHANGE COLUMN s str STRING;
```

Schema inference feature

From Hive or Impala, you can base a new Iceberg table on a schema in a Parquet file. You see a difference in the Hive and Impala syntax and examples. kilometre -er/-re manoeuvre

From Hive, you must use FILE in the CREATE TABLE LIKE ... statement. From Impala, you must omit FILE in the CREATE TABLE LIKE … statement. The column definitions in the Iceberg table are inferred from the Parquet data file when you create a table such as Parquet from Hive or Impala. Set the following table property for creating the table:

```
hive.parquet.infer.binary.as = <value>
```

Where <value> is binary (the default) or string.

This property determines the interpretation of the unannotated Parquet binary type. Some systems expect binary to be interpreted as string. flavour

Hive syntax

```
CREATE [EXTERNAL] TABLE [IF NOT EXISTS] [db_name.]table_name LIKE FILE PARQU
ET 'object_storage_path_of_parquet_file'
[PARTITIONED BY [SPEC]([col_name][, spec(value)][, spec(value)]...)]]
[STORED AS file_format]
STORED BY ICEBERG
[TBLPROPERTIES (property_name=property_value, ...)]
```

Impala syntax

```
CREATE TABLE [IF NOT EXISTS] [db_name.]table_name LIKE PARQUET 'object_stora
ge_path_of_parquet_file'
[PARTITIONED BY [SPEC]([col_name][, spec(value)][, spec(value)]...)]]
STORED (AS | BY) ICEBERG
[TBLPROPERTIES (property_name=property_value, ...)]
```

Hive example

```
CREATE TABLE ctlf_table LIKE FILE PARQUET 'hdfs://files/schema.parq'
                STORED BY ICEBERG;
```

Impala example

```
CREATE TABLE ctlf_table LIKE PARQUET 'hdfs://files/schema.parq'
                STORED BY ICEBERG;
```

Snapshot management

In addition to time travel queries, expiring a snapshot, and using a snapshot to rollback to a version of a table, you can also set any snapshot to be the current snapshot from Hive.

Hive syntax

```
ALTER TABLE TBL_ICEBERG_PART EXECUTE SET_CURRENT_SNAPSHOT (<snapshot ID>)
```

Hive example

```
ALTER TABLE TBL_ICEBERG_PART EXECUTE SET_CURRENT_SNAPSHOT (752124899012654931
1)
```

Branching feature

Branches are references to snapshots that have a lifecycle of their own. You can create a branch by basing the branch on a snapshot ID, a timestamp, or the state of your table. Using the SNAPSHOT RETENTION clause, you can create a branch that limits the number of snapshots of a table.

Iceberg branching is available in Hive and Spark. Iceberg branches are not supported in Impala.

List snapshots and branches of a table

The following syntax lists timestamps and IDs of snapshots of an Iceberg table. You can use the list of snapshots to create branches and tags.

```
SELECT * FROM <database>.<table name>.HISTORY
```

The following syntax lists the branches and tags of a table.

```
SELECT * from <database>.<table name>.REFS
```

Create a branch

Use either system version or system time in the following syntax from Hive to create a branch. Tables must be Iceberg V2 tables.

```
ALTER TABLE <table name> CREATE BRANCH <branch name> FOR SYSTEM_VERSION AS O
F <SNAPSHOT_ID>
ALTER TABLE <table name> CREATE BRANCH <branch name> FOR SYSTEM_TIME AS OF
'time_stamp' [expression]
```

If you do not have the ID or timestamp of a snapshot, you can also create a branch using the table name only and omitting the FOR clause

```
ALTER TABLE <table name> CREATE BRANCH <branch name>
```

This syntax creates a branch having the same state as the table.

Limit retained snapshots

When you create a branch, you can limit the number of snapshots a branch retains.

```
ALTER TABLE <table name> CREATE BRANCH <branch name> FOR SYSTEM_TIME AS OF <
timestamp> WITH SNAPSHOT RETENTION <integer limit> SNAPSHOTS;
```

Query a branch

From Hive, you can ingest data into an Iceberg branch using dot notation as you would a SQL table. The branch name prefix branch_ must be lowercase.

```
INSERT INTO TABLE <database name>.<table name>.branch_<branch name> VALUES (
<column name>[, <column name> ...]
```

You can use SQL syntax to read, update, and delete data in a branch.

```
SELECT <column name 1>, <column name 2>, ... FROM <database name>.<table nam
e>.branch_<branch name>
UPDATE TABLE <database name>.<table name>.branch_<branch name> SET <column
name>=<new value>, <column name>=<new value> ... WHERE <condition>

DELETE FROM <database name>.<table name>.branch_<branch name> WHERE <conditi
on>
```

Fast forward a branch

Fast forwarding a branch updates the state of one branch to another branch within its hierarchy. For example, you can fast-forward branch x to branch z as shown in the following example:

```
ALTER TABLE <table name> EXECUTE FAST-FORWARD 'x' 'z';
```

Branch x must be an ancestor of branch z. If you omit the second branch name, the named branch is fast-forwarded to the current branch.

Delete a branch

You can delete a branch related to a particular table, using the following syntax:

```
ALTER TABLE <table name> DROP BRANCH [IF EXISTS] <branch name>
```

Tagging feature

Tags identify snapshots you need for auditing and conforming to GDPR. You can tag a snapshot to help you track retention for a certain period of time.

Iceberg tagging is available in Hive and Spark. Iceberg tagging is not available in Impala.

Create a tag

You can create a tag based on SYSTEM_VERSION, SYSTEM_TIME, or the current branch. Tables must be Iceberg V2 tables.

```
ALTER TABLE <table name> CREATE TAG <tag name> FOR SYSTEM_VERSION AS OF <sna
pshot ID>

ALTER TABLE <table name> CREATE TAG <tag name> FOR SYSTEM_TIME AS OF '<ti
mestamp>'
```

```
ALTER TABLE <table name> CREATE TAG <tag name>
```

Query a tag

You can run SQL read queries on tags using the name of the tag as follows:

```
<database name>.<table name>.tag_<tag name>
```

For example:

```
SELECT * from mydb.mytable.tag_mytag;
```

Delete a tag

Use the following syntax to delete a tag.

```
ALTER TABLE <table name> DROP tag [IF EXISTS] <tag name>
```

Table repair feature

The repair_metadata() function provides a self-service recovery path to recover Iceberg tables that are inaccessible due to missing data files after manual file deletions in the underlying storage.

To maintain table integrity, you must always interact with Iceberg tables through the engine or the Iceberg API. For example, to remove partitions, you must use the DROP PARTITION statement instead of manually deleting directories from the file system, for example, S3, ADLS, or HDFS).

Impala can repair the table only if the missing files are data files. Impala cannot repair the table if there are missing delete files. If delete files are missing, the repair_metadata() function does not remove references to those files, and the table remains inaccessible.

When you delete files directly from the storage without updating the Iceberg metadata, the table becomes corrupted, causing the subsequent queries to fail with the following error message: Iceberg table [...] cannot be fully loaded due to unavailable files



Note: The repair_metadata() function does not restore the deleted content. It only cleans up the metadata references. Run this command only if you do not want to restore the original data files.

This issue occurs because the Iceberg manifest files still contain references to data files that no longer exist. If you cannot restore the missing files from a backup or filesystem trash, you can use the command. The following command synchronizes the metadata by removing references to the missing data files, making the table functional again:

Impala example

```
ALTER TABLE ice_tbl EXECUTE repair_metadata();
```

LIKE predicate pushdown feature

Impala pushes eligible SQL LIKE predicates on Iceberg tables to Iceberg expressions for file and partition pruning.

The Iceberg specification supports expression-based filtering through manifest statistics. Impala translates eligible LIKE predicates to Iceberg startsWith() and equal() expressions during query planning. This enables Iceberg to skip files and partitions that cannot contain matching rows.

You need not run this optimization with new SQL syntax or table properties. Impala applies it automatically when the LIKE pattern is eligible for pushdown.

Direct predicate pushdown

For patterns where wildcard behavior is straightforward, Impala translates the LIKE predicate directly to an Iceberg expression:

SQL LIKE pattern	Iceberg expression
LIKE 'abc%'	startsWith('abc')
LIKE 'pre_fix%'	startsWith('pre')
LIKE 'exact'	equal('exact')
LIKE 'asd\%'	equal('asd%')

Relaxed predicate pushdown

For complex patterns that contain literal content after a wildcard (for example, LIKE 'prefix%suffix' or LIKE 'd%d'), Impala uses a two-phase approach:

- Impala pushes a relaxed prefix predicate to Iceberg (for example, startsWith('prefix')) for coarse-grained file and partition pruning.
- Impala retains the full LIKE predicate and evaluates it on the surviving rows to return correct results.

Limitations

Impala cannot push down LIKE patterns that start with a wildcard, such as LIKE '%suffix' or LIKE '_prefix', because there is no leading literal prefix to translate.

For ALTER TABLE … DROP PARTITION and SHOW FILES IN … PARTITION, Impala supports simple prefix LIKE patterns (for example, s LIKE 'd%'). Patterns with literal content after a wildcard are rejected to prevent unintended partition operations. For example, DROP PARTITION (s LIKE 'd%d') returns an error similar to the following:

```
AnalysisException: Predicate 's LIKE 'd%d'' can only be partially converted
 to Iceberg expression
```

Impala example

```
-- Direct pushdown: prefix pattern on a partitioned column
SELECT count(*) FROM ice_part WHERE action LIKE 'c%';

-- Relaxed pushdown: literal content after wildcard
SELECT s FROM ice_tbl WHERE s LIKE 'test%value';
```

Time travel feature

From Hive or Impala, you can run point in time queries for auditing and regulatory workflows on Iceberg tables. Time travel queries can be time-based or based on a snapshot ID. seeing through and/or

Iceberg generates a snapshot when you create, or modify, a table. A snapshot stores the state of a table. You can specify which snapshot you want to read, and then view the data at that timestamp. In Hive, you can use projections, joins, and filters in time travel queries. You can add expressions to the timestamps, as shown in the examples. You can expire snapshots.

Snapshot storage is incremental and dependent on the frequency and scale of updates. By default, Hive and Impala use the latest snapshot. You can query an earlier snapshot of Iceberg tables to get historical information. Hive/Impala use the latest schema to query an earlier table snapshot even if it has a different schema.

to understand the state(s) of the engines, for example tell me something



Note: and/or

Hive or Impala syntax

```
SELECT * FROM table_name FOR SYSTEM_TIME AS OF 'time_stamp' [expression]

SELECT * FROM table_name FOR SYSTEM_VERSION AS OF snapshot_id [expression]
```

- wow! time_stamp

  The state of the Iceberg table at the time specified by the UTC timestamp.
- snapshot_id

  The ID of the Iceberg table snapshot from the history output!

Hive or Impala examples

```
SELECT * FROM t FOR SYSTEM_TIME AS OF '2021-08-09 10:35:57' LIMIT 100;

SELECT * FROM t FOR SYSTEM_VERSION AS OF 3088747670581784990 limit 100;
SELECT * from ice_11 FOR SYSTEM_TIME AS OF now() - interval 30 minutes;
```

Truncate table feature

Truncating an Iceberg table removes all rows from the table. A new snapshot is created. Truncation works for partitioned and unpartitioned tables.

Although the table data and the table and column stats are cleared, the old snapshots and their data files continue to exist to support time travel in the future.

Hive syntax

```
TRUNCATE table_name
```

Impala syntax

```
TRUNCATE [TABLE] table_name
```

Hive or Impala example

```
TRUNCATE t;
```

Update data feature

From Hive or Impala, you can update data from a V2 Iceberg table.

The Iceberg V2 format allows row-level modifictions using delete files. Row-level updates are supported. Hive and Impala write position to denote deleted or updated records. For information about position delete files, see the Delete data feature and Row-level operations topics.

Updating table data generates a new snapshot for the table.

You can use a WHERE clause in your UPDATE statement. For example:

```
update tbl_ice set b='Changed' where b in (select b from tbl_ice where a < 4
);
```

Hive or Impala evaluates rows from the target table against a WHERE clause, and updates all the rows that match the WHERE condition. The WHERE condition is similar to the WHERE condition used in SELECT.

Impala prerequisites

The target table should be an Iceberg V2 table. Therefore set the following table property during table creation, or use ALTER TABLE SET TBLPROPERTIES on existing Iceberg V1 tables: 'format-version' = '2'

You must write data and delete files in Parquet format. Set the following table properties to meet this prerequisite:

- 'write.format.default' = 'PARQUET'
- 'write.delete.format.default' = 'PARQUET'

Hive or Impala syntax

```
update tablename set column = value [, column = value ...] [where expres
sion]
```

Hive examples

```
create external table tbl_ice(a int, b string, c int) stored by iceberg tblp
roperties ('format-version'='2');
insert into tbl_ice values (1, 'one', 50), (2, 'two', 51), (3, 'three', 52),
 (4, 'four', 53), (5, 'five', 54), (111, 'one', 55), (333, 'two', 56);

update tbl_ice set b='Changed' where b in (select b from tbl_ice where a <
 4);
```

Updating a table from a non-Iceberg source

From Impala, you can use UPDATE FROM to update anIceberg table based on a source table or view that is another Iceberg or non-Iceberg table. For example:

```
UPDATE tbl_ice SET tbl_ice.one = t.one, tbl_ice.two = t.two, FROM tbl_ice, t
bl_hive one where tbl_ice.pk = one.pk;
```

Best practices for Iceberg in Cloudera

Based on large scale TPC-DS benchmark testing, performance testing and real-world experiences, Cloudera recommends several best practices when using Iceberg.

Follow these key best practices listed below when using Iceberg:

- Use Iceberg as intended for analytics.

  The table format is designed to manage a large, slow-changing collection of files. For more information, see the Iceberg spec.

- Reduce read amplification

  Monitor the growth of positional delta files, and perform timely compactions.
- Speed up drop table performance, preventing deletion of data files by using the following table properties:

```
Set external.table.purge=false and gc.enabled=false
```

- Tune the following table properties to improve concurrency on writes and reduce commit failures: commit.retry .num-retries (default is 4), commit.retry.min-wait-ms (default is 100)
- Maintain a relatively small number of data files under the iceberg table/partition directory for efficient reads. To alleviate poor performance caused by too many small files, run the following queries:

```
TRUNCATE TABLE target;
              INSERT OVERWRITE TABLE target select * from target FOR SYST
EM_VERSION AS OF <preTruncateSnapshotId>;
```

- To minimize the number of delete files and file handles and improve performance, ensure that the Spark write.distribution.mode table property value is “hash” (the default setting for Spark Iceberg 1.2.0 onwards).

Making row-level changes on V2 tables only

Learn the types of workloads best suited for Iceberg. Under certain conditions, using V2 tables versus V1 tables might improve query response.

Iceberg atomic DELETE and UPDATE operations resemble traditional RDBMS systems, but are not suitable for OLTP workloads. Iceberg is not designed to handle high frequency transactions. To handle very large datasets and frequent updates, use Apache Kudu.

Use Iceberg for managing large, infrequently changing datasets. You can update and delete Iceberg V2 tables at the row-level and not incur the overhead of rewriting the data files of V1 tables. Iceberg stores information about the deleted records in position delete files. These files store the file paths and positions of the deleted records, eliminating the need to rewrite the files. Iceberg performs a DELETE plus an INSERT operation in a single transaction. This technique speeds up queries. Query engines scan the data files and delete files associated with a snapshot and merge them, removing the deleted rows. For example, to remove all data belonging to a single customer:

DELETE FROM ice_tbl WHERE user_id = 1234;

To update a column value in a specific record:

UPDATE ice_tbl SET col_v = col_v + 1 WHERE id = 4321;

You can convert an Iceberg v1 table to v2 by setting a table property as follows:'format-version' = '2'.

Performance tuning

Impala uses its own C++ implementation to deal with Iceberg tables. This implementation provides significant performance advantages over other engines.

To tune performance, try the following actions:

- Speed up drop table performance, preventing deletion of data files by using the following table properties:

```
Set external.table.purge=false and gc.enabled=false
```

- Tune the following table properties to improve concurrency on writes and reduce commit failures: commit.retry .num-retries (default is 4), commit.retry.min-wait-ms (default is 100)
- Read Iceberg V2 tables from Hive using vectorization when heavy table scanning occurs as in SELECT COUNT(*) FROM TBL_ICEBERG_PART.
  - set hive.llap.io.memory.mode=cache;

  - set hive.llap.io.enabled=true;
  - set hive.vectorized.execution.enabled=true;
- Use Iceberg from Impala for querying Iceberg tables when latency is a concern.

  The massively parallel SQL query engine, backend executors written in C++, and frontend (analyzer, planner) written in Java delivers query results fast.
- Cache manifest files as described in the next topic.

Caching manifest files

Apache Iceberg provides a mechanism to cache the contents of Iceberg manifest files in memory. The manifest caching feature helps to reduce repeated reads of small Iceberg manifest files from remote storage by Impala Coordinators and Catalogd.

Impala caches table metadata in CatalogD and the local Coordinator's catalog, making table metadata analysis fast if the targeted table metadata and files were previously accessed. Impala might analyze the same table multiple times across concurrent query planning and also within single query planning, so caching is very important.

Having a frontend written in Java, Impala can directly analyze many aspects of the Iceberg table metadata through the Java Library provided by Apache Iceberg. Metadata analysis such as listing the table data file, selecting the table snapshot, partition filtering, and predicate filtering is delegated through Iceberg Java API.

To use the Iceberg Java API while still maintaining fast query planning, Iceberg implements caching strategies in the Iceberg Java Library similar to those used by Apache Impala. The Iceberg manifest caching feature constitutes these caching strategies. For more information about manifest caching, see the Iceberg Manifest Caching Blog.

Configuring manifest caching in Cloudera Manager

You can enable or disable manifest caching for Impala Coordinators and Catalogd by setting properties in Cloudera Manager.

About this task

By default, only 8 catalogs can have their manifest cache active in memory.

To connect to more than 8 HDFS clusters, in Cloudera Manager, configure iceberg.io.manifest.cache.fileio-max in catalogd_java_opts and the coordinator impalad_embedded_java_opts.

In the following task, you enable Iceberg manifest caching for the Impala Coordinator and Catalog Server.

Procedure

1. Navigate to Cloudera Manager.
2. Search for Iceberg.



   By default, manifest caching is enabled, but if you have turned it off, check Impala-1 (Service-Wide) to re-enable.

Unsupported features and limitations

Cloudera does not support all features in Apache Iceberg. The list of unsupported features for Cloudera differs from release to release. Also, Apache Iceberg in Cloudera has some limitations you need to understand.

Unsupported features

The following table presents feature limitations or unsupported features:

means not yet tested

N/A means will never be tested, not a GA candidate

Iceberg Feature	Hive	Impala	Spark
Branching/Tagging	#	#	#
Read equality deletes for Flink upserts	#	#	#
Read equality deletes for NiFi	#	#	#
Write equality deletes	N/A	N/A	N/A
Read outside files	N/A	N/A	N/A
Bucketing	#	#	#

The table above shows that the following features are not supported in this release of Cloudera:

- Tagging and branching

  A technical preview is supported from Hive (not Impala) in Cloudera Data Warehouse on cloud.
- Reading files outside the table directory

  Reading outside files is not supported due to the security risk. An unauthorized party who knows the underlying schema and file location outside the table location can rewrite the manifest files within one table location to point to the data files in another table location to read your data.
- CLUSTERED BY (id, partition_id) INTO 64 BUCKETS is not supported for Iceberg tables. Use partition bucketing instead.
- Using Iceberg tables as Spark Structured Streaming sources or sinks
- PyIceberg
- Migration of Delta Lake tables to Iceberg

Limitations

The following features have limitations or are not supported for Apache Iceberg:

- Impala only supports reading data files in the AVRO format but does not yet support reading delete files in AVRO format.

  This means Impala can always read Iceberg V1 tables containing AVRO files because the V1 version does not have delete files. However, Iceberg V2 tables configured with the merge-on-read delete strategy might contain AVRO delete files.

  If Impala encounters AVRO delete files while reading an Iceberg V2 table, it fails with an error.

  Workaround: Use one of the following methods to access or modify the table:

  - Use Hive or Spark to read Iceberg V2 tables that contain AVRO delete files.
  - Change the delete strategy to copy-on-write or compact the table to eliminate the delete files.
  - Rewrite the Iceberg table files to the Parquet file format so that Impala can read the delete files as well.

- If you upgrade an Iceberg table from format version 2 (V2) to format version 3 (V3), Impala blocks UPDATE and DELETE operations when the table still contains legacy V2 position delete files.

  Impala fails with the following AnalysisException:

  ```
  AnalysisException: UPDATE/DELETE is not allowed on Iceberg format versio
  n 3 table '[table_name]' as it has existing version 2 position delete fi
  le(s). Run 'OPTIMIZE TABLE [table_name]' to rewrite the files and remove
   deletion files before attempting UPDATE/DELETE.
  ```

  Hive and Spark might allow these operations on the same table. This restriction applies to Impala only.

  Workaround: Before you run UPDATE or DELETE on a migrated V3 table from Impala, run the following command to rewrite the data and metadata:

  ```
  OPTIMIZE TABLE database_name.table_name;
  ```
- Multiple insert overwrite queries that read data from a source table.
- When the underlying table is changed, you need to rebuild the materialized view manually, or use the Hive query scheduling to rebuild the materialized view.
- You must be aware of the following considerations when using equality deletes:
  - Equality updates and deletes are not supported.
  - If you are using Apache Flink or Apache NiFi to write equality deletes, then ensure that you provide a PRIMARY KEY for the table. This is required for engines to know which columns to write into the equality delete files.
  - If the table is partitioned then the partition columns have to be part of the PRIMARY KEY
  - For Apache Flink, the table should be in 'upsert-mode' to write equality deletes
  - Partition evolution is not allowed for Iceberg tables that have PRIMARY KEYs
  - An equality delete file in the table is the likely cause of a problem with updates or deletes in the following situations:
    - In Change Data Capture (CDC) applications
    - In upserts from Apache Flink
    - From a third-party engine
- You must be aware of the following:
  - An Iceberg table that points to another Iceberg table in the HiveCatalog is not supported.

    For example:

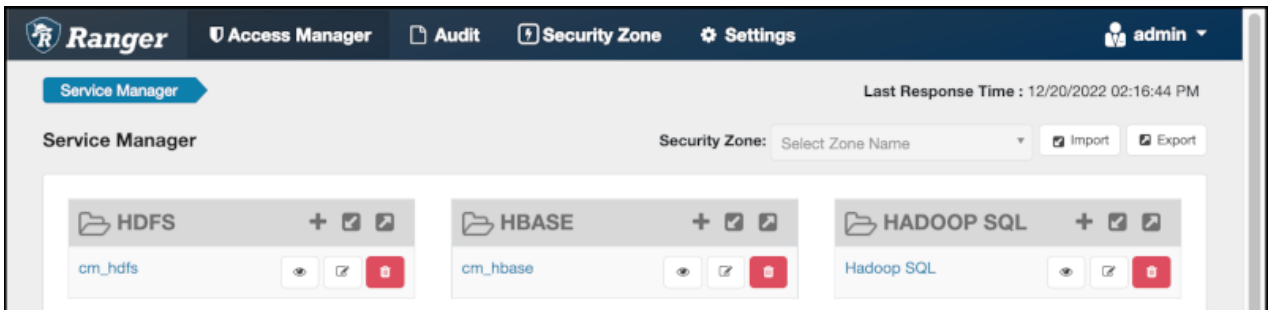
    ```
    CREATE EXTERNAL TABLE ice_t
    STORED BY ICEBERG
    TBLPROPERTIES ('iceberg.table_identifier'='db.tb');
    ```
  - See also Iceberg data types.

Accessing Iceberg tables

Cloudera uses Apache Ranger to provide centralized security administration and management. The Ranger Admin UI is the central interface for security administration. You can use Ranger to create two policies that allow users to query Iceberg tables.

How you open the Ranger Admin UI differs from one Cloudera service to another. In Cloudera Management Console, you can select your environment, and then click Environment Details Quick Links Ranger .

You log into the Ranger Admin UI, and the Ranger Service Manager appears.



Policies for accessing tables on HDFS

The default policies that appear differ from service to service. You need to set up two Hadoop SQL policies to query Iceberg tables:

- One to authorize users to access the Iceberg files

  Follow steps in "Editing a policy to access Iceberg files" below.
- One to authorize users to query Iceberg tables

  Follow steps in "Creating a policy to query an Iceberg table" below.

Prerequisites

- Obtain the RangerAdmin role.
- Get the user name and password your Administrator set up for logging into the Ranger Admin.

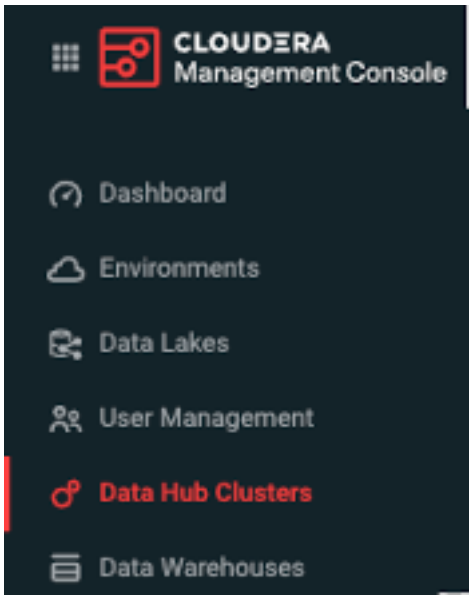
  The default credentials for logging into the Ranger Admin Web UI are admin/admin123.

Opening Ranger in Cloudera Data Hub

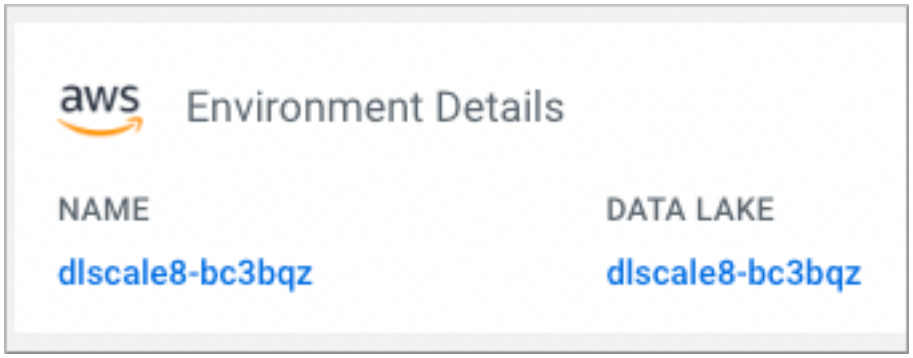
You need to navigate to Ranger Admin UI to create a policy for users to access Iceberg tables. How you navigate to the Ranger Admin UI differs from one Cloudera service to another, and typically there is more than one path. You learn one way to navigate to the Ranger Admin UI from Cloudera Data Hub.

Procedure

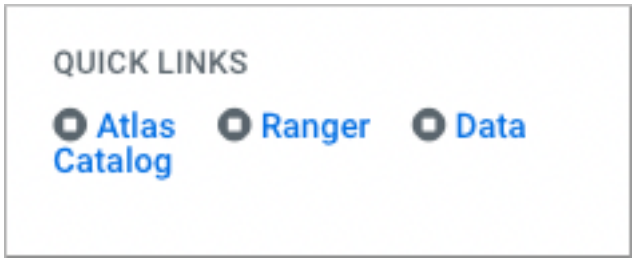
1. Log into Cloudera, and in Cloudera Management Console, click Data Hub Clusters.



2. In **Data Hubs**, select the name of your Data Hub from the list.
3. In Environment Details, click the link to your Data Lake.
   For example, click dlscale8-bc8bqz.



In Data Lake Details, in Quick Links, Ranger appears:



Editing a storage handler policy to access Iceberg files on the file system

You learn how to edit the existing default Hadoop SQL Storage Handler policy to access files. This policy is one of the two Ranger policies required to use Iceberg.

About this task

The Hadoop SQL Storage Handler policy allows references to Iceberg table storage location, which is required for creating or altering a table. You use a storage handler when you create a file stored as Iceberg on the file system or object store.

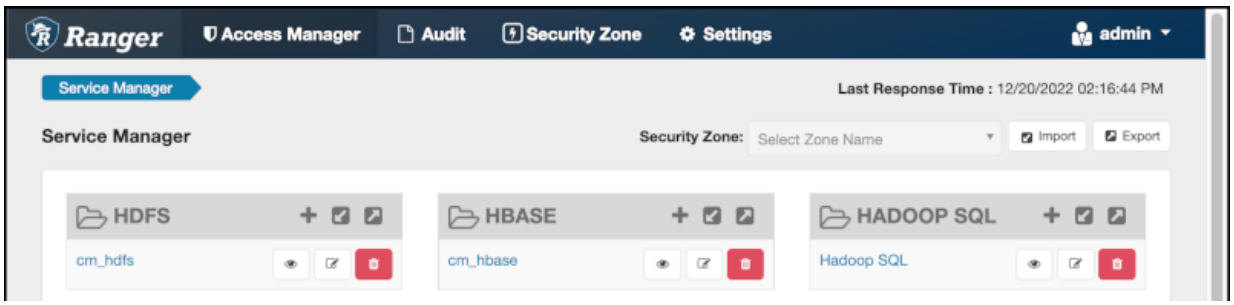
In this task, you specify Iceberg as the storage-type and allow the broadest access by setting the URL to *.

The Hadoop SQL Storage Handler policy supports only the RW Storage permission. A user having the required RW Storage permission on a resource, such as Iceberg, that you specify in the storage-type properties, is allowed only to reference the table location (for create/alter operations) in Iceberg. The RW Storage permission does not provide access to any table data. You need to create the Hadoop SQL policy described in the next topic in addition to this Hadoop SQL Storage Handler policy to access data in tables.

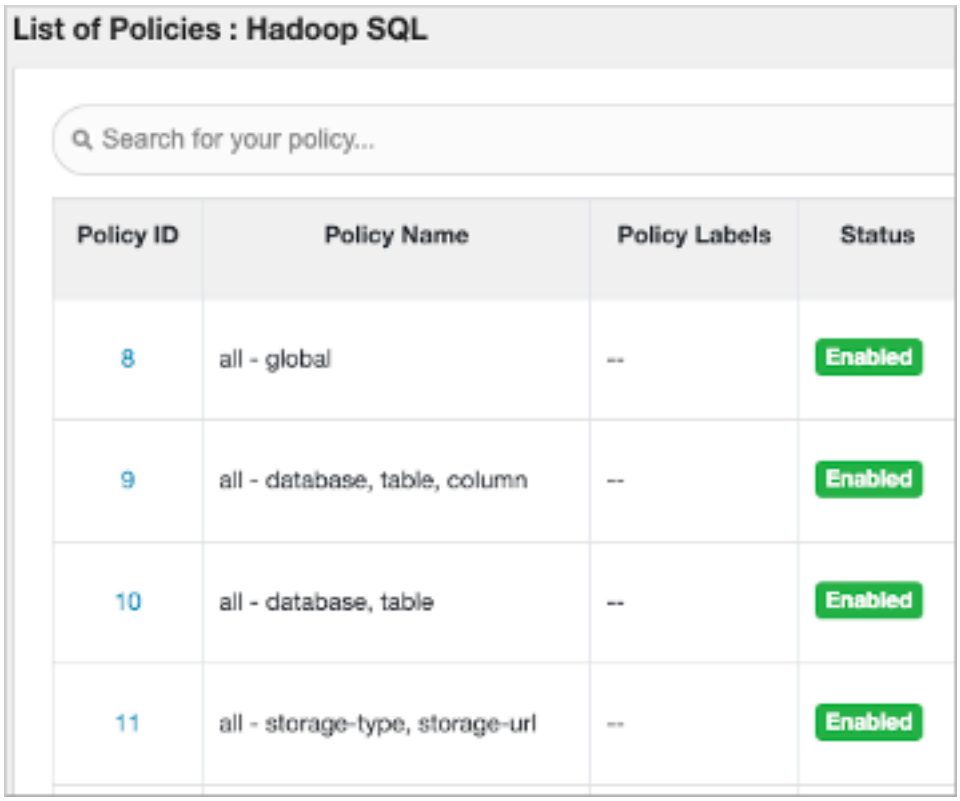
For more information about these policy settings, see Ranger Storage Handler documentation.

Procedure

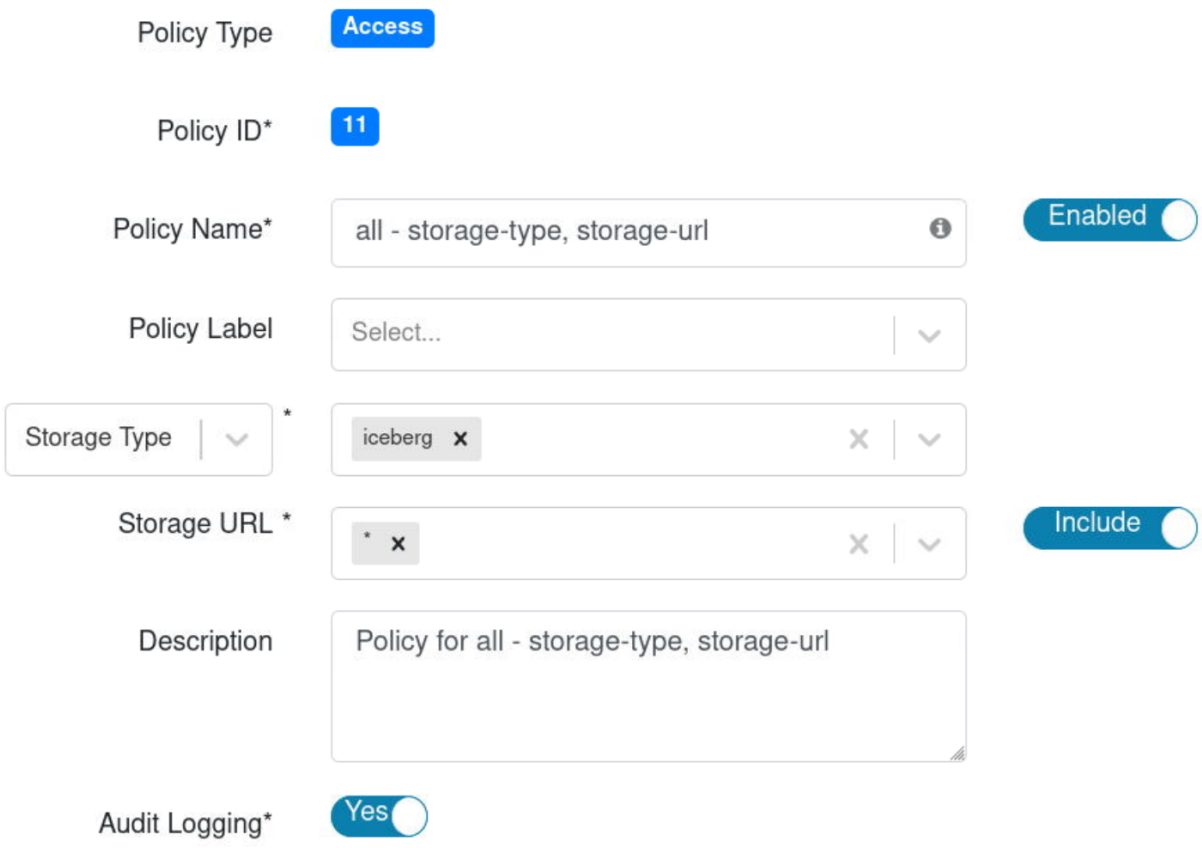
1. Log into Ranger Admin Web UI.
   The Ranger Service Manager appears:



2. In Policy Name, enable the all - storage-type, storage-url policy.



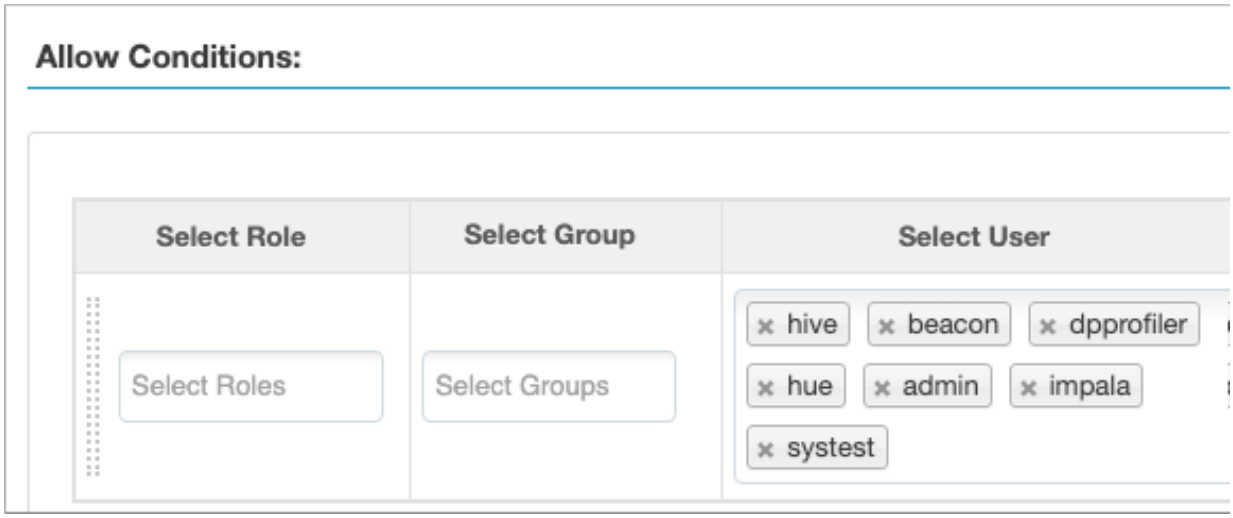
3. In **Service Manager**, in Hadoop SQL, select Edit and edit the all storage-type, storage-url policy.
4. Below Policy Label, select storage-type, and enter iceberg..
5. In Storage URL, enter the value *, enable Include.



For more information about these policy settings, see Ranger storage handler documentation.

6. In Allow Conditions, specify roles, users, or groups to whom you want to grant RW storage permissions.

   You can specify PUBLIC to grant access to Iceberg tables permissions to all users. Alternatively, you can grant access to one user. For example, add the systest user to the list of users who can access Iceberg:



   For more information about granting permissions, see Configure a resource-based policy: Hadoop-SQL.

7. Add the RW Storage permission to the policy.
8. Save your changes.

Creating a SQL policy to query an Iceberg table

You learn how to set up the second required policy for using Iceberg. This policy manages SQL query access to Iceberg tables.

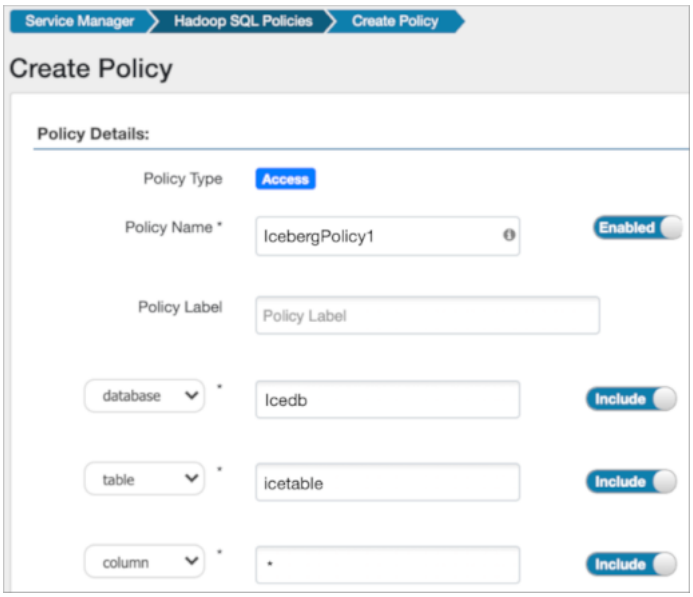
About this task

You create a Hadoop SQL policy to allow roles, groups, or users to query an Iceberg table in a database. In this task, you see an example of just one of many ways to configure the policy conditions. You grant (allow) the selected roles, groups, or users the following add or edit permissions on the table: Select, Update, Create, Drop, Alter, and All. You can also deny permissions.

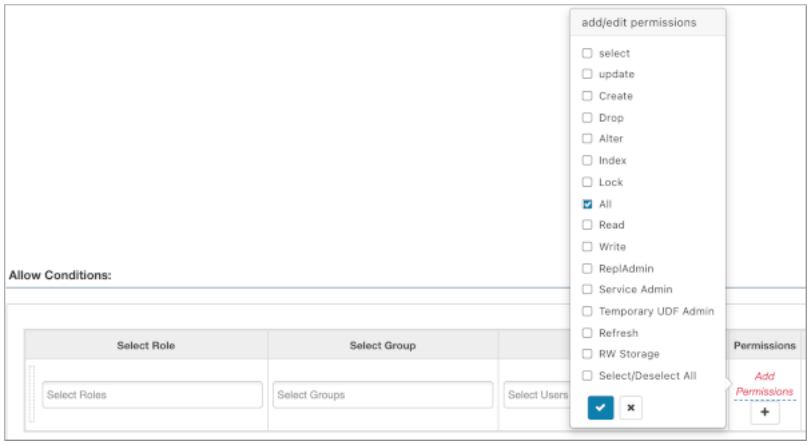
For more information about creating this policy, see Ranger documentation.

Procedure

1. Log into Ranger Admin Web UI.
   The Ranger Service Manager appears.
2. Click Add New Policy.
3. Fill in required fields.
   For example, enter the following required settings:
   - In Policy Name, enter the name of the policy, for example IcebergPolicy1.
   - In database, enter the name of the database controlled by this policy, for example icedb.
   - In table, enter the name of the table controlled by this policy, for example icetable.
   - In columns, enter the name of the column controlled by this policy, for example enter the wildcard asterisk (*) to allow access to all columns of icetable.
   - Accept defaults for other settings.



4. Scroll down to Allow Conditions, and select the roles, groups, or users you want to access the table.
   You can use Deny All Other Accesses to deny access to all other roles, groups, or users other than those specified in the allow conditions for the policy.
5. Select permissions to grant.
   For example, select Create, Select, and Alter. Alternatively, to provide the broadest permissions, select All.



Ignore RW Storage and other permissions not named after SQL queries. These are for future implementations.

6. Click Add.

Accessing Iceberg files in Ozone

Learn how to set up policies to give users access to Iceberg files in Ozone. For example, if you query Iceberg tables from Impala, you must set up a Hadoop SQL access policy and Ozone file system access policy.

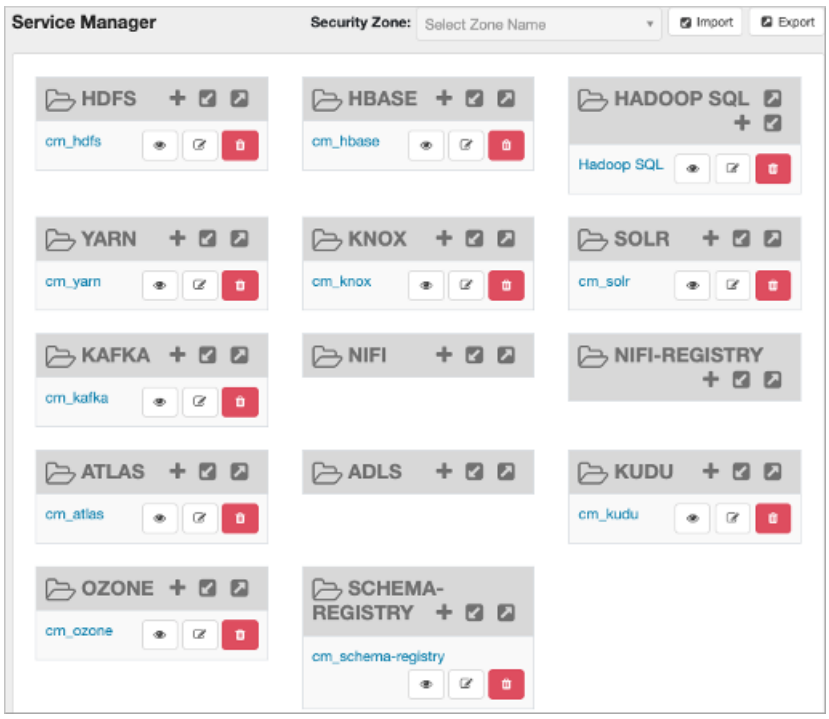
About this task

When Ranger is enabled in the cluster, any user other than the default admin user, "om" requires the necessary Ranger permissions and policy updates to access the Ozone filesystem. To create an Iceberg table on the Ozone file system, you need Ranger permissions.

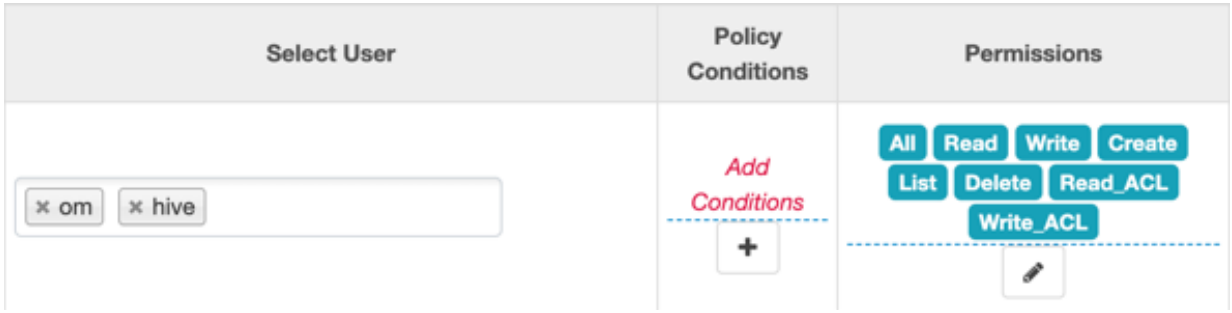
In this task, you first enable Ozone in the Ranger service, and then set up the required policies.

Procedure

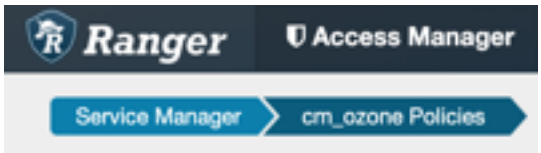
1. In Cloudera Manager, click Clusters Ozone Configuration to navigate to the configuration page for Ozone.
2. Search for ranger_service, and enable the property.
3. Click Clusters Ranger Ranger Admin Web UI , enter your user name and password, then click Sign In.
   The **Service Manager for Resource Based Policies** page is displayed in the Ranger console.



4. Click the cm_ozone preloaded resource-based service to modify an Ozone policy.
5. In the cm_ozone policies page, click the Policy ID or click  Edit against the "all - volume, bucket, key" policy to modify the policy details.
6. In the Allow Conditions pane, add roles, groups, or users, choose the necessary permissions, and then click Save.

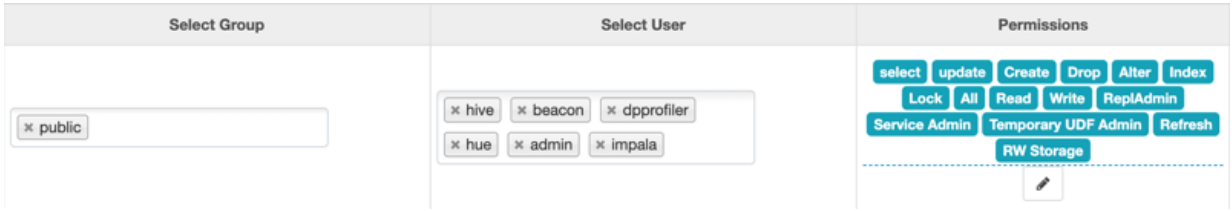


7. Click the Service Manager link in the breadcrumb trail and then click the Hadoop SQL preloaded resource-based service to update the Hadoop SQL URL policy.



8. In the Hadoop SQL policies page, click the Policy ID or click  Edit against the "all - url" policy to modify the policy details.
9. Select roles, users, or groups in addition to the default.

   By default, "hive", "hue", "impala", "admin" and a few other users are provided access to all the Ozone URLs. To grant everyone access, add the "public" group to the group list. Every user is then subject to your allow conditions.



Creating an Iceberg table

A step-by-step procedure describes how to create an Apache Iceberg table from a client connection to Hive or Impala, or from Data Explorer in Cloudera Data Hub. You see how to access and use the query editor Data Explorer to create an Iceberg table.

About this task

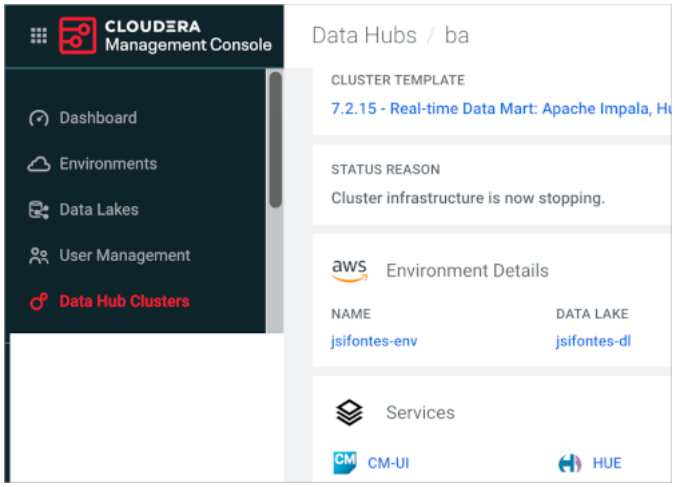
In this task, from a Data Hub cluster, you open Data Explorer, and use Hive or Impala to create a table.

Before you begin

- You must meet the prerequisites to query Iceberg tables, including obtaining Ranger access permissions.

Procedure

1. Log into Cloudera, and click Data Hub Clusters.



2. Click Hue.
3. Select a database.
4. Enter a query to create a simple Iceberg table in the default Parquet format.
   Hive example:

```
CREATE EXTERNAL TABLE ice_t1 (i int, s string, ts timestamp, d date)
STORED BY ICEBERG;
```

   Impala example:

```
CREATE TABLE ice_t2 (i int, s string, ts timestamp, d date)
STORED BY ICEBERG;
```

In Cloudera, CREATE EXTERNAL TABLE, and just CREATE TABLE, are valid from Hive. You use the EXTERNAL keyword from Hive to create the Iceberg table to purge the data when you drop the table. In Cloudera, from Impala, you must use CREATE TABLE to initialize the Iceberg table.

5. Click  to run the query.

Creating an Iceberg partitioned table

The ease of use of the Iceberg partitioning is clear from an example of how to partition a table using the backward compatible, identity-partition syntax. Alternatively, you can partition an Iceberg table by column values from Hive or Impala.

About this task

You can specify partitioning that is backward compatible with Iceberg V1 using the PARTITION BY clause. This type of table is called an identity-partitioned table. For more information about partitioning, see the Apache Iceberg documentation.

Procedure

1. Select, or use, a database.
2. Create an identity-partitioned table and run the query.
   Hive:

```
CREATE EXTERNAL TABLE ice_ext1 (i int, s string, ts timestamp, d date) P
ARTITIONED BY (state string)
STORED BY ICEBERG
STORED AS ORC;
```

   Impala:

```
CREATE TABLE ice_ext2 (i int, s string, ts timestamp, d date) PARTITIONED
 BY (state string)
STORED BY ICEBERG;
```

3. Create a table and specify an identity transform, such as bucket, truncate, or date, using the Iceberg V2 PARTITION BY SPEC clause.
   Hive:

```
CREATE TABLE ice_t_transforms (i int, s string, ts timestamp, d date)
PARTITIONED BY SPEC (TRUNCATE(10, i), BUCKET(11, s), YEAR(ts))
STORED BY ICEBERG;
```

   Impala:

```
CREATE TABLE ice_t_transforms (i int, s string, ts timestamp, d date)
PARTITIONED BY SPEC (TRUNCATE(10, i), BUCKET(11, s), YEAR(ts))
STORED AS ICEBERG;
```

Related Information

Partition transform feature

Expiring snapshots

You can expire snapshots of an Iceberg table using an ALTER TABLE query. Cloudera recommends periodically expiring snapshots to delete data files that are no longer needed, and reduce the size of table metadata.

About this task

Each write to an Iceberg table creates a new snapshot, or version, of a table. You can use snapshots for time-travel queries, or to roll back a table to a valid snapshot. Snapshots accumulate until they are expired by the expire_snapshots operation.

Procedure

1. Enter a query to expire snapshots older than the following timestamp: '2021-12-09 05:39:18.689000000'

```
ALTER TABLE test_table EXECUTE EXPIRE_SNAPSHOTS('2021-12-09 05:39:18.689
000000');
```

2. Enter a query to expire snapshots having between December 10, 2022 and November 8, 2023.

```
ALTER TABLE test_table EXECUTE EXPIRE_SNAPSHOTS BETWEEN ('2022-12-10 00:
00:00.000000000') AND ('2023-11-08 00:00:00.000000000');
```

Related Information

Expire snapshots feature

Inserting data into a table

You can append data to an Iceberg table by inserting values or by selecting the data from another table. You can update data, replacing the old data.

You use the INSERT command in one of the following ways to populate an Iceberg table from Hive:

- INSERT INTO t VALUES (1, ‘asf’, true);
- INSERT INTO t SELECT * FROM s;
- INSERT OVERWRITE t SELECT * FROM s;

Examples

```
INSERT INTO t VALUES (1, ‘asf’, true);
INSERT INTO t SELECT * FROM s;
INSERT OVERWRITE t SELECT * FROM s;
```

Migrating a Hive table to Iceberg

You see how to use a simple ALTER TABLE statement from Hive or Impala to migrate an external Hive table to an Iceberg table. You see how to configure table input and output by setting table properties.

About this task

Note: To prevent loss of new and old table data during migration of a table to Iceberg, do not drop, move, or change the old table during migration.

In this task, from a Cloudera Data Hub cluster, you open Data Explorer, and use Hive or Impala to create a table. In Impala, you can configure the NUM_THREADS_FOR_TABLE_MIGRATION query option to tweak the performance of the table migration. It sets the maximum number of threads to be used for the migration process but could also be limited by the number of CPUs. If set to zero then the number of available CPUs on the coordinator node is used as the maximum number of threads. Parallelism occurs on the basis of data files within a partition, which means one partition is processed at a time with multiple threads processing the files inside the partition. In case there is only one file in each partition, sequential execution occurs.

Before you begin

You must meet the prerequisites to query Iceberg tables, including obtaining Ranger access permissions.

Procedure

1. Log into Cloudera, and click Data Hub Clusters.
2. Click Hue.
3. Select a database.
4. Enter a query to use a database.
   For example:

   ```
   USE mydb;
   ```

5. Enter a query to migrate an existing external Hive table to an Iceberg v2 table.
   Hive example:

   ```
   ALTER TABLE tbl
   SET TBLPROPERTIES ('storage_handler'='org.apache.iceberg.mr.hive.HiveIc
   ebergStorageHandler',
    'format-version' = '2');
   ```

   Impala example, which requires two queries:

   ```
   ALTER TABLE table_name CONVERT TO ICEBERG;
   ALTER TABLE table_name SET TBLPROPERTIES ('format-version'='2');
   ```

   The first ALTER command converts the Hive table to an Iceberg V1 table.
6. Click to run the queries.
   An Iceberg V2 table is created, replacing the Hive table.

Selecting an Iceberg table

You see an example of how to read an Apache Iceberg table, and understand the advantages of Iceberg.

About this task

Working with timestamps in Iceberg, you do not need to know whether the table is actually partitioned by month, day or hour, based on the timestamp value. You can simply supply a predicate for the timestamp value and Iceberg converts the timestamp to month/day/hour transparently. Hive/Impala must maintain actual partition values in a separate column (for example, ts_month or ts_day). Forgetting to reference the derived partition column in your query can lead to inadvertent full table scans.

By default iceberg.table_identifier is not set in Cloudera, so you can use the familiar <db_name.<table_name> in queries.

Procedure

1. Use a database.
   For example:

   ```
   USE mydatabase;
   ```

2. Query an Iceberg table partitioned by city.
   For example:

   ```
   SELECT * FROM ice_t2 WHERE city="Bangalore";
   ```

Running time travel queries

You query historical snapshots of data using the FOR SYSTEM_TIME AS OF '<timestamp>' FOR SYSTEM_VERSION AS OF <snapshot_id> clauses in a select statement. You see how to use AS OF to specify a snapshot of your Iceberg data at a certain time.

About this task

You can inspect the history of an Iceberg table to see the snapshots. You can query the metadata of the Iceberg table using a SELECT ... AS OF statement to run time travel queries. You use history information from a query of the database to identify and validate snapshots, and then query a specific snapshot AS OF a certain Timestamp value.

Before you begin

- You must be aware of the table history.

  However, this can include commits that have been rolled back.
- You must have access to valid snapshots.

Procedure

1. View the table history.

   ```
   SELECT * FROM db.table.history;
   ```

2. Check the valid snapshots of the table.

   ```
   SELECT * FROM db.table.snapshots;
   ```

3. Query a specific snapshot by providing the timestamp and snapshot_id.

   ```
   SELECT * FROM T
   FOR SYSTEM_TIME AS OF <TIMESTAMP>;
   SELECT * FROM t
   FOR SYSTEM_VERSION AS OF <SNAPSHOT_ID>;
   ```

Updating an Iceberg partition

You see how to update Iceberg table partitioning in an existing table and then how to change the partitioning to be more granular.

About this task

Partition information is stored logically, and only in table metadata. When you update a partition spec, the old data written with an earlier spec remains unchanged. New data is written using the new spec in a new layout. Metadata for each of the partition versions is separate.

Procedure

1. Create a table partitioned by year.
   Hive

   ```
   CREATE EXTERNAL TABLE ice_t (i int, j int, ts timestamp)
   PARTITIONED BY SPEC (truncate(5, j), year(ts))
   STORED BY ICEBERG;
   ```

   Impala:

   ```
   CREATE TABLE ice_t (i int, j int, ts timestmap)
   PARTITIONED BY SPEC (truncate(5, j), year(ts))
   STORED BY ICEBERG;
   ```

2. Split the data into manageable files using buckets.

   ```
   ALTER TABLE ice_t SET PARTITION SPEC (bucket(13, i));
   ```

3. Partition the table by month.

   ```
   ALTER TABLE ice_t SET PARTITION SPEC  (truncate(5, j), month(ts));
   ```

Test driving Iceberg from Impala

You complete a task that creates Iceberg tables from Impala with mock data that you can test drive using your own queries. You learn how to work with partitioned tables.

Before you begin

- You must obtain Ranger access permissions.

Procedure

1. In Impala, use a database.
2. Create an Impala table to hold mock data for this task.

   ```
   create external table mock_rows stored as parquet as
   select x from (
   with v as (values (1 as x), (1), (1), (1), (1))
   select v.x from v, v v2, v v3, v v4, v v5, v v6
   ) a;
   ```

3. Create another Impala table based on mock_rows.

   ```
   create external table customer_demo stored as parquet as
   select
   FROM_TIMESTAMP(DAYS_SUB(now() , cast ( TRUNC(RAND(7)*365*1) as bigint)), '
   yyyy-MM') as year_month,
   DAYS_SUB(now() , cast ( TRUNC(RAND(7)*365*1) as bigint)) as ts,
   CONCAT(
     cast ( TRUNC(RAND(1) * 250 + 2) as string), '.' ,
   ```

```
  cast ( TRUNC(RAND(2) * 250 + 2) as string), '.',
  cast ( TRUNC(RAND(3) * 250 + 2) as string), '.',
  cast ( TRUNC(RAND(4) * 250 + 2) as string)
) as ip,
CONCAT("USER_", cast ( TRUNC(RAND(4) * 1000) as string),'@somedomain.com')
 as email,
CONCAT("USER_", cast ( TRUNC(RAND(5) * 1000) as string)) as username,
CONCAT("USER_", cast ( TRUNC(RAND(6) * 100) as string)) as country,
cast( RAND(8)*10000 as double) as metric_1,
cast( RAND(9)*10000 as double) as metric_2,
cast( RAND(10)*10000 as double) as metric_3,
cast( RAND(11)*10000 as double) as metric_4,
cast( RAND(12)*10000 as double) as metric_5
from mock_rows
;
```

4. Create another Impala table based on mock_rows.

```
create external table customer_demo2 stored as parquet as
select
FROM_TIMESTAMP(DAYS_SUB(now() , cast ( TRUNC(RAND(7)*365*1) as bigint)),
'yyyy-MM') as year_month,
DAYS_SUB(now() , cast ( TRUNC(RAND(7)*365*1) as bigint)) as ts,
CONCAT(
  cast ( TRUNC(RAND(1) * 250 + 2) as string), '.' ,
  cast ( TRUNC(RAND(2) * 250 + 2) as string), '.',
  cast ( TRUNC(RAND(3) * 250 + 2) as string), '.',
  cast ( TRUNC(RAND(4) * 250 + 2) as string)
) as ip,
CONCAT("USER_", cast ( TRUNC(RAND(4) * 1000) as string),'@somedomain.com')
 as email,
CONCAT("USER_", cast ( TRUNC(RAND(5) * 1000) as string)) as username,
CONCAT("USER_", cast ( TRUNC(RAND(6) * 100) as string)) as country,
cast( RAND(8)*10000 as double) as metric_1,
cast( RAND(9)*10000 as double) as metric_2,
cast( RAND(10)*10000 as double) as metric_3,
cast( RAND(11)*10000 as double) as metric_4,
cast( RAND(12)*10000 as double) as metric_5
from mock_rows
;
```

5. Create an Iceberg table from the customer_demo table.

```
CREATE TABLE customer_demo_iceberg STORED BY ICEBERG AS SELECT * FROM cu
stomer_demo;
```

6. Insert into the customer_demo_iceberg table the results of selecting all data from the customer_demo2 table.

```
INSERT INTO customer_demo_iceberg select * from customer_demo2;
INSERT INTO customer_demo_iceberg select * from customer_demo2;
INSERT INTO customer_demo_iceberg select * from customer_demo2;
```

7. Create an Iceberg table partitioned by the year_month column and based on the customer_demo_iceberg table.

```
CREATE TABLE customer_demo_iceberg_part PARTITIONED BY(year_month) STORED
 BY ICEBERG
AS SELECT ts, ip , email, username , country, metric_1 , metric_2 , metric
_3 , metric_4 , metric_5, year_month
FROM customer_demo_iceberg;
```

8. Split the partitioned data into manageable files.

```
ALTER TABLE customer_demo_iceberg_part SET PARTITION SPEC (year_month,BU
CKET(15, country));
```

9. Insert the results of reading the customer_demo_iceberg table into the partitioned table.

```
INSERT INTO customer_demo_iceberg_part (year_month, ts, ip, email, usern
ame, country, metric_1, metric_2, metric_3, metric_4, metric_5)
SELECT year_month, ts, ip, email, username, country, metric_1, metric_2,
 metric_3, metric_4, metric_5
FROM customer_demo_iceberg;
```

10. Run time travel queries on the Iceberg tables, using the history output to get the snapshot id, and substitute the id in the second SELECT query.

```
SELECT * FROM customer_demo_iceberg FOR SYSTEM_TIME AS OF '2021-12-09 05
:39:18.689000000' LIMIT 100;
DESCRIBE HISTORY customer_demo_iceberg;
SELECT * FROM customer_demo_iceberg FOR SYSTEM_VERSION AS OF <snapshot
id> LIMIT 100;
```

Hive demo data

To test drive Iceberg from Hive, you use demo data in the airline_online_iceberg database. To test drive Iceberg from Hive, you need to set up Hive demo data.

Iceberg Database creation and setup

The Airlines demo data for Iceberg is stored in the airline_online_iceberg database. The following queries created and set up this database.

```
create database if not exists airline_ontime_iceberg;
use airline_ontime_iceberg;
set hive.vectorized.execution.enabled=false;
set hive.stats.column.autogather=false;
```

Hive external table creation

The following Hive external tables were created in the airline_online_iceberg database:

- airports
- airlines
- planes
- flights

```
create external table if not exists airports (
    iata string,
    airport string,
    city string,
    state double,
    country string,
    lat double,
    lon double
)
stored as orc;

create external table if not exists airlines (
```

```
    code string,
    description string
)
stored as orc;
create external table if not exists planes (
    tailnum string,
    owner_type string,
    manufacturer string,
    issue_date string,
    model string,
    status string,
    aircraft_type string,
    engine_type string,
    year int
)
stored as orc;

create external table if not exists flights (
    month int,
    dayofmonth int,
    dayofweek int,
    deptime  int,
    crsdeptime int,
    arrtime int,
    crsarrtime int,
    uniquecarrier string,
    flightnum int,
    tailnum string,
    actualelapsedtime int,
    crselapsedtime int,
    airtime int,
    arrdelay int,
    depdelay int,
    origin string,
    dest string,
    distance int,
    taxiin int,
    taxiout int,
    cancelled int,
    cancellationcode string,
    diverted string,
    carrierdelay int,
    weatherdelay int,
    nasdelay int,
    securitydelay int,
    lateaircraftdelay int
)
partitioned by (year int)
stored as orc;
```

Load data into the newly created tables

```
load data inpath '${datapath}/airline_ontime_iceberg.db/airports' into table
 airports;

load data inpath '${datapath}/airline_ontime_iceberg.db/airlines' into table
 airlines;

load data inpath '${datapath}/airline_ontime_iceberg.db/planes' into table p
lanes;
```

```
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=1995' i
nto table flights partition (year=1995);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=1996'
into table flights partition (year=1996);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=1997'
 into table flights partition (year=1997);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=1998'
 into table flights partition (year=1998);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=1999' i
nto table flights partition (year=1999);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2000'
into table flights partition (year=2000);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2001'
 into table flights partition (year=2001);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2002'
 into table flights partition (year=2002);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2003' i
nto table flights partition (year=2003);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2004'
into table flights partition (year=2004);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2005'
 into table flights partition (year=2005);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2006'
 into table flights partition (year=2006);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2007' i
nto table flights partition (year=2007);
load data inpath '${datapath}/airline_ontime_iceberg.db/flights/year=2008'
into table flights partition (year=2008);
```

Convert these existing Hive external tables to Iceberg tables

```
ALTER TABLE planes ADD CONSTRAINT planes_pk PRIMARY KEY (tailnum) DISABLE NO
VALIDATE;
ALTER TABLE flights ADD CONSTRAINT planes_fk FOREIGN KEY (tailnum) REFEREN
CES planes(tailnum) DISABLE NOVALIDATE RELY;

ALTER TABLE airlines ADD CONSTRAINT airlines_pk PRIMARY KEY (code) DISABLE
 NOVALIDATE;
ALTER TABLE flights ADD CONSTRAINT airlines_fk FOREIGN KEY (uniquecarrier)
REFERENCES airlines(code) DISABLE NOVALIDATE RELY;

ALTER TABLE airports ADD CONSTRAINT airports_pk PRIMARY KEY (iata) DISABLE N
OVALIDATE;
ALTER TABLE flights ADD CONSTRAINT airports_orig_fk FOREIGN KEY (origin)
REFERENCES airports(iata) DISABLE NOVALIDATE RELY;
ALTER TABLE flights ADD CONSTRAINT airports_dest_fk FOREIGN KEY (dest) RE
FERENCES airports(iata) DISABLE NOVALIDATE RELY;

ALTER TABLE airports SET TBLPROPERTIES ('storage_handler'='org.apache.iceb
erg.mr.hive.HiveIcebergStorageHandler');
ALTER TABLE airlines SET TBLPROPERTIES ('storage_handler'='org.apache.icebe
rg.mr.hive.HiveIcebergStorageHandler');
ALTER TABLE planes SET TBLPROPERTIES ('storage_handler'='org.apache.iceberg.
mr.hive.HiveIcebergStorageHandler');
ALTER TABLE flights SET TBLPROPERTIES ('storage_handler'='org.apache.iceber
g.mr.hive.HiveIcebergStorageHandler');
```

Test driving Iceberg from Hive

You learn how to access the Hive demo data, which you can use to get hands-on experience running Iceberg queries.

About this task

Query sample airline demo data in Data Explorer.

Before you begin

- You must meet the prerequisites to query Iceberg tables.
- You obtained the Ranger permissions to run Hive queries.

Procedure

1. Connect to Hive running in a Cloudera Data Hub cluster.
2. Run the queries in the previous topic, "Hive demo data" to set up the following databases: airline_ontime_iceberg, airline_ontime_orc, airline_ontime_parquet.
3. Use the airline_ontime_iceberg database.
4. Take a look at the tables in the airline_ontime_iceberg database.

   ```
   USE airline_ontime_iceberg;
   SHOW TABLES;
   ```

   Flights is the fact table. It has 100M rows and three dirmensions, ariline, airports, and planes. This records flights for more than 10 years in the US, and includes the following details:

   - origin
   - destination
   - delay
   - air time
5. Query the demo data from Hive.
   For example, find the flights that departed each year, by IATA code, airport, city, state, and country. Find the average departure delay.

   ```
   SELECT f.month, a.iata, a.airport, a.city, a.state, a.country
   FROM flights f,
   airports a
   WHERE f.origin = a.iata
   GROUP BY
   f.month,
   a.iata,
   a.airport,
   a.city,
   a.state,
   a.country
   HAVING COUNT(*) > 10000
   ORDER BY AVG(f.DepDelay) DESC
   LIMIT 10;
   ```

   Output appears as follows:

   ```
   +----------+---------+-----------------------------------+---------------
   +----------+------------+
   | f.month  | a.iata  |            a.airport              |    a.city
    | a.state  | a.country  |
   ```

```
+----------+---------+-----------------------------------+---------------
+----------+------------+
| 12       | ORD     | Chicago O'Hare International      | Chicago
 | NULL     | USA        |
| 6        | EWR     | Newark Intl                       | Newark
 | NULL     | USA        |
| 7        | JFK     | John F Kennedy Intl               | New York
 | NULL     | USA        |
| 6        | IAD     | Washington Dulles International   | Chantilly
 | NULL     | USA        |
| 7        | EWR     | Newark Intl                       | Newark
 | NULL     | USA        |
| 6        | PHL     | Philadelphia Intl                 | Philadelphia
 | NULL     | USA        |
| 1        | ORD     | Chicago O'Hare International      | Chicago
 | NULL     | USA        |
| 6        | ORD     | Chicago O'Hare International      | Chicago
 | NULL     | USA        |
| 7        | ATL     | William B Hartsfield-Atlanta Intl | Atlanta
 | NULL     | USA        |
| 12       | MDW     | Chicago Midway                    | Chicago
 | NULL     | USA        |
+----------+---------+-----------------------------------+---------------
+----------+------------+
10 rows selected (103.812 seconds)
```

6. Split the partitioned data into manageable files.

```
ALTER TABLE airports SET PARTITION SPEC (iata,BUCKET(15, country));
```

Iceberg data types

References include Iceberg data types and a table of equivalent SQL data types by Hive/Impala SQL engine types.

Iceberg supported data types

Table 1:

Iceberg data type	SQL data type	Hive	Impala
binary		BINARY	BINARY
boolean	BOOLEAN	BOOLEAN	BOOLEAN
date	DATE	DATE	DATE
decimal(P, S)	DECIMAL(P, S)	DECIMAL (P, S)	DECIMAL (P, S)
double	DOUBLE	DOUBLE	DOUBLE
fixed(L)		BINARY	Not supported
float	FLOAT	FLOAT	FLOAT
int	TINYINT, SMALLINT, INT	INTEGER	INTEGER
list	ARRAY	ARRAY	Read only
long	BIGINT	BIGINT	BIGINT
map	MAP	MAP	Read only
string	VARCHAR, CHAR	STRING	STRING
struct	STRUCT	STRUCT	Read only

Iceberg data type	SQL data type	Hive	Impala
time		STRING	Not supported
timestamp	TIMESTAMP	TIMESTAMP	TIMESTAMP (see limitation below)
timestamptz	TIMESTAMP WITH LOCAL TIME ZONE	Use TIMESTAMP WITH LOCAL TIMEZONE for handling these in queries	Read timestamptz into TIMESTAMP values Writing not supported
uuid	none	STRING Writing to Parquet is not supported	Not supported
variant	VARIANT	VARIANT	Not supported

Data type limitations

An implicit conversion to an Iceberg type occurs only if there is an exact match; otherwise, a cast is needed. For example, to insert a VARCHAR(N) column into an Iceberg table you need a cast to the VARCHAR type as Iceberg does not support the VARCHAR(N) type. To insert a SMALLINT or TINYINT into an Iceberg table, you need a cast to the INT type as Iceberg does not support these types.

Iceberg supports two timestamp types:

- timestamp (without timezone)
- timestamptz (with timezone)

Starting from Spark 3.4 onwards, Spark SQL supports a timestamp with local timezone (TIMESTAMP_LTZ) type and a timestamp without timezone (TIMESTAMP_NTZ) type, with TIMESTAMP defaulting to the TIMESTAMP_LTZ type. However, this can be configured by setting the spark.sql.timestampType (the default value is TIMESTAMP_LTZ).

When creating an Iceberg table using Spark SQL, if spark.sql.timestampType is set to TIMESTAMP_LTZ, TIMESTAMP is mapped to Iceberg's timestampz type. If spark.sql.timestampType is set to TIMESTAMP_NTZ, then TIMESTAMP is mapped to Iceberg's timestamp type.

Impala is unable to write to Iceberg tables with timestamptz columns. For interoperability, when creating Iceberg tables from Spark, you can use the Spark configuration, spark.sql.timestampType=TIMESTAMP_NTZ.

Note that the timestamp and timestamptz types have different semantics.

Unsupported data types

Impala does not support the following Iceberg data types:

- TIMESTAMPTZ (only read support)
- TIMESTAMP in tables in AVRO format
- FIXED
- UUID

Iceberg table properties

The Cloudera environment for querying tables from Hive overrides some Iceberg table properties. You learn which table properties are supported for querying tables from Impala.

Iceberg documentation describes all the properties for configuring tables. This documentation focuses on key properties for working with Iceberg tables in Cloudera.

Iceberg supports concurrent writes by default. You can tune Iceberg v2 table properties for concurrent writes. You set the following properties if you plan to have concurrent writers on Iceberg v2 tables:

- commit.retry.min-wait-ms
- commit.retry.num-retries

Cloudera supports adding the Parquet compression type using table properties. For more information, see Iceberg documentation about Compression Types.

You can use the Alter Table feature to set a property. From Hive, the following Iceberg table property overrides are in effect:

- iceberg.mr.split.size overrides read.split.target-size.
- read.split.open-file-cost is overridden.

You can tune Iceberg v2 table properties for concurrent writes. From Impala, the following subset of Iceberg table properties are supported:

- history.expire.min-snapshots-to-keep

  Valid values: integers. Default = 1
- write.format.default

  Valid value: Parquet
- write.metadata.delete-after-commit.enabled

  Valid values: true or false.
- write.metadata.previous-versions-max

  Valid values: integers. Default = 100.
- write.parquet.compression-codec

  Valid values: GZIP, LZ4, NONE, SNAPPY (default value), ZSTD
- write.parquet.compression-level

  Valid values: 1 - 22. Default = 3
- write.parquet.row-group-size-bytes

  Valid values: 8388608 (or 8 MB) - 2146435072 (or 2047MB). Overiden by PARQUET_FILE_SIZE.
- write.parquet.page-size-bytes

  Valid values: 65536 (or 64KB) - 1073741824 (or 1GB).
- write.parquet.dict-size-bytes

  Valid values: 65536 (or 64KB) - 1073741824 (or 1GB)

Cloudera Lakehouse Optimizer for Iceberg table maintenance

The Cloudera Lakehouse Optimizer service in Cloudera Base on premises 7.3.2 or higher versions cluster provides automated Iceberg table maintenance, using Spark jobs, for Iceberg tables in Cloudera Open Data Lakehouse. The service simplifies table management, improves query performance, and reduces operational costs.

Cloudera Lakehouse Optimizer for Iceberg table maintenance

The Cloudera Lakehouse Optimizer service in Cloudera Base on premises 7.3.2 or higher versions cluster provides automated Iceberg table maintenance, using Spark jobs, for Iceberg tables in Cloudera Open Data Lakehouse. The service simplifies table management, improves query performance, and reduces operational costs.

You can add the Cloudera Lakehouse Optimizer service to an existing Cloudera Base on premises 7.3.2 or higher versions cluster in Cloudera Manager 7.13.2 or higher versions, or you can create a dedicated cluster for Cloudera Lakehouse Optimizer and then add the service. You must ensure that the cluster contains all the required services. After you finish configuring the service, you can use the Cloudera Lakehouse Optimizer service REST APIs to define the Cloudera Lakehouse Optimizer policies, perform policy management, and run other Iceberg table optimization operations.

Cloudera Lakehouse Optimizer improves the Cloudera Open Data Lakehouse performance in the following ways:

- Simplifying the Iceberg table management by automating the optimization and maintenance activities, such as data file compaction and snapshot expiration.
- Reducing the total cost of ownership (TCO) by lowering the storage and compute costs.
- Maintaining the health of Iceberg tables to ensure high operational efficiency.
- Improving the query execution time on the Iceberg tables.

You create Cloudera Lakehouse Optimizer policies to perform Iceberg table maintenance. You can manage and monitor these policies as necessary, depending on your assigned role.

You can define the policies, perform policy management, and Iceberg table optimization operations using REST APIs. You can choose to trigger the table maintenance based on events, or schedule it to run at regular intervals. You can also perform manual (ad hoc) maintenance when required. Ensure that you dry run the policy before manual evaluation. During the dry run process, Cloudera Lakehouse Optimizer only generates the table maintenance actions but does not initiate any maintenance actions.

A Cloudera Lakehouse Optimizer policy consists of a JEXL script and JSON file, collectively referred to as policy resources. The JEXL script is mandatory, and the JSON file is optional. The default ClouderaAdaptive policy consists of the ClouderaAdaptive.jexl script and the ClouderaAdaptive.json file. Cloudera Lakehouse Optimizer uses the Clou deraAdaptive policy as a template to help you create policies.

Cloudera Lakehouse Optimizer features

Cloudera Lakehouse Optimizer supports several features. Some features are available only through REST APIs. You can choose to create Cloudera Lakehouse Optimizer policies in the UI or use REST APIs depending on your requirements.

Important: Ensure that you do not have any other optimization services enabled alongside Cloudera Lakehouse Optimizer, such as AWS S3 Table Optimization, to avoid conflicts during the optimization process.

The following table lists the supported features and the supported methods to use these features:

Table 2: Supported feature list

Features	Available methods to use the feature	Description
Event-based policy	• UI – at table and namespace level • REST API – at table, namespace, and catalog level	Schedules the policies to be evaluated when an HMS event is triggered, such as an insert, update or delete operation on the table. You can create only one version of the policy definition at the catalog level in the UI. However, you can create multiple versions of the policy definition at catalog, namespace, or table level using REST APIs. For example, when you create policy P1 in the UI, the definition is defined at the catalog level. However, using REST APIs you can create another definition for P1 at namespace level or table level.

Features	Available methods to use the feature	Description
Schedule-based policy	• UI – at table and namespace level • REST API – at table, namespace, and catalog level	Schedules policies to be evaluated at regular intervals. You can create only one version of the policy definition at the catalog level in UI. However, you can create multiple versions of the policy definition at catalog, namespace, or table level using REST APIs. For more information, see Creating a Cloudera Lakehouse Optimizer policy in Lakehouse Optimizer UI on page 105 and Defining Cloudera Lakehouse Optimizer resources using REST APIs on page 121.
Manual (ad hoc) evaluation	• REST API	Runs the policies manually to optimize the Iceberg tables when required. For more information, see Performing manual Iceberg table maintenance using Cloudera Lakehouse Optimizer REST APIs on page 123. **Note:** Ensure that you perform a dry run on the policy before you run it manually.
Dry-run policies	• REST API	Generates the table maintenance actions but does not initiate any maintenance actions. Dry run the policies to ensure they run effectively without failure. For more information, see Preparing and defining policies using REST APIs.
Small file compaction options include: • Target file size • Minimum number of input files • Delete file threshold • Maximum concurrent file group rewrites • Enable partial progress • Maximum number of commits during partial progress • Use starting sequence number of snapshot • Rewrite all	• UI • REST API	Automates the Iceberg data file compaction maintenance actions. In Apache Iceberg documentation, this procedure is called *rewrite_data_files*, and it supports Table, Strategy (binpack or sort), sort_order (zorder, sortDirection, NullOrder), options, and where arguments which are also supported by Cloudera Lakehouse Optimizer.
Orphan file removal includes: • Delete older than	• UI • REST API	Automates the Iceberg orphan file removal maintenance action.
Snapshot expiration options include: • Maximum snapshot age • Retain last • Expire snapshot ID • Clean expired files	• UI • REST API	Automates the Iceberg snapshot management maintenance actions.
Rewrite manifest options include: • Target file size • Use caching	• UI • REST API	Automates the Iceberg manifest rewrite maintenance actions.

Features	Available methods to use the feature	Description
Positional delete rewrite options include: • Rewrite job order • Enable partial progress • Maximum number of commits during partial progress • Minimum number of input files • Maximum concurrent group rewrites • Target file size	• UI • REST API	Automates the Iceberg positional delete rewrite maintenance actions.
Data file compression options include: • Compression method • Compression level • Filter	• REST API	Re-encodes Iceberg data files that match a SQL filter with a more efficient compression codec to reduce storage costs without changing query results. A filter is required when compression is enabled. Create a separate compression policy from compaction, associate it at the namespace level, and run it on an ad hoc basis or on a monthly cadence because compression targets cold data. Individual tables opt in with ALTER TABLE … SET TBLPROPERTIES. For more information, see Data file compression for Cloudera Lakehouse Optimizer.
Pause and resume table maintenance manually	• UI • REST API	Pauses table maintenance. The table maintenance is paused in the following scenarios: • You manually paused the table maintenance. • The recurring failures, during the execution phase of the policy, exceeded the retry value. **Note:** When you pause maintenance for a table or namespace, all the associated policies for that table or namespace are not evaluated. Therefore, you must be cautious before you pause maintenance for a table or a namespace. For more information, see Pausing and resuming table maintenance on page 132.
High availability	Cloudera Manager configuration	Runs multiple Cloudera Lakehouse Optimizer service instances with automatic leader election and failover. At any time, one instance is the leader that accepts API requests and performs state-mutating operations. Standby instances take over automatically when the leader fails. For more information, see High availability for Cloudera Lakehouse Optimizer and Configuring high availability for Cloudera Lakehouse Optimizer.
Maintenance task prioritization	• UI - at catalog level • REST API - at catalog, namespace, and table level	Allows you to define a priority level for maintenance tasks that are configured for a policy. Higher-priority tasks are scheduled before lower-priority tasks when multiple policies compete for cluster resources. You can set task priority when you create or edit a policy in the Cloudera Lakehouse Optimizer UI or define the priority value in the policy JSON constants when you use REST APIs.

Features	Available methods to use the feature	Description
State persistence	Cloudera Manager configuration (shared PostgreSQL database)	Persists scheduler and task state—including dispatched tasks, event-based schedules, and in-flight Livy task submission data—in a shared database so maintenance work survives restarts and high availability failover. For more information, see State persistence for Cloudera Lakehouse Optimizer and Cloudera Lakehouse Optimizer persistence configuration properties.
Puffin statistics	REST API	Allows you to create a custom policy that computes NDV statistics for Iceberg table columns and writes them to a Puffin statistics file registered in the table metadata. Query engines that read Iceberg Puffin files can use these estimates for cost-based optimization. For more information, see Puffin statistics for Cloudera Lakehouse Optimizer.
Auditing	Cloudera Manager configuration	Provides an audit trail of user and system activities on Cloudera Lakehouse Optimizer, including policy creation, editing, deletion, association, namespace operations, authentication events, and table maintenance actions. Audit events are written as NDJSON files to a configurable filesystem location. For more information, see Auditing for Cloudera Lakehouse Optimizer.
Cloudera Lakehouse Optimizer event logging	REST API	Ingests the maintenance task metadata, also called an event, into the sys.clo_events Iceberg table. You can use the table to analyze the event logs, use it for troubleshooting purposes and for root cause analysis, and to generate reports. For more information, see Viewing logs for Cloudera Lakehouse Optimizer on page 134.
Monitoring policy jobs	• UI • REST API • Cloudera Consumption dashboard	Monitor the policy jobs using one of the following methods: • View the latest status for the recent tasks that ran for the table or policy on the UI. • Use the `GET /tasks` or `GET /tasks/id/{id}` APIs. • Monitor the Spark jobs on the **Cloudera Consumption** dashboard. For more information, see Viewing table maintenance status on page 133 and Monitoring table maintenance tasks on Cloudera Observability dashboard.
Manual (ad hoc) evaluation	• UI (Available in 7.3.2 and higher versions) • REST API	Manually run the policies to optimize the Iceberg tables when required.
Policy dry-run	• UI (Available in 7.3.2 and higher versions) • REST API	Dry run existing policies to ensure they run effectively without failure.

Features	Available methods to use the feature	Description
Backup policies and association	REST API	Backs up all the existing policies and associations to a TAR file. You can restore it to another Data Hub, when required. You can use this feature when you want to delete the current Cloudera Lakehouse Optimizer Data Hub and provision another Data Hub. For more information, see Cloudera Lakehouse Optimizer REST APIs on page 126.
Fine-grained access to namespaces	(Available in 7.3.2 and higher versions)	Creates Ranger policies and provides the required access to groups or users at namespace level.

The following additional features are also available:

- Viewing the metrics such as data read and data written for each task on the **Cloudera Consumption** dashboard.
- Viewing the real-time analysis of the infrastructure, jobs, users, and services for the Data Hub hosting the Cloudera Lakehouse Optimizer service in Cloudera Management Console.

Related Information

Lakehouse Optimizer REST APIs

Best practices to use Cloudera Lakehouse Optimizer

Cloudera recommends following a few best practices to ensure error-free Iceberg table maintenance while using Cloudera Lakehouse Optimizer.

Important: Ensure that you do not have any other optimization services enabled alongside Cloudera Lakehouse Optimizer, such as AWS S3 Table Optimization, to avoid conflicts during the optimization process.

The following are the recommended best practices:

- Enabling autoscaling for compute nodes.
- Using ClouderaAdaptive as a template to create your policies.
- Creating only one policy per table to avoid policy conflict.
- Creating an event-based schedule for frequent evaluation if your environment setup continuously generates small files.
- Using policy names that show the actions the policies run, the tables they apply for, or both. This best practice enhances policy management and governance.
- Ensuring the Cloudera Lakehouse Optimizer policy file or policy template size does not exceed 10 MB.
- Scheduling policy runs so that they do not interfere with the Iceberg-related replication tasks.

  When you have Iceberg replication policies running on the Iceberg tables that are also slated for Cloudera Lakehouse Optimizer table maintenance, you must ensure that the schedules for the replication policies and Cloudera Lakehouse Optimizer policies do not overlap. This is because these concurrent policy jobs interfere with each other, and one or both of the policies might fail or create issues.
- Reviewing and cross-checking the arguments and table associations before you use or trigger the orphan file removal, rewrite positional delete, and snapshot expiration maintenance operations. Cloudera recommends this best practice because these operations might delete data.
- Using a where clause to filter the data files when the number of data files to be compacted is very large. The where clause is part of the rewrite data file argument.
- Enabling partial progress for compaction to prevent out-of-memory issues.
- Performing a dry run on the policies before you trigger manual evaluations.
- Performing manual table maintenance on a static Iceberg table using REST APIs instead of including it in a Cloudera Lakehouse Optimizer policy.

Static tables are tables that have not undergone any modifications, such as insert, update, delete operations, or schema evolution, since the time of table creation.

- Using event-based policies to prevent excessive evaluation for tables that are not updated frequently.

Use cases for Cloudera Lakehouse Optimizer

Cloudera Lakehouse Optimizer is a service in Cloudera on cloud Management Console that automates Iceberg table maintenance for Open Data Lakehouse users, leveraging on all of the optimization actions available with Iceberg. You can use Cloudera Lakehouse Optimizer for various use cases.

Some of the use cases where you can automate your Iceberg table maintenance tasks include:

- Removing older and unnecessary snapshots from the table's metadata. This action helps to keep the metadata compact and the query performance optimal.
- Compacting small files into larger ones. This action optimizes storage and improves read performance.
- Rewriting the manifest files to remove entries for deleted data files. This action improves the metadata scan performance.
- Removing the data files that are no longer referenced by the table's metadata. This action reclaims storage space.
- Compacting small positional delete files into larger ones, and filtering out the positional delete records that refer to data files that are no longer available. This action reclaims storage space.

Understand and prepare to use Cloudera Lakehouse Optimizer

Before you use Cloudera Lakehouse Optimizer, you must understand how a policy works, the policy resources, and the prerequisites you must complete before you create a policy.

How Cloudera Lakehouse Optimizer performs table maintenance

Cloudera Lakehouse Optimizer and its components perform several operations in the background after you create or define a policy to initiate, evaluate, and complete the Iceberg table maintenance.

About this task

After you create a policy, Cloudera Lakehouse Optimizer performs the following steps:

Procedure

1. Initiates the schedule for each table based on the policies attached to the resource, such as catalog, namespace, and table.
2. Consolidates the policies depending on the CRON schedule in all the active policies, and creates a schedule entry for all the policies that share the same CRON trigger.
3. Initiates the *evaluation phase* when a CRON trigger is activated. The evaluation phase consists of the following steps:
   a) Initiating the evaluation of the policy script per table.
   b) Determining the sequence of maintenance actions per table where the policy script:
      1. compares the current Iceberg metadata and table statistics to the expected statistics to evaluate and determine the table maintenance actions,
      2. generates a list of (potentially empty) maintenance actions, and then
      3. queues up the generated maintenance actions.

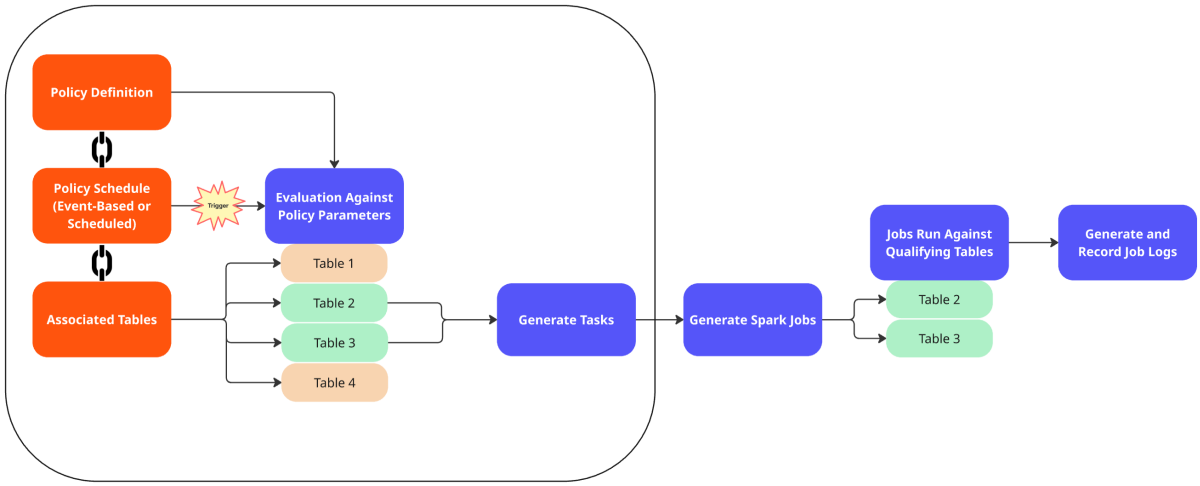
      This list of actions is called a task, and the task has a task ID.
4. Sends the maintenance actions to the Spark Engine through Livy in the correct order of sequence. The Spark engine performs each maintenance action as a Spark job.
5. Supervises the task execution, and collects all the relevant task execution information in the events logs.

Cloudera Lakehouse Optimizer captures every maintenance action, Spark job, and the job results as task metadata. The task metadata includes the following elements:

- The task ID
- The maintenance actions and its results
- The Spark computation metrics for the Spark jobs
- The current task status

The task metadata is stored in the sys.clo_events table.

The following diagram shows the high-level steps for Iceberg table maintenance by Cloudera Lakehouse Optimizer:



What to do next

Cloudera Lakehouse Optimizer persists dispatched tasks and event-based triggers in a shared database. If the service restarts, it reloads pending tasks and triggers and resumes scheduling. For more information, see State persistence for Cloudera Lakehouse Optimizer.

Related Information

Cloudera Lakehouse Optimizer - Refining Optimizations through Apache Iceberg Table Properties blog

What are Cloudera Lakehouse Optimizer policy resources

A Cloudera Lakehouse Optimizer policy consists of a JEXL script and JSON file, collectively referred to as policy resources. The JEXL script is mandatory, and the JSON file is optional. The default ClouderaAdaptive policy consists of the ClouderaAdaptive.jexl script and the ClouderaAdaptive.json file.

JEXL script

The Cloudera Lakehouse Optimizer Data Hub contains the default ClouderaAdaptive.jexl file that is a policy script. You can use the script as is. Otherwise, you can use the default JEXL file as a template to add or modify the contents, and then upload the file as a new JEXL file.

Important: You can define any maintenance action and condition in the JEXL script to trigger the required maintenance action.

The following snippet shows a sample JEXL script:

```
// A sample jexl-based policy definition to evaluate the need for expire sna
pshot and generate action argument
#pragma dlm.statistics true
// Stores the list of generated actions
const actions = [...]
```

```
// Fetch iceberg stats
let stats = statistics($table);
if ($constants.expireSnapshot.enabled) {
    let min = $table['dlm.expireSnapshot.retainLast'] ?? $constants.expireS
napshot.retainLast ?? 2;
    let snapshots = stats.numberOfSnapshots;
    let snapshotTimestampDelta = stats.snapshotTimestampDelta ?? 3600000;
    let deltaDurationThreshold = $table['dlm.expireSnapshot.snapshotsDurati
onDeltaMax'] ?? $constants.expireSnapshot.snapshotsDurationDeltaMax ?? 0;
    if (snapshots > min || snapshotTimestampDelta > deltaDurationThreshold)
 {
        // create an expire snapshot action
        actions.add(dlm:expireSnapshots($table, $constants));
    }
}
// return the list of actions
actions
```

The JEXL script contains the following variables, out of which, some variables are predefined:

Variables	Description
$table	Iceberg table object for which the policy is being evaluated. For example, $table[‘dlmproperty’] refers to the Iceberg table property named dlmproperty in the specified table.
$constants	Set of action argument values and defaults. For example, $constants.expireSnapshot.retainLast refers to the value for the retainLast argument for the expireSnapshot action.
$catalog	Iceberg catalog object.

The policy script supports retrieving the values from the table property or the policy constants. For example, let filesCountDrop =$table['dlm.rewriteDataFiles.filesCountDrop'] ?? $constants.rewriteDataFiles.filesCountDrop ?? 0.15;.

For more information about modifying the JEXL script contents, see JEXL syntax.

JSON file

The JSON file is an optional component. Cloudera Lakehouse Optimizer uses the threshold values in the file during the evaluation phase to compare the current and expected stats. The action arguments are supported by a corresponding Spark action.

Use the following list of action arguments in the JSON file as required for your use case:

```
{
  "expireSnapshot": {
    "expireOlderThan": 432000000,
    "retainLast": 5,
    "cleanExpiredFiles": true
  },
  "rewriteManifest" : {
    "useCaching": true,
    "fileCountMax": 100,
    "manifestFileSize": 8388608,
    "smallFileRatioMax": 0.5
  },
  "rewriteDataFiles" : {
    "targetFileSize": 536870912,
    "maxConcurrentRewriteFileGroups": 5,
    "minInputFiles": 5,
    "partialProgressEnabled": true,
    "partialProgressMaxCommits": 10,
```

```
    "deleteFileThreshold": 2000000,
    "useStartingSequenceNumber": false,
    "rewriteAll": false
  },
  "rewritePositionDelete" : {
    "enabled" : false,
    "targetFileSize": 67108864,
    "maxConcurrentGroupRewrite": 5,
    "minInputFiles": 6
    "partialProgressMaxCommits": 10,
    "partialProgressEnabled": true
  },
  "deleteOrphanFiles" : {
    "olderThan": 259200000
  },
  "priority": "MEDIUM",
  "description": "An example policy constant",
  "cron": "0 4 * ? * *"
}
```

For data file compression on Cloudera Base on premises, include a compressDataFiles block in the JSON file:

```
"compressDataFiles": {
  "enabled": true,
  "method": "zstd",
  "level": 12,
  "filter": "1 = 0"
}
```

For more information, see Data file compression for Cloudera Lakehouse Optimizer.

The following table lists the action arguments to specify in the JSON file, which are then sent to the Spark engine to use during the maintenance actions:

Action argument	Value	Description
		expireSnapshot
enabled	Default is true	Determines whether to evaluate th
cleanExpiredFiles	Default is true	Removes the expired snapshots pe
expireOlderThan	Default is 120 * 3600 * 1000 ms, that is 5 days. Minimum is 10 seconds	Deletes the snapshot when the sna For example, a snapshot is deleted
retainLast	Default is 5 Minimum is 1	Deletes the last snapshot when the For example, by default the first s
expireSnapshotId	No default value	Expires the specified snapshot.
		rewriteManifest
enabled	Default is true	Determines whether to evaluate th
useCaching	Default is true	Uses cache during the rewrite mar
targetFileSize	Default is 8388608 bytes	Specifies the target manifest file s
		rewriteDataFiles
enabled	Default is true	Determines whether to evaluate th
targetFileSize	Default is 512 MB Minimum is 1 KB Maximum is 64 GB	Determines the target output file s

Action argument	Value	Description
maxConcurrentRewriteFileGroups	Default is 5 Minimum is 1 Maximum is 1000	Defines the maximum number of
minInputFiles	Default is 5 Minimum is 1	Rewrites a file group when the file
partialProgressMaxCommits	Default is 10 Minimum is 1	Defines the maximum number of
deleteFileThreshold	Default is 2000000 Minimum is 1	Defines the minimum number of
partialProgressEnabled	Default is false.	Defines the maximum number of rewrite operation is in progress. If
use-starting-sequence-number	Default is false.	Specifies the sequence number of
rewrite-all	Default is false.	Force rewrites all the files overrid
		deleteOrphanFiles
enabled	Default is true	Determines whether to evaluate th
olderThan in ms	Default is 72 * 3600 * 1000 that is 3 days. Minimum is 10 in seconds.	Removes orphan files created befo
		rewritePositionDelete
enabled	Default is true	Determines whether to evaluate th
targetFileSize	Default is 64 MB Minimum is 1 KB	Determines the target output file s
maxConcurrentGroupRewrite	Default is 5 Minimum is 1 Maximum is 1000	Defines the maximum number of
minInputFiles	Default is 5 Minimum is 1	Rewrites a file group when the file
partial-progress.max-commits	Default is 10	Defines the maximum number of rewrite operation is in progress.
partialProgressEnabled	Default is true	Enables committing groups of file
rewrite-job-order	No default value	Forces the rewrite job order based You can choose one of the followi • bytes (asc) rewrites the smalle • bytes (desc) rewrites the large • files (asc) rewrites the job gro • files (desc) rewrites the job gr • none rewrites the job groups i
		computeTableStats
enabled	No default; must be true to run the action	Opt-in gate for the compute-table-
columns	No default; omit for all columns	Array of column names to comput schema. Maximum of 512 column
snapshotId	No default; omit for current snapshot	Snapshot ID to compute statistics
		compressDataFiles
enabled	Default is false	Must be true to run the data file co

Action argument	Value	Description
filter	No default; required when enabled	SQL WHERE clause that scopes v RTIES. Cloudera Lakehouse Opti
method	Default is the table format default	Compression codec. For supporte
level	Default is the codec default	Codec-specific compression level
validateFilterColumns	Default is true	When true, Cloudera Lakehouse C
		priority
priority	Default is Medium Valid values are Critical, Urgent, High, Medium, Low, and Minimal.	Defines the priority level for main when multiple policies compete fo

Puffin statistics policy example

Puffin statistics require a separate custom policy. Do not add computeTableStats to the default ClouderaAdaptive policy constants.

Use a dedicated JEXL script and JSON constants file. The action runs only when computeTableStats.enabled is true in the JSON constants (the default is false for this action).

JEXL

```
// Dry-run policy: compute table stats using the same merge rules as the for
mer ClouderaAdaptive.jexl block.
// Opt-in via computeTableStats.enabled in policy constants (default false
per PolicyConstants for this action).
#pragma dlm.cron "0 0 13 ? * *"
#pragma dlm.statistics false

const actions = [...]

if ($constants.computeTableStats.enabled) {
    const computeStats = dlm:computeTableStats($table);

    let colsTable = $table['dlm.computeTableStats.columns'];
    if (colsTable != null && colsTable != '') {
        computeStats.addColumns(colsTable);
    } else if (colsTable == null && $constants.computeTableStats.columns !=
 null) {
        computeStats.addColumns($constants.computeTableStats.columns);
    }

    computeStats.validate($table);
    let resolvedSnapshotId = $table['dlm.computeTableStats.snapshotId'] ??
 $constants.computeTableStats.snapshotId;
    if (resolvedSnapshotId != null) {
        computeStats.snapshot(resolvedSnapshotId);
    }

    actions.add(computeStats);
}

actions
```

JSON

```
{
  "computeTableStats": {
    "enabled": true,
    "columns": ["id", "data"],
```

```
    "snapshotId": 424242
  },
  "description": "Constants for compute table stats"
}
```

Omit snapshotId to use the current snapshot. Upload the JEXL and JSON files when you define the policy resources. For more information, see Defining Cloudera Lakehouse Optimizer resources using REST APIs.

ClouderaAdaptive-specific resources

Use ClouderaAdaptive resources to create your policy. The resources include the ClouderaAdaptive.jexl script and the ClouderaAdaptive.json file.

ClouderaAdaptive.jexl

The Lakehouse Optimizer Data Hub contains the default ClouderaAdaptive.jexl file which is a policy script.



Important: Define any maintenance action and condition in the JEXL script to trigger the required maintenance action.

The following sections show:

- Curl command to fetch the default ClouderaAdaptive.jexl file contents
- Default ClouderaAdaptive.jexl file contents
- How to get the table statistics
- Using the policy constants to specify the maintenance type and the default action in the default JEXL file
- View the table properties in the default JEXL file

Add or modify the script elements as necessary. Each action builder statement generates a maintenance action.

Curl command to fetch the default ClouderaAdaptive.jexl file contents

```
curl --location 'https://[***LAKEHOUSE OPTIMIZER ENDPOINT***]/api/v1/poli
cies/resource?uri=dlm%3A%2F%2Ftps%3Adefault%2FClouderaAdaptive'
        \ --header 'Authorization: Bearer ey'
```

Default JEXL file contents

The following sample script displays the default ClouderaAdaptive.jexl file contents:

```
/*
* Description : Policy that evaluates if a certain maintenance action is to
 be scheduled based on iceberg stats.
*
* Scheduling
*
* Cron expressions are used to define the schedule for policy evaluation. T
his follows the Quartz cron syntax.
* 0 0/10 * ? * * will trigger the policy evaluation every 10 minutes.
* Refer this for more details : https://www.quartz-scheduler.org/documenta
tion/quartz-2.3.0/tutorials/crontrigger.html
*
* Iceberg statistics
*
* #pragma dlm-statistics : If this is set to true stats will pre-computed
before the evaluation.
* statistics(Table, force) : Method defined in the jexl policy context to
 retrieve the table stats from cache.
*
* Arguments
*
* $table : Iceberg table object. This is used to retrieve the table proper
ties.
```

```
* $constants : Policy constants applicable on this table.
* stats : Iceberg table stats used for decision making.
* dlm : Builder instance used to create ActionBuilder objects.
*
* Evaluation of actions
*
* Expire Snapshot : Expires the iceberg table snapshots based on the crit
eria, trigger this action if number of snapshots is greater than the
* dlm.expireSnapshot.retainLast defined in the policy constants or if the d
uration between current and oldest snapshot is exceeding the dlm.expireSnaps
hot.snapshotsDurationDeltaMax (milliseconds).
*
* Data file Compaction/Rewrite data file : Compacts small data files and rem
oves delete files.
* This action is triggered if there is an expected significant drop in the
 number of files.
* The compaction operation is guided by statistics and the target file size
 and small files to determine whether compaction should be performed.
* The threshold for small files can be stored in table properties as (write.
bin-packing.min-file-count) or minInputFiles from json file.
* The threshold for the reduction or drop can be either stored as an Iceberg
 table property (dlm.rewriteDataFiles.filesCountDrop)
* or as a policy property ($constants.rewriteDataFiles.filesCountDrop). If
 none of these are specified, the
* value 0.15 is used; if a reduction in number of files greater that 15% is
 expected, compact the table.
* Note that the ratio is calculated based on the target file size write.targ
et-file-size-bytes.
*
* Rewrite manifest/Compact manifests : Compacts manifest files, trigger this
 action based on table('write.manifest.target-size-bytes') / json (manifes
tFileSize) property  if
* 1) the manifest file count exceeds the table property dlm.rewriteManifest.
smallFileRatioMax.
* 2) the table property dlm.rewriteManifest.fileCountMax / json (fileCoun
tMax).
*
* This dlm.rewriteManifest.smallFileRatioMax ratio should by default is 0.5,
 if kept lower that means rewrite manifest will aggressively act on table,
if kept
* higher then rewrite manifest will lazily act on table.
*
* Delete orphan files : Delete the files ( manifest, metadata.json, data/
delete files ) which are not tracked by any snapshot, trigger this action we
ekly or monthly.
*
* Data file compression (Cloudera Base on premises only) : Re-encodes data
 files that match the configured filter with a more efficient compression c
odec.
* Requires a non-empty filter when enabled. Table properties override policy
 constants.
* Create a separate compression policy from compaction. Run compression on
 an ad hoc basis or on a monthly cadence because compression targets cold da
ta.
*
*/


// Set to true if iceberg table stats pre-computation is required.
#pragma dlm.statistics true
#pragma dlm.description "The template evaluates whether a certain action m
ust be executed based on the computed statistics that are derived from the I
ceberg table metadata. You can also configure the action arguments using the
 policy constants and the Iceberg table properties."
```

```
// Stores the list of generated actions
const actions = [...]


// Fetch iceberg stats
let stats = statistics($table);


// Evaluate compaction
if ($constants.rewriteDataFiles.enabled) {
   let filesCountDrop = $table['dlm.rewriteDataFiles.filesCountDrop'] ?? $
constants.rewriteDataFiles.filesCountDrop ?? 0.15;
   // Target file size used to compute the reduction is defined in the stat
 computation module.
   // use specified target file size or default if no policy specific
   let targetFileSize = $table['write.target-file-size-bytes'] ?? $constant
s.rewriteDataFiles.targetFileSize ?? 512*1024*1024;
   let minInputFiles = $table['write.bin-packing.min-file-count'] ?? $consta
nts.rewriteDataFiles.minInputFiles ?? 5;
   stats = stats.compactionInfo.fetchCompactionInfo($table, stats, targetFi
leSize, filesCountDrop, minInputFiles)
   if (stats.compactionInfo.eligiblePartitionsCount > 0) {


      let partialProgressEnabled = $table['dlm.rewriteDataFiles.partialProgr
essEnabled'] ?? $constants.rewriteDataFiles.partialProgressEnabled ?? false;
      let partialProgressMaxCommits = $table['dlm.rewriteDataFiles.partialP
rogressMaxCommits'] ?? $constants.rewriteDataFiles.partialProgressMaxCommits
 ?? 10;
      let useStartingSequenceNumber = $table['dlm.rewriteDataFiles.useStarti
ngSequenceNumber'] ?? $constants.rewriteDataFiles.useStartingSequenceNumber
?? false;
      let rewriteAll = $table['dlm.rewriteDataFiles.rewriteAll'] ?? $const
ants.rewriteDataFiles.rewriteAll ?? false;
      let maxConcurrentRewriteFileGroups = $table['dlm.rewriteDataFiles.m
axConcurrentRewriteFileGroups'] ?? $constants.rewriteDataFiles.maxConcurrent
RewriteFileGroups ?? 5;
      let deleteFileThreshold = $table['dlm.rewriteDataFiles.deleteFileThres
hold'] ?? $constants.rewriteDataFiles.deleteFileThreshold ?? 2000000;
      let maxFileGroupSize = $table['dlm.rewriteDataFiles.maxFileGroupSize
'] ?? $constants.rewriteDataFiles.maxFileGroupSize ?? 3221225472;


       const compaction = dlm:rewriteDataFiles($table)
           .targetFileSize(targetFileSize)
           .partialProgressEnabled(partialProgressEnabled)
           .partialProgressMaxCommits(partialProgressMaxCommits)
           .useStartingSequenceNumber(useStartingSequenceNumber)
           .rewriteAll(rewriteAll)
           .maxConcurrentRewriteFileGroups(maxConcurrentRewriteFileGroups)
           .minInputFiles(minInputFiles)
           .deleteFileThreshold(deleteFileThreshold)
           .maxFileGroupSize(maxFileGroupSize);

       // Comma separated column names , Eg : a,b,c
       let zOrderColumns = $table['dlm.rewriteDataFiles.zOrderColumns'] ??
 $constants.rewriteDataFiles.zOrderColumns;
       // default strategy is binpack and default sort order is table's sort
 order
       if (zOrderColumns != null) {
           compaction.zOrder(zOrderColumns.split(","));
       }
```

```
       let where  = $table['dlm.rewriteDataFiles.where'] ?? $constants.rew
riteDataFiles.where;
       if (where != null) {
         compaction.where(where);
       } else {
         // rely on filter expression given by CLO. Use threshold-based p
artition filter to avoid performance issues with large filter expressions
         let maxPartitionsPerChunk = $table['dlm.rewriteDataFiles.maxParti
tionsPerChunk'] ?? $constants.rewriteDataFiles.maxPartitionsPerChunk ?? 100;
         let partitionFilterThreshold = $table['dlm.rewriteDataFiles.partit
ionFilterThreshold'] ?? $constants.rewriteDataFiles.partitionFilterThreshold
 ?? 0.7;
         let partitionFilterExpr = stats.compactionInfo.createPartitionFilt
erExpression($table, maxPartitionsPerChunk, partitionFilterThreshold);
         if(partitionFilterExpr != null)
          compaction.partitionFilter(partitionFilterExpr);
       }

       // Sort items (columnName, orderAsc, nullFirst) , Eg : (a false tru
e),(b true false)
       let sort = $table['dlm.rewriteDataFiles.sort'] ?? $constants.rewrit
eDataFiles.sort;
       if (sort != null) {
           let sortItemArray = sort.split(",")
           for (let item : sortItemArray) {
               let sortItem  = item.substring(1, size(item) - 1);
               let arr = sortItem.split(" ");
               compaction.sort(arr[0], arr[1] == 'true', arr[2] == 'true');
           }
       }
       actions.add(compaction);
   }
}

// Evaluate rewrite manifest
if ($constants.rewriteManifest.enabled) {
   let manifestMax = $table['dlm.rewriteManifest.fileCountMax'] ?? $constan
ts.rewriteManifest.fileCountMax ?? 1000;
   let manifestFileSize = $table['write.manifest.target-size-bytes'] ?? $co
nstants.rewriteManifest.manifestFileSize ?? 8388608;
   stats = stats.manifestInfo.fetchManifestInfo($table, stats, manifestFile
Size)


   let manifestCount = stats.manifestInfo.manifestFilesCount;
   let smallManifestRatio = stats.manifestInfo.smallManifestRatio;


   let smallRatioThreshold =
      $table['dlm.rewriteManifest.smallFileRatioMax']
      ?? $constants.rewriteManifest.smallFileRatioMax
      ?? 0.5;


   if (manifestCount >= manifestMax && smallManifestRatio >= smallRatioThres
hold) {
      actions.add(dlm:rewriteManifests($table, $constants));
   }
}

// Delete orphan files
if ($constants.deleteOrphanFiles.enabled) {
   actions.add(dlm:deleteOrphanFiles($table, $constants));
```

```
}


// Rewrite positional delete files
if ($constants.rewritePositionDelete.enabled) {
   actions.add(dlm:rewritePositionDeletes($table, $constants));
}


// Evaluate expire snapshot
if ($constants.expireSnapshot.enabled) {
   let min = $table['dlm.expireSnapshot.retainLast'] ?? $constants.expireSna
pshot.retainLast ?? 2; // if null 1
   let snapshots = stats.numberOfSnapshots;
   let snapshotTimestampDelta = stats.snapshotTimestampDelta ?? 3600000; //
if null 0
   let deltaDurationThreshold = $table['dlm.expireSnapshot.snapshotsDurati
onDeltaMax'] ?? $constants.expireSnapshot.snapshotsDurationDeltaMax ?? 0;
   if (snapshots > min || snapshotTimestampDelta > deltaDurationThreshold) {
       // create an expire snapshot action
       actions.add(dlm:expireSnapshots($table, $constants));
   }
}


// Evaluate data file compression (runs last)
if ($constants.compressDataFiles.enabled) {
   actions.add(dlm:compressDataFiles($table, $constants));
}

// return the list of actions
actions
```

On Cloudera Base on premises, the default JEXL file also evaluates compressDataFiles when it is enabled in the associated policy constants.

Table statistics

- The #pragma dlm.statistics true statement ensures the pre-computation of statistics before the evaluation phase.
- The const stats = statistics($table) method gets the table statistics.

Tip: You can also view the table statistics using the `GET /tables/{tableId}/stats` API.

Important: ClouderaAdaptive policy does not generate computeTableStats action. To write Puffin statistics files to Iceberg table metadata, create a separate policy that uses the computeTableStats action. For more information, see Puffin statistics for Cloudera Lakehouse Optimizer.

Important: The default ClouderaAdaptive.jexl file on Cloudera Base on premises can evaluate data file compression when compressDataFiles is enabled. Create a separate compression policy instead of combining compression with compaction in the same policy. Run compression on a temporary basis or on a monthly cadence because compression targets cold data. For more information, see Data file compression for Cloudera Lakehouse Optimizer.

Using the policy constants to specify the maintenance type and the default action in the default JEXL file

Policy constants are maintenance operation types with their default action argument values.

Action builders can be used to generate the required action with the arguments defined in the policy constants. For example, dlm:expireSnapshots($table, $constants).

The policy script supports retrieving the values from the table property or the policy constants. For example, let filesCountDrop =$table['dlm.rewriteDataFiles.filesCountDrop'] ?? $constants.rewriteDataFiles.filesCountDrop ?? 0.15; .

For more information about modifying the JEXL script contents, see JEXL syntax.

Viewing the table properties in the default JEXL file

If you have configured the Iceberg table properties using DDL commands, use the table property values in the script.



Important: The table properties are prioritized over the policy constants.

ClouderaAdaptive.json

The following sections show:

- Curl command to fetch the default ClouderaAdaptive.json file contents
- Default ClouderaAdaptive.json file contents
- Action arguments in the default JSON file
- CRON expression to schedule the table maintenance

Curl command to fetch the contents of the default ClouderaAdaptive.json file

```
curl --location 'https://[***LAKEHOUSE OPTIMIZER ENDPOINT***]/api/v1/
policies/resource?uri=dlm%3A%2F%2Ftpp%3Adefault%2FClouderaAdaptive'
        \
--header 'Authorization: Bearer ey'
```

Use the default action arguments and values as-is. If your use case requires more arguments in addition to the default action arguments or you want to tune the existing default arguments, then use the default JSON file as a template to add or modify the arguments, upload the modified file as a new JSON file, and reschedule the namespace.

ClouderaAdaptive.json file contents

The following snippet displays the default ClouderaAdaptive.json file contents:

```
{
  "expireSnapshot": {
    "expireOlderThan": 432000000,
    "retainLast": 50,
    "snapshotsDurationDeltaMax": 432000000
  },
  "rewriteManifest" : {
    "useCaching": true,
    "fileCountMax": 100,
    "manifestFileSize": 8388608,
    "smallFileRatioMax": 0.5
  },
  "rewriteDataFiles" : {
    "targetFileSize": 536870912,
    "filesCountDrop": 0.15,
    "minInputFiles": 5,
    "partialProgressEnabled": false
  },
  "rewritePositionDelete" : {
    "enabled" : true,
    "targetFileSize": 67108864,
    "maxConcurrentGroupRewrite": 10,
    "minInputFiles": 6
  },
  "deleteOrphanFiles" : {
    "olderThan": 259200000
  },
  "label": "Cloudera Adaptive policy"
}
```

Action arguments in the default JSON file

The following table lists the action arguments that are specific to the default ClouderaAdaptive JSON file, their default values and ranges, and their description:

Table 3: Action arguments specific to default JSON file

Action argument	Value	Can be overwritten by specified
snapshotsDurationDeltaMax	Default is 600000 ms	snapshotsDurationDeltaMax
fileCountMax	Default is 5	fileCountMax
manifestFileSize	Default is 8388608	write.manifest.target-size-bytes
smallFileRatioMax	Default is 0.5	smallFileRatioMax
filesCountDrop	Default is 0.15	filesCountDrop
zOrderColumns	-	zOrderColumns
where	-	where
sort	-	sort

CRON expression to schedule the table maintenance

The default ClouderaAdaptive.json file does not have a schedule. Upload a JSON file with the required CRON expression to define the schedule of the evaluation phase. Alternatively, you can perform manual table maintenance.

For example, the "cron": "0 0 * * * ?" expression runs the evaluation phase every hour. For more information about using CRON expressions, see CRON expression generator.

The following sample curl uploads a policy constant for ClouderaAdaptive at catalog level:

```
curl --location --request PUT 'https://[***LAKEHOUSE OPTIMIZER ENDPOINT***]/
api/v1/policies/resource?uri=dlm%3A%2F%2Ftpp%2Fhive%2FClouderaAdaptive' \
--header 'Authorization: Bearer ey' \
--form 'resource=@"/Users/test/Desktop/ClouderaAdaptive.json"
```

In this scenario, the ClouderaAdaptive.json file only contains the following CRON expression:

```
 {
 "cron": "0 2 * * *"
}
```

Important: After you upload a new resource, ensure that you reschedule the namespace.

High availability for Cloudera Lakehouse Optimizer

High availability (HA) for Cloudera Lakehouse Optimizer ensures that the service is running when a Cloudera Lakehouse Optimizer instance fails during planned maintenance. In an HA deployment, you can run multiple Cloudera Lakehouse Optimizer service instances within the same cluster or Data Hub, and the system elects one leader instance to accept requests and modify shared state.

At any time, exactly one Cloudera Lakehouse Optimizer instance is the *leader*. The leader is the only instance that accepts API requests and performs state-mutating operations. Standby instances wait to take over if the leader fails.

Leader election is coordinated through ZooKeeper, and database-level *fencing tokens* prevent split-brain writes if a former leader tries to continue writing after losing leadership.

How HA protects against split-brain

When multiple Cloudera Lakehouse Optimizer instances run in parallel, HA uses three layers of protection:

1. **Leader election (ZooKeeper)**: ZooKeeper elects one leader at a time. If the leader fails, a new leader is elected automatically.
2. **In-memory pre-check**: Before a state-mutating operation, the instance verifies that it is still the leader. If leadership was lost, the request is rejected immediately.
3. **Database-level fencing**: Each leader receives a unique fencing token. Every database write includes this token. The database rejects writes from a stale token, even if a former leader attempts to write after failover.

Operational modes

Cloudera Lakehouse Optimizer can run in the following modes:

Table 4: Cloudera Lakehouse Optimizer HA operational modes

Mode	Description	When it occurs
Full HA mode	Multiple Cloudera Lakehouse Optimizer instances participate in leader election through ZooKeeper.	HA is enabled and all Cloudera Lakehouse Optimizer instances can reach ZooKeeper at startup.
Degraded mode	Cloudera Lakehouse Optimizer started with HA enabled but cannot reach ZooKeeper and runs as a standalone instance without per-write fencing protection.	ZooKeeper is unreachable when Cloudera Lakehouse Optimizer starts, but HA was configured.
Standalone mode	A single Cloudera Lakehouse Optimizer instance runs without leader election or fencing.	HA is not configured, or only one Cloudera Lakehouse Optimizer instance is deployed.



Important: If Cloudera Lakehouse Optimizer starts in degraded mode because ZooKeeper is unavailable, restore ZooKeeper and restart Cloudera Lakehouse Optimizer to return to complete HA protection. Degraded mode is intended for short-term outages only.

Failover behavior

When the leader fails, ZooKeeper detects the failure, elects a new leader, and the new leader begins accepting requests. Failover is automatic and typically completes in less than 30 seconds. During failover, passive instances can return HTTP 503 Service Unavailable or redirect clients to the leader, depending on your configuration.

When clients access Cloudera Lakehouse Optimizer through Apache Knox, the elected leader can update the Knox service descriptor so Knox routes traffic to the current leader automatically.

Shared database and state persistence

In an HA deployment, all Cloudera Lakehouse Optimizer instances share the same database that stores persisted scheduler and task state. Only the elected leader writes to the database. On failover, the new leader reads current state from the shared database without reloading from individual instances. For more information, see State persistence for Cloudera Lakehouse Optimizer.



Important: When you enable state persistence and HA together, configure /var/metering on each HA node. State persistence introduces a metering dependency on the HA node.

Configuring high availability for Cloudera Lakehouse Optimizer

Configure high availability (HA) for Cloudera Lakehouse Optimizer in Cloudera Manager by deploying multiple Cloudera Lakehouse Optimizer service instances and setting HA properties on each instance.

Before you begin

Before you configure HA, ensure the following requirements are met:

- ZooKeeper is running and healthy on at least three hosts in the cluster or Data Hub.
- Multiple Cloudera Lakehouse Optimizer service instances deployed on separate hosts within the same cluster or Data Hub.
- All Cloudera Lakehouse Optimizer instances connect to the same shared database (PostgreSQL or MySQL).
- All Cloudera Lakehouse Optimizer instances can reach ZooKeeper and the shared database.
- Apache Knox is configured if you use Knox for REST API access.



Tip: Provisioning one Cloudera Lakehouse Optimizer Data Hub for each environment is separate from HA. HA runs multiple Cloudera Lakehouse Optimizer service instances within that single Data Hub or cluster.

About this task

Configure HA on every Cloudera Lakehouse Optimizer instance in the cluster or Data Hub. All instances must use the same ZooKeeper connection string, leader latch path, and shared database settings.

To configure HA on a Cloudera Lakehouse Optimizer instance:

Procedure

1. On Cloudera Base on premises, download, distribute, and activate the Cloudera Lakehouse Optimizer parcel, and deploy multiple service instances.
   a) Download the Cloudera Lakehouse Optimizer parcel from Cloudera.
   b) In Cloudera Manager Hosts Parcels , configure the remote parcel repository, download the parcel, distribute it to the cluster hosts, and activate it.

      For instructions, see Managing Parcels.
   c) Add the Cloudera Lakehouse Optimizer service to an existing Cloudera Base on premises cluster, or create a dedicated cluster and add the service.

      Ensure that the cluster contains the required dependent services, including ZooKeeper and Knox.
   d) In Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* , assign Cloudera Lakehouse Optimizer roles to at least two separate hosts so that multiple service instances are running.
2. Navigate to Configuration conf/dlm-client.properties_role_safety_valve .
   a) Set dlm.ha.enabled to true.
   b) Set dlm.ha.zookeeper.connect to the ZooKeeper connection string for your cluster. For example, zk1.example.com:2181,zk2.example.com:2181,zk3.example.com:2181.
   c) Set dlm.ha.zookeeper.latch.path to the ZooKeeper path for leader election. For example, /dlm/leader.

   If you access Cloudera Lakehouse Optimizer through Apache Knox, configure the Knox discovery properties so the elected leader updates the Knox service descriptor. For the complete list of HA properties, see Cloudera Lakehouse Optimizer high availability configuration properties.
3. Optionally, set dlm.ha.advertise.address on each instance if the auto-detected API URL is not reachable by other Cloudera Lakehouse Optimizer instances or API clients.
4. Restart the Cloudera Lakehouse Optimizer service on all instances.
5. Verify HA status using the `GET /config/health` API with scope=full.

   For instructions, see Verifying service health and viewing default Lakehouse Optimizer policy.

   When HA is enabled, the health check response includes a High Availability entry that reports whether the instance is the leader or a standby.

What to do next

After you configure HA:

- Test failover by stopping the leader instance and verifying that a standby instance becomes the leader.
- Monitor Cloudera Lakehouse Optimizer logs for leader election messages.

- Restore ZooKeeper and restart Cloudera Lakehouse Optimizer if an instance starts in degraded mode.

Cloudera Lakehouse Optimizer high availability configuration properties

Use the following properties to configure high availability (HA) for Cloudera Lakehouse Optimizer.

Table 5: HA configuration properties

Property	Description	Default	Example
dlm.ha.advertise.address	The base URL (scheme, host, and port) that this Cloudera Lakehouse Optimizer instance advertises to peers for leader redirection. Use this property when the auto-detected API URL is not reachable by other Cloudera Lakehouse Optimizer instances or API clients.	Auto-detected from the API server configuration and local hostname	https://dlm-node1.example.com:8443
dlm.ha.enabled	Enables HA mode. When set to true, Cloudera Lakehouse Optimizer participates in ZooKeeper leader election and uses database fencing tokens for state-mutating operations.	false	true
dlm.ha.knox.discovery.path	The ZooKeeper path of the Knox descriptor JSON managed by Cloudera Manager. When set, the elected leader reads this descriptor, updates the Cloudera Lakehouse Optimizer service URL in the services array, and writes the modified JSON back so Knox routes traffic to the current leader.	Not set	/knox/config/descriptors/cdp-proxy.json
dlm.ha.knox.service.name	The Knox service name in the descriptor services array. Must match the name Knox exposes for Cloudera Lakehouse Optimizer.	DLM	DLM
dlm.ha.knox.service.version	The Knox service version in the descriptor services array. Used when Cloudera Lakehouse Optimizer is not yet present in the descriptor.	1.0.0	1.0.0
dlm.ha.passive.redirect	Controls passive-node behavior. When set to true, passive instances redirect API clients to the leader with HTTP 307 Temporary Redirect. When set to false, passive instances return HTTP 503 Service Unavailable.	false	true
dlm.ha.retry-after-seconds	The value for the HTTP Retry-After header returned to clients when a passive instance cannot serve the request.	5	5
dlm.ha.zookeeper.connect	The ZooKeeper connection string used for leader election. All Cloudera Lakehouse Optimizer instances must use the same connection string.	Not set	zk1:2181,zk2:2181,zk3:2181
dlm.ha.zookeeper.delay	The retry delay in milliseconds when the leader election service fails to connect to ZooKeeper.	1000	1000
dlm.ha.zookeeper.latch.path	The ZooKeeper path used for leader election.	Not set	/dlm/leader
dlm.ha.zookeeper.retries	The number of retries when the leader election service fails to acquire the leader latch.	3	3

State persistence for Cloudera Lakehouse Optimizer

State persistence ensures that Cloudera Lakehouse Optimizer operational state such as scheduled tasks, triggers, and maintenance jobs handles service restarts, failures, and network issues. When you restart a Cloudera Lakehouse Optimizer instance, it resumes from persisted state in the shared database.

Cloudera Lakehouse Optimizer stores scheduler and task state in a shared PostgreSQL database. On startup, the service installs or migrates the database schema if needed, loads persisted triggers and tasks, and resumes scheduling. Cloudera Lakehouse Optimizer does not start without a reachable database.

Configuration

Configure the database connection for Cloudera Lakehouse Optimizer in Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* Configuration . Cloudera Lakehouse Optimizer uses this database connection to persist dispatched tasks, event-based triggers, trigger policies, and in-flight Livy task submission data.

Currently, only PostgreSQL is supported. Set dlm_db_type to postgres. Configure the connection using individual host, port, database, and credential properties, or provide a JDBC URL with user and password credentials. For the complete list of properties, see Cloudera Lakehouse Optimizer persistence configuration properties.

Important: All Cloudera Lakehouse Optimizer instances in a high availability deployment must connect to the same shared database. For more information, see Configuring high availability for Cloudera Lakehouse Optimizer.

What gets persisted

Cloudera Lakehouse Optimizer persists the following state in the database:

- **Dispatched tasks**: Maintenance jobs that have been scheduled, including tasks that have not yet started executing. Dispatched tasks that are not yet executing survive a service restart.
- **Event-based triggers**: Custom event-based schedules, next fire times, and associated policies.
- **Trigger policies**: Maintenance policies attached to each trigger.
- **In-flight Livy task submissions**: Livy task submission data and session information for executing tasks, therefore Cloudera Lakehouse Optimizer can restore state after a restart and retrieve results from Livy.

Schedule-based (cron) triggers are rebuilt from Cloudera Lakehouse Optimizer configuration on restart and are not persisted long term.

For the database tables that store persisted state, see Cloudera Lakehouse Optimizer persistence database schema.

How persistence works

Rather than bulk-loading entire state on every change, Cloudera Lakehouse Optimizer uses incremental database operations. Each save, update, or delete runs in a database transaction.

On startup, Cloudera Lakehouse Optimizer checks the database schema version and applies any pending schema migrations automatically. Schema migrations run forward only. Back up the database before you upgrade Cloudera Lakehouse Optimizer.

State lifecycle

On startup: The database schema is installed or migrated if needed, existing triggers and tasks are loaded, and the scheduler begins processing.

During runtime: Tasks and triggers are saved, updated, and deleted in the database because they are scheduled, executed, and completed.

On restart: Pending tasks and triggers are reloaded from the database and scheduling resumes from persisted state.

Integration with high availability

In a high availability (HA) deployment, all Cloudera Lakehouse Optimizer instances share the same database. Only the elected leader writes scheduler state. Standby instances read state but do not modify it. On failover, the new leader reads current state from the database without reloading from individual instances.

Important: High availability for Cloudera Lakehouse Optimizer is in beta. When you enable state persistence and HA together, configure /var/metering on each HA node. State persistence introduces a metering dependency on the HA node.

For more information, see High availability for Cloudera Lakehouse Optimizer.

Limitations

- Cloudera Lakehouse Optimizer does not provide export or import tools for scheduler state. Use native database backup tools such as pg_dump.
- Database must be reachable at startup.
- Database corruption or extended unavailability requires restore from backup or manual intervention.
- Only PostgreSQL is supported for the persistence database.

Cloudera Lakehouse Optimizer persistence configuration properties

Use the following properties to configure state persistence for Cloudera Lakehouse Optimizer.

Table 6: Persistence configuration properties

Property	Description	Default	Example
dlm_db_type	Database type for state persistence. Currently, only PostgreSQL is supported.	Not set	postgres
dlm_db_name	Name of the PostgreSQL database that stores persisted scheduler and task state.	Not set	dlm_persistence
dlm_db_host	Hostname or IP address of the PostgreSQL server.	Not set	postgres.example.com
dlm_db_port	Port number of the PostgreSQL server.	5432	5432
dlm_db_user	Database user name for the PostgreSQL connection.	Not set	dlm_user
dlm_db_password	Password for dlm_db_user.	Not set	********
dlm_db_jdbc_url	Alternative to individual host, port, and database properties. When you use a JDBC URL, also set dlm_db_user and dlm_db_password.	Not set	jdbc:postgresql://postgres.example.com:5432/dlm_persistence

Cloudera Lakehouse Optimizer persistence database schema

Cloudera Lakehouse Optimizer stores persisted scheduler and task state in a shared PostgreSQL database. The following tables store state persistence data:

Table 7: Persistence database tables

Table	Purpose
dispatched_tasks	Scheduled and in-flight maintenance jobs, including task status.
simple_triggers	Event-based trigger schedules.
cron_triggers	Static cron trigger definitions. Recreated from configuration on restart.
trigger_policies	Policy assignments for triggers.
clo_fencing_token	High availability fencing token storage. Used when HA is enabled.
clo_node_ownership	Standalone mode node coordination.
clo_schema_version	Schema version tracking for automatic database migrations.
tasks	Executed tasks
tasks_actions	Executed task actions

Auditing for Cloudera Lakehouse Optimizer

Cloudera Lakehouse Optimizer auditing provides an audit trail of user and system activities on the service, including policy changes, namespace operations, authentication events, and table maintenance actions. Audit events are written as NDJSON files to a configurable filesystem location.

Cloudera Lakehouse Optimizer emits an audit event for every state-changing REST API call and every authentication attempt. Events are buffered in memory and periodically written to audit log files on a Hadoop-compatible filesystem such as HDFS, S3, GCS, Azure Blob Storage, or local FS.

What is audited

Cloudera Lakehouse Optimizer audits the following categories of activity:

- **Policies**: Create, update, delete, view, subscribe, unsubscribe, evaluate, dry-run, pause, and resume operations including failed or denied requests.
- **Namespaces**: Declare, reschedule, delete, pause, resume, list, and view operations.
- **Tables**: Pause, unpause, list, statistics, and registry eviction operations.
- **Tasks**: View, list, and event-ingestion trigger operations.
- **Configuration**: Reconfigure, cache clear, health check, and backup or restore operations.
- **UI**: Read-only browse operations performed through the Cloudera Lakehouse Optimizer UI.

Requests denied by validation or authorization checks are recorded with a FAILURE or DENY outcome before the service returns a response.

Audit event format

Each audit event is a single-line JSON object (NDJSON). The following table describes the audit event fields

Table 8: Audit event fields

Field	Description
timestamp, hostName, serviceName	Set when the audit event is published.
userName, clientIp, role	Read from the authenticated user session. For login and authentication-failure events, these values are supplied manually.
action	Identifies the REST operation that was attempted (for example, CREATE_POLICY_DEFINITION or LOGIN).
outcome	SUCCESS, FAILURE, or DENY.
message	Human-readable description of the event.
Resource fields	policies, namespaces, tables, catalog, resourceUri, and taskId when applicable.

Audit log files

Audit files are written to <basePath>/<prefix>.<sequence>, for example hdfs:///user/dlm/audit-logs/clo-audit.log.42. On HDFS and local filesystems, the default file permission is 775. Files roll over based on configured time and event-count thresholds. On startup, Cloudera Lakehouse Optimizer scans existing audit files and resumes writing to the latest file when appropriate.

Cloudera Lakehouse Optimizer does not delete old audit files. Configure external retention such as HDFS quotas, object-store lifecycle policies, or a log-shipping pipeline.

Important: The default audit URI uses a path such as /tmp, which is not suitable for production. Override dlm.audit.fs.uri to a directory owned by the Cloudera Lakehouse Optimizer service principal. For configuration instructions, see Configuring auditing for Cloudera Lakehouse Optimizer.

Supported filesystems

Cloudera Lakehouse Optimizer supports local, HDFS, and object-store locations through the Hadoop FileSystem API. HDFS supports native append and POSIX file permissions. Object stores such as S3 do not support native append or POSIX permissions; Cloudera Lakehouse Optimizer falls back to read-and-rewrite and relies on bucket policies for access control. Configure bucket versioning, object lock, and restrictive IAM policies for tamper-evident audit storage on object stores.

Audit pipeline behavior

If audit initialization fails, Cloudera Lakehouse Optimizer continues to serve requests. Subsequent audit events are dropped, and the service logs a rate-limited warning with the drop count. Monitor the application log for audit pipeline errors.

For the complete list of auditing configuration properties, see Cloudera Lakehouse Optimizer auditing configuration properties.

Configuring auditing for Cloudera Lakehouse Optimizer

You can configure auditing for Cloudera Lakehouse Optimizer in Cloudera Manager by setting auditing properties on the Cloudera Lakehouse Optimizer service.

Before you begin

Ensure that the target filesystem URI is reachable from all Cloudera Lakehouse Optimizer service hosts and that the Cloudera Lakehouse Optimizer service principal has write access to the audit directory.

About this task

Configure auditing on every Cloudera Lakehouse Optimizer instance in the cluster or Data Hub.

To configure auditing on a Cloudera Lakehouse Optimizer instance:

Procedure

1. In Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* Configuration , set the following auditing properties:
   a) Set dlm.audit.enabled to true to enable auditing.
   b) Set dlm.audit.fs.uri to a production audit directory that is owned by the Cloudera Lakehouse Optimizer service principal.

      For example, hdfs:///user/dlm/audit-logs.
   c) Optionally, configure dlm.audit.push.interval.ms, dlm.audit.fs.rollover.time.ms, dlm.audit.fs.max.events.per.file, and dlm.audit.file.name.prefix based on your operational requirements.
   d) Optionally, configure dlm.audit.fs.file.permission and dlm.audit.fs.dir.permission for HDFS or local filesystem deployments.

   For the complete list of auditing properties, see Cloudera Lakehouse Optimizer auditing configuration properties.
2. Restart the Cloudera Lakehouse Optimizer service.
3. Verify that audit log files are created at the configured URI after you perform a Cloudera Lakehouse Optimizer operation.

What to do next

After you configure auditing:

- Monitor the application log for audit pipeline warnings if events are dropped.
- Configure external retention or log shipping for rolled audit files.
- On object stores, configure bucket versioning and restrictive IAM policies for tamper protection.

Cloudera Lakehouse Optimizer auditing configuration properties

Use the following properties to configure auditing for Cloudera Lakehouse Optimizer:

Table 9: Auditing configuration properties

Property	Description	Default	Example
dlm.audit.enabled	Enables or disables Cloudera Lakehouse Optimizer auditing. When set to false, audit events are not written.	true	true

Property	Description	Default	Example
dlm.audit.fs.uri	The Hadoop-compatible URI where audit log files are written. Override the default path for production deployments.	hdfs:///tmp/clo-audit-logs	hdfs:///user/dlm/audit-logs
dlm.audit.publisher	The audit publisher type. Only fs is supported.	fs	fs
dlm.audit.push.interval.ms	The interval in milliseconds at which buffered audit events are written to the audit log. Events can also be written when the buffer reaches capacity.	600000 (10 minutes)	600000
dlm.audit.fs.rollover.time.ms	The time-based rollover threshold in milliseconds. When this interval elapses, Cloudera Lakehouse Optimizer starts a new audit log file.	86400000 (24 hours)	86400000
dlm.audit.fs.max.events.per.file	The maximum number of audit events per file before rollover. Set to 0 to disable event-count rollover.	65536	65536
dlm.audit.file.name.prefix	The filename prefix for audit log files. Cloudera Lakehouse Optimizer appends a period and sequence number. Do not include a trailing period in the property value.	clo-audit.log	clo-audit.log
dlm.audit.fs.file.permission	The POSIX permission mode (octal) applied to each audit file on HDFS and local filesystems.	775	775
dlm.audit.fs.dir.permission	The POSIX permission mode (octal) applied to the audit base directory on HDFS and local filesystems.	750	750

Lakehouse Optimizer UI

You can create Cloudera Lakehouse Optimizer policies, view the Iceberg tables that are under maintenance, view the policy details, and view the maintenance status for each table on the Lakehouse Optimizer UI. You can also pause and resume table maintenance.

Deploying and configuring Cloudera Lakehouse Optimizer UI

On Cloudera Base on premises, you can deploy the Lakehouse UI parcel in Cloudera Manager, add the Lakehouse UI service to your cluster, and configure the Knox Gateway URL to access the Lakehouse Optimizer UI.

Before you begin

Before you deploy the Lakehouse UI parcel, ensure that the following requirements are met:

- Cloudera Manager is installed and you can log in as a cluster administrator.
- Apache Knox is installed, running, and accessible on your cluster.
- The cluster contains the dependent services required by Cloudera Lakehouse Optimizer, including HDFS, Hive, Spark 3, YARN, Ranger, and Livy for Spark 3.

About this task

To deploy and configure the Lakehouse Optimizer UI on Cloudera Base on premises:

Procedure

1. In Cloudera Manager Parcels Parcel Repositories & Network Settings , add the Lakehouse UI parcel repository URL.
   a) In Remote Parcel Repository URLs field, enter the following URL:
      https://archive.cloudera.com/p/lakehouse-ui/1.0.0/parcels/
   b) Click Save & Verify Configuration.
      Cloudera Manager validates the repository and downloads the parcel manifest.
2. For the LAKEHOUSE_UI parcel, click Download, Distribute, and Activate in sequence.
   The parcel status changes to Distributed, Activated.

3. Add the Lakehouse UI service to your cluster.
   a) On the Cloudera Manager Home page, select the three dots menu, and click Add Service.
   b) Select Lakehouse UI and click Continue.
4. On the **Select Dependencies** page, review the dependent services and click Continue.
5. Assign the Lakehouse UI Server role to a host.
   a) On the **Assign Roles** page, assign the Lakehouse UI Server role to a host in your cluster.
   b) Click Continue.
6. Configure the Knox Gateway Base URL.
   a) On the **Review Changes** page, set Knox Gateway Base URL (knox_host) to the base URL of your Knox Gateway instance.

      For example, https://[***knox-host.example.com***]:8443.
   b) Click Continue.
7. Complete the Lakehouse UI service setup.
   a) On the **Command Details** page, wait for the First Run Command to finish. The command deploys Knox artifacts for Lakehouse UI.
   b) Click Continue.
   c) On the **Summary** page, review the installation status and click Finish.
8. Start the Lakehouse UI service if it is not already running.
   a) In Cloudera Manager Clusters Lakehouse UI , select the three dots menu, and click Start if the service is stopped.
9. Access the Lakehouse Optimizer UI through Knox Gateway.
   a) In a browser, open the Lakehouse Optimizer UI using the following Knox Gateway URL:

      https://[***knox-host.example.com***]:8443/gateway/cdp-proxy/lakehouse-ui/
   b) Alternatively, in Cloudera Manager Clusters Lakehouse UI , open the Lakehouse UI quick link.

What to do next

After you deploy the Lakehouse UI parcel and access the UI, complete the prerequisites in Preparing to create Cloudera Lakehouse Optimizer policies in UI before you create policies. For more information about managing parcels, see Managing Parcels.

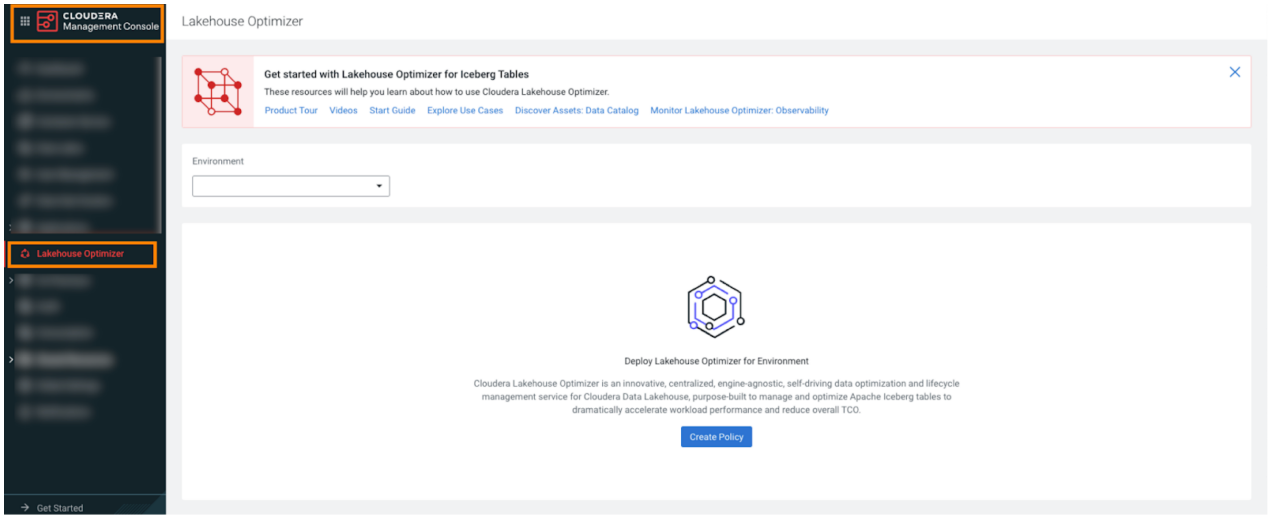
Access Lakehouse Optimizer UI

Access the Lakehouse Optimizer page on "Cloudera on cloud Management Console". You can create Cloudera Lakehouse Optimizer policies, associate them with namespaces, and view the policy details in the Lakehouse Optimizer UI.

Note: Ensure to also have the *EnvironmentUser* role to access the Lakehouse Optimizer UI.

The following image displays the Lakehouse Optimizer page when you visit the page for the first time:

Figure 1: Lakehouse Optimizer page



When you access the Lakehouse Optimizer UI for the first time, the following elements are displayed:

- **Get Started banner**: Provides links to various pages.
- **Environment**: Choose the environment where you configured the Lakehouse Optimizer Data Hub. The current **Status** of the environment and **Quick links** to other services are displayed.

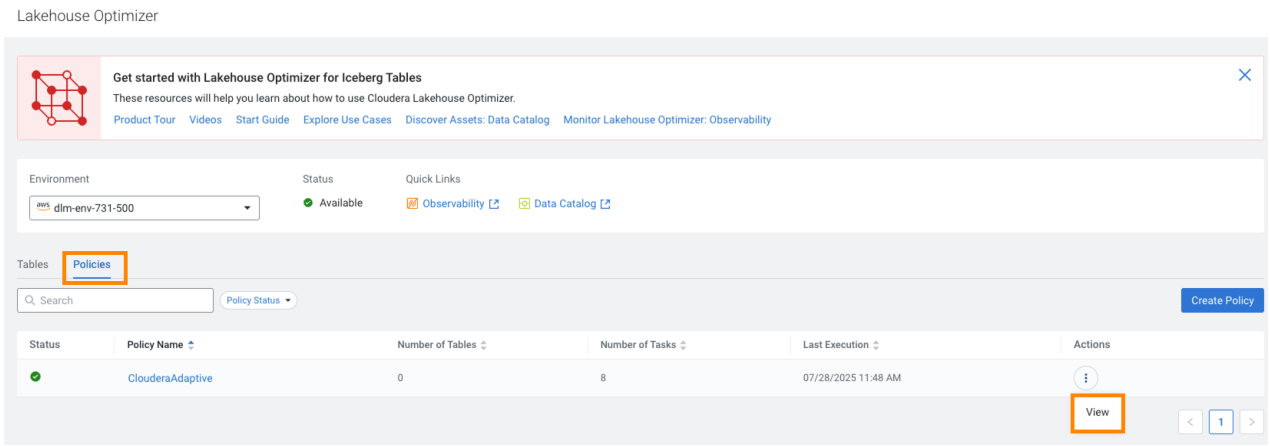
After you choose an environment, the following elements are displayed:

- **Create Policy**: Click to create a new Cloudera Lakehouse Optimizer policy. For more information, see Creating a Cloudera Lakehouse Optimizer policy in Lakehouse Optimizer UI.
- **Policies**: Click to view the list of existing policies for the chosen environment.
- **Tables**: Click to view the tables in the chosen namespace. Available only in Cloudera on cloud 7.3.1.500 and higher versions.
- **Associations**: Click to view the tables associations and namespace associations. Available in Cloudera on cloud 7.3.2 and higher versions.

Policies tab

The **Policies** tab displays the list of existing policies for the chosen environment. Search for a policy, or filter the policies based on the **Policy Status**.

The following image displays the **Policies** tab on the Lakehouse Optimize page:



When you open the **Policies** tab for the first time, before you create a policy, the default policy named ClouderaAdaptive is displayed. Lakehouse Optimizer uses the default policy as a template to create the policies.

The **Policies** tab displays the following details:

- The current **Status** icon for the table maintenance activity of the tables associated with the policy.

- The **Number of Tables** that are associated with the policy.
- The **Number of Tasks** or maintenance tasks that ran as a result of policy evaluation on the associated tables.
- The **Last Execution** or the last policy run timestamp.
- The  icon for each policy displays the actions menu. Select the View action to see the policy details on the View Policy page.

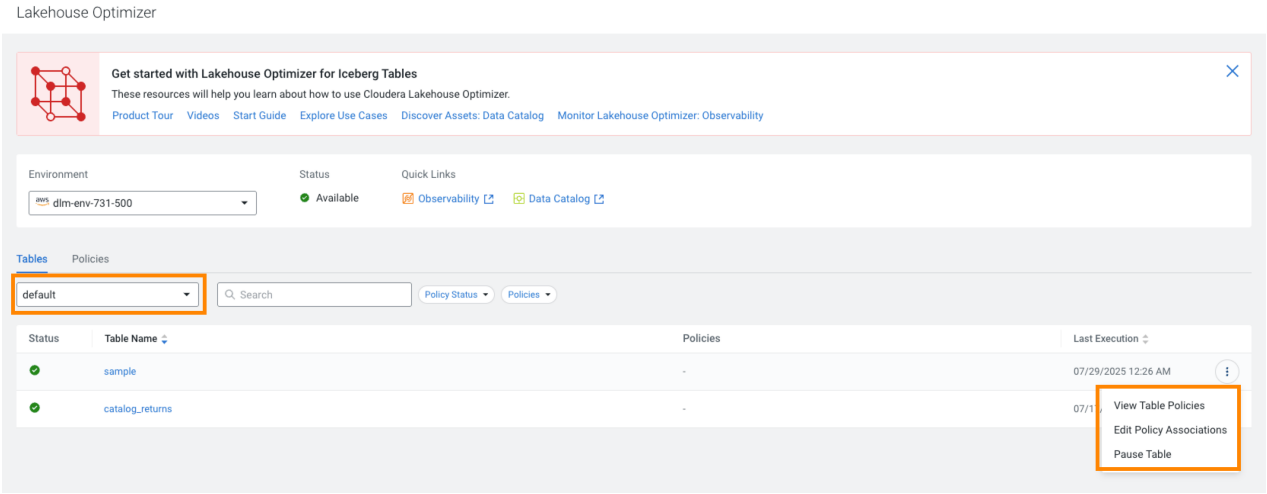
Tables tab

In Cloudera on cloud 7.3.1.500 and higher versions, the Tables tab in Cloudera Lakehouse Optimizer provides the list of all the Iceberg tables in the selected namespace, including tables undergoing table maintenance and tables in Pause state, and the table maintenance details.

The namespaces that you declare for Cloudera Lakehouse Optimizer populate the namespaces drop-down list, allowing you to choose the required namespace to view its tables.

The following image shows the **Tables** tab on the Lakehouse Optimizer page, which lists the tables in the default namespace:

Figure 2: Tables tab



The Tables tab displays the following details:

- The current table maintenance task **Status** icon.
- The **Table Name**.
- The **Policies** that the table is associated with.
- The **Last Execution** timestamp for each table.

You can select the following actions, when you click  for a table:

- **View Table Policies** that shows the list of policies associated with the table on the **Table Policies** side panel. The panel is also displayed when you click the table name on the **Tables** tab.
- **Edit Policy Associations** that opens the **Edit Policy Associations** modal window. The option is also displayed on the **Table Policies** side panel.
- **Pause Table** that pauses the maintenance tasks that are scheduled to be run on the table. You can click Unpause Table later to pause a table when required.

> **Important:** When you pause maintenance for a table or namespace, all the policies for that table or namespace are paused automatically. Therefore, you must be cautious before you pause maintenance for a table or a namespace.

The following table describes the status icons displayed on the Tables tab:

Table 10: Status icon descriptions

Status icon	Status description
	The last policy run for the table completed successfully.
	The last policy run for the table failed.
	One of the following conditions is true: • You paused the table maintenance. • Cloudera Lakehouse Optimizer paused the table maintenance because the recurring failures during the execution phase for the policy exceeded the retry value. For instructions about setting the retry value, see Pause and resume table maintenance.
	No tasks have run on the table.

Associations tab

In Cloudera on cloud 7.3.2 and higher versions, the Associations tab replaces the Tables tab and provides Tables Associations and Namespaces Associations subtabs for managing maintenance tasks and policy links.

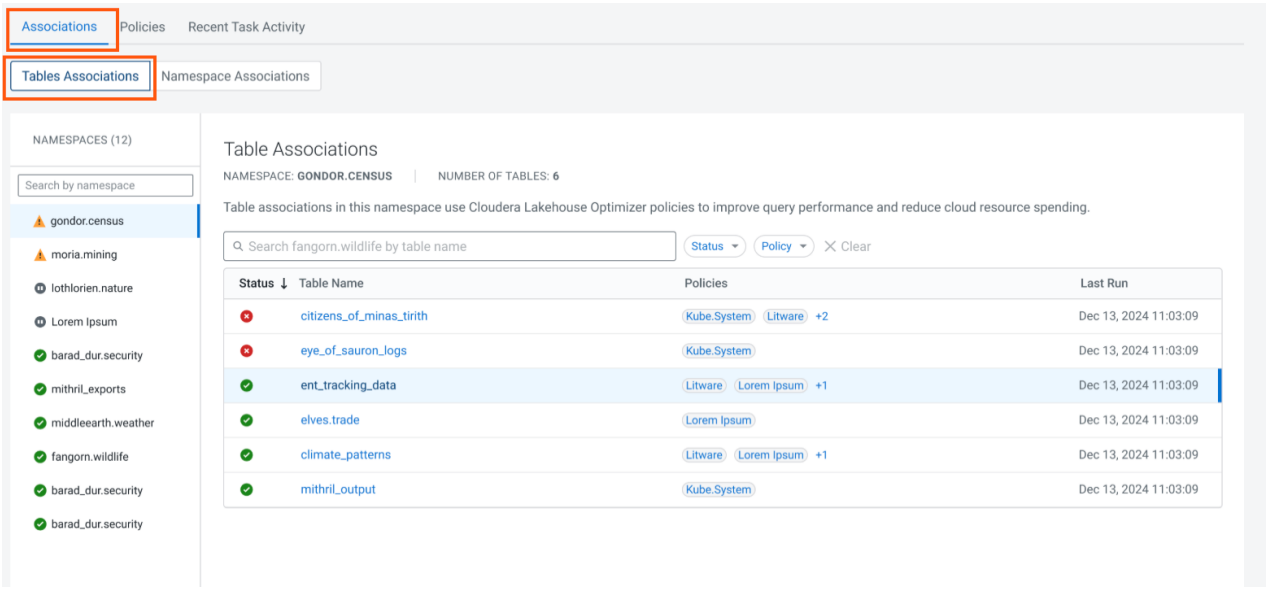
Tables Associations sub tab

The tab displays all the available namespaces. Select a NAMESPACE to view the total NUMBER OF TABLES in the namespace and the following details for each table in the namespace:

- The current Status icon of the maintenance task
- The Table Name. Click to open the **Table Policies** side panel.
- Policies that the table is associated with. Click to open the **View Policy** page.
- Last Run timestamp

The following image shows the **Table Associations** tab:

Figure 3: Table Associations tab



Optionally, you can search for tables using keywords, task Status, or Policy name.

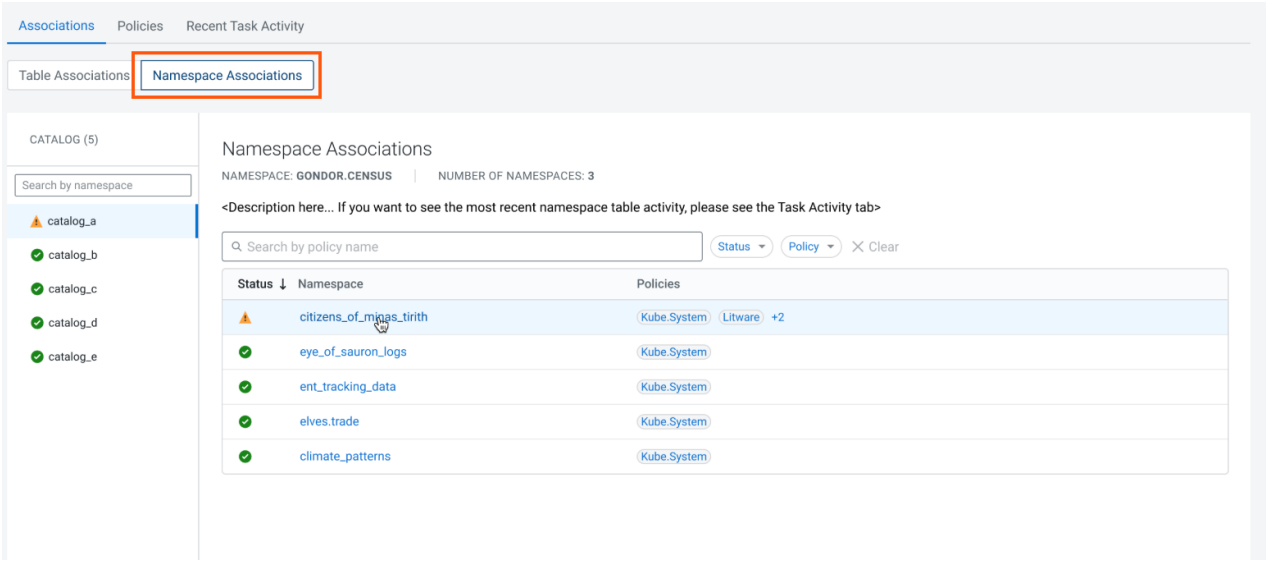
Namespace Associations sub tab

The tab displays all the available catalogs. Select one or more catalogs, or search for the required catalog to view the total **NUMBER OF NAMESPACES** in the catalogs, and the following details about the namespace associations in it:

- The current Status icon of the maintenance tasks in the namespace
- The Namespace name
- Policies that are associated with the namespace. Click to open the **View Policy** page.

The following image shows the Namespace Associations tab:

Figure 4: Namespace Associations tab



Optionally, you can search for the policy names using keywords, task Status, or Policy name.

Recent Task Activity tab

The Recent Task Activity tab lists a maximum of 100 recent tasks run by Cloudera Lakehouse Optimizer.

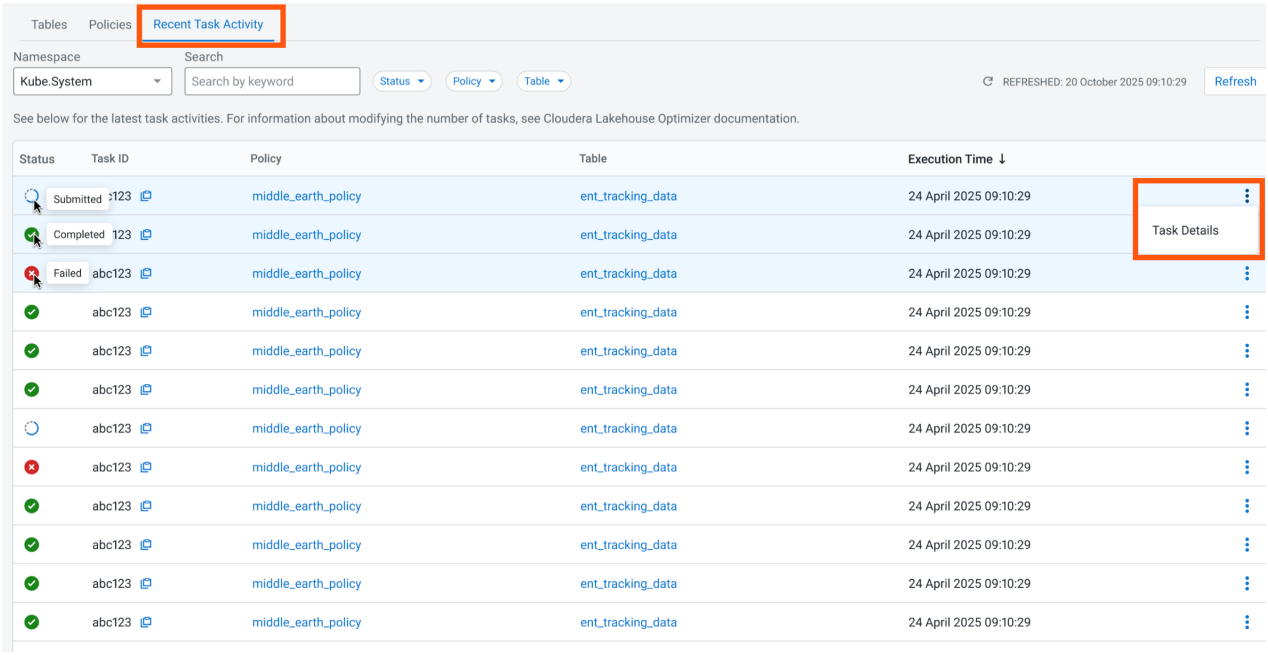
The tab displays the Status icon, Task ID that you can copy, Policy name, Table name, and the Execution Time for each task. You can go to Actions Task Details to view more details about the task in the **Task Table Details** side panel.

The Recent Task Activity tab provides the following options:

- You can select a Namespace to view the recent maintenance tasks run on the Iceberg tables in it.
- You can enter a policy name or table name as a keyword in the Search field to view its task details.
- Select either Status, Policy, or Table or a combination of these to display the required task details.

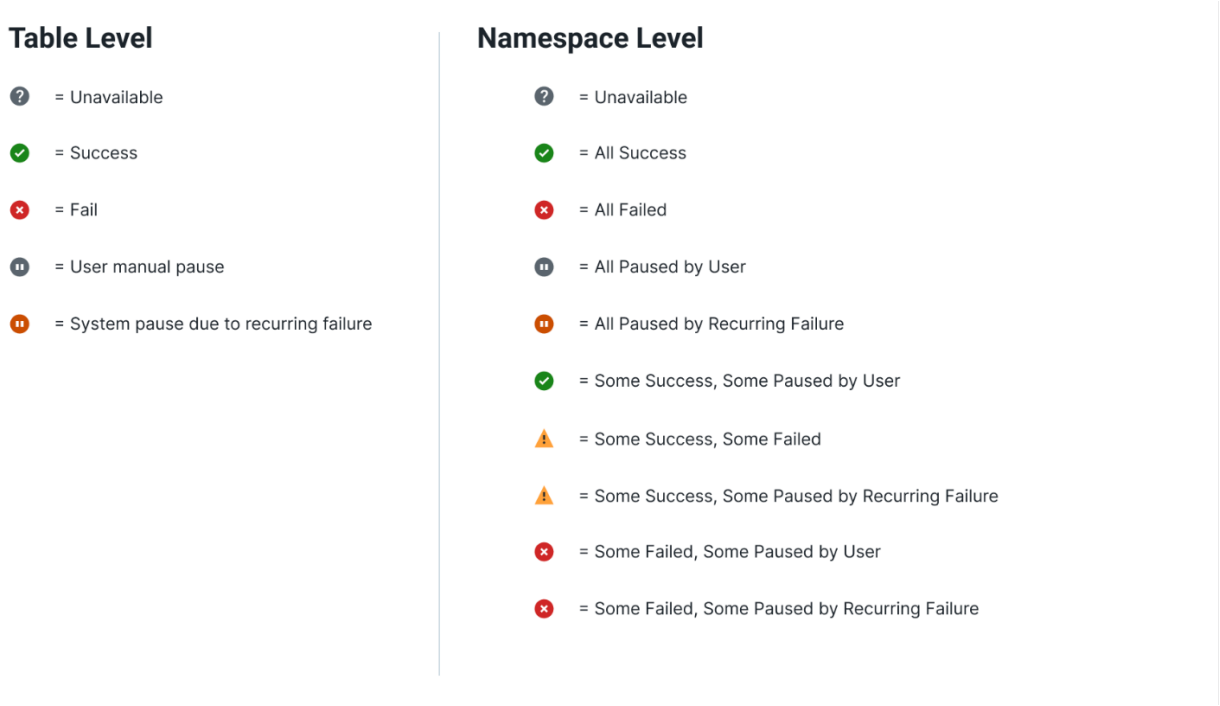
The following page shows the Recent Task Activity tab:

Figure 5: Recent Task Activity tab



The following image shows all the Status icons that are available at table level and namespace level on the Recent Task Activity tab. These icons provide detailed information to help you understand the current status of the task and to take appropriate action to mitigate any issues.

Figure 6: Status icons at table and namespace level



Understanding the side panels in Lakehouse Optimizer UI

Several modal windows or side panels are available in the Lakehouse Optimizer UI that provide more information about the relevant components.

Table Policies

The side panel is displayed when you click one of the following options:

- A table name on the Tables tab or Tables Associations subtab.
- The Tables tab *[***TABLE NAME***]* Actions View Table Policies option.
- A table on the **View Policy** page.

The side panel displays the following details:

- The table name and the list of policies it is associated with.
- The Edit Policy Associations button that opens the **Edit Policy Associations** modal window.
- Policy details, including the Policy Name, the Policy Status icon, the Number of Tasks that have run on the table, the Last Execution timestamp, and the Policy Actions menu.

  The Policy Actions menu in Cloudera on cloud 7.3.2 and higher versions displays the following options:

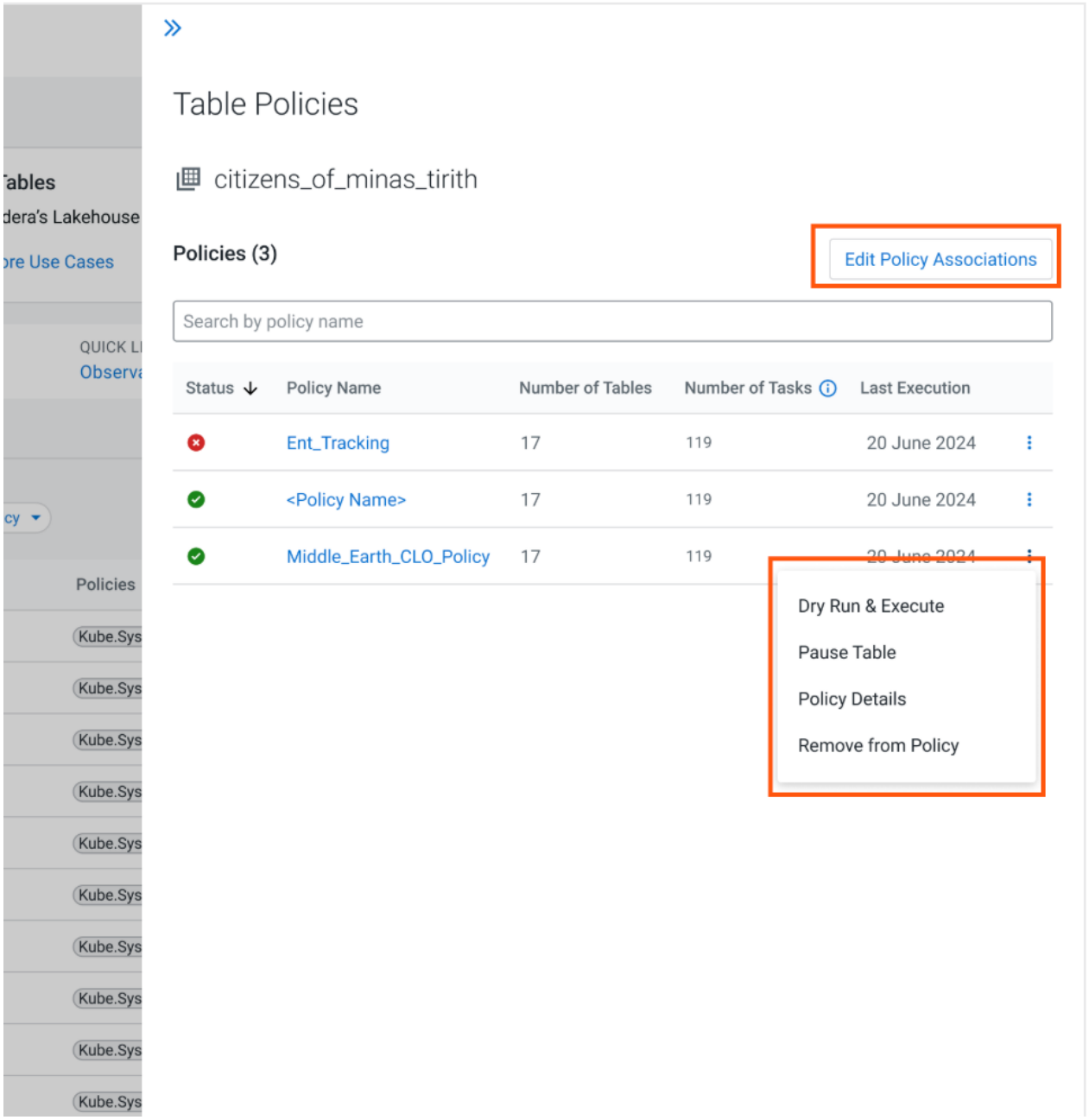
  - The Dry Run & Execute action that opens the **Dry Run & Execution** modal window or side panel.
  - The Pause Table action to pause the maintenance tasks on the table.

    Important: When you pause maintenance for a table or namespace, all the policies for that table or namespace are paused automatically. Therefore, you must be cautious before you pause maintenance for a table or a namespace.
  - The Policy Details action that opens the **View Policy** page.
  - The Remove From Policy action that removes the table from the chosen policy.

  In Cloudera on cloud 7.3.1.500 and higher versions, the Policy Actions menu displays the View Policy action that opens the **View Policy** page, and the Remove From Policy action that removes the table from the chosen policy.

The following image shows the **Table Policies** side panel:

Figure 7: Table Policies side panel



Edit Policy Associations

The side panel is displayed when you click one of the following options:

- The Tables tab *[***TABLE NAME***]* Actions Edit Policy Associations option.
- The Associations tab Tables Associations sub tab *[***TABLE NAME***]* Actions View Table Policies Edit Policy Associations option.
- The Actions Edit Associations option on the **View Policy** page

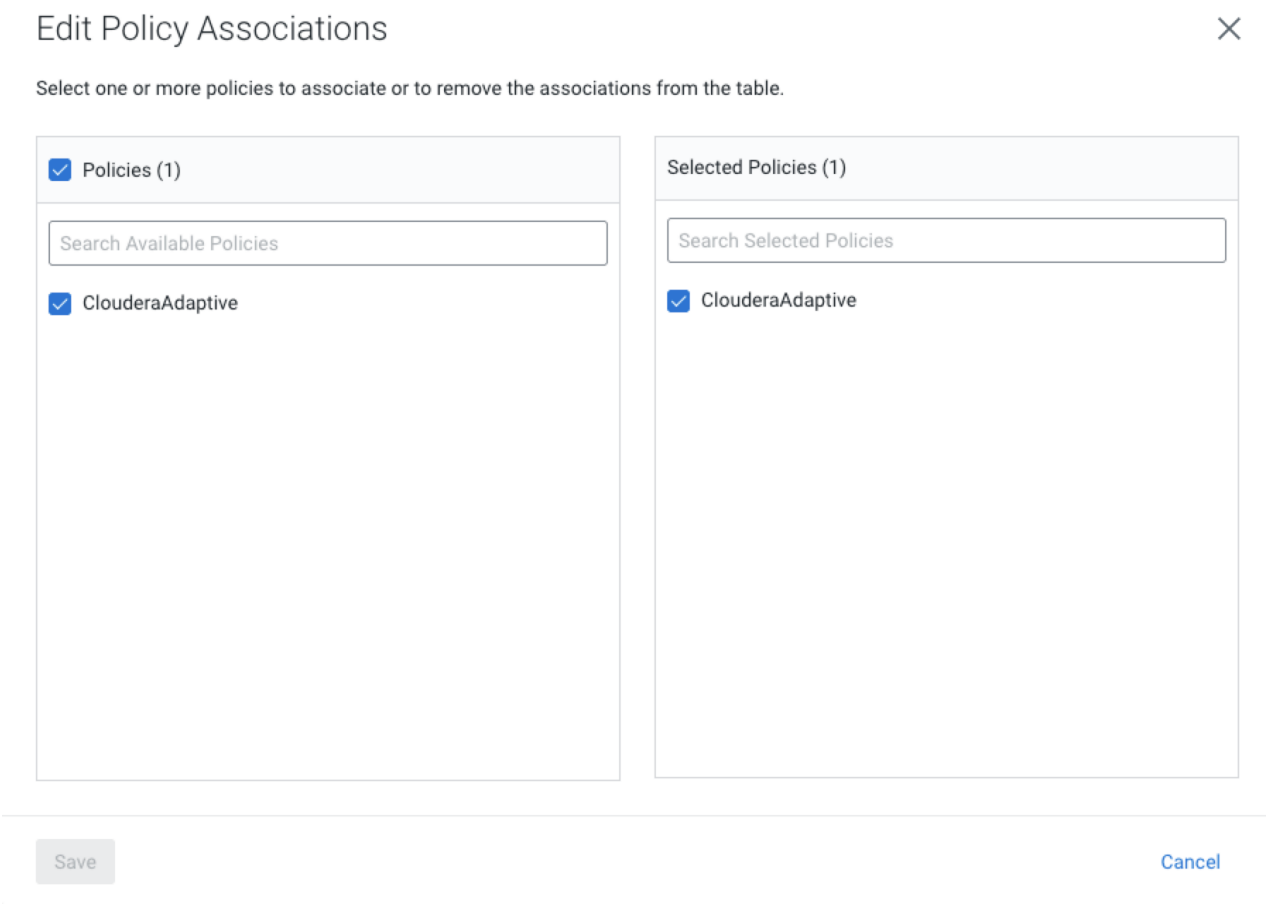
On the **Edit Policy Associations** modal window, you can perform the following actions:

- Select one or more policies to associate with the table.
- Remove policy associations from the table.
- Search for available policies to create associate with the table.
- Search for policies currently associated with the table.

- Save the changes.

The following image shows the **Edit Policy Associations** modal window:

Figure 8: Edit Policy Associations modal window



Namespace Details

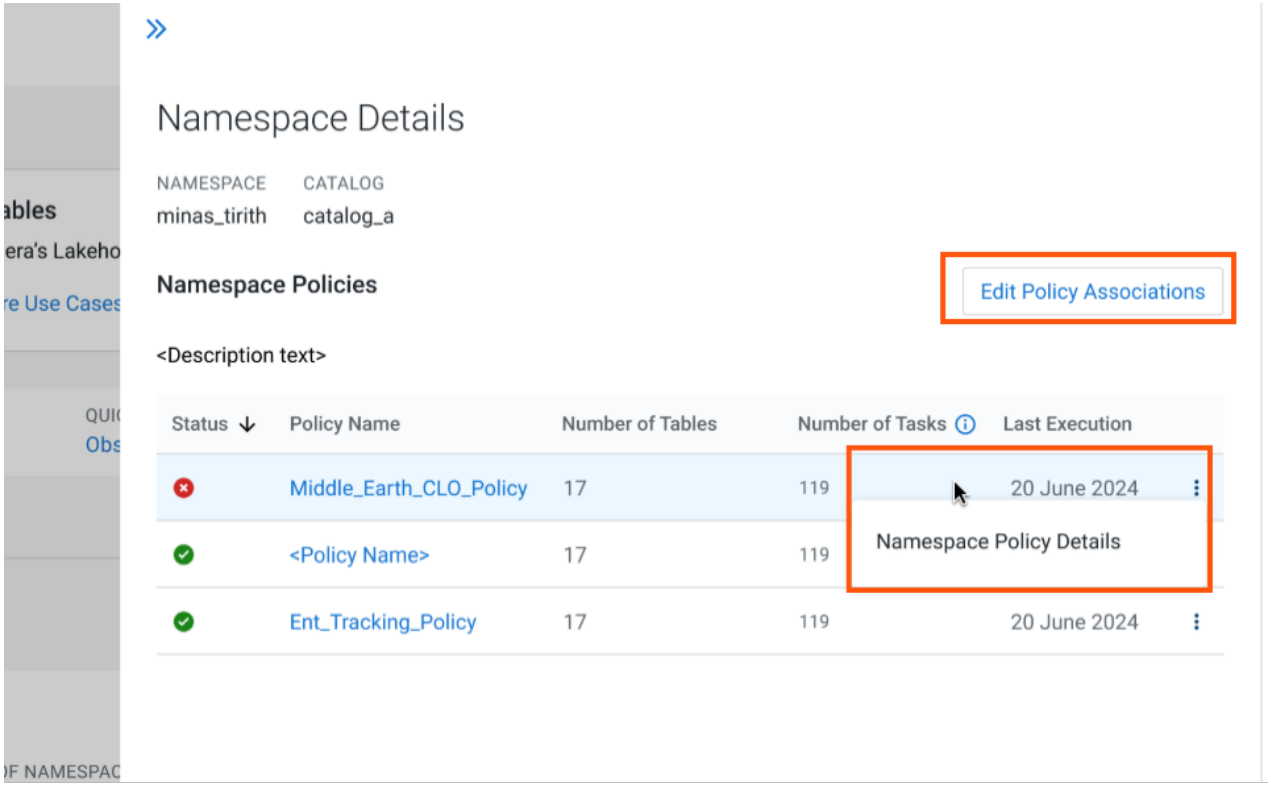
The side panel is displayed when you click a namespace on the Namespace Associations subtab.

The side panel displays the following details:

- The namespace name, catalog name, and the list of policies it is associated with.
- The Edit Policy Associations option that opens the **Edit Policy Associations** modal window.
- The namespace details, including the Status icon, the Policy Name, the Number of Tables in the policy, the Number of Tasks that have run on the policy, and the Last Execution timestamp.
- The  icon displaying the Namespace Policy Details option that opens the **View Policy** page.

The following image shows the **Namespace Details** side panel:

Figure 9: Namespace Details side panel

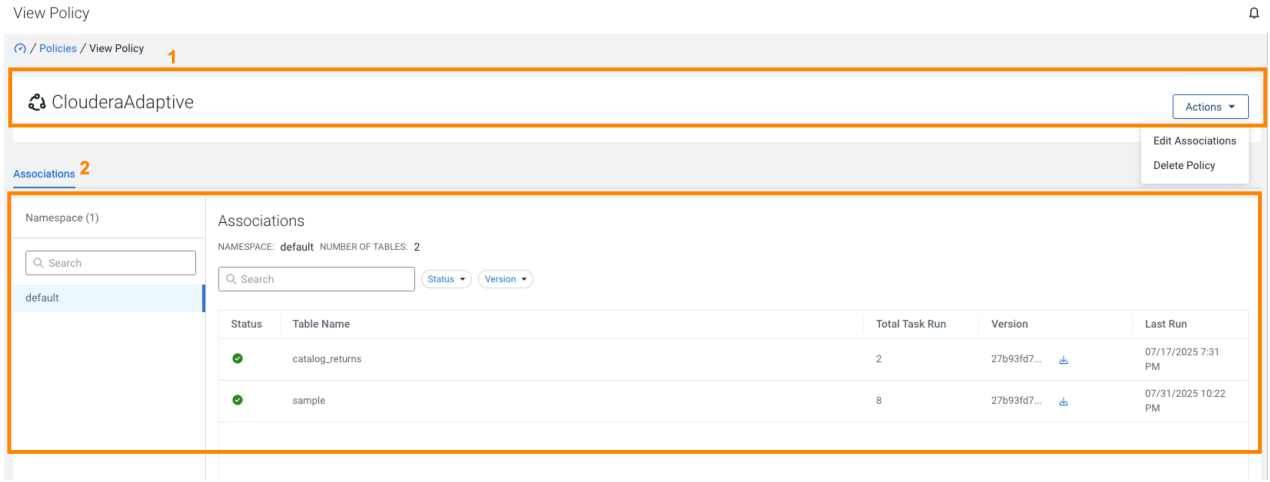


View Policy page

The View Policy page is displayed after you click a policy on the Policies tab, or when you click View Policy for a policy on the Table Policies side panel.

The following image shows the **ClouderaAdaptive** policy page:

Figure 10: ClouderaAdaptive policy page



The **View Policy** page displays the following sections:

1. The **Policy banner** section displaying the policy name and the Actions menu.

   Click the Actions menu to access the following actions:

   - Edit Associations – This action opens the **Edit Associations** side panel which displays the Policy name, the Namespaces, and the tables associated with the policy. You can choose one or more namespaces to associate with the policy.

- Delete Policy – This action removes the policy permanently. Ensure that you have a backup plan to maintain the tables that were associated with this policy.

2. The **Associations** section contains the following panels:

- The **Namespace** panel lists all the namespaces that contain the tables associated with the policy, which might be an entire namespace in a policy or a few tables in the namespace.

  Search for a namespace name to verify whether the namespace or the tables in the namespace are associated with the policy, or to view more details about the namespace in the **Associations** panel.
- The **Associations** panel displays all the tables in the selected namespace that are associated with the policy; the NAMESPACE name, and the NUMBER OF TABLES in the namespace that are associated with the policy.

  The **Associations** panel includes the following table details:

  - The current task maintenance Status icon.

    The Status icon colors indicate the following statuses:

    - Green, when the last policy run for the table completed successfully
    - Red, when the last policy run for the table failed
    - Grey, when the table maintenance has been paused either by the user manually or because the recurring failures during the execution phase for the policy exceeded the retry value
  - The Table Name
  - The Total Task Run
  - The Version

    Click the download icon to download the policy definitions applicable to that table.
  - The Last Run timestamp.

  You can enter the namespace name in the Search field, or filter the list based on the Status or the Version to view more details about the namespace-policy associations in the **Associations** panel.

Preparing to create Cloudera Lakehouse Optimizer policies in UI

Before you create the Cloudera Lakehouse Optimizer policies, the administrator must complete a few prerequisites. The administrator must perform the necessary steps to install the Cloudera Lakehouse Optimizer Data Hub and complete the other required configuration steps to initiate Iceberg table maintenance. You can use the Lakehouse Optimizer UI or REST APIs to create and maintain the policies.

Before you begin

Consider the best practices before you create or define the policy.

About this task

The following section explains the steps to complete before you create Cloudera Lakehouse Optimizer policies using the Lakehouse Optimizer UI.

Procedure

1. Ensure that you do not have any other optimization services enabled alongside Cloudera Lakehouse Optimizer, such as AWS S3 Table Optimization, to avoid conflicts during the optimization process.
2. The CDP Environment administrator must perform the following steps to install and configure Cloudera Lakehouse Optimizer:
   a) Contact your Cloudera account team to enable the Cloudera Lakehouse Optimizer service for your Cloudera Open Data Lakehouse environment.
   b) Ensure that the environment has the following minimum configuration to support Cloudera Lakehouse Optimizer:

   - The AWS environment must have 1x m5.4xLarge (Master Node); 2x r5d.xLarge (Worker Nodes); r5d.xLarge (Compute Nodes - 0 by default).

- The Azure environment must have 1x Standard_D16d_v5 (Master Node)- 2x Standard_E8ds_v5 (Worker Nodes)- 0x Standard_E8ds_v5 (Compute Nodes - 0 by default).

Important: Enable autoscaling on the compute node before you use Cloudera Lakehouse Optimizer.

c) Modify the default values for Spark Executor Memory and Spark Driver Memory. The default values are 8 GB and 4 GB respectively. The default memory settings might be enough for a majority of the use cases. However, for heavy workloads you might want to increase these values.

To modify the spark.driver.memory and spark.executor.memory settings, go to the Cloudera Manager Clusters *CLOUDERA_LAKEHOUSE_OPTIMIZER* Configuration conf/dlm-client.properties_role_safety_valve property.

3. Cloudera Lakehouse Optimizer administrators must perform the following actions before creating the policies:
   a) Verify whether you can access the Lakehouse Optimizer page in Cloudera Management Console.
   b) Verify whether the ClouderaAdaptive default policy is available on the Cloudera Management Console Lakehouse Optimizer Policies tab.

What to do next

Create an event-based policy or a schedule-based policy to initiate Iceberg table maintenance. For more information, see Creating a Cloudera Lakehouse Optimizer policy in Lakehouse Optimizer UI.

Creating a Cloudera Lakehouse Optimizer policy in Lakehouse Optimizer UI

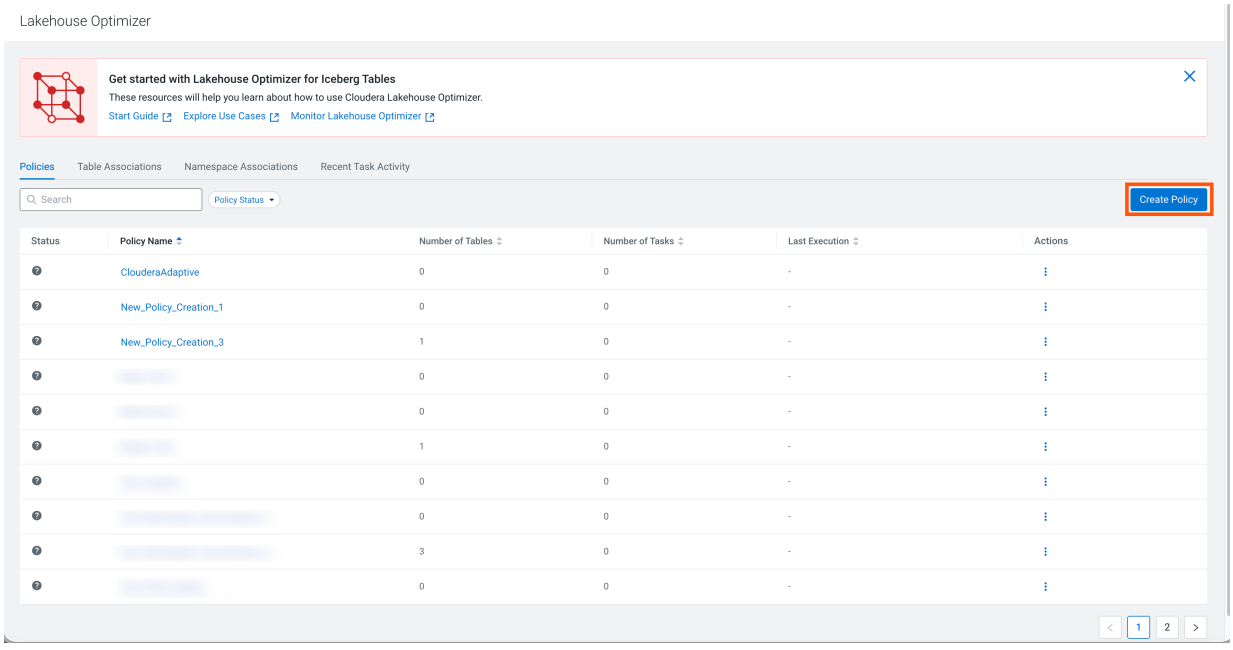
After you complete the prerequisites, you can create the Cloudera Lakehouse Optimizer policies in the Lakehouse Optimizer UI. You can create one or more policies for a table to perform Iceberg table maintenance.

About this task

To create a policy in the Lakehouse Optimizer UI, perform the following steps:

Procedure

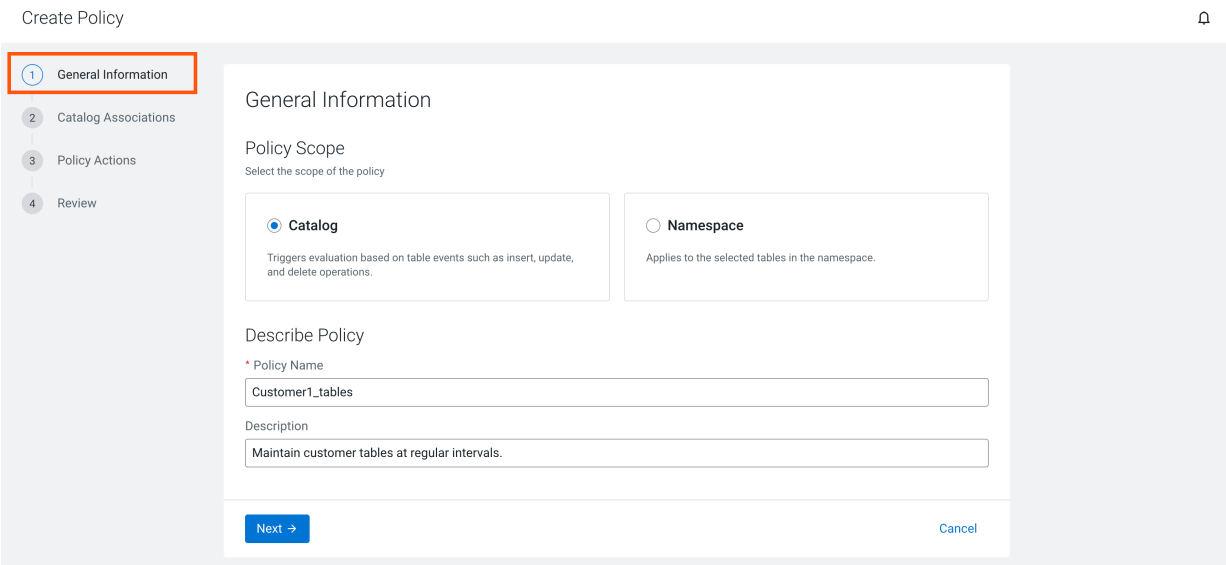
1. Ensure that you have completed the steps in Preparing to create Cloudera Lakehouse Optimizer policies in UI on page 104.
2. Go to the Cloudera Management Console Lakehouse Optimizer page.
3. Click Create Policy to open the **Create Policy** wizard.



4. Enter the following details on the **General Information** tab:

Option	Description
Catalog	Choose the option to apply this policy to all the tables in all the selected namespaces.
Namespace	Choose the option to apply this policy to all the selected tables in the specified namespace.
Policy Name	Enter a unique policy name so that you can identify it later when required. Enter an underscore instead of spaces in the policy name. **Note:** If you enter a space between words in the policy name, the whitespace is automatically replaced with an underscore later on during the policy creation process.
Description	Optional. Enter a brief description that describes the table maintenance goal, and any other information that might help you and other users.

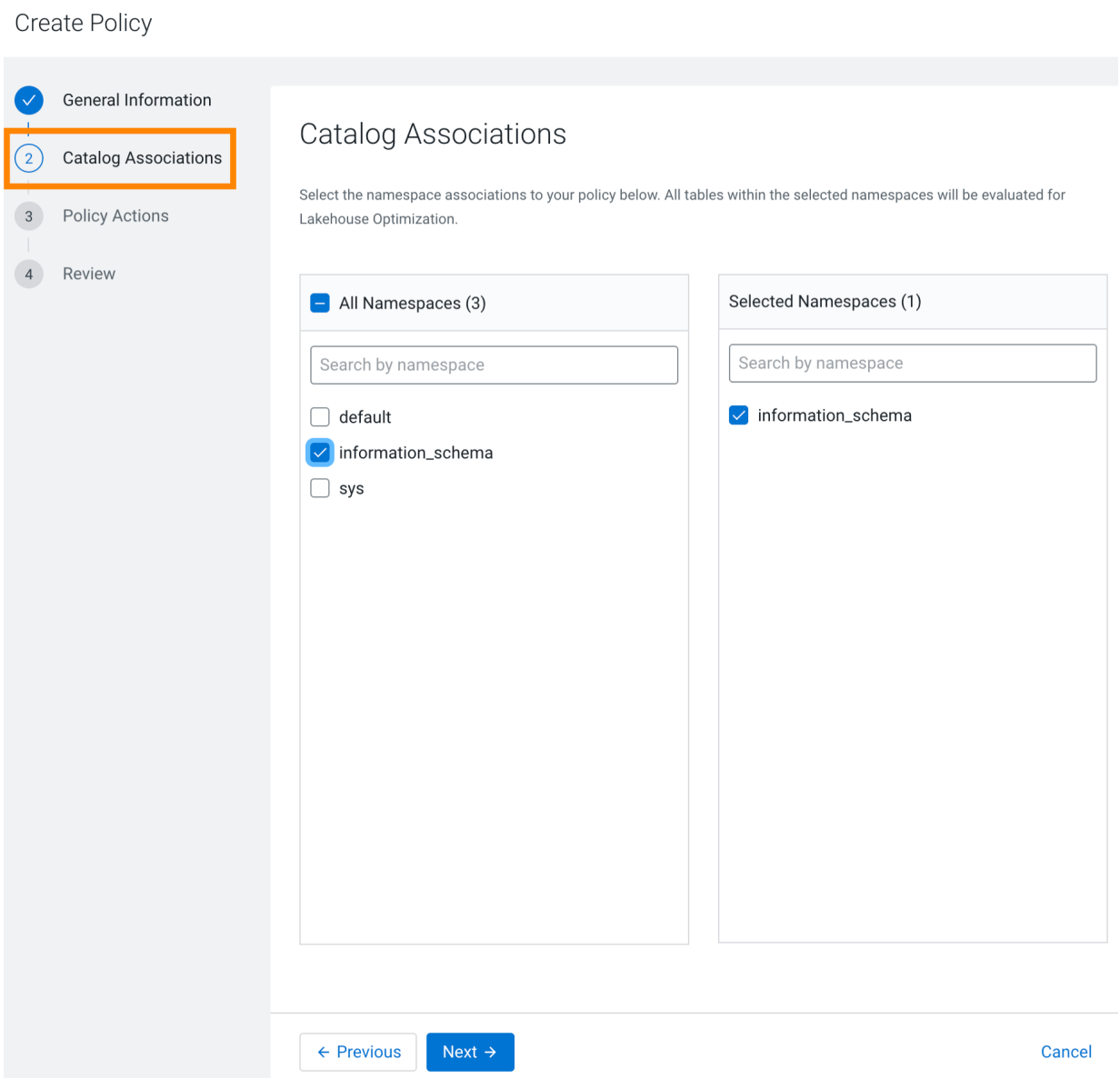
The following image shows the **General Information** tab:



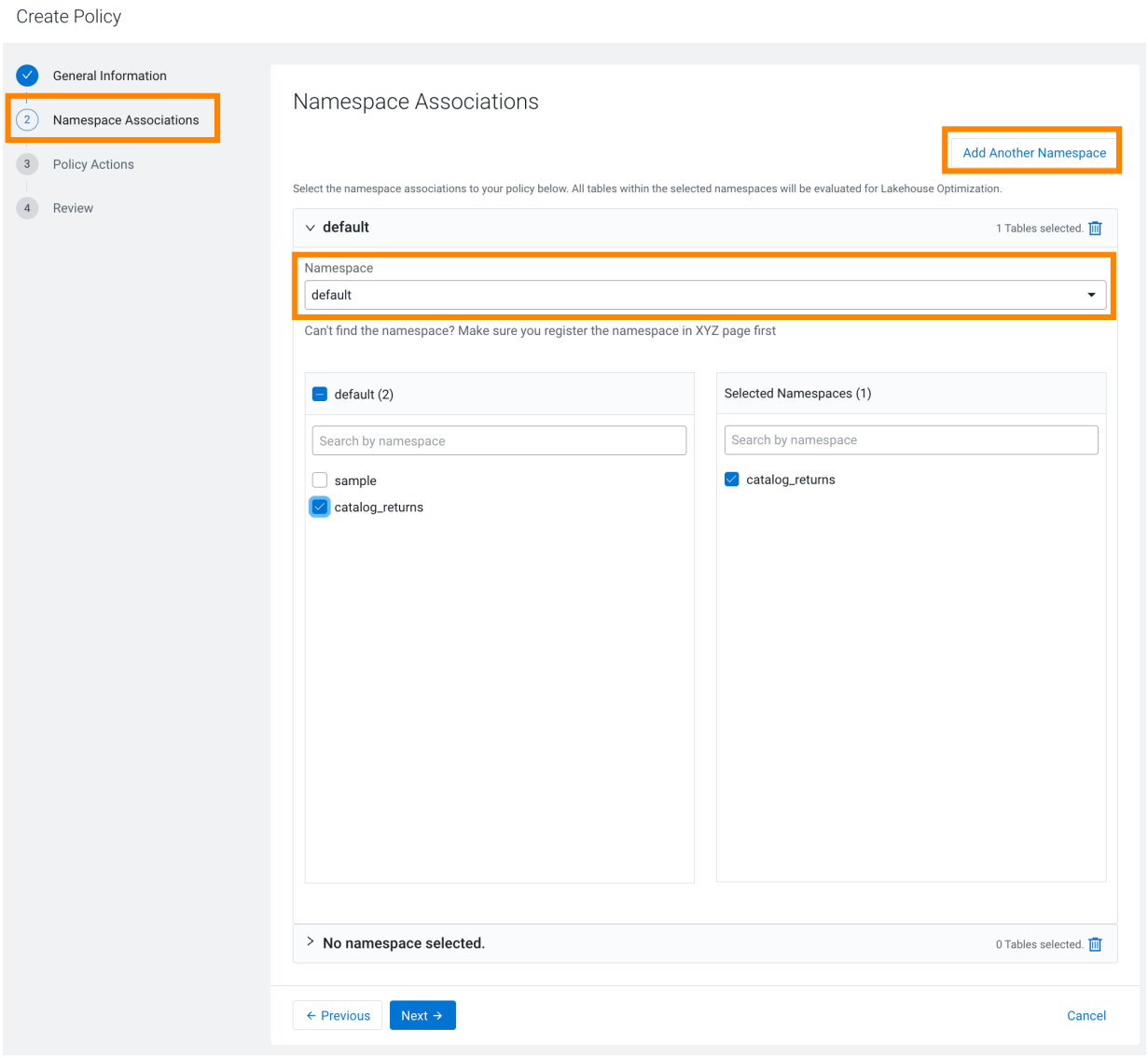
5. Click Next.
6. Enter the details in the next tab; the tab details appear depending on whether you chose Catalog or Namespace on the **General Information** tab:

Choice chosen on the General Information tab	Options to select or choose
Catalog	Choose one or more namespaces to associate with the policy on the **Catalog Associations** tab.
Namespace	Choose or **Add Another Namespace** to associate with the policy. You can choose the tables in the namespace to associate with the policy.

The following image shows the **Catalog Associations** tab and the tab details:



The following image shows the **Namespace Associations** tab and the tab details:



7. Click Next.
8. Enter the following details on the **Policy Actions** tab:

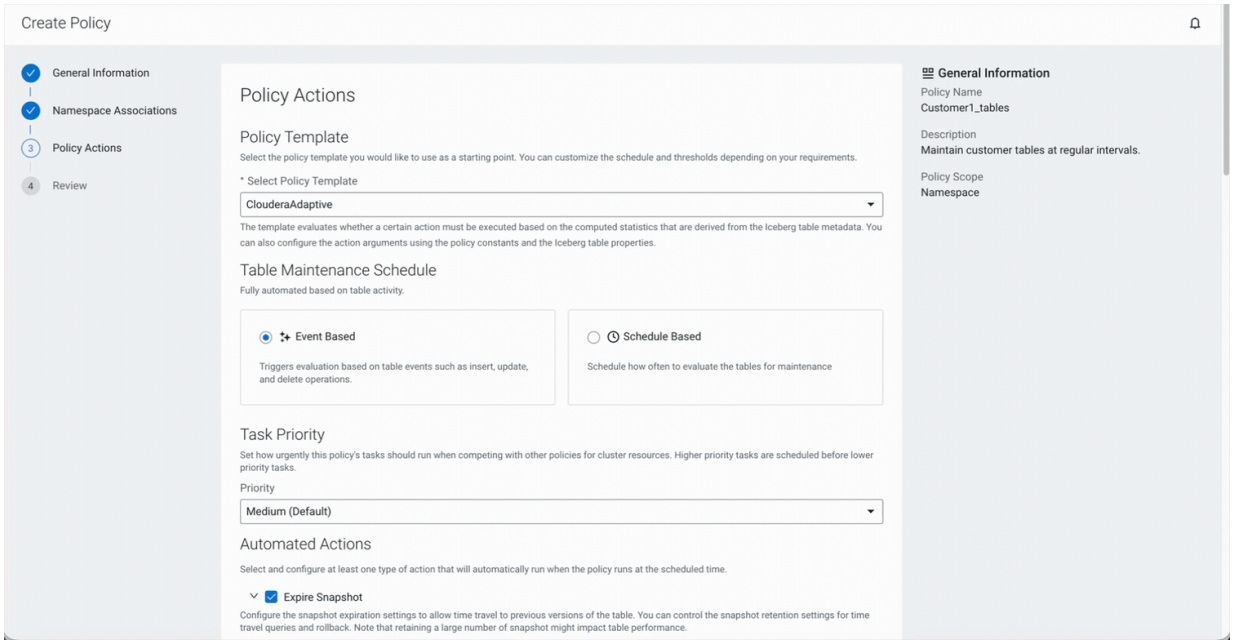
Option	Description
Select Policy	Choose *ClouderaAdaptive*. This is the default policy that you use as a template to create the new policy.
Table Maintenance Schedule	Choose Event Based to trigger the policy evaluation when the HMS events, such as insert, delete, and update operations complete.

Option	Description
	Choose Schedule Based to trigger the policy evaluation at regular intervals. You can choose the following options to run the policy: You can choose the following options to run the policy: • Select Frequency as Daily, Weekly on <current week day>, Monthly on the first current <week day>, Yearly on <today's date>, Every weekday (Monday to Friday), or Custom CRON scheduler. You can generate a CRON expression using the https://www.freeformatter.com/cron-expression-generator-quartz.html page. For example, the 0 */30 * ? * * expression schedules the policy to run every 30 minutes; the 0 0 */4 ? * * expression schedules the policy to run for every 4 hours; the 0 0 0 * * ? expression schedules the policy to run every day at midnight. • Choose the Time Zone. • Select the Hour. • Select the Minutes. **Note:** When you use multiple policies for a table, there might be scenarios where two concurrent tasks on a single table might run at the same time which might create issues. Therefore, ensure that you use only one policy for a table, and no two policies are scheduled to run at the same time. Select and configure at least one of the maintenance tasks.
Task Priority	Set how urgently a policy's maintenance tasks must run when they compete with other policies for cluster resources. Higher priority tasks are scheduled before lower priority tasks. You can select one of the following options from the Priority list: • Critical • Urgent • High • Medium (Default): This option is selected by default. • Low • Minimal
Automated Actions	Choose Expire Snapshot, and configure the following options to ensure compact metadata and optimal query performance. This action allows you to time travel to a previous version of the table, and allows you to control the snapshot retention settings for time travel queries and rollback. **Note:** Retaining too many snapshots impact the table performance. Therefore, ensure that you retain only the required number of snapshots. • Enter the Expire Older Than value in milliseconds. Cloudera Lakehouse Optimizer expires the snapshot older than the set value. The default value is 432000000 ms (7200 minutes). • Enter the Retain Last value to specify the number of snapshots to retain. Cloudera Lakehouse Optimizer deletes the older snapshots when the set value is reached. The default value is 5. For example, by default the first snapshot gets deleted automatically after the sixth snapshot is created. • Optional. Enter the snapshot ID in the Expire Snapshot with ID field. Cloudera Lakehouse Optimizer expires the specified snapshot. There is no default value. • Toggle Clean Expired File to remove the expired snapshots permanently. The default value is true.

Option	Description
	Choose Compaction settings, and configure the following options to improve query performance. This action compacts small files into larger ones, thereby optimizing storage and improving read performance. • Enter the Target File Size in bytes. Cloudera Lakehouse Optimizer merges the small files into a larger file of the specified size. The default value is 536870912 bytes. Improves the query performance because of larger optimally sized files. • Enter the Max Concurrent Rewrite File Groups value to specify the maximum number of file groups to be simultaneously rewritten. The default value is 5. Improves the compaction performance and rewrite operations.  **Note:** A higher value speeds up the rewrite operation but consumes a lot of resources. Ensure that you choose an optimal value for your environment. • Enter the Minimum Input Files value to specify the minimum number of files in a file group. Any file group exceeding the set number of files is rewritten into a larger data file regardless of other criteria. The default value is 5. Optimizes data management and improves query performance. • Enter the Partial Progress Max Commits value to specify the maximum number of commits that are allowed during the rewrite operation. This ensures that the changes are committed and the snapshots are created even while the rewrite operation is in progress. The default value is 10. Ensures that the longer rewrite operation is chunked, and the commits are performed as the operation proceeds.  **Note:** This action might lead to a large number of snapshots, therefore Cloudera recommends specifying an optimal value. • Enter the Delete File Threshold value to specify the minimum number of delete files that must be associated with a data file, after which those delete files are rewritten into a larger data file. The default value is 2000000. Optimizes the read operations. • Choose Partial Progress Enabled to ensure that the completed parts of the rewrite operation are committed. The default value is false. Improves performance, fault tolerance, and prevents out of memory (OOM) issues. • Choose Use Starting Sequence Number to use the sequence number of the snapshot at compaction operation start time instead of the newly produced snapshot. The default value is false. Optimizes the performance. • Choose Rewrite All to force rewrite all the files overriding other options. The default value is false. Ensures full compaction of the tables.
	Choose Rewrite Manifest, and configure the following options to consolidate multiple small manifest files (metadata files) into a larger manifest file, remove the deleted data references, and ensure that the manifest files are efficiently structured which in turn improves the metadata scan performance. • Enter the Target File Size value to specify the target manifest file size in bytes. The default value is 8388608. Improves query performance and optimizes metadata management. • Choose Use Caching to cache the manifest files in memory. The default value is true. Improves query planning and performance.

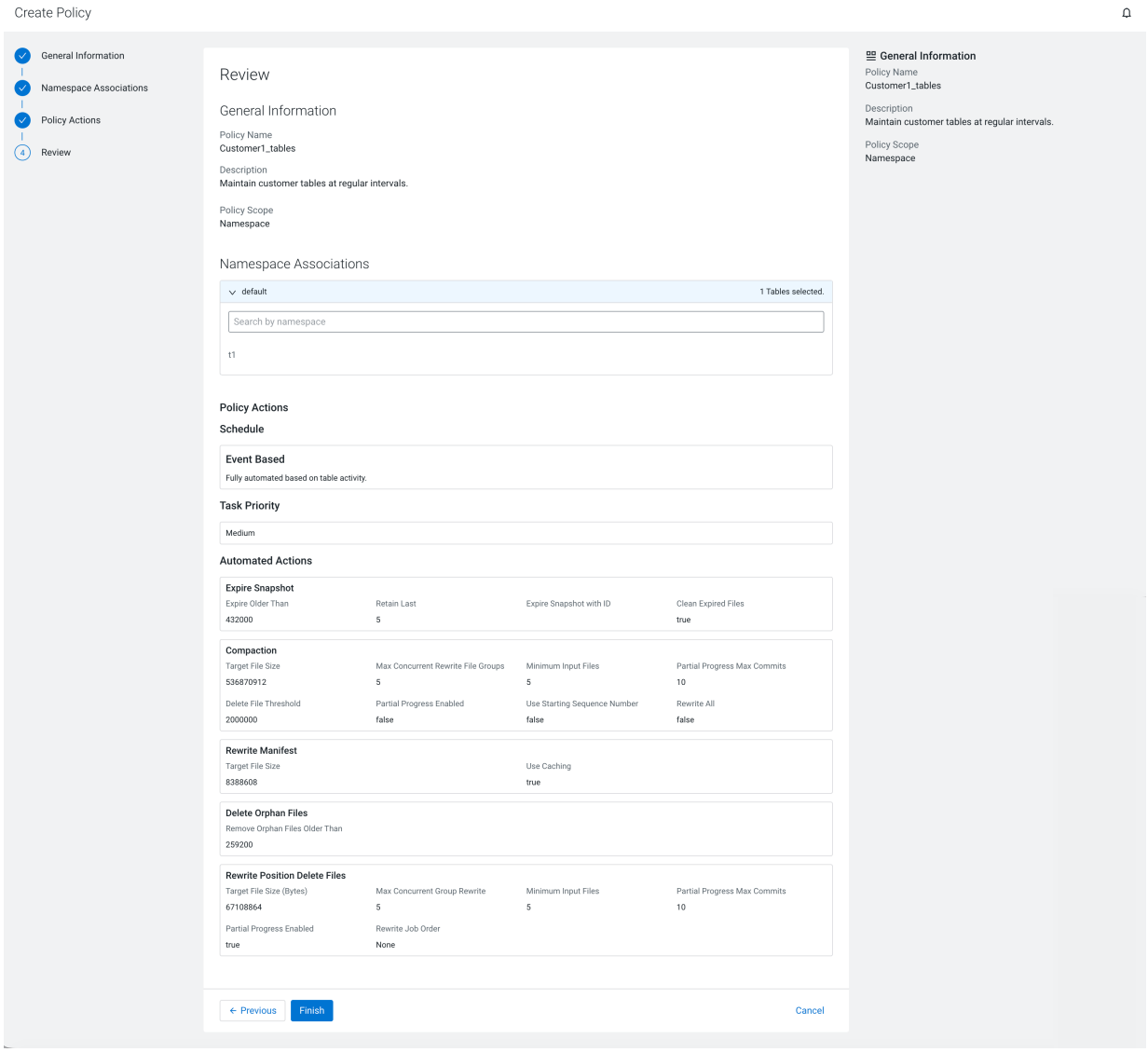
Option	Description
	Choose Delete Orphan Files, and configure the Remove Orphan Files Older Than value to identify and remove the data files that are no longer referenced by any table metadata. These files are called orphan files, and might exist due to failed writes, manual deletions, or schema changes. Enter the Remove Orphan Files Older Than value to specify the time duration in milliseconds to remove the orphan files that meet the criteria. This action reclaims storage space, maintains data integrity, and ensures the storage footprint of Iceberg tables are as efficient as possible.
	Choose Rewrite Position Delete Files, and configure the following options to optimize the size and layout of the positional delete files in the Iceberg tables by compacting small positional delete files into larger ones, and filtering out positional delete records that refer to data files that are no longer available. • Enter the Target File Size value to specify the target positional delete file size in bytes. Cloudera Lakehouse Optimizer merges the small positional delete files into a larger file of the specified size. The default value is 67108864 bytes. • Enter the Max Concurrent Group Rewrite value to specify the maximum number of file groups to be simultaneously rewritten. The default value is 5. Reduces rewrite operation time. **Note:** A higher value speeds up the rewrite operation but consumes a lot of resources. Ensure that you choose an optimal value for your environment. • Enter the Minimum Input Files value to specify the minimum number of files in a file group. Any file group exceeding this number of files is rewritten in a larger data file regardless of other criteria. The default value is 5. Optimizes data management. • Enter the Partial Progress Max Commits value to specify the maximum number of commits that are allowed during the rewrite operation. This ensures that the changes are committed and the snapshots are created even while the rewrite operation is in progress. The default value is 10. Ensures that the longer rewrite operation is chunked, and the commits are performed as the operation proceeds. **Note:** This action might lead to a large number of snapshots, therefore Cloudera recommends specifying an optimal value. • Choose Partial Progress Enabled to ensure that the completed parts of the rewrite operation are committed. The default value is true. Improves performance and fault tolerance. • Choose one of the following Rewrite Job Older options to force the rewrite job order based on the chosen value: • bytes (asc) rewrites the smallest job groups first. • bytes (desc) rewrites the largest job groups first. • files (asc) rewrites the job groups with the least number of files first. • files (desc) rewrites the job groups with the most files first. • none rewrites the job groups in the order they were planned. The default value is none.

The following GIF shows the **Policy Actions** tab and the tab details:



9. Click Next.
10. Verify the settings on the **Review** tab.

    The following image shows the **Review** tab:



11. Click Finish.

Results

Depending on the **Table Maintenance Schedule**, Cloudera Lakehouse Optimizer initiates the table maintenance tasks on the chosen tables.

What to do next

To monitor the policies, use one of the following methods:

- View the policy run details on the Lakehouse Optimizer UI Policies tab.
- View the chosen tables' maintenance details on the Lakehouse Optimizer UI Tables tab.
- Verify whether the table maintenance completed successfully on the Lakehouse Optimizer UI Tables tab.
- Monitor the policy jobs as Spark jobs on the Cloudera Observability dashboard. For more information, see Monitoring table maintenance tasks on Cloudera Observability dashboard.
- View the various methods that are available to monitor the table maintenance tasks.

Additionally, you can pause and resume table maintenance on the **Policies** tab to manage policies.

Important: When you pause maintenance for a table or namespace, all the policies for that table or namespace are paused automatically. Therefore, you must be cautious before you pause maintenance for a table or a namespace.

Performing dry run on Cloudera Lakehouse policies in Lakehouse Optimizer UI

In Cloudera on cloud 7.3.2 and higher versions, dry run Cloudera Lakehouse Optimizer policies to preview generated maintenance tasks and verify expected outcomes before initiating a manual or scheduled execution.

Procedure

1. Go to the Cloudera Management Console Lakehouse Optimizer page.
2. Select one of the following methods to open the **Dry Run and Execution** side panel:
   - Click  for a table on the Table Associations tab, and then click the Dry Run & Execute action.
   - Click  for a table on the Table Associations tab, and then click the View Policy action. On the **View Policy** page, click  for a table, and then click the Dry Run & Execute action.
3. Select one of the following options depending on your use case:
   - Dry Run Only –- This option generates the maintenance tasks for the table, and performs a dry run to ensure they run without failure. No action is performed on the table.

     After you choose this option, Cloudera Lakehouse Optimizer initiates the dry run of the policy during which it fetches and evaluates the policy definition on the table to generate the maintenance actions. The output is provided in the side panel in JSON format.

     Click JSON Details to view the output. Optionally, you can copy the JSON details, and save these details in a file on your machine for future reference. After reviewing the generated output, you can decide whether to run the policy on the table, or to modify the policy to meet your requirements. You can close the side panel after the dry run is complete or cancelled.

     Note: If you close the side panel during the dry run process, the dry run action is aborted.
   - Dry Run & Execute –- This option performs a dry run and allows you to execute the policy immediately after reviewing the results.
     a. After you click Dry Run & Execute, Cloudera Lakehouse Optimizer generates the maintenance tasks and initiates a dry run, providing the output in the side panel. Click JSON Details to view the output.
     b. After you review the output, click Execute Table Policy. Cloudera Lakehouse Optimizer then initiates the policy run and displays the run details, which include the policy run status, task ID, and task details.
     c. Optionally, you can click See All Recent Task Activity to view more details about the policy run."

     Note: The actions generated during the dry run and the actions that are generated during the execution step might differ if the policy conditions change. For example, if there were changes made to the table or if there was a customization in the Iceberg table parameters between the two steps.
4. Click Done to close the side panel.

Cloudera Lakehouse Optimizer REST APIs

You can use Cloudera Lakehouse Optimizer REST APIs to create, maintain, and monitor Cloudera Lakehouse Optimizer policies. You can perform several maintenance operations on the Iceberg tables using REST APIs.

Preparing and defining Cloudera Lakehouse Optimizer policies using REST APIs

Before you create or define the Cloudera Lakehouse Optimizer policies using REST APIs, the administrator must complete the necessary prerequisites. After defining the policy, the administrator must onboard the required namespace, associate the required tables to the policy, and reschedule the namespace.

Before you begin

Consider the best practices before you create or define the policy.

About this task

The following section explains the prerequisites, defining the policy, and post-requisites that you must complete using REST APIs for Cloudera Lakehouse Optimizer to initiate table maintenance.

Procedure

1. Contact your Cloudera account team to enable the Cloudera Lakehouse Optimizer service for your Cloudera Open Data Lakehouse environment.
2. Use an existing Cloudera Base on premises cluster, or create a dedicated Cloudera Lakehouse Optimizer cluster.

   Ensure that the cluster contains the Zookeeper, HDFS, Hive, YARN, Spark3_on_yarn, Oozie, Hue, Livy_for_spark3, Ranger, Knox, and Meteringv2 services.

   Add the Cloudera Lakehouse Optimizer service to the cluster.
3. Assign roles to Cloudera Lakehouse Optimizer users. For instructions, see Configuring roles for Lakehouse Optimizer users.
4. If required, enable the Ranger service for Cloudera Lakehouse Optimizer, and then create the Ranger policies to provide the fine-grained access to a user or group. For instructions, see Providing fine-grained access to namespaces using Ranger.
5. Modify the default values for Spark Executor Memory and Spark Driver Memory. The default values are 8 GB and 4 GB respectively. The default memory settings might be enough for a majority of the use cases. However, for heavy workloads you might want to increase these values. For more information, see Troubleshooting: Configuring the Spark engine to optimize the rewrite compaction job.

   To modify the spark.driver.memory and spark.executor.memory settings, go to the Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* Configuration conf/dlm-client.properties_role_safety_valve property.
6. Optionally, configure custom YARN queues to use for Cloudera Lakehouse Optimizer workloads. This ensures that the policy runs do not overlap with other workloads.
7. For all node groups, configure the yarn.scheduler.maximum-allocation-mb and yarn.nodemanager.resource.memory-mb advanced configuration snippets on the Cloudera Manager Clusters *YARN SERVICE* Configuration tab to more than 8 GB depending on your requirements.
8. By default, the dlm_admin, dlm_operator, or dlm_monitor groups are available that are assigned to the Cloudera Lakehouse Optimizer administrator, operator, or monitor role respectively. You can add one or more comma-separated groups to perform the administrator, operator, or monitor roles.

   To view the default groups or to add more groups, perform the following steps:

   a. Go to the Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* Configuration tab.
   b. For the administrator role, search for the DLM Security Role Admin property. By default, the property contains the dlm_admin group. You can append more groups that are comma separated. For example, dlm_admin, clo_admin.
   c. For the operator role, search for the DLM Security Role Operator property. By default, the property contains the dlm_operator group. You can append more groups that are comma separated.

      For the monitor role, search for the DLM Security Role Monitor property. By default, the property contains the dlm_monitor group. You can append more groups that are comma separated.
9. Complete the trusted-proxy authentication prerequisites for Cloudera Lakehouse Optimizer REST APIs on Cloudera Base on premises.

For more information, see Prerequisites for Cloudera Lakehouse Optimizer REST API authentication (trusted proxy).

10. Cloudera Lakehouse Optimizer administrators must perform the following actions to verify whether the REST APIs are accessible and the service is available:
    a) Verify whether you can access Lakehouse Optimizer REST APIs after you generate an Apache Knox token. For more information, see Generating tokens and access Lakehouse Optimizer.
    b) Perform a health check, and then verify whether the ClouderaAdaptive default policy is available. For more information, see Verifying Lakehouse Optimizer health and policy script.
11. Cloudera Lakehouse Optimizer administrators must perform the following actions to initiate the Iceberg table maintenance activity:
    a) Onboard a namespace using the `PUT /namespaces/{namespace}` API. The API informs Cloudera Lakehouse Optimizer to include the associated tables in the namespace for maintenance.
    b) Define or create the policy. For more information, see Defining Lakehouse Optimizer resources.
    c) Associate or subscribe the tables in the onboarded namespace to the required policy using the `POST /policies/{policyName}/tables/{tableName}/subs` API. The action modifies or appends the mapping of the Iceberg tables to the policies in the association file.

       You must run the `PUT /policies/{policyName}/tables/{tableName}/subs` API for subsequent associations. For more information, see Understanding table-policy associations.

       Note: A table can be associated with a set of policies using the dlm_policies table property. For example, `ALTER TABLE t1 SET TBLPROPERTIES ('dlm_policies' = 'p1');` where p1 is the policy name. After you update the table properties, run a manual evaluation on the IcebergPropertiesTable policy. You must then run the `POST /policies/IcebergPropertiesTable/tables/t1/evaluation` API, and reschedule the namespace.
    d) Reschedule the namespace to update the policies for all the tables in an existing namespace using the `PATCH /namespaces/{namespace}` API.
    e) Optionally, dry run the policy, during which Cloudera Lakehouse Optimizer only generates the table maintenance actions and does not initiate the maintenance actions. For instructions, see Step 2 of Performing manual table maintenance.

Results

Wait for Cloudera Lakehouse Optimizer to initiate the table maintenance based on the CRON schedule in the policies.

Optionally, you can also initiate a manual table maintenance, when required. Ensure that you dry run the policy before manual maintenance. For more information, see Performing manual table maintenance.

What to do next

To monitor the policies, view the various methods that are available to manage and monitor the table maintenance tasks.

Configuring roles for Cloudera Lakehouse Optimizer users

Depending on your role, you can add the service to Cloudera Manager, create the Cloudera Lakehouse Optimizer policies, and monitor the policies. Cloudera Lakehouse Optimizer REST APIs uses Knox to infer authentication and group membership.

To access the Cloudera Lakehouse Optimizer features and REST APIs, you might require one or more of the following roles:

Role	Description	Default values
Administrators	Can access and use all the features including all the REST APIs.	dlm_admin
Operators	Have limited access to REST APIs. They can run certain tasks, and have no privileges to modify the service configuration.	dlm_operator

Role	Description	Default values
Monitors	Can only observe the health and status of the service and its running tasks.	dlm_monitor

The **CLO Security Role Admin**, **CLO Security Role Operator**, and **CLO Security Role Monitor** properties on the Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* Configuration tab contain the group names that define the user roles. By default, the properties contain the dlm_admin, dlm_operator, and dlm_monitor values respectively.

Note: You can change the default names with the names of the groups in your cluster, and you can also add multiple comma-separated group names.

Ensure that you are in the administrator, operator, or monitor group to use the Cloudera Lakehouse Optimizer REST APIs.

Providing fine-grained access to namespaces using Ranger

Provide administrator, operator, or monitor role access for a user or a group at namespace level. Enable the Ranger service for Cloudera Lakehouse Optimizer, and then create the Ranger policies to provide the fine-grained access to a user or group.

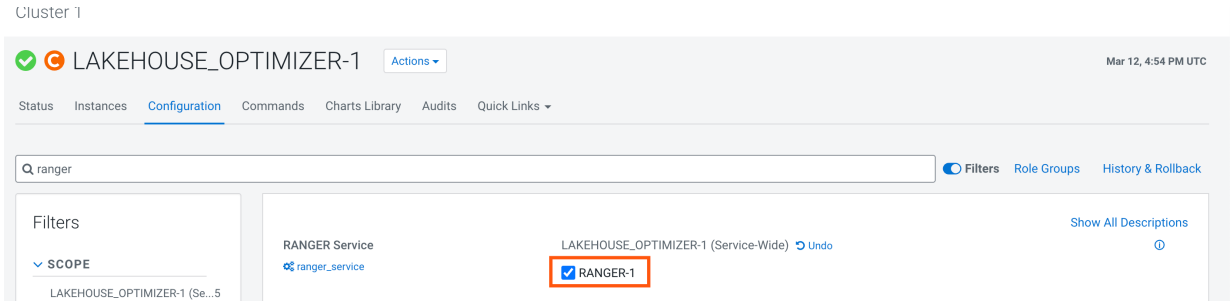
About this task

Namespace-level permissions supersede universal permissions. You can assign the namespace administrator permissions to one or more groups, and then assign the required Ranger policies to those groups. For example, if clo_user1 group is assigned the all-database Ranger policy, the users within that group have access to all the Cloudera Lakehouse Optimizer policies. To deny access to a user in the clo_user1 group, you remove the user from the clo_user1 group, and then configure deny permission for the user.

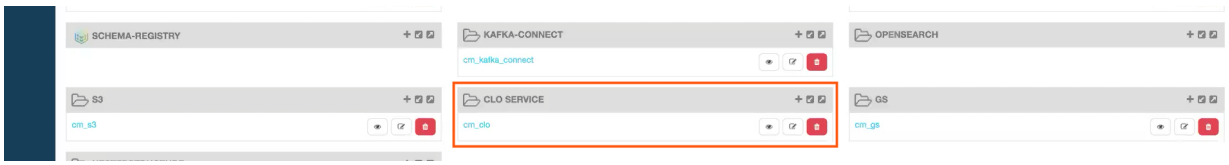
Procedure

1. Verify whether the Ranger service is enabled for Lakehouse Optimizer.
   a) Go to the Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* Configuration tab.
2. Search for Ranger.

   The Ranger field must be selected as shown in the following image:

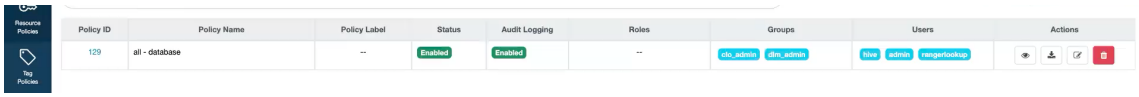


   The CLO SERVICE (cm_clo) resource is displayed in Ranger as shown in the following image:



Tip:

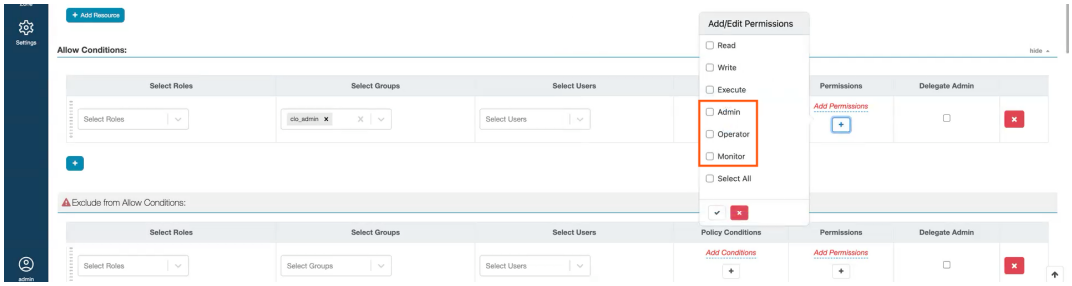
The default all-database Ranger policy contains the dlm_admin group and the users in this group are superusers. The following sample image shows dlm_admin and clo_admin groups along with hive, admin, and rangerlookup users as superusers:



You can create Ranger policies to provide fine-grained access to groups and users based on your requirements.

3. Create the required Ranger policy in Ranger.
   a) Go to Cloudera Manager Clusters Ranger Ranger Admin Web UI .
      The Ranger UI is displayed in a new tab.
   b) Go to cm_clo Add New Policy .
   c) Enter the following details in the Create Policy wizard:
      1. Enter a unique Policy Name.
      2. Optionally, enter a Description.
      3. Select one or more names from the CLO Namespace Name list.
      4. In the **Allow Conditions** section, select the Group, User, or both, and then select the Permissions.

         Important: Cloudera highly recommends selecting Admin, Operator, or Monitor permissions in the Add/Edit Permissions list. This is because an incorrect or incompatible combination of permissions might result in unpredictable behaviors.



      5. Deny permissions to a specific group or user in the **Deny Conditions** section as described in the **Allow Conditions** instruction step.
      6. Click Save.

Caution:

If the Ranger field in the Cloudera Manager Clusters *[***CLOUDERA LAKEHOUSE OPTIMIZER***]* Configuration tab is not selected, Cloudera Lakehouse Optimizer does not implement fine-grained permissions when running the Cloudera Lakehouse Optimizer policies.

Results

Cloudera Lakehouse Optimizer checks the permissions before it runs a Cloudera Lakehouse Optimizer policy.

Prerequisites for Cloudera Lakehouse Optimizer REST API authentication (trusted proxy)

On Cloudera Base on premises, trusted-proxy configuration, identity forwarding, and a valid signing key are required before Cloudera Lakehouse Optimizer REST API calls succeed through Apache Knox.

The topic describes adding users in the dlm_admin group, generating a Knox token, and calling the API with a Bearer token. Ensure to the meet the following prerequisites to run API calls including the sample health check, failing which, might return a 403 Forbidden or 401 Unauthorized errors through Knox with no corresponding error in dlm.log.

Trusted-proxy prerequisites

- hadoop.proxyuser.knox.hosts and hadoop.proxyuser.knox.groups must be present in the cluster configuration that the Cloudera Lakehouse Optimizer server loads, such as core-site.xml or dlm-cluster.properties. With dlm.security.trusted.proxy=true, Cloudera Lakehouse Optimizer builds its trusted-proxy allow list from hadoop.proxyuser.knox.hosts. If absent, the list is empty and every proxied request is rejected with the Rejecting doAs … from untrusted sender message.

  Important: The core-site.xml that Cloudera Lakehouse Optimizer loads is the same cluster-level file that Cloudera Manager provides. Cloudera Lakehouse Optimizer adds the core-site.xml file as a resource from the hadoop-confservice directory to its classpath. For example, configuration.addResource("core-site.xml") and configuration.addResource("hdfs-site.xml").
- The end-user identity must be forwarded as the doAs query parameter (?doAs=<user>). Groups must be forwarded in the X-Knox-Actor-Groups-1 header. Cloudera Lakehouse Optimizer maps the group to the role through dlm.security.role.admin (default mapping: dlm_admin # admin). If the header is missing, Cloudera Lakehouse Optimizer logs doAs user <u>, groups null and returns 403 Forbidden.
- dlm.security.authkey.path is optional. It applies only when Cloudera Lakehouse Optimizer must be authenticatd directly with JWTs instead of through an Apache Knox gateway. Set the property to a JWK (JSON) file that contains an RSA key pair, including the private key. If the property is not set or the file is missing or unreadable, Cloudera Lakehouse Optimizer generates a temporary in-memory key. Tokens are invalidated on restart and are not shared between HA nodes. If the file exists but cannot be parsed or contains only a public key, Cloudera Lakehouse Optimizer logs Failed to start security and continues running, but API calls are rejected (typically 403 Forbidden).

Knox gateway path and query/header contract

When you call REST APIs through Knox, use the Knox-proxied path (for example, https://<ENDPOINT>/cdp-proxy-token/clo/api/v1/config/health) with the Authorization: Bearer header. In addition, Knox must forward the end-user identity as the doAs query parameter and group membership in the X-Knox-Actor-Groups-1 header.

Without the trusted-proxy contract, the following call returns 403 Forbidden or 401 Unauthorized:

```
curl 'https://<ENDPOINT>/cdp-proxy-token/clo/api/v1/config/health' -H 'Autho
rization: Bearer <JWT>'
```

Working call with the prerequisites met:

```
curl "http://<clo-host>:8085/clo/api/v1/config/health?doAs=clodemo" -H "X-Kn
ox-Actor-Groups-1: dlm_admin"
```

Troubleshooting REST API authentication

Condition	Cause
403 Forbidden with no entry in dlm.log	Security failed to initialize (authentication key)
Rejecting doAs … from untrusted sender	Missing hadoop.proxyuser.knox.* properties
doAs user X, groups null	Missing X-Knox-Actor-Groups-1 header
401 Unauthorized with Couldn't find bearer token	Identity not forwarded as ?doAs=

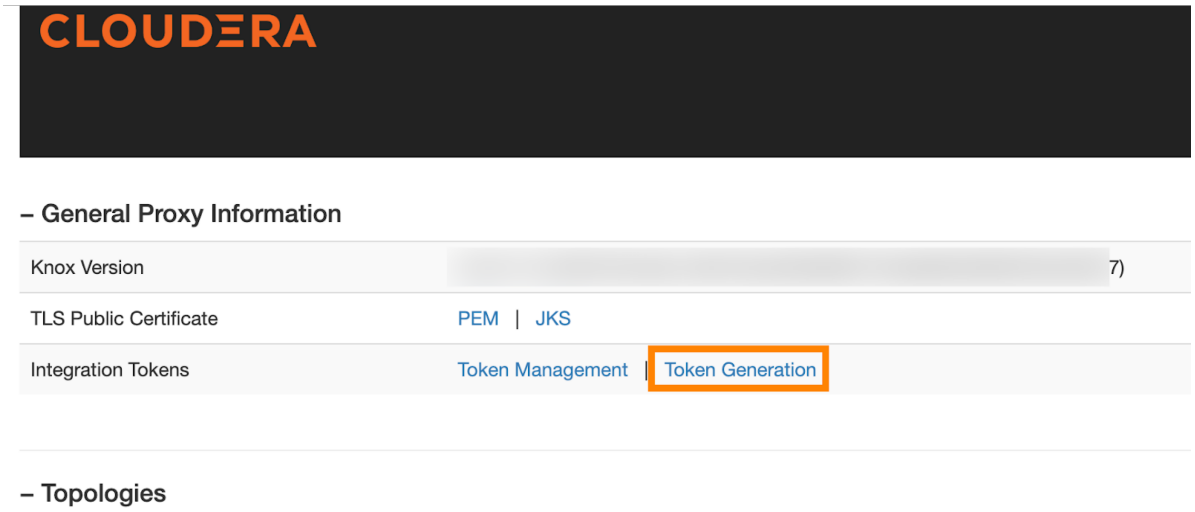
Generating tokens and accessing Cloudera Lakehouse Optimizer REST APIs

Generate JWT tokens through Apache Knox to authorize and access Cloudera Lakehouse Optimizer REST APIs for service-level administrative tasks.

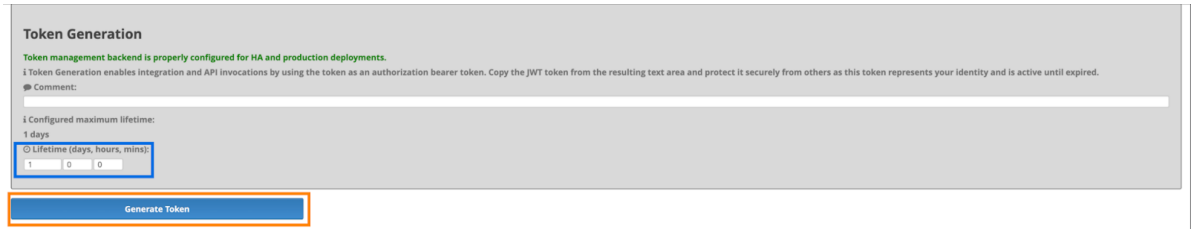
Procedure

1. Go to Cloudera Manager Clusters Knox service Web UI Knox Gateway Home .
2. Enter the credentials to access the Knox gateway.

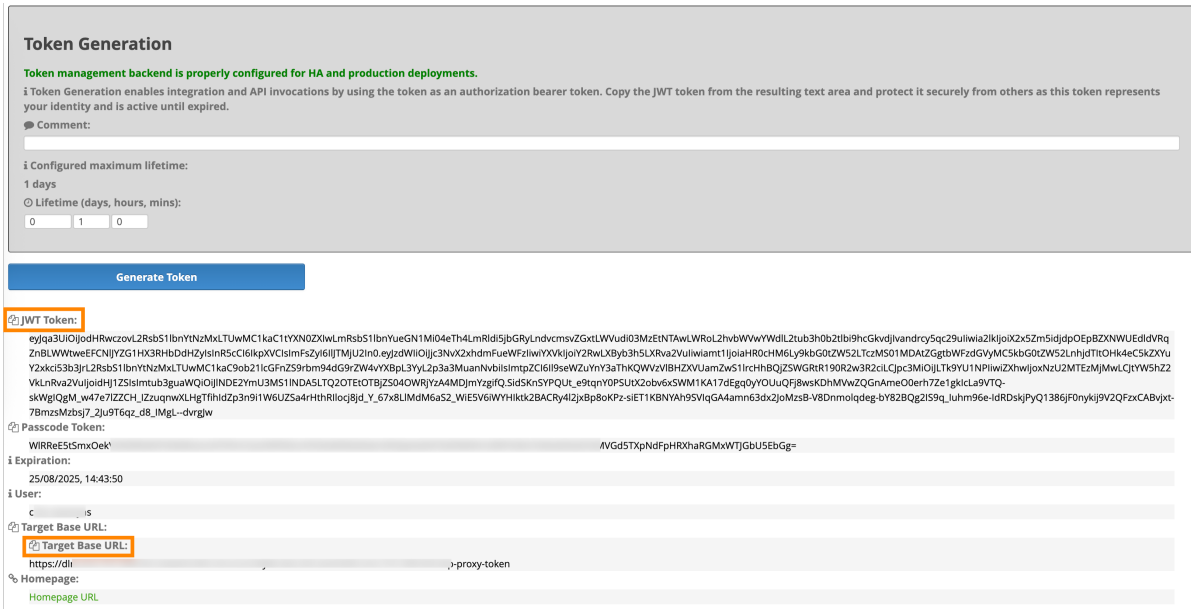
3. Obtain the token from Apache Knox to use the Cloudera Lakehouse Optimizer REST APIs.
   a) Click Cloudera Management Console Data Hub Clusters *[***LAKEHOUSE OPTIMIZER DATA HUB***]* .
   b) Click Token Integration in the **Services** section.
   Apache Knox appears in a new browser tab.
   c) Click Token Generation in the **General Proxy Information** section as shown in the following screenshot:



   The **Token Generation page** appears in a new browser tab.
   d) Enter the required number of days, hours, and minutes for the token in the Lifetime (days, hours, mins) field on the **Token Generation** page, and then click Generate Token as shown in the following screenshot:



   The following screenshot shows the generated token details:



   The **JWT Token** and the **Target Base URL** values are generated. Use the **JWT Token** to authorize the REST APIs, and the **Target Base URL** as the base URL for REST APIs.

4. Access the REST APIs by clicking the Apache Knox browser tab Topologies API Services Cloudera Lakehouse Optimizer - API link. Alternatively, use the **cdp-proxy-token** section to get the base url to call the REST APIs using the curl commands (with JWT).

   Note: Some Cloudera Lakehouse Optimizer configuration parameters might use the dlm prefix.

Verifying service health and viewing default Cloudera Lakehouse Optimizer policy

After you verify whether you can access Cloudera Lakehouse Optimizer REST APIs, Cloudera recommends following the best practice of performing a health check of the service, and then verifying whether the ClouderaAdaptive default policy is available in the Cloudera Lakehouse Optimizer Data Hub.

Procedure

1. Ensure that the Cloudera Lakehouse Optimizer Data Hub and its components are available and healthy using the `GET /config/health` API.

   ```
   curl --location 'https://[***LAKEHOUSE OPTIMIZER ENDPOINT***]/cdp-proxy-
   token/clo/api/v1/config/health' \ --header 'Authorization: Bearer [**JWT
    TOKEN**]'
   ```

   By default, Cloudera Lakehouse Optimizer performs a partial health check. To perform a full health check, append scope.

   ```
   curl --location 'https://[***LAKEHOUSE OPTIMIZER ENDPOINT***]/cdp-proxy-t
   oken/clo/api/v1/config/health?scope=full' \ --header 'Authorization: Bea
   rer ey'
   ```

   Tip: Cloudera Lakehouse Optimizer monitors the service components, service tasks, and maintenance actions which include:

   - Recent running tasks, task details, and task status.
   - Livy connection status.
   - Number of triggers available in the scheduler, the policy evaluation count, the ignore task count, and the paused table count.
   - Count of tasks sent to Livy and the pending tasks count.

   When high availability (HA) is enabled, the full health check response also includes a High Availability entry that reports the HA status of the instance. The entry indicates whether the instance is the leader or a standby, and whether leader election is in progress. For more information, see High availability for Cloudera Lakehouse Optimizer.

2. Verify whether the ClouderaAdaptive default policy is available in the Cloudera Lakehouse Optimizer Data Hub using the `GET /policies/resource` API. The API fetches the default ClouderaAdaptive policy details.

   ```
   curl --location 'https://[***LAKEHOUSE OPTIMIZER ENDPOINT***]/api/v1/poli
   cies/resource?uri=clo%3A%2F%2Ftps%3Adefault%2FClouderaAdaptive' \ --header
    'Authorization: Bearer ey'
   ```

   For information about the URI to use in the `GET /policies/resource` API, see Constructing the URI to fetch, upload, and update policy resources.

Defining Cloudera Lakehouse Optimizer resources using REST APIs

When you use REST APIs to create or define a Cloudera Lakehouse Optimizer policy, use the default policy called "ClouderaAdaptive" as-is, or create a new policy. When you create a new policy or override an existing policy, you define the policy in a new JEXL file, alternatively add the required constants in a new JSON file, and then upload the resources.

About this task

Perform the following steps to create a new policy or to override an existing policy by defining the policy in the new JEXL file. Alternatively you can add the constants in the new JSON file.

Procedure

1. Create the JEXL file, and optionally the JSON file. You can use the default policy resources as a template to create the files, or you can create the files from scratch.

   For more information about the JEXL file and JSON file, see Understanding policy resources.

   Note: To create a Puffin statistics policy, include the computeTableStats action in the JEXL script and JSON constants file. For more information, see JEXL script, JSON file, and Puffin statistics for Cloudera Lakehouse Optimizer.

   Important:

   - You cannot update the default *ClouderaAdaptive* policy resources at default scope or level. However, override it at the catalog or namespace scope.

     For example, if you want to override the definition of the default policy:

     a. Define another ClouderaAdaptive.jexl and ClouderaAdaptive.json files.
     b. Upload the definition at the catalog or namespace level using the `PUT/policy/resource` API.
     c. Reschedule the namespace using the `PATCH /namespaces/{namespace}` API.
   - Create more than one policy for an Iceberg table, namespace, or catalog depending on your requirements.
   - Ensure that the JEXL script and the JSON file share the same name, and the name is unique within a scope. The scope or resource scope refers to catalog level, namespace level, or table level. For example, daily.jexl and daily.json can be defined at table level.
   - Define more than one policy at table, namespace, or catalog scope to associate with an Iceberg table.
2. Upload the resources using the `PUT /policies/resource` API.

   Use the `PATCH /policies/resource` API to upload an updated policy resource.

   Use the following curl command to upload the resources:

   ```
   curl -X PUT --location '[***LAKEHOUSE OPTIMIZER ENDPOINT***]/cdp-proxy-
   token/clo/api/v1/policies/resource' --header 'Authorization: Bearer ey
   ```

   Note: To create an event-based policy using REST APIs, ensure that the CRON expression is 0 0 0 1 1 ? 2199 for the policy. When an event on the table completes, Cloudera Lakehouse Optimizer checks all the event-based policies for that table and triggers the evaluation phase for that table. The policies without the 0 0 0 1 1 ? 2199 CRON expression are considered to be a schedule-based policy.

   To construct an URI to fetch, upload, and update the policy resources, see URI to fetch, upload, and update policy resources when using REST APIs on page 122

URI to fetch, upload, and update policy resources when using REST APIs

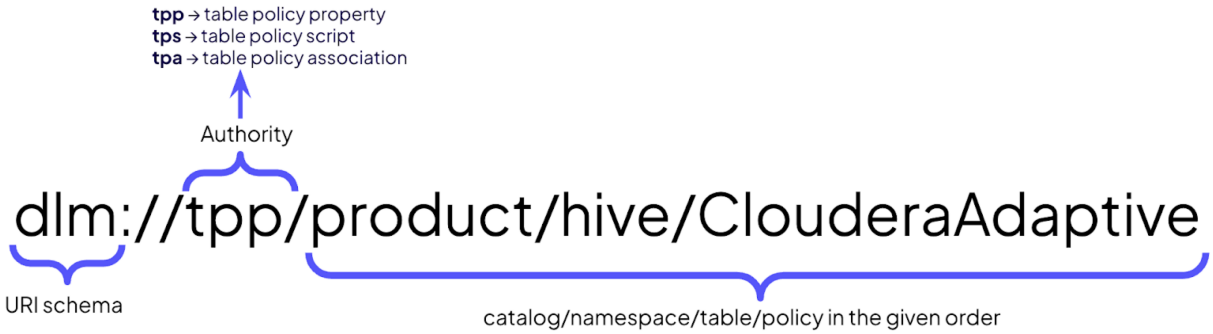
The Cloudera Lakehouse Optimizer resources use URI for identification. The URI is used to fetch, upload, or update the policy resources when you use REST APIs.

To construct a URI, for example the dlm://tpp/product/hive/ClouderaAdaptive URI, you must provide the following elements:

1. The dlm URI scheme.
2. An authority, such as tpp for table policy property, tps for table policy script, or tpa for table policy association.
3. The catalog / namespace / table / policy in the given order. This determines the level at which the script and properties have been defined, and also determines the scope of the resource. The default resource is dlm://tps:default/ClouderaAdaptive.

Note: For tpa, the scope is the namespace. For example, dlm://tpa/tpcds/hive identifies the table and policy association for the hive namespace in the tpcds catalog.

The following image shows the different components of a sample URI:



Examples:

- The dlm://tpp/product/hive/ClouderaAdaptive URI identifies the ClouderaAdaptive policy properties (JSON) defined for all the tables in the hive namespace in the product catalog.
- The dlm://tpp/product/hive/customers/ClouderaAdaptive URI identifies the properties defined for the customers table in the hive namespace of the product catalog.

To understand the policy resources and the policy structure, see Understanding policy resources.

Understanding table-policy associations

Cloudera Lakehouse Optimizer maintains a table-policy association JSON file for each namespace.

After you define the policies using REST APIs for the first time, you must run the `POST /policies/{policyName}/tables/{tableName}/subs` API so that the Cloudera Lakehouse Optimizer modifies or appends the mapping of the Iceberg tables to the policies in the association file. You must run the `PUT /policies/{policyName}/tables/{tableName}/subs` API for subsequent associations.

Alternatively, to do a bulk-update of table-policy associations, you must download the existing file which you can update and then upload. If the file does not exist, create a JSON file, update it as required, and upload it using the `PUT /policies/resource` API, or use the curl --location --request PUT 'https://*[*** *LAKEHOUSE OPTIMIZER ENDPOINT* ***]*/api/v1/policies/resource' \ --form 'dlm://tpa/hive/n1=@"/path/to/file" command.

The following example shows an entry in the association file:

```
{
“Catalog_sales” : [“snapshot”, “manifests_orphan”, “compact”] //
Catalog_sales is the Iceberg table name.
}
```

Performing manual Iceberg table maintenance using Cloudera Lakehouse Optimizer REST APIs

You can use REST APIs to perform manual table maintenance as needed depending on your assigned role.

Procedure

1. Get the current statistics for the required table using the GET /tables/{tableId}/stats API.

   Perform this step to view the current table statistics to compare them with the expected statistics after the maintenance actions are complete.
2. Dry run the policy to determine whether the generated maintenance actions are as expected using the `PUT /policies/{policyName}/tables/{tableName}/dryrun` API. The API evaluates the statistics and then generates only a set of maintenance actions.

3. Initiate the evaluation phase using the `POST /policies/{policyName}/tables/{tableName}/evaluation` API.

   The 200 response shows a unique task ID that you can use to monitor the task and the task metadata.
4. Monitor the status of the task using the GET /tasks or GET /tasks/id/{id} API.

   Tip: To fetch the task metadata for the task in the submitted state for a specific policy and table name, use the `GET /tasks/policies/{policyName}/tables/{tableName}/submitted` API.
5. Verify whether the task completed the maintenance actions as expected using the `GET /tables/{tableId}/stats` API.
6. For a Puffin statistics policy, verify that the table metadata includes a new entry in statisticsFiles pointing to the generated Puffin file.

What to do next

You can use this workflow to run a Puffin statistics policy manually. For more information, see Puffin statistics for Cloudera Lakehouse Optimizer.

Puffin statistics for Cloudera Lakehouse Optimizer

You can create a custom Cloudera Lakehouse Optimizer policy that computes NDV (number of distinct values) statistics for Iceberg table columns and writes them to a Puffin statistics file registered in the table metadata.

The COMPUTE_TABLE_STATS action reads a target Iceberg table, computes an NDV (theta-sketch) blob for each requested column, and writes those blobs to a Puffin statistics file. Cloudera Lakehouse Optimizer registers the file in the table metadata statisticsFiles list in metadata.json.

Query engines that read Iceberg Puffin files such as Spark cost-based optimizer, Hive CBO, Impala, and Trino can use the NDV estimates to plan joins and aggregations.

Puffin statistics require a custom policy with a JEXL script and an optional JSON constants file that enable the comp uteTableStats action.

When the policy job completes, Cloudera Lakehouse Optimizer registers a Puffin statistics file in the Iceberg table metadata. Query engines configured to read Puffin files can use the NDV estimates for cost-based optimization.

Important: In 7.3.2.10000 and later versions, Puffin statistics policies are available through REST APIs only. You cannot configure this feature in the Cloudera Lakehouse Optimizer UI.

Data file compression for Cloudera Lakehouse Optimizer

Cloudera Lakehouse Optimizer can automatically re-encode Iceberg data files that match a SQL filter with a more efficient compression codec. This reduces storage costs without changing query results.

Data file compression is distinct from data file compaction (rewriteDataFiles). Compaction is statistics-driven and merges small files. Compression rewrites every file that matches the configured filter using a bin-pack strategy and changes the encoding of existing files regardless of file size or count.

Create a separate compression policy. Do not combine data file compression with compaction or other maintenance actions in the same policy. Because compression targets cold data, run the compression policy on an ad hoc basis when possible, or on a monthly cadence.

Configure data file compression by adding a compressDataFiles block to the policy constants JSON file and associating the policy at the namespace level. Individual tables opt in by setting the filter and other compression properties with ALTER TABLE … SET TBLPROPERTIES.

For compressDataFiles policy constants, supported codecs, per-table overrides, and configuration examples, see JSON file.

Configure data file compression at the namespace level

Associate a compression policy at the namespace level and use a default filter that matches no data until individual tables opt in. Set filter to 1 = 0 in the policy constants JSON file:

```
"compressDataFiles": {
  "enabled": true,
  "method": "zstd",
  "level": 12,
  "filter": "1 = 0"
}
```

This value passes validation and produces a no-op for any table that does not have a table-level filter override. Iceberg evaluates the filter against file metadata, finds no matching files, and exits without rewriting anything.

Opt in tables individually by setting the filter with ALTER TABLE … SET TBLPROPERTIES:

```
ALTER TABLE catalog.db.events SET TBLPROPERTIES (
  'dlm.compressDataFiles.filter' = 'event_date < ''2024-01-01'''
);
```



Important: The compression action is still dispatched to Spark for every associated table on each scheduled run, including tables that do not have a table-level filter override. For namespaces with many tables that will never opt in, scope the policy to a narrower set of tables instead of using 1 = 0 as a permanent default.

Supported codecs for data file compression

Table 11: Compression codecs

Codec	Level range	Notes
zstd	1–22	Best compression ratio; recommended for cold data.
gzip	1–9	Wide compatibility.
brotli	0–11	—
snappy	—	No level; optimizes for speed.
lz4	—	No level.
lzo	—	No level.
zlib	—	No level.
uncompressed	—	Removes compression.
none	—	Alias for uncompressed.

Codec availability depends on the table file format (Parquet, ORC, or Avro). Cloudera Lakehouse Optimizer accepts any value from this list; format-specific filtering happens at execution time in Spark.

Per-table overrides for data file compression

Set the following Iceberg table properties with ALTER TABLE … SET TBLPROPERTIES. Table properties override namespace-level policy constants.

Table 12: compressDataFiles table properties

Table property	Overrides	Description
dlm.compressDataFiles.method	method	Overrides the compression codec from policy constants.
dlm.compressDataFiles.level	level	Overrides the compression level from policy constants.
dlm.compressDataFiles.filter	filter	Overrides the SQL filter from policy constants.

Table property	Overrides	Description
dlm.compressDataFiles.validateFilterColumns	validateFilterColumns	Set to false to skip pre-flight column validation for this table.

Column validation for data file compression

By default, Cloudera Lakehouse Optimizer validates that all column names referenced in the filter exist in the table schema before dispatching the task. If a column is missing, the action is skipped for that table and an error is logged.

Disable this pre-flight check when the filter is known-correct but references constructs the column extractor cannot parse, such as quoted identifiers or complex sub-expressions. Spark still evaluates the filter at runtime. Set validateFilterColumns to false in policy constants or set dlm.compressDataFiles.validateFilterColumns to false on the table.

Cloudera Lakehouse Optimizer REST APIs

Cloudera Lakehouse Optimizer provides REST APIs that you can use to create, manage, and monitor the Cloudera Lakehouse Optimizer policies. You can also use REST APIs to onboard namespaces, pause and resume maintenance, perform manual table maintenance, view task status and metadata, and fetch the details of different policy-related operations.

You can use the following REST APIs, when necessary, to perform the maintenance tasks:

REST API	Required parameters	Role required	Description
			config
`GET /config/backup`	`name string (query)`	• Administrator	Creates a back
`GET /config/health`	`scope string (query)`	• Administrator • Operator • Monitor	Fetches the ful By default, the When high ava standby. The c • Active - L • Passive - S • No leader
`POST /config/reconfigure`	-	• Administrator	Fetches the lat
`POST /config/restore`	`simulate boolean (query)`	• Administrator	Restores the ba
			namespaces
`GET /namespaces/{namespace}/policies`	`namespace string (path)`	• Administrator • Operator • Monitor	Enter the requi

REST API	Required parameters	Role required	Description
`GET /namespaces/active`	-	• Administrator • Operator • Monitor	Fetches all the
`GET /namespaces`	`fetch boolean (query)`	• Administrator • Operator • Monitor	Fetches all the maintained by
`GET /namespaces/{namespace}`	`namespace string (path)`	• Administrator • Operator • Monitor	Enter the requi
`PUT /namespaces/{namespace}`	`namespace string (path)`	Administrator	Enter the requi Optimizer.
`DELETE /namespaces/{namespace}`	`namespace string (path)`	Administrator	Enter the requi
`PATCH /namespaces/{namespace}`	`namespace string (path)`	• Administrator • Operator	Enter the requi
`PUT /namespaces/{namespace}/paused`	`namespace string (path)`	• Administrator • Operator	Enter the requi namespace fro
`DELETE /namespaces/{namespace}/paused`	`namespace string (path)`	• Administrator • Operator	Enter the requi namespace to t

REST API	Required parameters	Role required	Description
			policies
`POST /policies/{policyname}/association`	`policyName string (path)`	• Administrator • Operator	Enter the polic list of dot-sepa For example, v **catalog a**, and

`GET /policies/{policyName}/scope/{resourceScope}/def`	`policyName string (path)`
	`resourceScope string (path)`
	`solve string (query)`
	`base64 boolean (query)`
		• Administrator
		• Operator
			Enter the polic specified resou

`PUT /policies/{policyName}/scope/{resourceScope}/def`	`policyName string (path)`
	`resourceScope string (path)`
	`base64 boolean (query)`
		Administrator	Enter the polic specified resou

`POST /policies/{policyName}/scope/{resourceScope}/def`	`policyName string (path)`
	`resourceScope string (path)`
	`base64 boolean (query)`
		Administrator	Enter the polic

`DELETE /policies/{policyName}/scope/{resourceScope}/def`	`policyName string (path)` `resourceScope string (path)`	Administrator	Enter the polic
`GET /policies/resource`	`uri string (query)`	• Administrator • Operator • Monitor	Enter the URI For example, t The dlm://tpp/ respectively. tps is the autho Similarly, the

REST API	Required parameters	Role required	Description
`DELETE /policies/paused`	-	• Administrator • Operator	Removes all th Cloudera Lake
`GET /policies/{policyVersion}/paused`	`policyVersion string (path)`	• Administrator • Operator • Monitor	Enter the polic For example, t
`DELETE /policies/{policyVersion}/ paused`	`policyVersion string (path)`	• Administrator • Operator	Enter the polic Cloudera Lake

`GET /policies/{policyName}/tables/ {tableName}/desc`	`policyName string (path)`
	
	`tableName string (path)`
		• Administrator
		• Operator
		• Monitor
			Enter the polic

`POST /policies/{policyName}/catalogs/ {catalogName}/subs`	`policyName string (path)`
	
	`catalogName string (path)`
		• Administrator
		• Operator
			Enter the polic
			across all name

`DELETE /policies/{policyName}/ catalogs/{catalogName}/subs`	`policyName string (path)`
	
	`catalogName string (path)`
		• Administrator
		• Operator
			Enter the polic
			all the tables in

REST API	Required parameters	Role required	Description
			tables
`GET /tables/{tableId}/paused`	`tableId string (path)`	• Administrator • Operator • Monitor	Enter the table
`PUT /tables/{tableId}/paused`	`tableId string (path)`	• Administrator • Operator	Enter the table
`DELETE /tables/{tableId}/paused`	`tableId string (path)`	• Administrator • Operator	Enter the table Cloudera Lake
`GET /tables/{tableId}/policies`	`tableId string (path)`	• Administrator • Operator • Monitor	Enter the table

`GET /tables/{tableId}/stats`	`tableId string (path)`
	
	`force boolean (query)`
		• Administrator
		• Operator
		• Monitor
			Enter the table
			You can also s `hive.defau`
			If the specified

`GET /tables/paused`	-	• Administrator • Operator • Monitor	Fetches the list
`DELETE /tables/paused`	-	• Administrator • Operator	Removes all th Cloudera Lake
			task details
`GET /tasks/ingestlog`	-	• Administrator • Operator • Monitor	Fetches the eve

REST API	Required parameters	Role required	Description
`POST /tasks/ingestlog`	-	• Administrator • Operator	Triggers the Li

`GET /tasks/id/{id}`	`id`
	`string`
	`(path)`
		• Administrator
		• Operator
		• Monitor
			Enter the task

`GET /tasks`	`policy`
	`string`
	`(query)`
	
	`table`
	`string`
	`(query)`
	
	`status`
	`string`
	`(query)`
	
	`sortBy`
	`string`
	`(query)`
	
	`count`
	`string`
	`(query)`
		• Administrator
		• Operator
		• Monitor
			Fetches the list Configuration
			Optionally, yo
			• the policy
			• the table n
			• the task st
			• the sortBy
			• the count t

`GET /tasks/policies/{policyName}/tables/{tableName}/submitted`	`policyName`
	`string`
	`(path)`
	
	`tableName`
	`string`
	`(path)`
		• Administrator
		• Operator
		• Monitor
			Enter the polic

Manage and monitor Cloudera Lakehouse Optimizer table maintenance tasks

You can perform several policy managing activities, such as pausing and resuming table maintenance, associating policies, and fetching maintenance task details. You can also perform several monitoring tasks, such as verifying whether the tasks completed successfully, monitoring the Spark jobs on the Cloudera Observability dashboard, and viewing logs to troubleshoot issues.

For information about monitoring the Spark jobs on the Cloudera Observability dashboard, see Monitoring table maintenance tasks on Cloudera Observability dashboard.

Associating policies and fetching details using Cloudera Lakehouse Optimizer REST APIs

You can append or remove a policy association, fetch the list of policies associated with a specific policy, specific table, or at catalog level, fetch the policy names associated with all the tables in a specific namespace using Cloudera Lakehouse Optimizer REST APIs.

Procedure

- Append or remove the association of a policy to the table in the association file using the following APIs:
  - Define the association using the `POST /policies/{policyName}/tables/{tableName}/subs` API.
  - Add or update the association using the `PUT /policies/{policyName}/tables/{tableName}/subs` API.
  - Delete the association using the `DELETE /policies/{policyName}/tables/{tableName}/subs` API.
- List the table names associated with a specific policy using the `GET /policies/{policyName}/tables` API.
- List the policy names associated with the specified table using the `GET /tables/{tableId}/policies` API.
- List the policies associated with all the tables in a catalog using the `GET /policies/active` API.
- Retrieve the policy names associated with all the tables in the specified namespace using the `GET /namespaces/{namespace}/policies` API.
- List all the active onboarded namespaces available for table maintenance using the `GET /namespaces/active` API. When you set the fetch parameter for the `GET /namespaces` API to true, the API fetches all the namespaces in the catalog. When you set it to false, the API fetches only the declared namespaces to be maintained by Cloudera Lakehouse Optimizer.

Pausing and resuming table maintenance

You can pause or suspend the maintenance of a table, when required. For example, when there are a lot of writes or bulk ingestion. Cloudera Lakehouse Optimizer also pauses the table maintenance when certain criteria are met. You can resume the table maintenance when required.

About this task

The following steps show how to pause or resume table maintenance, how to list the policies in the paused list, and how to retrieve the reason for placing a table or namespace in the paused list.

Procedure

- Pause table maintenance using one of the following methods:
  - Go to the Cloudera Management Console Lakehouse Optimizer UI Tables tab, click  for a table, and then click Pause Table.
  - Use one of the following REST APIs:
    - The `PUT /tables/{tableId}/paused` API pauses the maintenance of an Iceberg table.
    - The `PUT /namespaces/{namespace}/paused` API pauses the maintenance of Iceberg tables in the specified namespace.

  When you or Cloudera Lakehouse Optimizer places a table or namespace in the paused list, Cloudera Lakehouse Optimizer ignores the tables and namespaces in the paused list during the evaluation phase.

  Important: When you pause maintenance for a table or namespace, all the policies for that table or namespace are paused automatically. Therefore, you must be cautious before you pause maintenance for a table or a namespace.

Important:

- Cloudera Lakehouse Optimizer places a policy in the paused-policy list if the policy fails multiple times during the *evaluation* phase.
- Cloudera Lakehouse Optimizer places a table in the paused-table list if any policy for this table fails multiple times during the *execution* phase. This action is performed when the number of recurring failures exceed the set retry value.

  You can configure the number of retries using the Cloudera Manager Clusters *CLOUDERA_LAKEHOUSE_OPTIMIZER* Configuration dlm.table.tolerable.errors.count property. The default value is 10.

- Resume table maintenance using one of the following methods:
  - Go to the Cloudera Management Console Lakehouse Optimizer UI Tables tab, click  for a table, and then click Unpause Table.
  - Resume table maintenance by using one of the listed REST APIs:
    - Removes a table from the paused list using the `DELETE /tables/{tableId}/paused` API.
    - Removes the tables in the specified namespace from the paused list using the `DELETE /namespaces/{namespace}/paused` API.
    - Removes all the policies from the paused list using the `DELETE /policies/paused` API.
    - Removes the specified policy from the paused list using the `DELETE /policies/{policyVersion}/paused` API.
    - Removes all the tables from the paused list using the `DELETE /tables/paused` API.

  When you remove the table, policy, or namespace from the paused list, the evaluation phase resumes for the table maintenance depending on the CRON schedule set in the policy.
- List the policies in the paused list using the `GET /policies/paused` API.
- Retrieve the reason for pausing the policy by using one of the following APIs:
  - The `GET /policies/{policyVersion}/paused` API gets the reason for the specified paused policy.
  - The `GET /tables/{tableId}/paused` API gets the reason for the specified paused table.

Viewing table maintenance status

You can verify the table maintenance action status on the Lakehouse Optimizer UI or using REST APIs.

Procedure

1. Use one of the following methods to verify the table maintenance status on the Cloudera Management Console Lakehouse Optimizer UI :
   - Go to the Policies tab, and click Actions View for the required policy.

     The policy details appear on the View Policy page, and the **Status** column shows the current status for the table.
   - On the Tables tab, find the required table. The **Status** column shows the current status for the table.
2. Perform the following steps to verify the table maintenance status using REST APIs:
   a) Fetch the **submitted** task details for a specific policy and tables using the `GET /tasks/policies/{policyName}/tables/{tableName}/submitted` API.
   b) Monitor the status of all the tasks using the `GET /tasks` API. You can fetch the task metadata of a specific task ID using the `GET /tasks/id/{id}` API.

   The following sample snippet shows the response for the 019348e0-2d6f-7997-b117-8441d8d757a6 taskID.

```
{
    "status": 200,
    "action": "READ",
```

```
    "message": "Success",
    "result": [
        {
            "taskId": "019348e0-2d6f-7997-b117-8441d8d757a6",
            "createAt": "2024-11-20T09:21:11.534Z",
            "createdBy": 1,
            "tablePath": "hive.default.itest",
            "policyName": "ClouderaAdaptive",
            "policyVersion": "0be2a330-0282-56d6-bbc4-67c96ad7e9c0",
            "currentAction": "EXPIRE_SNAPSHOT",
            "actions": {
                "EXPIRE_SNAPSHOT": {
                    "result": {
                        "deletedDataFilesCount": 0.0,
                        "deletedEqualityDeleteFilesCount": 0.0,
                        "deletedPositionDeleteFilesCount": 0.0,
                        "deletedManifestsCount": 0.0,
                        "deletedManifestListsCount": 4.0,
                        "deletedStatisticsFilesCount": 0.0
                    },
                    "sparkMetrics": {
                        "jvmGcTime": 4637.0,
                        "totalBytesRead": 65983.0,
                        "totalBytesWritten": 0.0,
                        "taskCount": 610.0,
                        "executorCpuTime": 3.4583323352E10
                    },
                    "sparkRuntime": {
                        "applicationId": "application_1732084186181_00
15",
                        "jobIds": [
                            9.0,
                            10.0,
                            11.0,
                            12.0,
                            13.0,
                            14.0
                        ]
                    },
                    "status": "Success"
                }
            },
            "status": "SUBMITTED"
        }
    ]
}
```


c) Verify whether the statistics improved after the maintenance actions completed using the `GET /tables/{tableId}/stats` API.

Viewing logs for Cloudera Lakehouse Optimizer

You can view the service log files and event logs for Cloudera Lakehouse Optimizer.

Procedure

- **Event logs**

  By default, the Cloudera Lakehouse Optimizer Data Hub contains a read-only Iceberg table named sys.clo_events in the default namespace to host the task metadata or event logs which enables historical trend analysis through queries. Cloudera Lakehouse Optimizer automatically logs the maintenance task metadata into this table every 20 minutes. You can query the table for root cause analysis, troubleshooting, and reporting purposes.

  You can perform the following operations related to this table:

- Configure the ingestion schedule in the Cloudera Manager Clusters *CLOUDERA_LAKEHOUSE_OPTIMIZER* Configuration dlm.task.cleanup.and.ingestion.interval.seconds advanced configuration snippet.
- Change the namespace from sys to another namespace in the Cloudera Manager Clusters *CLOUDERA_LAKEHOUSE_OPTIMIZER* Configuration conf/dlm-client.properties_role_safety_valve dlm.event.log.namespace=*[***ENTER NAMESPACE NAME***]* property.
- Change the table name from task_events to another table name in the Cloudera Manager Clusters *CLOUDERA_LAKEHOUSE_OPTIMIZER* Configuration dlm.task.event.log.iceberg.table advanced configuration snippet.
- Run queries on the table. For example, the `select * from sys.clo_events where tablename='demo_table';` query fetches the event log details for the demo_table Iceberg table in the default namespace; the `select count(*) as maintenance_tasks, [***TABLE NAME***] from sys.clo_events group by [***TABLE NAME***];` query fetches the number of maintenance tasks that ran on the specified table.

- **Service log files**

  The service log files are available in Cloudera Manager. Perform the following steps to view the service logs.

  a) Click CLO Server on the Cloudera Manager Clusters *CLOUDERA_LAKEHOUSE_OPTIMIZER* Instances page.
  b) Click the Log Files drop-down, and then click Role Log File to view the complete service log.

Troubleshooting: Configuring the Spark engine to optimize the rewrite compaction job

Optimize RewriteDatafiles (compaction) task in a Cloudera Lakehouse Optimizer policy by configuring the Spark engine, identifying and removing an overloaded YARN Node Manager, and tuning policy parameters.

Problem

The Cloudera Lakehouse Optimizer policy for the RewriteDatafiles (compaction) task failed on a very large Iceberg table with the java.lang.RuntimeException: Failed while running parallel task error caused by the java.lang.OutOfMemoryError: Java heap space error.

The Iceberg table in this use case has the following details:

- Snapshots = 9,603
- Total size = ~708 GB
- Data files = 2,728,148 (numerous very small files)
- Partitions = 100 buckets

Cause

The RewriteDatafiles task failed because of the following cascade of errors:

- The table contained millions of small data files, which generated a large number of manifest files. During the compaction process, the Spark driver had to scan and decompress these manifests to plan the file groups. This process created a driver heap memory causing the out-of-memory (OOM) error.
- One of the YARN nodes was running additional heavy services (HDFS NameNode, HBase RegionServer) that consumed a lot of memory. This caused a cluster resource crunch, severely reducing the effective memory available for Spark and YARN workloads.

Solution

Procedure

1. Optimize driver memory by increasing the following parameters on the Cloudera Manager Clusters Cloudera Lakehouse Optimizer Configuration conf/dlm-client.properties_role_safety_valve property:

   - spark.driver.memory=16g

- spark.executor.cores=5

Increase in driver memory resulted in successful scan of table metadata and successful planning and rewriting of file groups. Increase in executor cores resulted in increased concurrent tasks and improved compaction task speed.

2. Clean up YARN Node Managers by removing the overloaded NodeManager from the YARN pool for Cloudera Lakehouse Optimizer service.
3. Tune the following rewriteDatafiles policy parameters in the Cloudera Lakehouse Optimizer policy to keep each file group reasonably large for efficiency, and to avoid overwhelming the resources with too many concurrent file groups:
   - max-file-group-size-bytes = 3g
   - maxConcurrentRewriteFileGroups = 3

Outcome

Applying these configurations has the following results:

- The service successfully scanned Iceberg metadata and planned file groups.
- The service started rewriting file groups.
- In this specific scenario, the optimization reduced the number of files from ~2.2M to ~0.7M.