Cloudera Runtime 7.3.1

Cloudera Iceberg REST Catalog for Data Sharing Overview

Date published: 2024-12-11
Date modified: 2025-11-06

CLOUDƎRA

Legal Notice

Contents

Cloudera Iceberg REST Catalog overview

Cloudera Iceberg REST Catalog is a server-side implementation based on Apache Iceberg REST Catalog Open API, extended with enterprise-grade security and governance through Apache Ranger and Apache Knox, enabling REST-enabled third-party tools to manage Iceberg table metadata.

Apache Iceberg is a table format for huge analytics datasets in the cloud that defines how metadata is stored and data files are organized. Iceberg is also a library that compute engines can use to read or write a table.

Cloudera supports a data lakehouse architecture by pre-integrating and unifying the capabilities of Data Warehouses and Data Lakes to support data engineering, business intelligence, and machine learning—all on a single platform. Cloudera's support for an open data lakehouse brings high-performance, self-service reporting and analytics to your business, simplifying data management for both data practitioners and administrators. Cloudera's open data lakehouse is built on Apache Iceberg, which makes it easy to manage operational metadata.

Table operations like creating, dropping, or renaming tables are handled by a catalog. An Iceberg catalog helps query engines to manage and organize collections of tables, which are usually grouped into namespaces. Iceberg catalog implementations can be of various types, such as REST, HiveCatalog, JDBC, HadoopCatalog, Nessie, and so on. For more information about Iceberg catalogs, see the Apache Iceberg Catalogs documentation.

Cloudera supports the Iceberg REST catalog server implementation, which enables the server to expose Iceberg table metadata to REST-enabled third-party compute engines and tools like Spark, Trino, Snowflake, AWS Athena, AWS EMR Spark, and Databricks.

Cloudera Iceberg REST Catalog is a server-side catalog based on the Apache Iceberg REST Catalog Open API specification and is exposed through REST APIs to manage Iceberg tables. The REST catalog provides API endpoints to perform table management tasks, such as creating, listing, updating, or deleting tables. It also allows users to access and manage Iceberg table metadata.
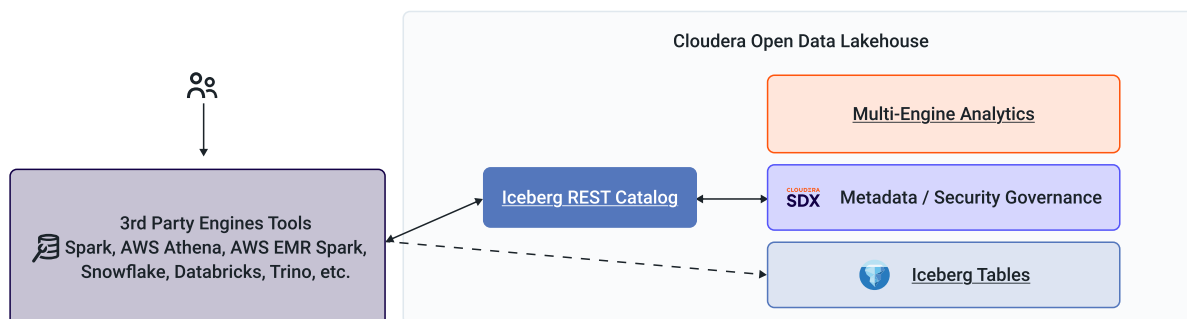
As part of the current offering, the Iceberg REST Catalog service is provided as an embedded service that runs within the same Java Virtual Machine (JVM) as the Hive Metastore (HMS). The scope of the metadata that the catalog can serve is limited to what the host HMS can serve. The REST catalog can be made available to third-party engines by using the OAuth authentication mechanism defined by the Knox Token management system, and you can use Apache Ranger policies to govern the Iceberg data that is accessed.

The REST service is deployed within the HMS through an embedded jetty engine. The configuration for the REST Catalog server is within the host HMS, and any changes made to the configurations will affect the REST service.

Related Information

Setting up Cloudera Iceberg REST Catalog for data sharing

How data sharing with Cloudera Iceberg REST Catalog works

Iceberg REST Catalog API specification

Data sharing with Cloudera Iceberg REST Catalog

Cloudera Iceberg REST Catalog enables secure metadata-level access to Iceberg tables from external engines that support the Iceberg REST API. It can serve as the foundation for secure data sharing, when combined with proper access control and network configuration.

Note: This feature is available as technical preview and is under entitlement. To obtain the required entitlement, contact your Cloudera Account Representative.

Cloudera Iceberg REST Catalog lets you share Cloudera on cloud data, specifically Iceberg tables, with external users (clients) who are outside of Cloudera environments. As a Cloudera user, you can share Iceberg table data with your clients. These clients can then access the data using third-party engines like Databricks or Snowflake that support the Iceberg REST catalog.

Cloudera Iceberg REST Catalog is implemented based on the Iceberg REST Catalog API specification. You can make data available to your clients by using the OAuth authentication mechanism defined by the Knox Token management system and using Apache Ranger policies for defining authorization rules for Iceberg tables.

Benefits of data sharing with Cloudera Iceberg REST Catalog

- Enables cross-platform data sharing. For example, you can share data from a Cloudera data lake with clients who use other data platforms like Databricks, Amazon EMR, Snowflake, Splunk, and more.
- Provides zero-copy Cloudera live data, meaning no replication is needed, and there is no latency.
- Offers centralized governance of data using Apache Ranger and Atlas
- Secure collaboration between customers and their clients
- Creates a foundation for building a marketplace to share datasets, notebooks, and Cloudera AI models.

Scope and assumptions

- In this release, only read access to Iceberg tables is supported.
- Clients accessing the data can use non-Cloudera engines that understand the Iceberg table format.
- Table access audit in Ranger is supported.
- Currently, this feature is supported only on AWS S3 storage in Cloudera on cloud environments.
- Ranger column masking and row-level filtering is not supported. You can restrict the access only to complete tables in a Data Share.
- The High Availability feature is not supported for Cloudera Data Sharing powered by REST Catalog.

Related Information

Cloudera Iceberg REST Catalog

Setting up Cloudera Iceberg REST Catalog for data sharing

How data sharing with Cloudera Iceberg REST Catalog works