CDP Public Cloud / Data Catalog

Cloudera Data Catalog Overview

Date published: 2019-11-14
Date modified: 2026-03-31



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

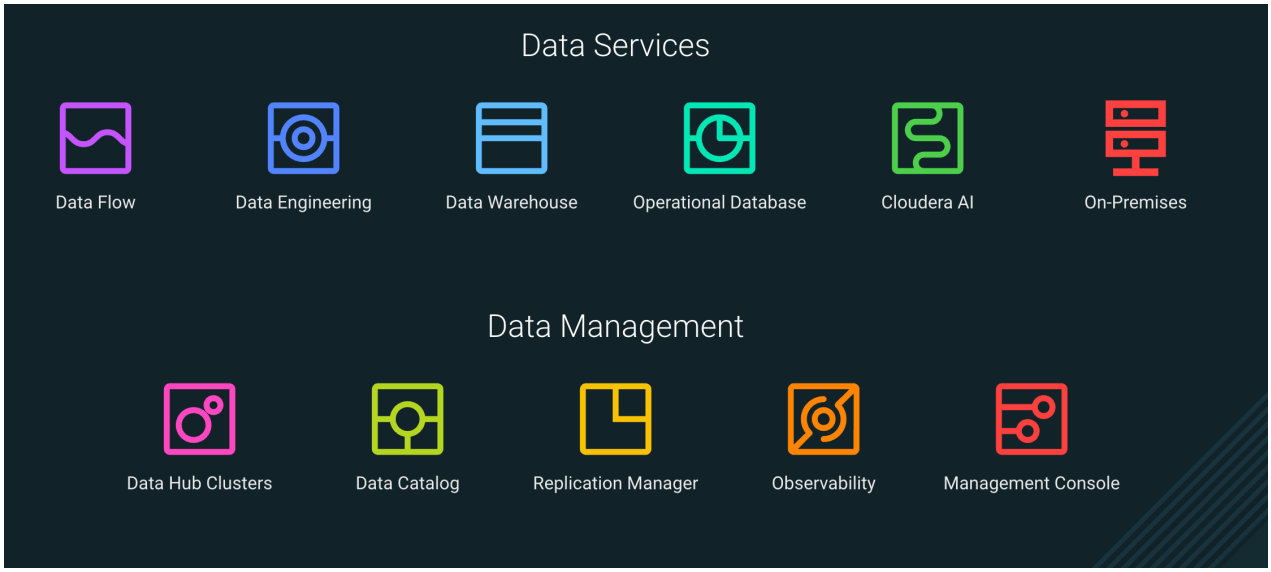
About Cloudera Data Catalog....4
Cloudera Data Catalog Dashboard....5
Profiler architecture in VM-based environments....7
Profiler architecture in Compute Cluster enabled environment....8
Cloudera Data Catalog terminology....9
Datasets overview....10
Data Sharing in Cloudera Data Catalog overview....11
Data Sharing user interface in Cloudera Data Catalog....12
Search overview....17
Bookmarks overview....19
Atlas tags overview....24

Before you start....24
Prerequisites to access Cloudera Data Catalog....25
Providing role access....25

About Cloudera Data Catalog

Cloudera Data Catalog is a service within Cloudera that enables you to understand, manage, secure, and govern data assets across the enterprise.
Cloudera Data Catalog helps you understand data in your Data Lake. You can locate relevant data of interest based on various parameters. Using Cloudera Data Catalog, you can understand how data is interpreted for use, how it is created and modified, and how data access is secured and protected.

Note: Cloudera Data Catalog is supported in the EU and APAC regional Control Plane. You can ensure that you manage, secure, collaborate, and govern data assets across multiple clusters and environments within the EU region or APAC region where your organization operates as per the data protection regulatory requirements.



Cloudera Data Catalog enables data stewards across the enterprise to work with data assets in the following ways:

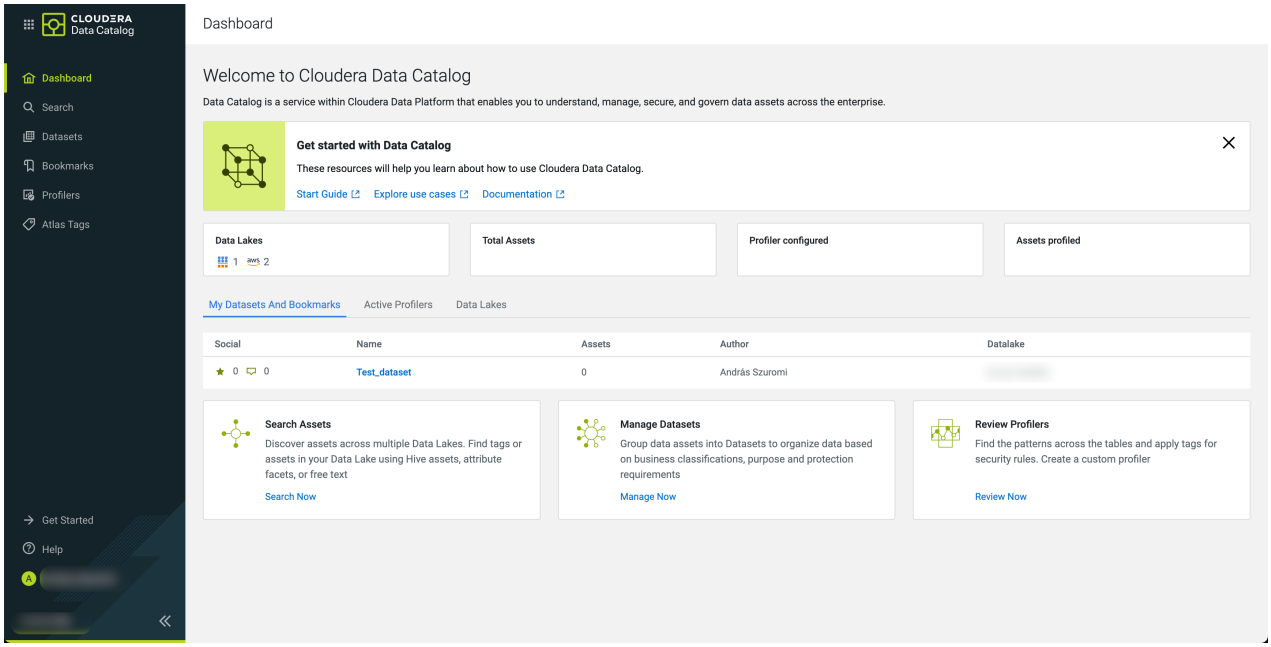
- Organize and curate data globally:
  - Organize data based on business classifications, purpose, protections needed, etc. For more information, see:
    - Creating a classification from Asset Details
    - Creating a classification from Search page
  - Promote responsible collaboration across enterprise data workers. For more information, see Bookmarks overview.
- Understand where relevant data is located:
  - Catalog and search to locate relevant data of interest (sensitive data, commonly used, high risk data, etc.).
  - Understand what types of sensitive personal data exists and where it is located
  - For more information, see:
    - VM-based environments:The Cluster Sensitivity Profiler
    - Compute cluster enabled environments: The Data Compliance Profiler
- Understand how data is interpreted for use:
  - View basic descriptions: schema, classifications (business cataloging), and encodings
  - View statistical models and parameters
  - View user annotations, wrangling scripts, view definitions etc.
  - For more information, see Viewing Data Asset details.

    - For more information, see Navigation in Asset Details.
    - VM-based environments: The Hive Column Profiler / Compute cluster enabled environments: The Statistics Collector Profiler
- Understand how data is created and modified:
    - - Visualize upstream lineage and downstream impact
        - Understand how schema or data evolve
        - View and understand data supply chain (pipelines, versioning, and evolution)
    - For more information, see Navigation support for hive entities within Lineage.
- Understand how data access is secured, protected, and audited:
    - - Understand who can see which data and metadata (for example, based on business classifications) and under what conditions (security policies, data protection, anonymization)
        - View who has accessed what data from a forensic audit or compliance perspective
        - Visualize access patterns and identify anomalies
    - For more information, see
        - VM-based environments: The Ranger Audit Profiler
        - Compute cluster enabled environments: The Activity Profiler
        - Viewing Ranger access audits
        - Viewing Atlas entity audits
        - Viewing Ranger policies

Cloudera Data Catalog Dashboard

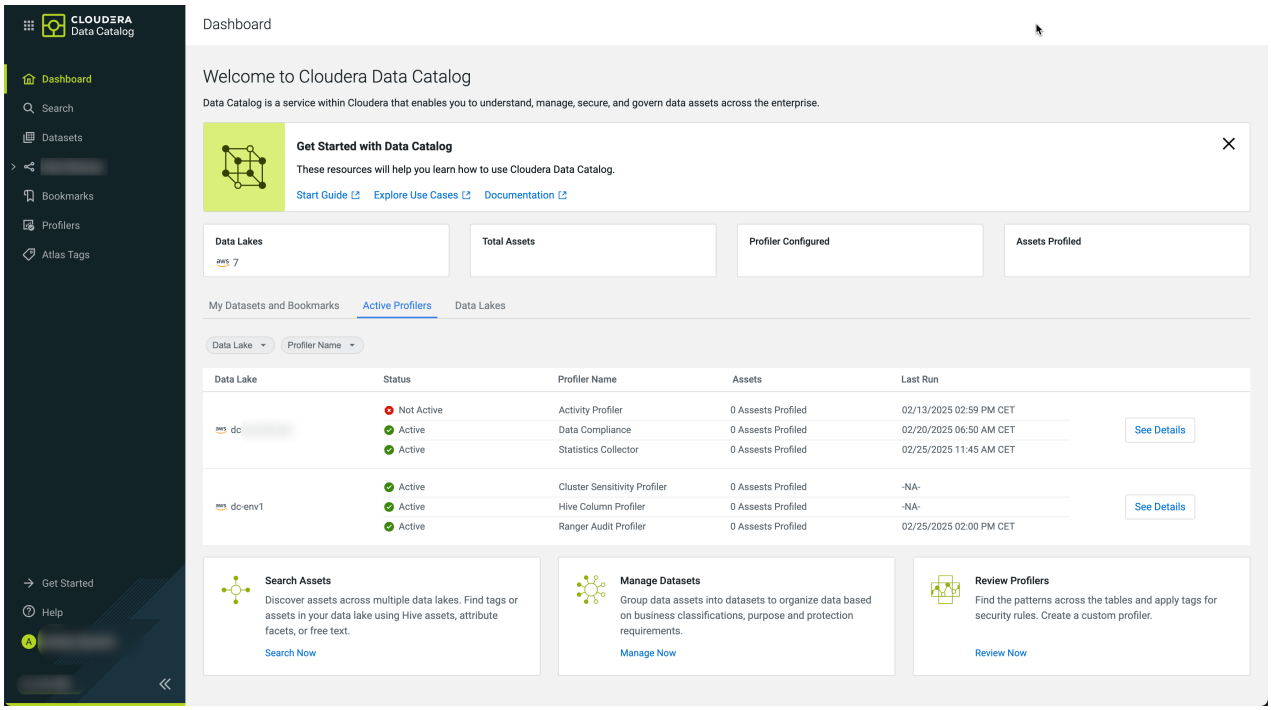
The Dashboard provides quick access to vital service information at a glance, in the form of visual, actionable navigation for multiple operations. The user-friendly navigation enables viewing, filtering, and acting upon data quickly and in a simple manner.

Data Stewards can view the **Dashboard** at a glance, and also focus on the most important tasks, enabling faster decision making as well as immediate action. The application lets you perform multiple actions for different types of content that helps in visualizing information with ease.



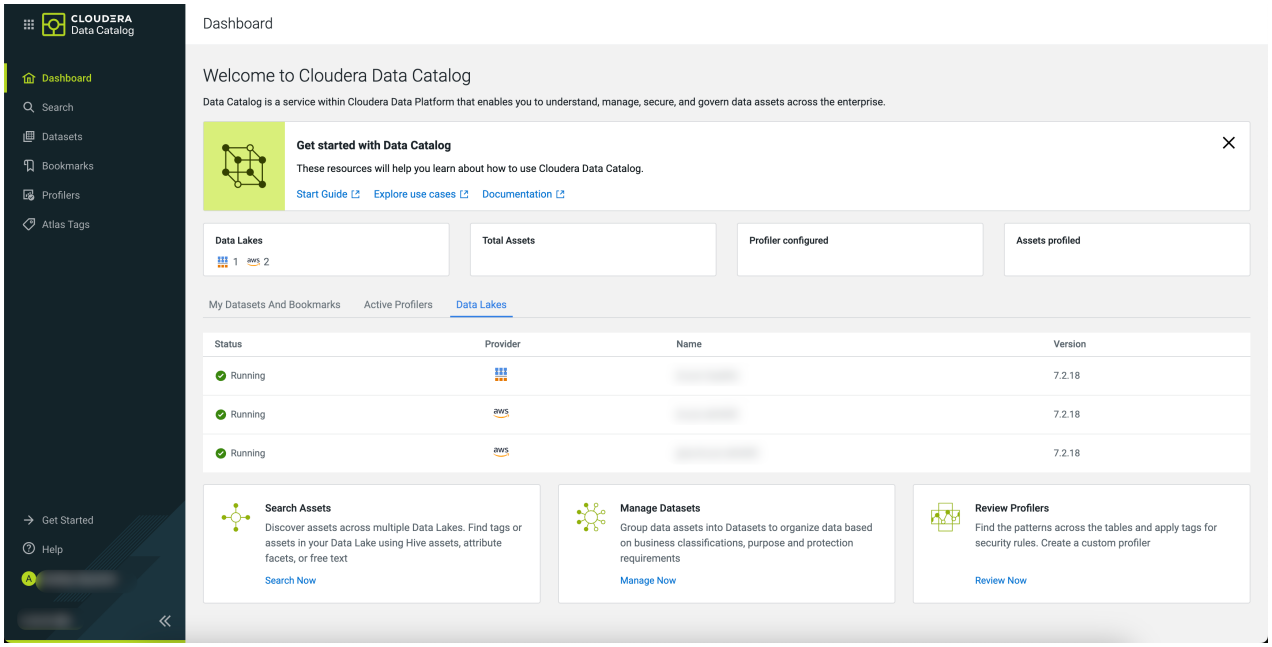
The displayed cards and tabs are fully interactive, with clickable areas for easy navigation to relevant parts of applications. Users can access individual sections and narrow down the information displayed. For example, you can manage the datasets created and bookmarked by you. For more information, see:

- Creating datasets
- Managing datasets
- Bookmarks overview



The **Dashboard** page contains information on your profilers, the total number of assets that are profiled, along with the assets that are scanned for data. Clicking See Details leads you to your individual profilers:

- VM-based environments:The Cluster Sensitivity Profiler / Compute cluster enabled environments: The Data Compliance Profiler
- VM-based environments: The Hive Column Profiler / Compute cluster enabled environments: The Statistics Collector Profiler
- VM-based environments: The Ranger Audit Profiler / Compute cluster enabled environments: The Activity Profiler



Additionally, you can check the status of your Data Lakes. For more information, see Introduction to Data Lakes.

Profiler architecture in VM-based environments

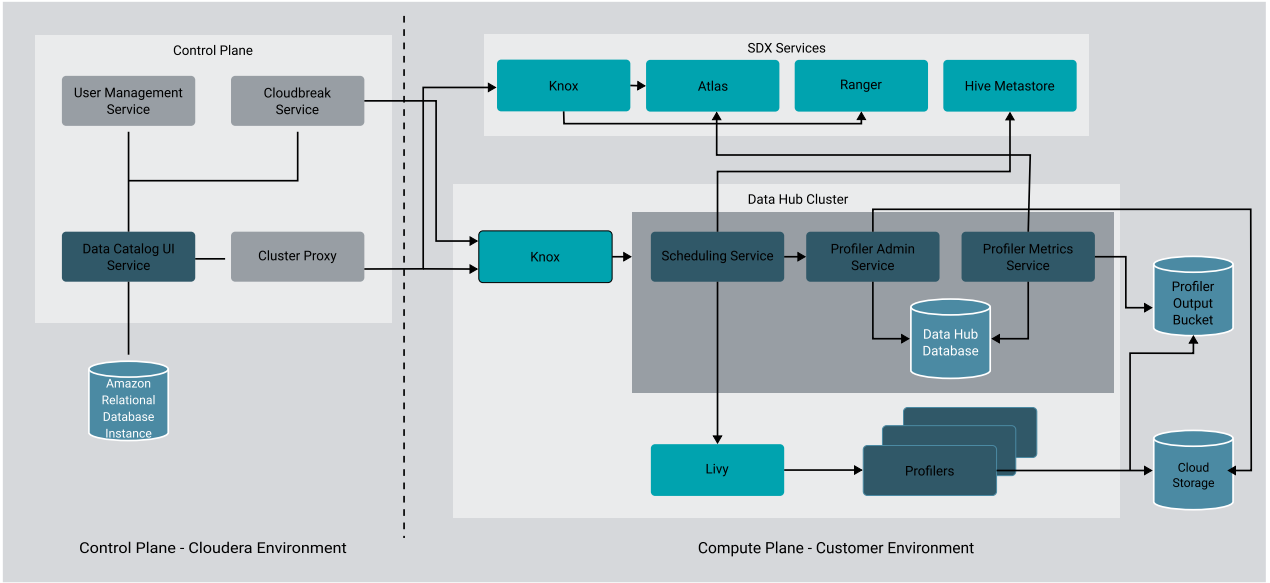
In a VM-based environment, the Cloudera Data Catalog Profiler architecture uses a Cloudera Data Hub workload cluster.

Note:

The VM-based architecture (using the Cluster) is deprecated from the 3.0.0 release but remains available until 7.2.18 is supported (Sept 2025). Therefore, based profilers will also not be available in versions after 7.2.18. Only Compute Cluster enabled environment will be able to run profilers after version 7.2.18.

For more information, see Cloudera Support lifecycle policy.

Figure 1: VM-based profiler architecture



After registering a VM-based environment, you have to launch a Cloudera Data Hub cluster for each Data Lake to provide the resources and services required for a profiler workload. This can be handled by Cloudera Data Catalog. For more information, see Launching profilers in VM based environments.

Note: In comparison to a Kubernetes pod in a Compute Cluster enabled environment, a Cloudera Data Hub workload cluster reserves compute resources even when a profiler task is not running. Also, more services are required to be included in the Cloudera Data Hub cluster template in contrast to the default compute cluster:

Zookeeper

For configuration information, naming, synchronization and group services over large clusters in distributed systems

Yarn

For resource management

1. Cloudera Data Hub uses the internal service called Cloudbreak to start the necessary services in the Profiler Cloudera Data Hub cluster. It is also used to access data about profilers and the Data Lake. In comparison, the Cluster Proxy provides the connection between the Cloudera Data Hub UI service and the rest of the Cloudera Data Catalog services.
2. An additional Amazon Relational Database (PostgreSQL) is used to store data required for the profiling process, such as, Cluster Sensitivity Profiler Rules, profiler-Data Lake mappings and datasets.
3. Knox is used to authenticate services between your and Cloudera's environment
4. Livy is used together with a dedicated Scheduler Service to start the individual profiler instances with Spark jobs.
5. The Cloudera Data Hub cluster manages the different services responsible for the profiling.
   a. Profiler Admin service is similar to an interface for Profilers. It allows Cloudera Data Hub to fetch information from the workload Cloudera Data Hub about scheduled jobs, profiler configurations and so on.

      Profiler Metrics is responsible for the metrics calculation and synching it to Cloudera Data Hub database and Atlas.
   b. The profilers use a cloud storage called Profiler output bucket as a temporary storage to aggregate all their collected data, such as profiler snapshots, which help to continue the profiling by saving interim data.
6. The final profiler results are stored in an attached cloud storage.

Related Information

Cloudera Data Hub

Introduction to Data Lakes

Cloudera Management Console

Profiler architecture in Compute Cluster enabled environment

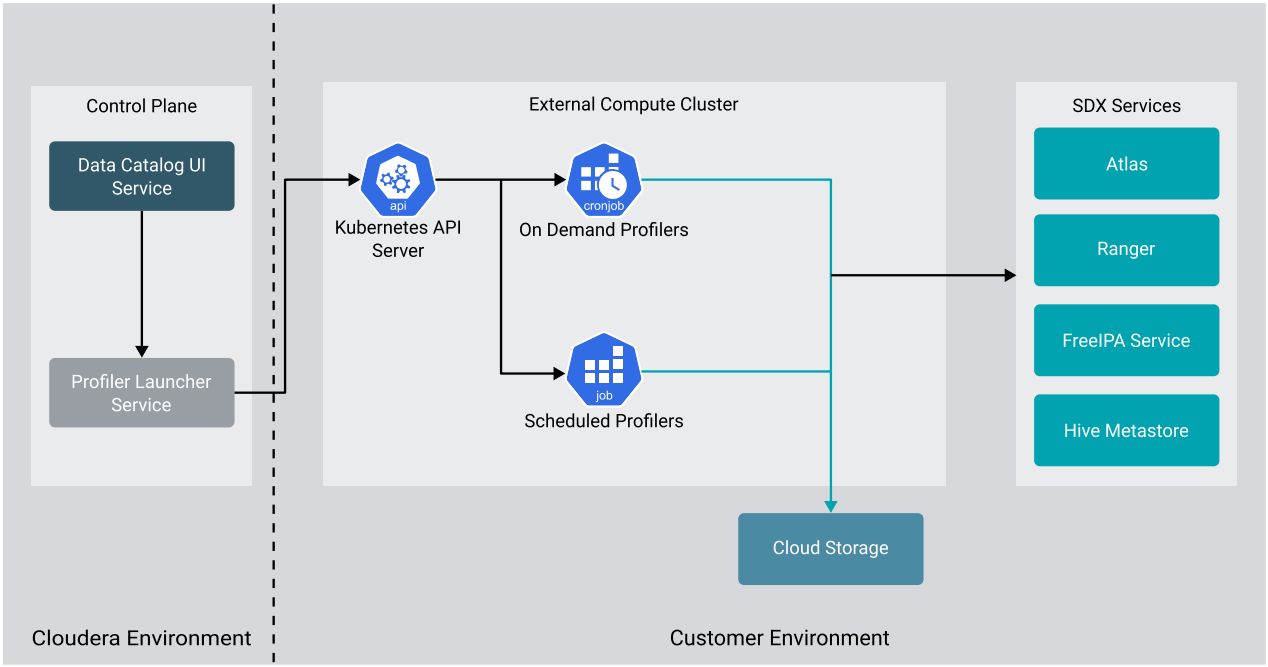
Next to the Cloudera Data Hub based profiler cluster, Cloudera Data Catalog offers the possibility to run profilers as a containerized service in a standardized Kubernetes base cluster called Externalized Compute Cluster. This consumes far less resources and provides auto-scaling.

Note:

The VM-based architecture (using the Cluster) is deprecated from the 3.0.0 release but remains available until 7.2.18 is supported (Sept 2025). Therefore, based profilers will also not be available in versions after 7.2.18. Only Compute Cluster enabled environment will be able to run profilers after version 7.2.18.

For more information, see Cloudera Support lifecycle policy.

Figure 2: Profiler architecture in Compute Cluster enabled environment



1. Once the container-ready environment is set up, a default Kubernetes cluster (Externalized Compute Cluster) is also created in this environment.

   Note: The Kubernetes jobs and API server offers the same API and UI interface capabilities for you as the VM-based Cloudera Data Hub, therefore, there is no difference in use.

2. The Profiler Launcher Service (PLS) internal to Cloudera Data Catalog schedules Kubernetes jobs, cron jobs in the compute cluster using HTTP API calls. PLS creates one Kubernetes CRON job for each profiler you select, so the number of CRON jobs scheduled on the compute cluster matches the number of launched profilers.
3. Once the time of the schedule is reached the Kubernetes job will launch a pod that will start profiling a Data Lake or ranger audit logs. The configuration for the jobs are received via the Cloudera Data Catalog API.
4. Using these settings the profiler connects to a Data Lake, identifies all the assets present in the Data Lake then starts profiling.
5. The results will be synced to Atlas and Cloudera Data Catalog using their respective APIs.

Related Information

Cloudera Data Hub
Cloudera Management Console
Using Compute Clusters in AWS environments
Using Compute Clusters in Azure environments

Cloudera Data Catalog terminology

An overview of terminology used in Cloudera Data Catalog.

Profiler

Enables the Cloudera Data Catalog service to gather and view information about different relevant characteristics of data such as shape, distribution, quality, and sensitivity which are important to understand and use the data effectively. For example, view the distribution between males and females in column Gender, or min/max/mean/null values in a column named avg_income. Profiled data is generated on a periodic basis from the profilers, which run at regularly scheduled intervals. Works with data sourced from Apache Ranger Audit Logs, Apache Atlas Metadata Store, and Hive.

Data Lake

A trusted and governed data repository that stores, processes, and provides access to many kinds of enterprise data to support data discovery, data preparation, analytics, insights, and predictive analytics. In the context of Cloudera, a Data Lake can be realized in practice with an Cloudera Manager enabled Cloudera Shared Data Experience cluster that runs Apache Atlas for metadata and governance services, Apache Knox, and Apache Ranger for security services.

Data Asset

A data asset is a physical asset located in the Cloudera ecosystem such as a Hive table which contains business or technical data. A data asset could include a specific instance of an Apache Hive database, table, or column. An asset can belong to multiple asset collections. Data assets are equivalent to “entities” in Apache Atlas.

Datasets

Datasets allow users of Cloudera Data Catalog to manage and govern various kinds of data objects as a single unit through a unified interface. Asset collections help organize and curate information about many assets based on many facets including data content and metadata, such as size/schema/tags/alterations, lineage, and impact on processes and downstream objects in addition to the display of security and governance policies.

Related Information

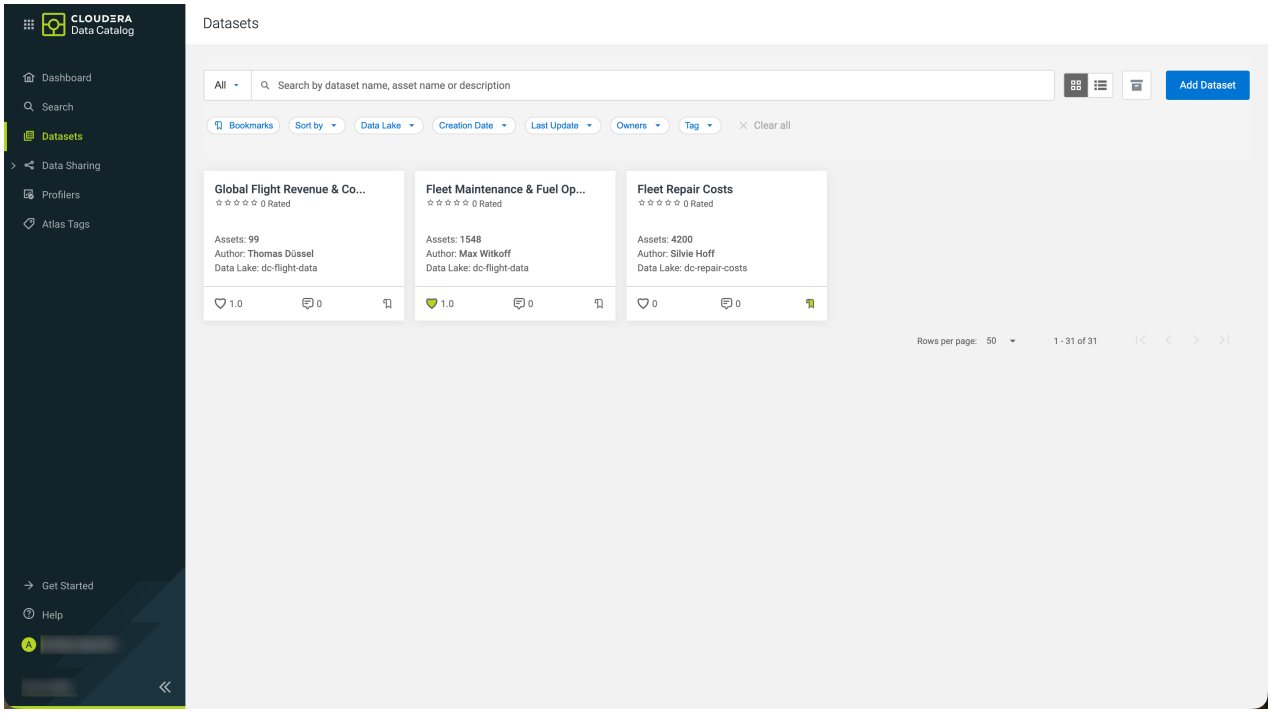
About Cloudera Data Catalog

Introduction to Data Lakes

Datasets overview

A dataset is a group of assets that fit a set of search criteria so that you can manage and administer them collectively for specific business purposes.

Figure 3: Datasets



Asset collections enable you to perform the following tasks when working with your data:

- Organize

  Group data assets into datasets based on business classifications, purpose, protections, relevance, ownership etc.

Example use cases:

- A company may mark all data that needs to be compliant with GDPR, for example, with the tag "PII" and use Ranger policies to control access to these assets.
- A data steward may mark all data related to sales with tags as sales_transactions, sales_targets, customer_segments and include them into a common dataset for easier discovery.
- Ownership by different teams can be signified by datasets.

- Search

  Find tags or assets in your Data Lake using Hive assets, attribute facets, or free text.

  Advanced asset search uses facets of technical and business metadata about the assets, such as those captured in Apache Atlas, to help users define and build collections of interest.
- Understand

  Audit data asset security and use for anomaly detection, forensic audit and compliance, and proper control mechanisms.

You can edit datasets after you create them and the assets contained within the collection will be updated. CRUD (Create, Read, Update, Delete) is supported for datasets.

Related Information

Managing datasets

Bookmarks overview

Data Sharing in Cloudera Data Catalog overview

Data Sharing in Cloudera Data Catalog enables secure, self-service access to Iceberg tables for external users using the Cloudera Iceberg REST Catalog. It simplifies data sharing by using the Cloudera Data Catalog user interface.

Important:

This feature is available under the entitlement CDP_DATA_SHARE_ADMIN. To obtain the required entitlement, contact your Cloudera Account Representative.

Cloudera Data Catalog provides a self-service, secure way to provide read-only access to Iceberg tables to external users building on Cloudera Iceberg REST Catalog, including compute engines supporting the Iceberg REST API, such as Snowflake, Databricks, Amazon EMR, Amazon Athena, and Amazon Redshift. Cloudera Data Sharing avoids duplicating data to external storage which enhances security, reduces storage costs and processing time especially by dynamic tables. For example a manufacturer (the Data Provider) can share manufacturing and inventory data to resellers or suppliers (Data Consumers). By using the Iceberg format, ETL processes can be avoided regardless of cloud environment providers and data ecosystems. Additionally, the Cloudera Iceberg REST Catalog uses time-bound Knox IDs and tokens (Client ID and Secret and access token) to manage security of the Data Sharing. Ranger audit reports are shown in Cloudera Data Catalog for Data Share updates and access events.

Cloudera Data Catalog greatly simplifies Data Sharing by providing centralized user interface and asset discovery (particularly useful when setting up Data Shares). It also enables adding metadata in the form keywords (to Data Shares) and Atlas classifications (to assets) to improve searchability. The user interface also provides a quick overview of audit reports for Data Share updates and access events.

Note:

- To enable Data Sharing in Cloudera Data Catalog, ensure that Configuring Hive Metastore as a REST Catalog is complete.
- Data Sharing in Cloudera Data Catalog is only supported for AWS environments. To use Data Sharing in Azure environments, see Creating a Data Share.

Who are Data Providers?

Designate Cloudera users as Data Providers to manage the responsibility of sharing data stored in Cloudera with external users or across organizations. They are usually data owners or data stewards. They will have the responsibility to designate Iceberg tables to be shared. Additionally, they have the responsibility to generate secure, timed credentials (Client ID and Secret) for users, and to control how long the data can be accessed.

Data Providers must have the DataShareAdmin resource role, limited for a specific cloud environment.

Although there can be more Data Providers in an environment, they have a federated view of all Data Shares across Data Lakes.



Note: Users with the DataShareAdmin resource role have access to all Hive resources of that environment.

Who are Data Consumers?

Data Providers can designate external users as Data Consumers by setting up their Knox Client IDs and Secrets. These credentials have to be shared with them so that they can enter them into their Iceberg REST Catalog compatible compute engines. For example, they can be users of a supply chain department or a third-party vendor consuming the manufacturing data to optimize their own operations.

How does Data Sharing work?

After the Data Providers set up Data Consumers by creating Knox credentials (Client ID, Secret), they create Data Shares, logical units of grouped Iceberg tables using the Cloudera Data Catalog user interface of CDP CLI. During Data Share creation, they assign the assets to be shared and Data Consumers. Behind the scenes Knox authorizes the users and manages the lifecycle of the credentials (expiration, revocation, and regeneration). Also, it routes the incoming read and authentication requests.

Ranger creates groups and the read-only policies for the designated Iceberg tables for the authenticated external users. Additionally, Ranger handles the asset and access auditing.

Hive provides the storage layer and serves the read request within the framework of Iceberg REST Catalog API.

Related Information

Cloudera Iceberg REST Catalog overview
Data sharing with Cloudera Iceberg REST Catalog
Creating external users
Creating a new Data Share
Modifying Data Shares
Revoking user credentials
Regenerating user credentials
Deleting a Data Share
Verifying external access to a Data Share
Troubleshooting Data Sharing after restoring a backup to a new Data Lake

Data Sharing user interface in Cloudera Data Catalog

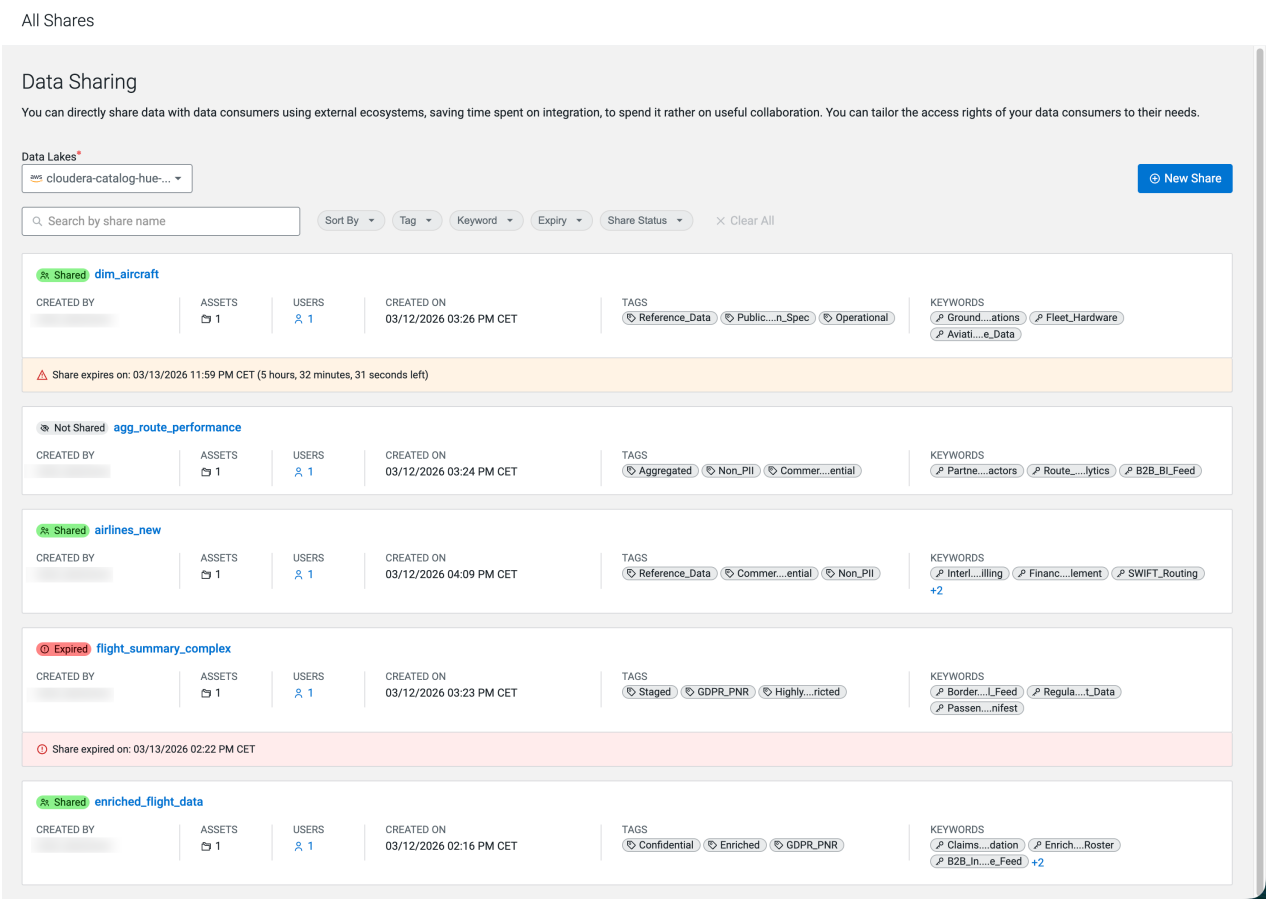
The Data Sharing user interface in Cloudera Data Catalog enables Data Providers to manage logical data shares and external user access. Key features include lifecycle management of data shares, detailed share metrics, and secure credential handling for external Data Consumers.

Data Providers interact with the following pages in Cloudera Data Catalog:

All Shares

The **All Shares** page serves as the central command hub where Data Providers can search, discover, and manage the lifecycles of their logical data shares. As Data Providers have a federated view, they can manage shares across multiple data lakes within a specific environment.

Figure 4: All Shares for Data Sharing



Share Creation and Modification

Admins can initiate the creation of a new Data Share, defining its name, summary, terms and conditions, and time-bound validity period. They can also modify existing shares (which temporarily places them in an unshared state) to add or remove assets and users.

Filtering and Discovery

The page provides filtering to easily locate shares. Admins can search by share name, or filter by specific Atlas Tags, custom Keywords (for example, "finance" or "B2B_Operational"), expiration status, and the current share status (for example, Shared or Not Shared).

Metrics

The main list view displays critical information for each share, including the number of assets included (up to a limit of 15 per share), the number of external users with access, and exact creation and expiration dates.

Share Details

Clicking into a specific share opens a detailed view with dedicated tabs for:

Assets

Lists the specific Iceberg tables included in the share.

Users

Lists the external consumers authorized to view the data.

Metadata Audits

Tracks administrative changes made to the share itself, such as adding a new table or changing the summary.

Figure 5: Accessing metadata audit details

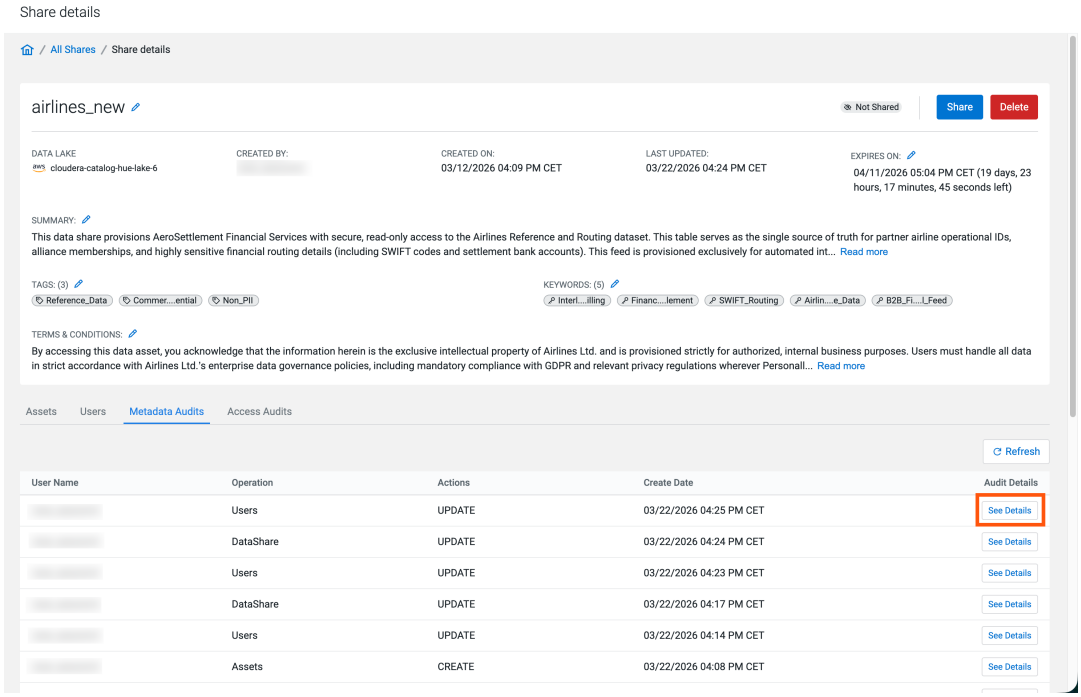
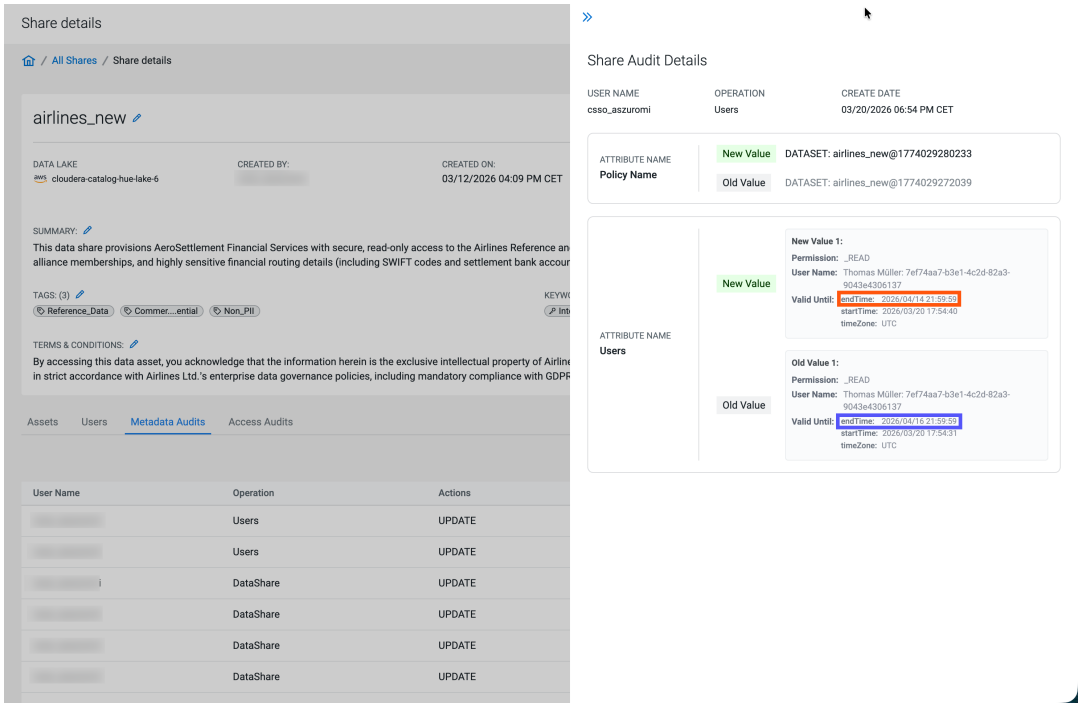
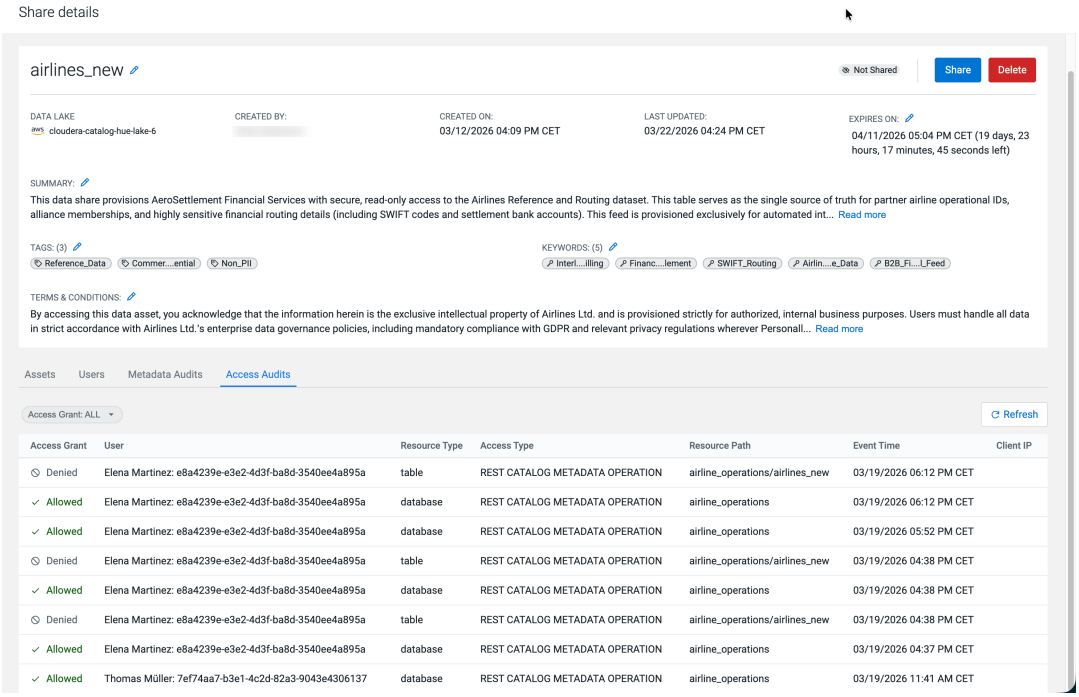


Figure 6: Metadata audit of updating the user access time for a Data Share



Access Audits

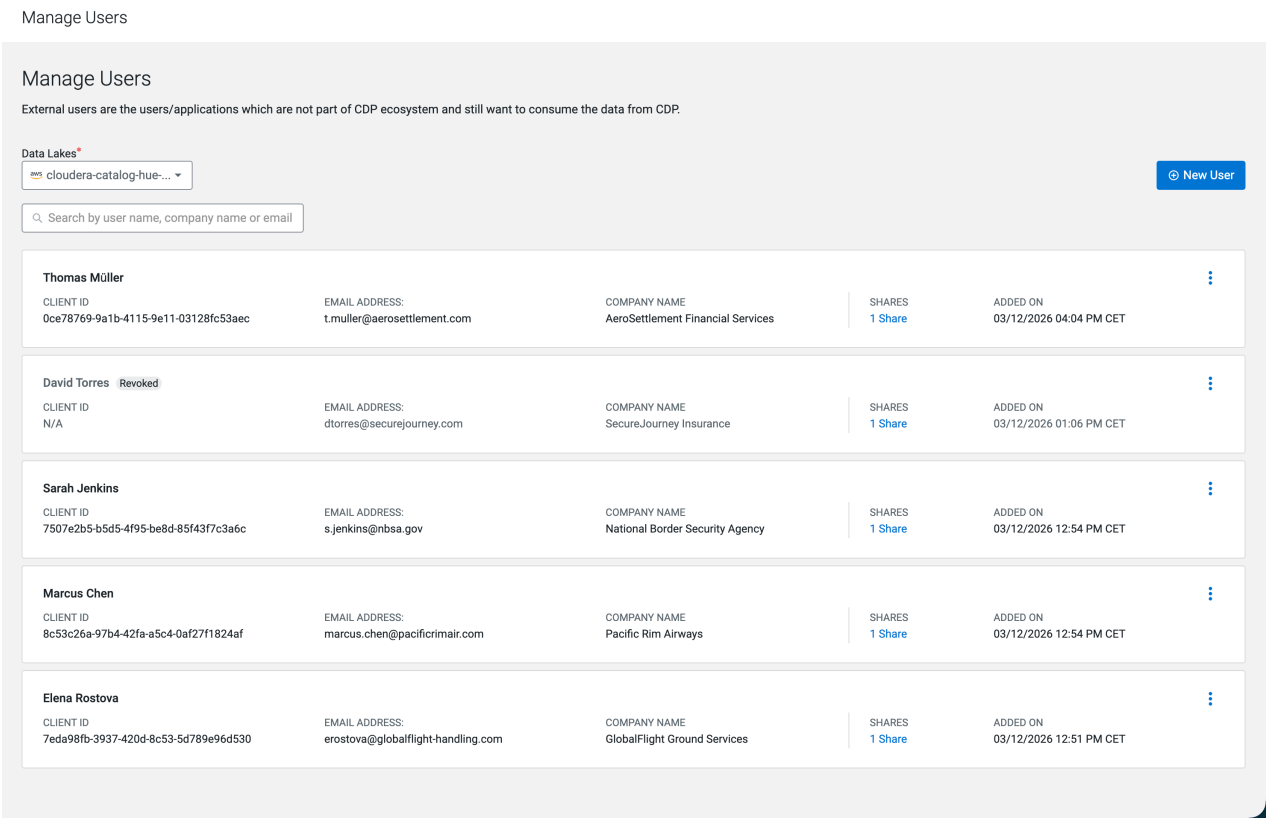
Displays logs of actual data access events, showing exactly which user queried which table and whether the API request was allowed or denied by Ranger.



Manage Users

The **Manage Users** page is dedicated to handling the identities and credentials of external Data Consumers. Because these consumers operate outside of the Cloudera ecosystem (using third-party engines like Snowflake or Databricks), this page acts as the bridge to securely grant them access.

Figure 7: List of external users



Key capabilities of this interface include:

External User Registration

Admins can register a new user by inputting their name, email address, and company name.

Secure Credential Generation

Upon registering a new user, the system automatically uses Knox to generate OAuth2 credentials (a Client ID and Secret). The UI immediately prompts the Share Admin to download these credentials as a CSV file.

Note: This is a one-time download; the secret is never displayed again, and the admin is responsible for securely distributing this file to the external consumer.

Credential Lifecycle Management

To ensure strict security, admins can use an action menu next to any user to fully control their access:

Revoke Credential

Immediately disables the user's access to any shares. The user remains in the system for auditing purposes, and the UI displays a Revoked status next to their name.

Regenerate Credential

Invalidates the user's old token and issues a brand new Client ID and Secret. This is highly useful if a consumer loses their credential file or if a security compromise is suspected.

Delete User

Permanently removes the user and revokes their access.

Note: These capabilities are also available via the CDP CLI. For more information, see the Related Information.

Related Information

Creating a Data Share with CDP CLI
CDP CLI datacatalog command

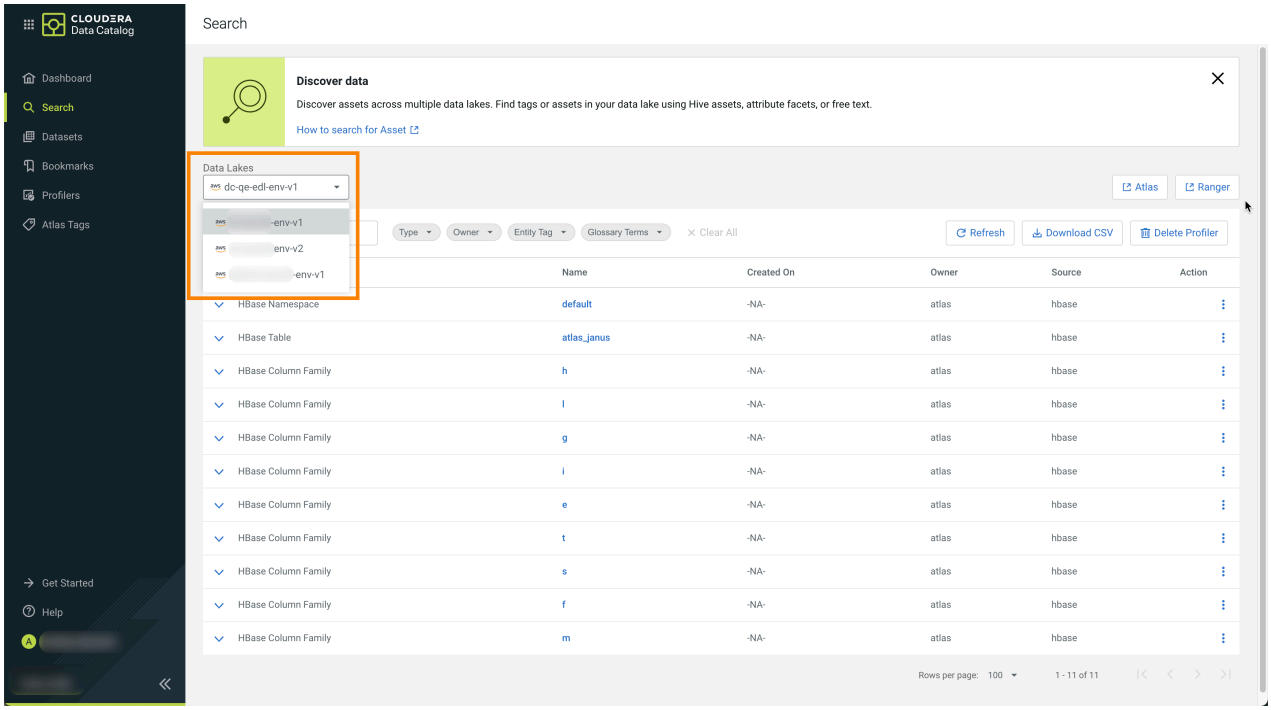
Search overview

On the Cloudera Data Catalog **Search** page, select a Data Lake and enter a search string in the search box to view all the assets with details that contain the search string.

When you enter the search terms in **Search**, you are looking up names, types, descriptions, and other metadata collected by Cloudera Data Catalog. The search index includes metadata, not data, about your environment and cluster data assets and operations. You can make the search more powerful by associating your own information (business metadata) to the stored assets.

Accessing Data Lakes

Users have access to the Data Lakes based on the permissions that are granted. You can choose from the available Data Lakes by selecting the appropriate radio button.

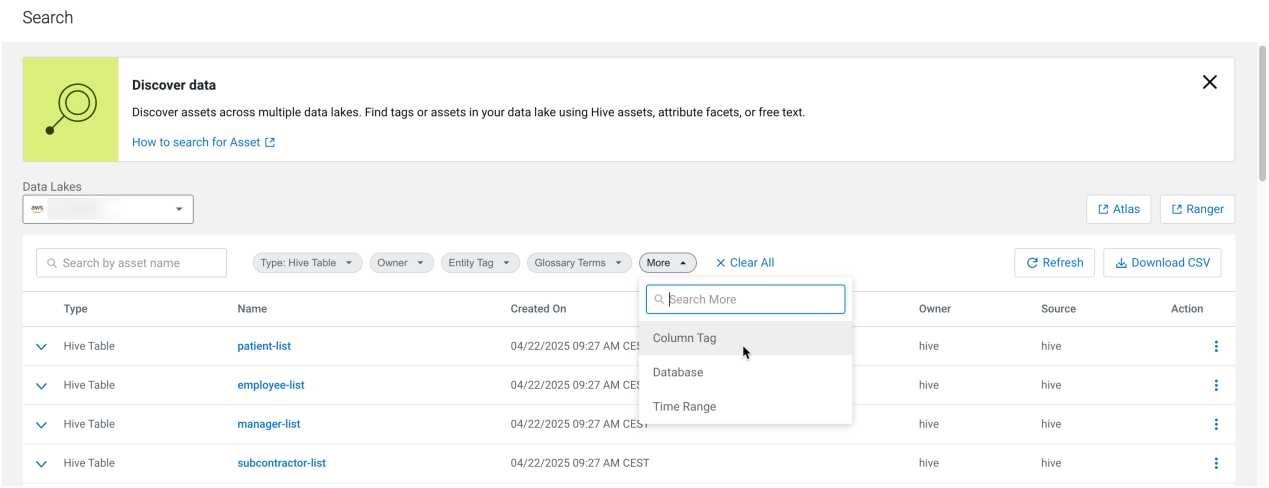


Note:

- You can search the assets of one Data Lake at a time.
- For the selected Data Lake, click the Atlas and Ranger links to go to the respective base cluster services in a new browser tab.

Using search filters

Use the search filter to fine-tune your results. By selecting a type, additional filter options become available and irrelevant filters are hidden. For example, after selecting Hive Table, the Column Tag, Database and Time Range filters are shown under More.

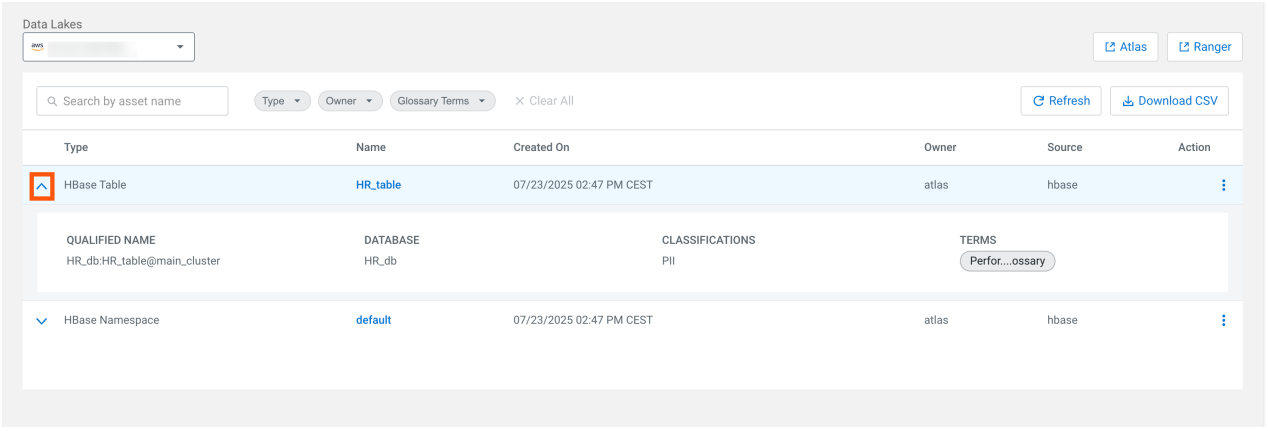


Viewing Asset Details

Clicking the icon for a search result shows the following key data about an asset:

- **Qualified name** – Qualified names are a unique identifier in Cloudera Data Catalog, identifying the asset with its context.

  A Hive table has the following qualified name patterns: *[***DATABASE NAME***].[***TABLE NAME***]@[***CLUSTER NAME***]*
- **Database**
- **Classifications** – Atlas tags
- **Terms**



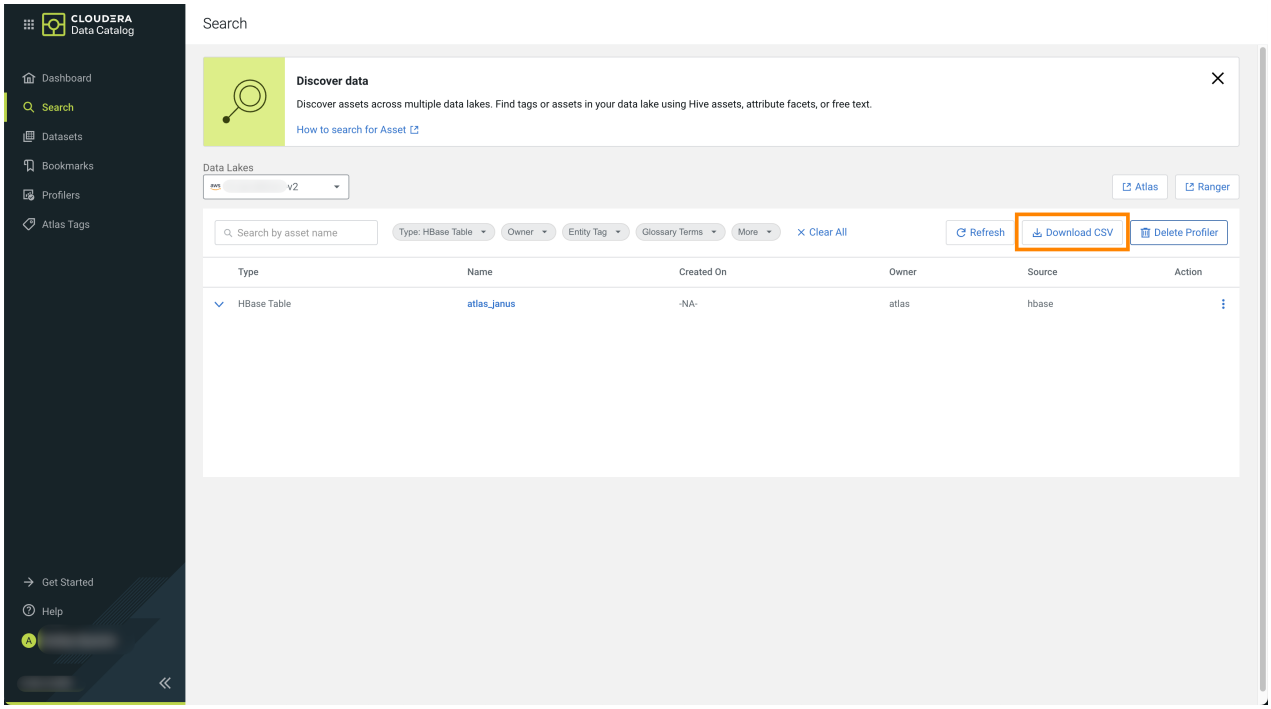
Clicking the Name of the entity will open its **Asset Details**.

Downloading search results as CSV files

You can also download the search result for the current query with the selected Data Lake. The feature allows you to download up to 10,000 rows for the current search query.

The CSV file format does not conform to any specific order or continuation in the downloaded results. For example, you can download 10,000 assets and download the results for the same query again, then the two downloaded CSV files might not contain the search results in the same order.

Click Download CSV to start your download.



Related Information

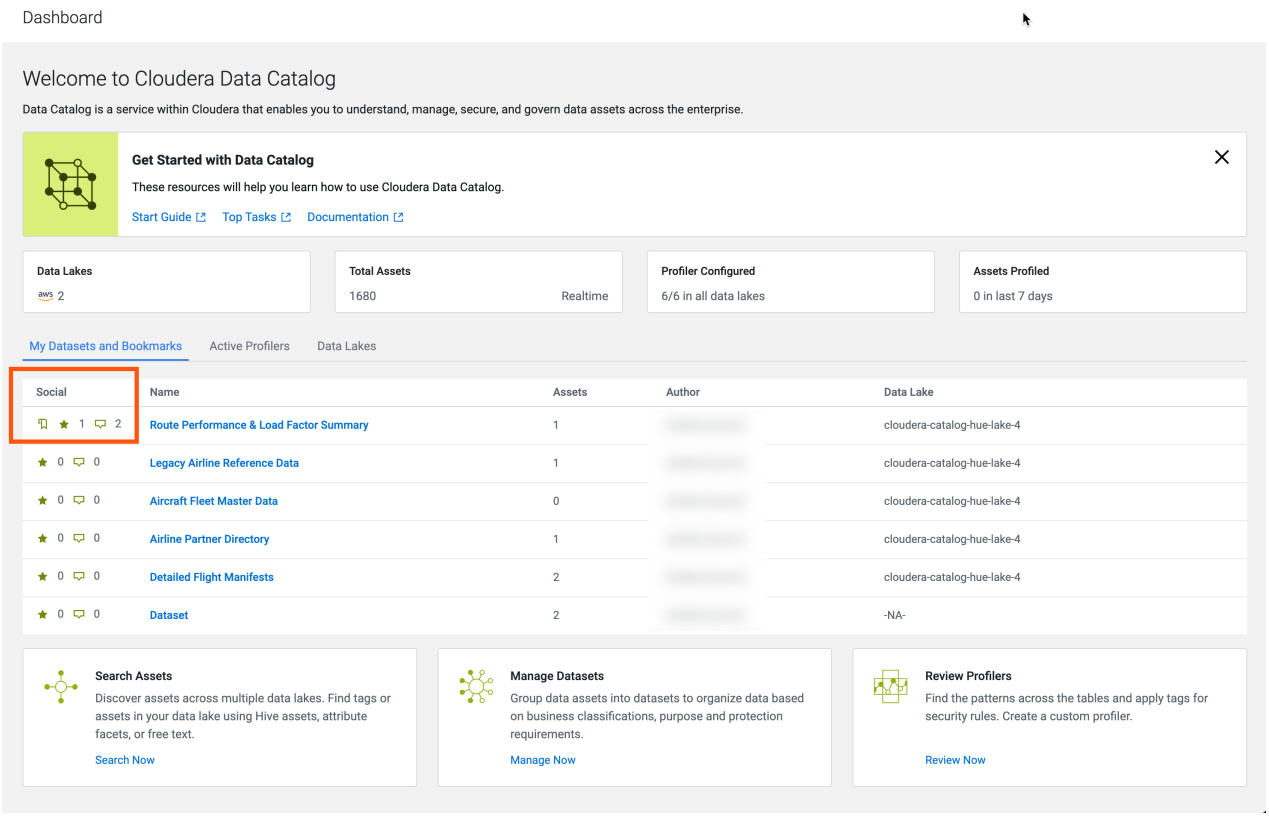
Search filters
Searching for assets using Atlas glossaries
Additional search options for asset types
Accessing tables based on Ranger policies
Viewing Data Asset details

Bookmarks overview

Dataset Details pages show collaboration metrics, keywords, authorship, collaborator lists, Created at and Updated at timestamps, asset name counts, moderated discussion, bookmarks, likes, and ratings.

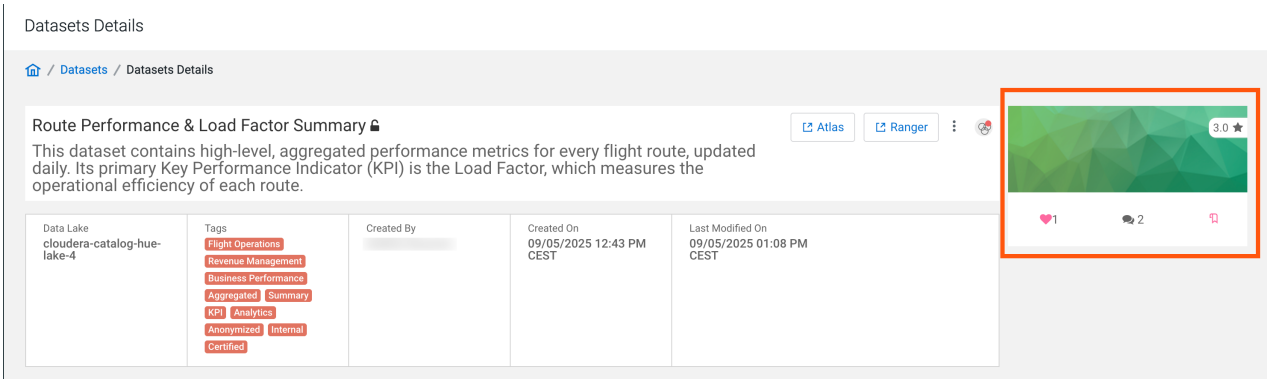
You can rate datasets and view the average rating of a dataset. This can help other users to find datasets with higher ratings easily. You can also add your knowledge and insights about the asset collection by adding comments. Other users can respond to your comments or add their comments about each data asset collection.

Figure 8: Dataset collaboration metrics



The dataset **Details** layout shows the following: asset names beside counts, authoring and collaborator summaries, tagging and keyword facets that feed full-text discovery, moderated comments, bookmarks, votes, ratings, likes, Created at timestamps, Updated at timestamps, and archive indicators when Data Lakes sunset.

Figure 9: Dataset Details collaboration panel



The collaboration panel highlights the quantitative signals stewards skim first:

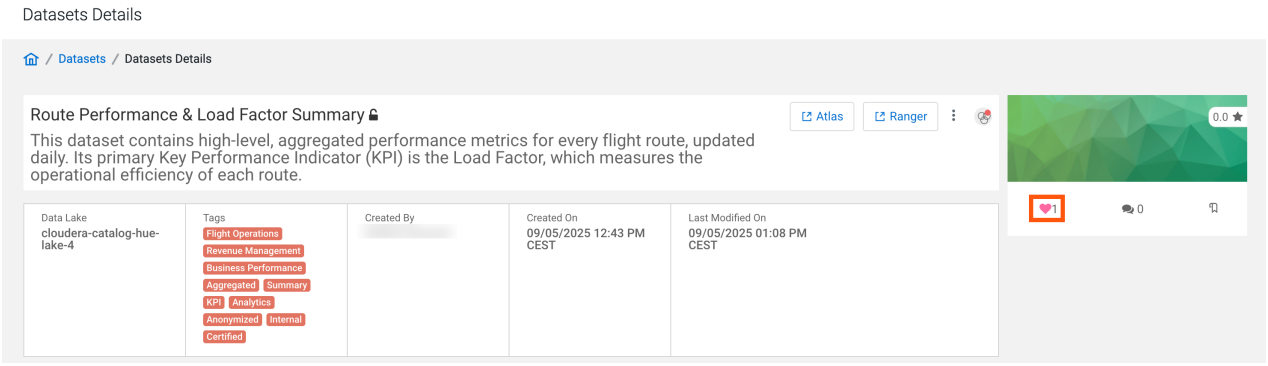
- Average rating and cumulative vote totals for the dataset
- Total number of likes
- Comment counts
- The bookmark icon indicating if the dataset is bookmarked by the current user or not

You can perform the following collaboration actions for each dataset.

Like a dataset

You can let other users know that you like a dataset. The ♥ icon on the dataset **Details** page displays the total number of likes received by this dataset helping you to find it faster.

Figure 10: Like a dataset



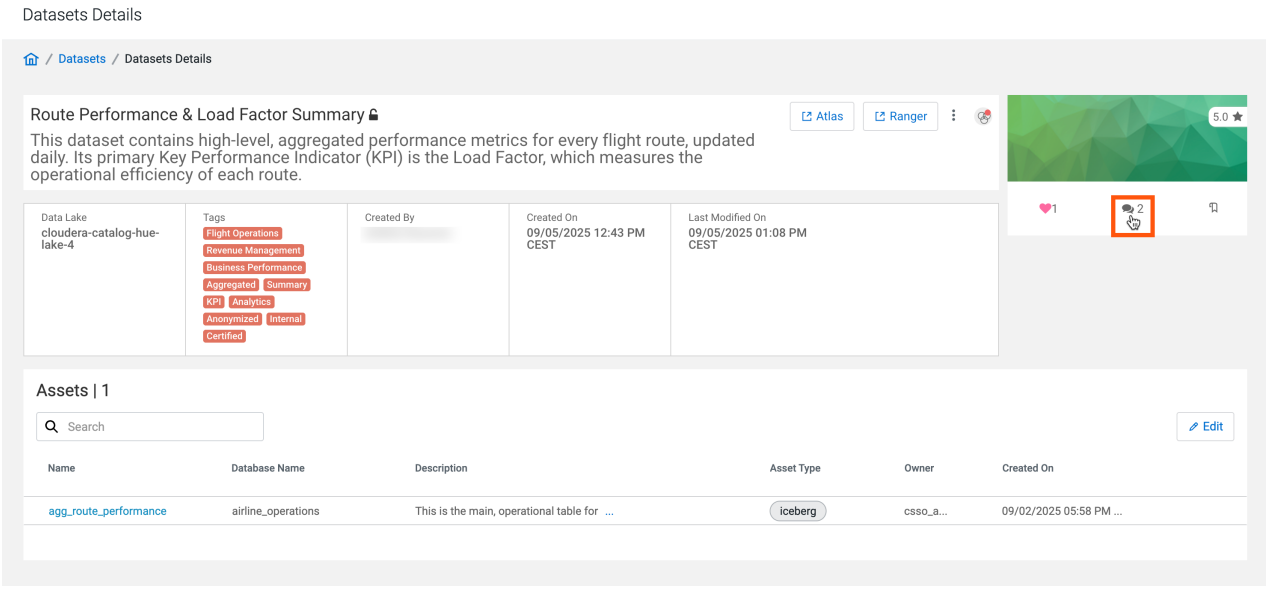
Click the ♥ icon to add the dataset to your list of liked collections.

Comment on or discuss a dataset

You might want to share your knowledge or insights about this dataset with other users. Cloudera Data Catalog allows you to collaborate with other users by adding comments.

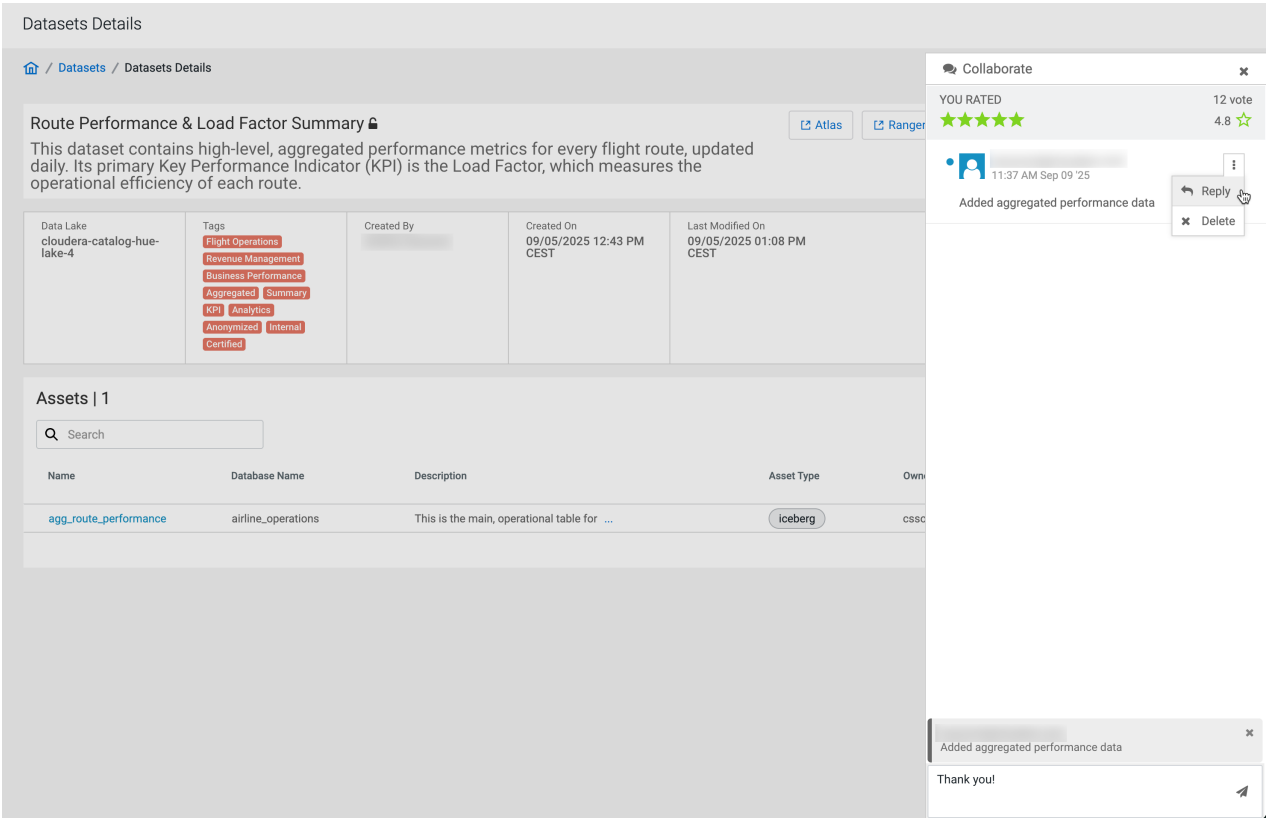
Click the 💬 icon to add a comment about this dataset.

Figure 11: Add a comment



The **Collaborate** tab expands. Click ⋮ icon to reply to an existing comment. You can continue to add comments for each dataset.

Figure 12: Reply to a comment

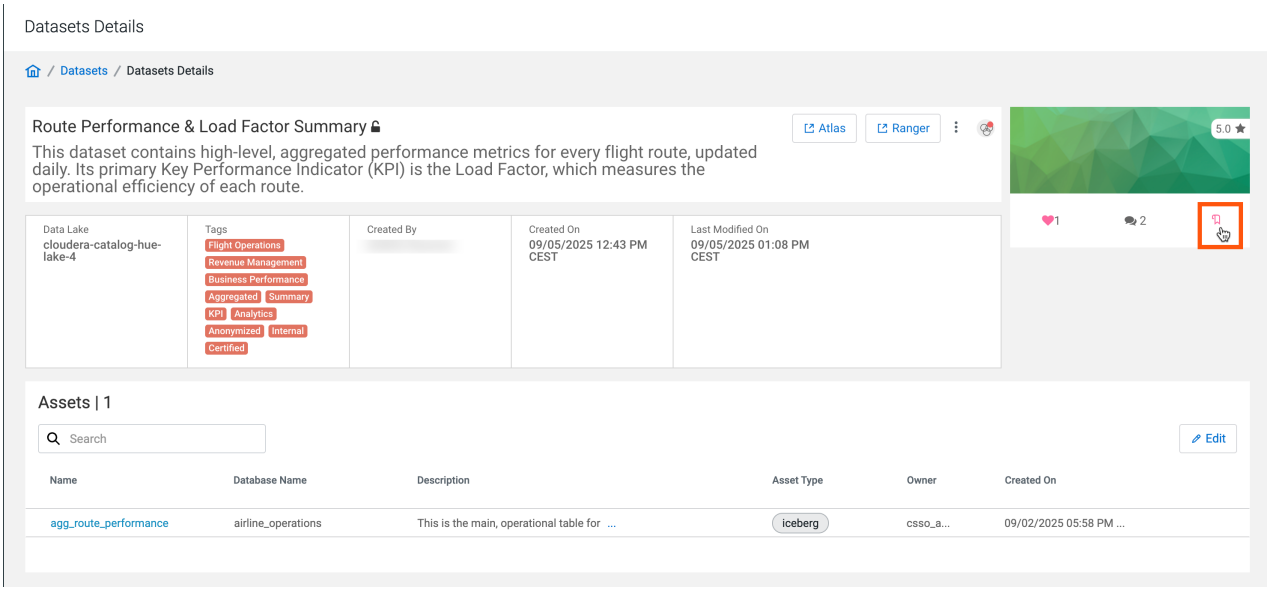


Bookmark the dataset

In addition to sharing with other users, you can also bookmark datasets for easy access in the future.

Click the icon to add the dataset to your list of bookmarks.

Figure 13: Bookmark action



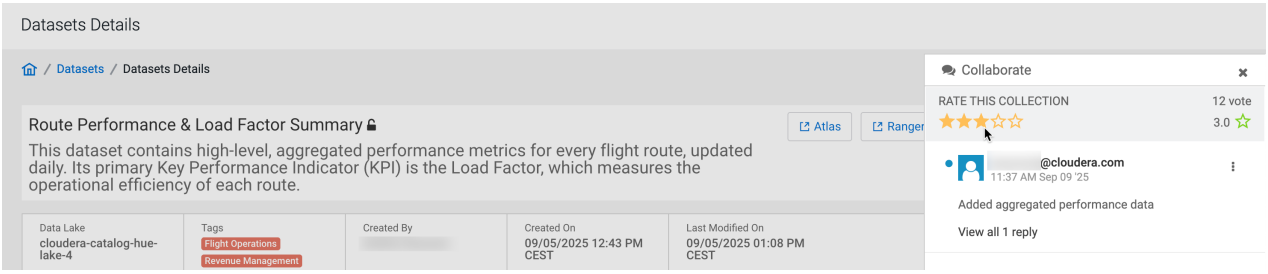
This dataset will appear in the list of bookmarks when you click the Bookmarks link in the left navigation menu.

Rate the dataset

You can also rate the datasets on a scale of one to five. Click the icon to enable rating, then click the icons to rate the open dataset. The **Collaborate** tab expands.

Click the stars to provide your own rating.

Figure 14: Rate a dataset

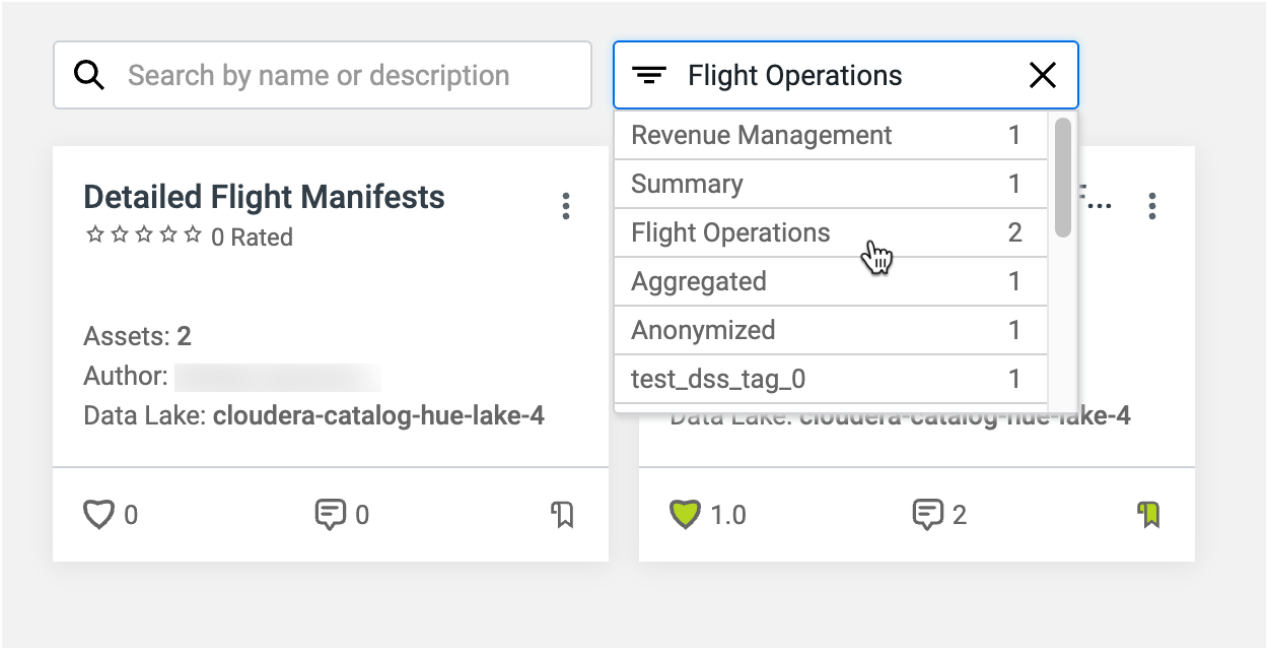


The rating on the **Datasets** page shows the average of the rating provided by various users. The **Rating** section also displays the number of votes given for this dataset.

View the tags on a dataset

You can add tags while creating a dataset. You can filter your datasets based on these tags.

Figure 15: Filter by tags



Note: These tags do not synchronize to Atlas.

Related Information

Managing datasets

Atlas tags overview

Atlas tags enhance searchability by serving as metadata labels, either created manually or applied automatically through profiling tools.

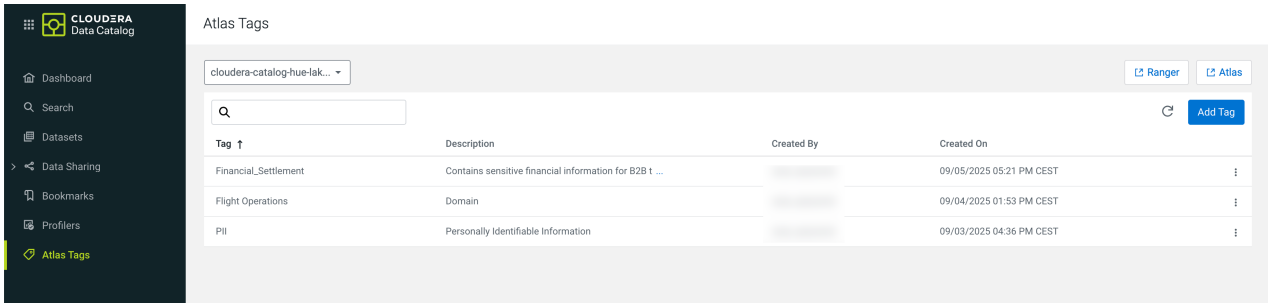
Atlas tags can created directly from Cloudera Data Catalog which represent Apache Atlas classifications. You can use these as metadata labels to improve searching assets. These tags or classifications can also be automatically applied by using the relevant profiler:

Compute Cluster enabled environments

Data Compliance Profiler

VM Based environments

Cluster Sensitivity Profiler



Related Information

Working with Atlas classifications and labels
Adding attributes to classifications
Creating tag rules in compute cluster environments
Creating tag rules in VM based environments

Before you start

Before you access Cloudera Data Catalog on cloud, you must deploy the service.

Before deploying Cloudera Data Catalog, make sure you have reviewed and compiled with the requirements in the installation guide for your environment.

Cloudera Data Catalog is supported on Compute Cluster enabled and VM-based environments. For more information, see the related documents.

For VM-based environment

- Creating and managing Cloudera deployments
- Register your first environment

For Compute Cluster enabled environment

- AWS environments:
  - Using Compute clusters
  - Enabling default Compute Cluster for an existing environment
- Azure environments:
  - Using Compute clusters
  - Enabling default Compute Cluster for an existing environment

Prerequisites to access Cloudera Data Catalog

To access the Cloudera Data Catalog, you must have the required credentials.

Follow these instructions to provide the required access to the Cloudera Data Catalog users.

Cloudera Data Catalog users must have either an EnvironmentAdmin or EnvironmentUser role assigned for the specific environment they need to access. For more information, see the related documents.

Note: You must be a PowerUser to launch and delete profilers. However, users without this role can still use the rest of the features of Cloudera Data Catalog.

The PowerUser must provide access to subscribers who plan to use Cloudera Data Catalog, either as an EnvironmentAdmin or an EnvironmentUser.

EnvironmentAdmin	EnvironmentUser
Can perform similar actions as EnvironmentUser.	Can create Dataset and related actions (Add assets, remove assets, tag assets, tag asset columns, and few others) for data lakes.
Additionally, in Cloudera Management Console, can perform the following: • Delete Environments. • Stop Environments. • Upgrade Data Lake.	Connect to Atlas and Ranger entities for Data Lakes for which there is access.
	Search for assets on the search page.
	Bookmark any dataset (even with no Data Lakes or Data Lakes to which access is not available).
	Access profilers.
	Create custom tags.
	Launch profilers on a Data Lake from the search page. Ability to launch the Workload cluster (Must have the Power User role assigned).
	Use Filters on the **Search** page of Cloudera Data Catalog to filter results.
	Launch profilers on a Data Lake from the search page.

Note: These roles can be assigned as the following types:

Account roles

An account role grants a user, machine user, or group permissions to access or perform tasks on all resources within the Cloudera tenant.

Resource roles

A resource role grants a user, machine user, or group permissions to access or perform tasks on a specific resource (such as a specific environment or a specific Cloudera Data Hub cluster).

For more information, see Understanding account roles and resource roles.

Additionally, using Cloudera Manager, you must configure Apache Atlas and Apache Ranger services. Use the following instructions to complete the process.

Related Information

Account roles

Resource roles

Providing role access

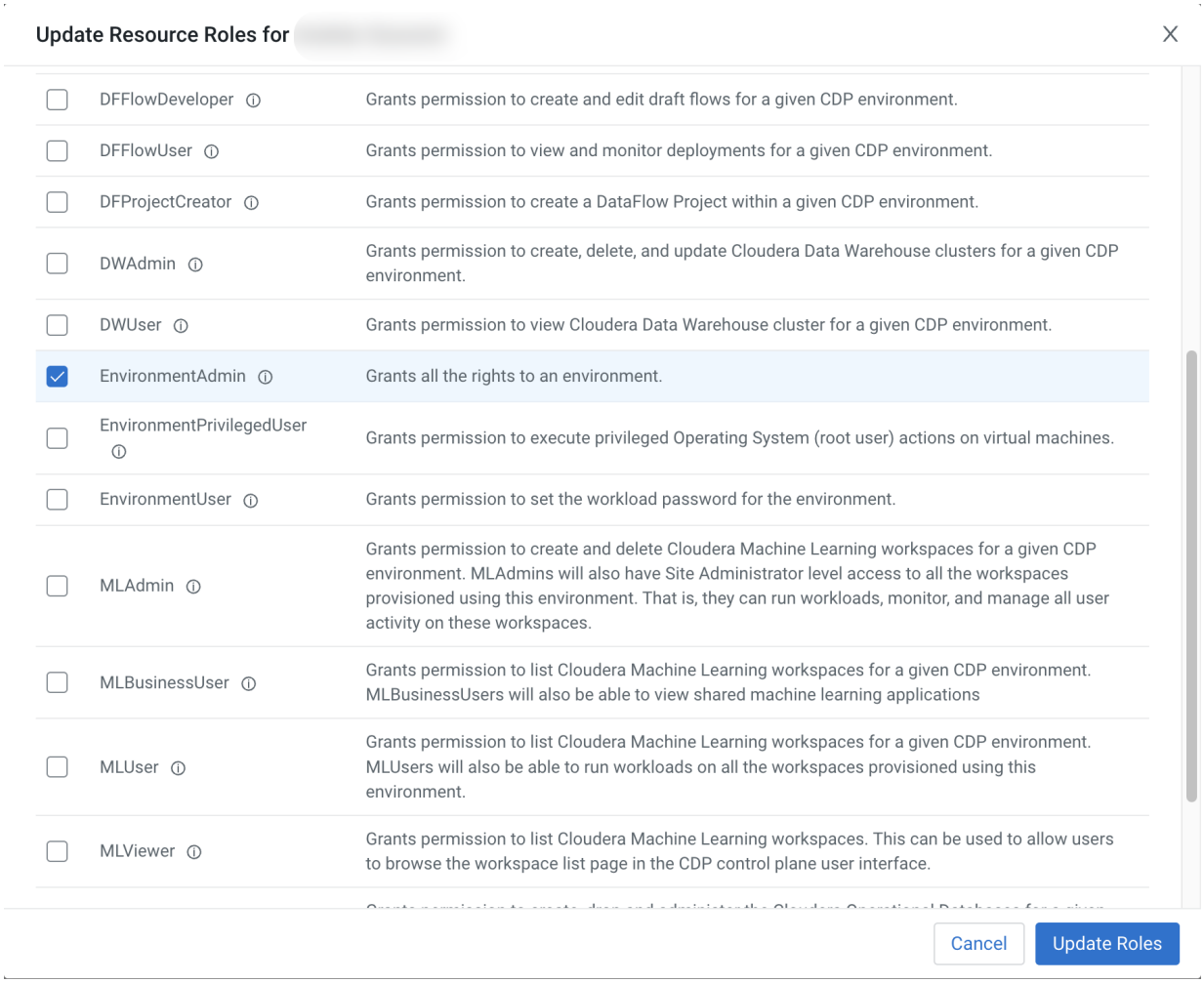
You must have either the EnvironmentAdmin or the EnvironmentUser role access to use Cloudera Data Catalog.

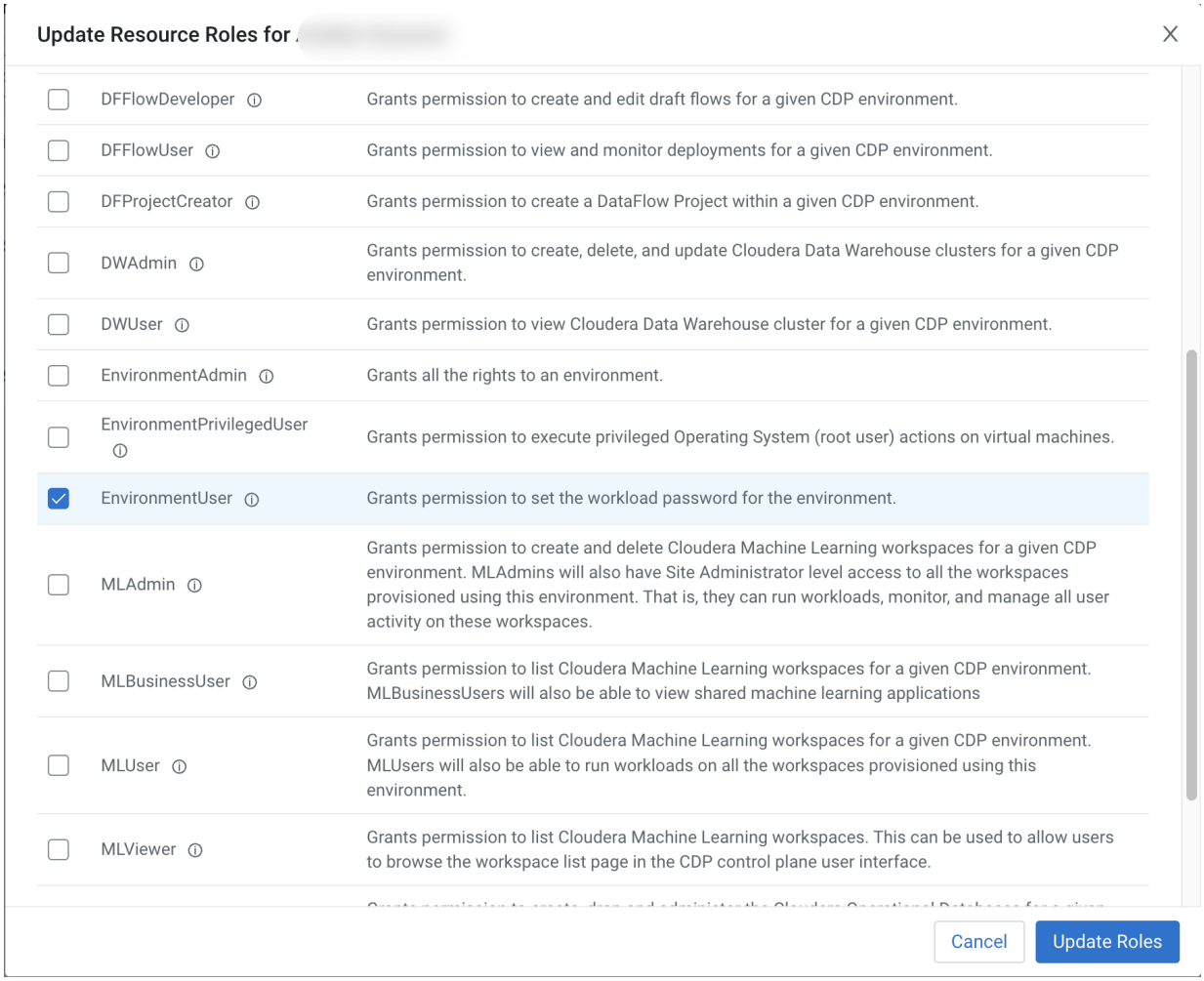
About this task

Compared to EnvironmentUsers, EnvironmentAdmins can manage environments and Data Lakes. For detailed differences between EnvironmentAdmin and EnvironmentUser roles, see Prerequisites to access Cloudera Data Catalog. For more information, see the related documents. The following procedure shows how to assign the EnvironmentAdmin role as a environment resource role. This means that the role will be valid only for a specific environment.

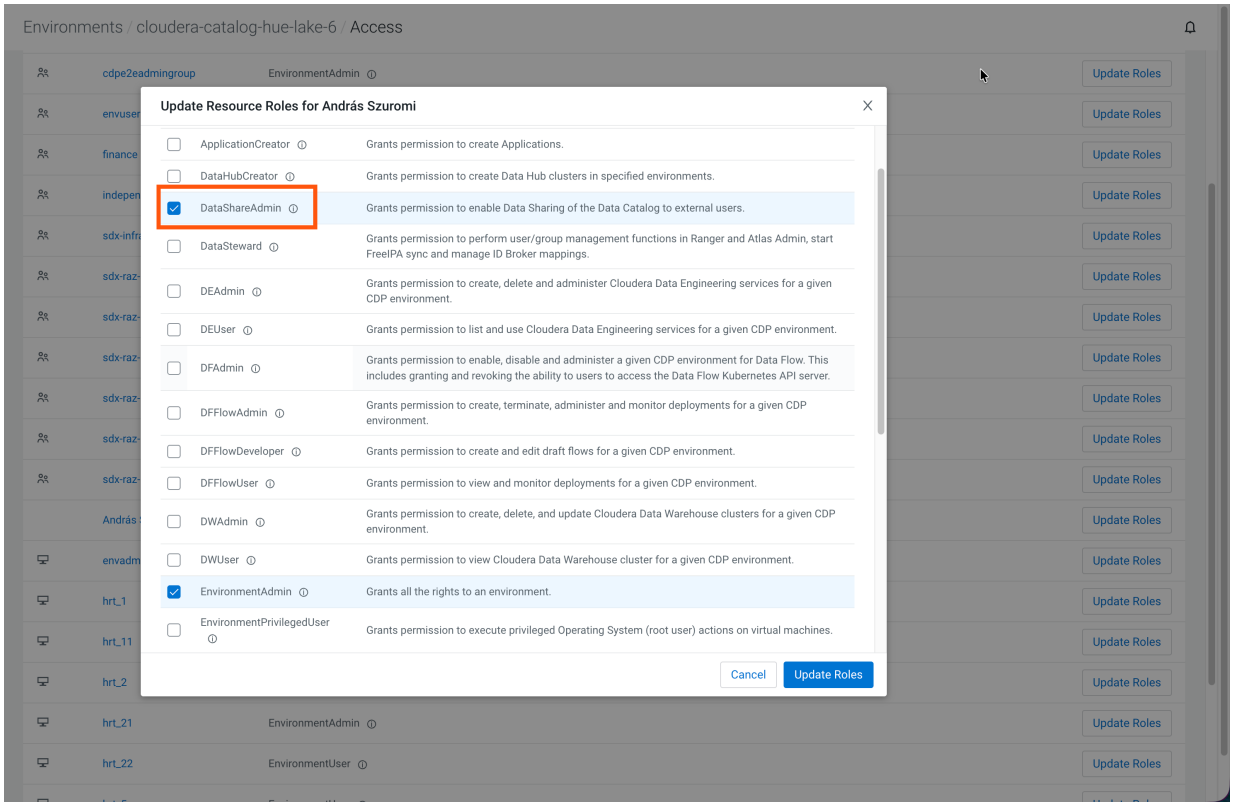
Procedure

1. From Cloudera Management Console Environments > Select an environment > select the Actions drop-down > Manage Access.
2. Search for the user who requires Cloudera Data Catalog access > Select the check-box, either EnvironmentAdmin or EnvironmentUser > Click Update Roles.





3. To use Data Sharing in Cloudera Data Catalog, you have to enable the DataShareAdmin role for your Data Providers. For more information, see Data Sharing in Cloudera Data Catalog overview on page 11.



4. Navigate back to the **Clusters** page and select Actions > select Synchronize Users
   Allow the sync operation to complete and the changes to take effect.

Related Information

Missing authorization for viewing assets
Understanding account roles and resource roles
Managing users and machine users in Cloudera