CDP Public Cloud / Data Catalog

Cloudera Data Catalog Overview

Date published: 2019-11-14
Date modified: 2025-10-17

https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2025. 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.

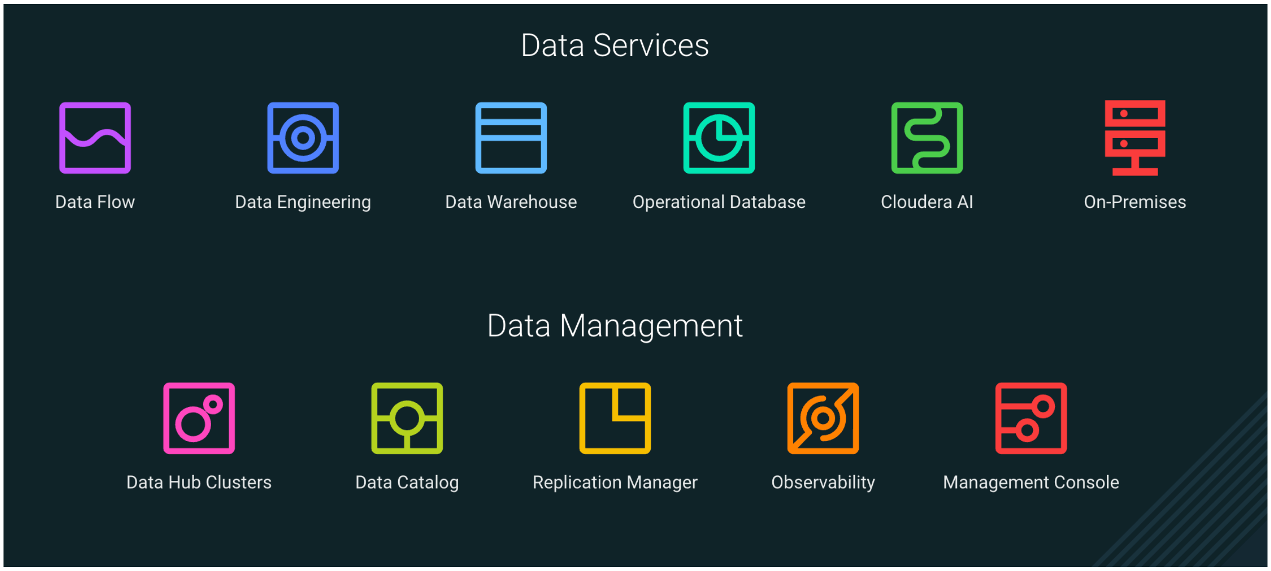
Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES 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

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 lying 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 Collaborate with other users.
- 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

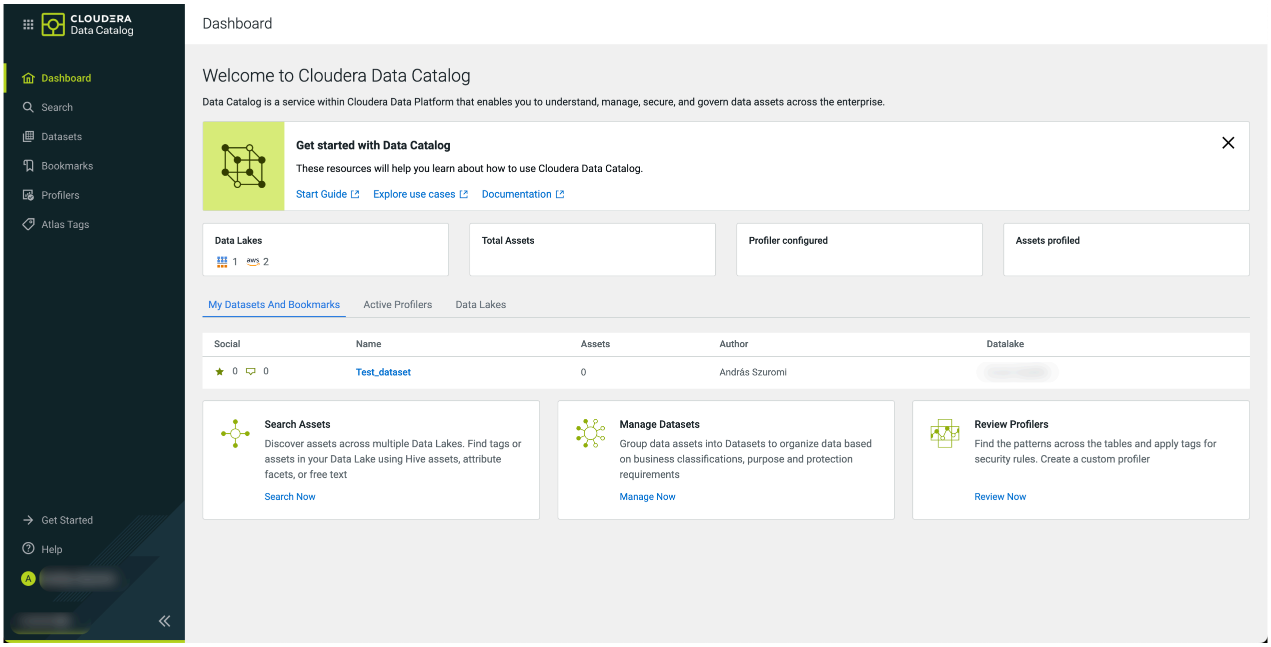
Related Information

Cloudera Data Catalog Terminology

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.



Dashboard

Welcome to Cloudera Data Catalog

Data Catalog is a service within Cloudera Data Platform that enables you to understand, manage, secure, and govern data assets across the enterprise.

Get started with Data Catalog

These resources will help you learn about how to use Cloudera Data Catalog.

Start Guide · Explore use cases · Documentation

Data Lakes	Total Assets	Profiler configured	Assets profiled
1 aws 2			

My Datasets And Bookmarks · Active Profilers · Data Lakes

Social	Name	Assets	Author	Datalake
★ 0 0	Test_dataset	0	András Szuromi	[redacted]

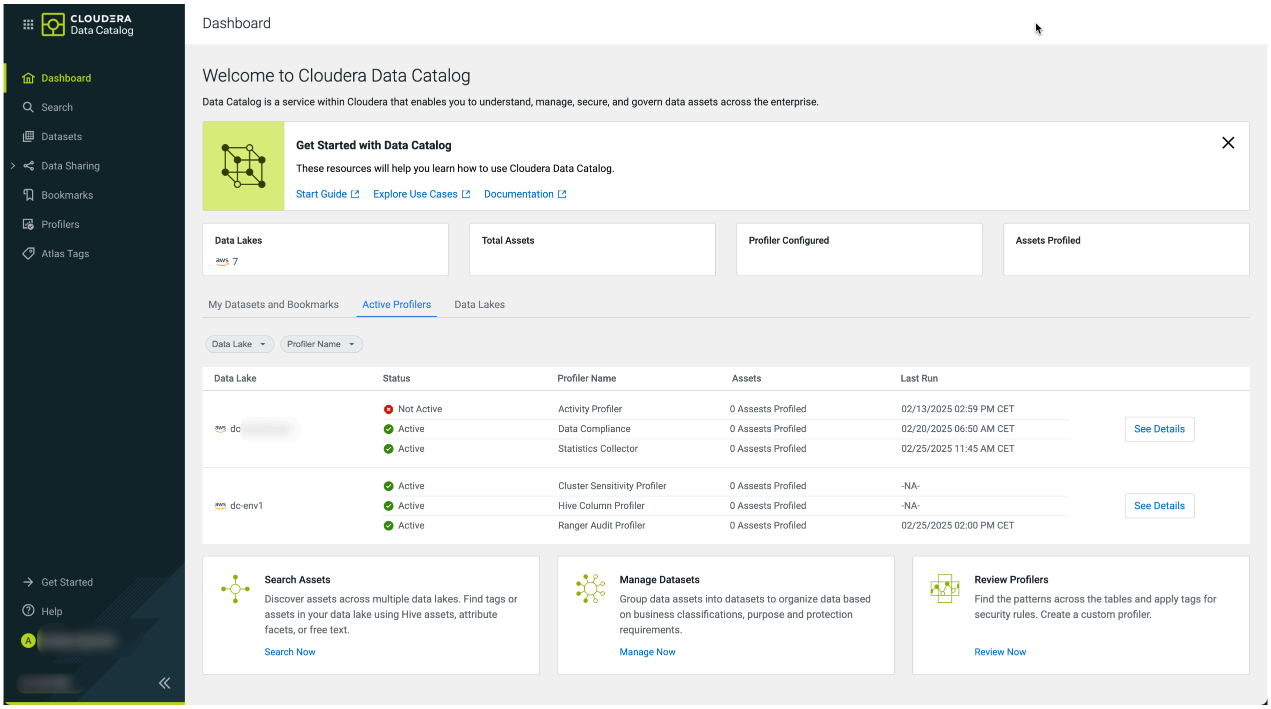
Search Assets
Discover assets across multiple Data Lakes. Find tags or assets in your Data Lake using Hive assets, attribute facets, or free text
Search Now

Manage Datasets
Group data assets into Datasets to organize data based on business classifications, purpose and protection requirements
Manage Now

Review Profilers
Find the patterns across the tables and apply tags for security rules. Create a custom profiler
Review Now

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



Dashboard

Welcome to Cloudera Data Catalog

Data Catalog is a service within Cloudera that enables you to understand, manage, secure, and govern data assets across the enterprise.

Get Started with Data Catalog

These resources will help you learn how to use Cloudera Data Catalog.

Start Guide · Explore Use Cases · Documentation

Data Lakes	Total Assets	Profiler Configured	Assets Profiled
aws 7			

My Datasets and Bookmarks · Active Profilers · Data Lakes

Data Lake ▾ Profiler Name ▾

Data Lake	Status	Profiler Name	Assets	Last Run	
aws dc[redacted]	Not Active	Activity Profiler	0 Assests Profiled	02/13/2025 02:59 PM CET	See Details
	Active	Data Compliance	0 Assests Profiled	02/20/2025 06:50 AM CET	
	Active	Statistics Collector	0 Assests Profiled	02/25/2025 11:45 AM CET	
aws dc-env1	Active	Cluster Sensitivity Profiler	0 Assests Profiled	-NA-	See Details
	Active	Hive Column Profiler	0 Assests Profiled	-NA-	
	Active	Ranger Audit Profiler	0 Assests Profiled	02/25/2025 02:00 PM CET	

Search Assets
Discover assets across multiple data lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text.
Search Now

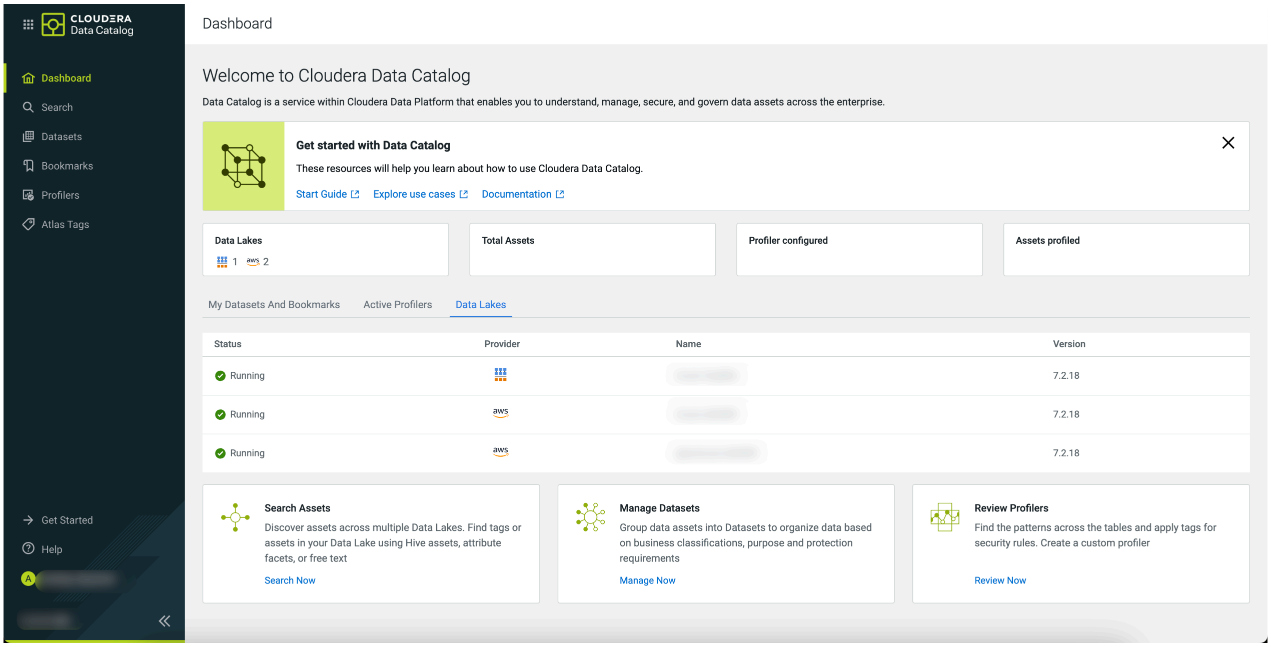
Manage Datasets
Group data assets into datasets to organize data based on business classifications, purpose and protection requirements.
Manage Now

Review Profilers
Find the patterns across the tables and apply tags for security rules. Create a custom profiler.
Review Now

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



Dashboard

Welcome to Cloudera Data Catalog

Data Catalog is a service within Cloudera Data Platform that enables you to understand, manage, secure, and govern data assets across the enterprise.

Get started with Data Catalog

These resources will help you learn about how to use Cloudera Data Catalog.

Start Guide Explore use cases Documentation

Data Lakes: 1 aws 2

Total Assets

Profiler configured

Assets profiled

My Datasets And Bookmarks | Active Profilers | Data Lakes

Status	Provider	Name	Version
Running		[redacted]	7.2.18
Running	aws	[redacted]	7.2.18
Running	aws	[redacted]	7.2.18

Search Assets

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

Search Now

Manage Datasets

Group data assets into Datasets to organize data based on business classifications, purpose and protection requirements

Manage Now

Review Profilers

Find the patterns across the tables and apply tags for security rules. Create a custom profiler

Review Now

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, Custom 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. Each type of a profiler has its own Kubernetes cron-jobs for handling scheduled 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

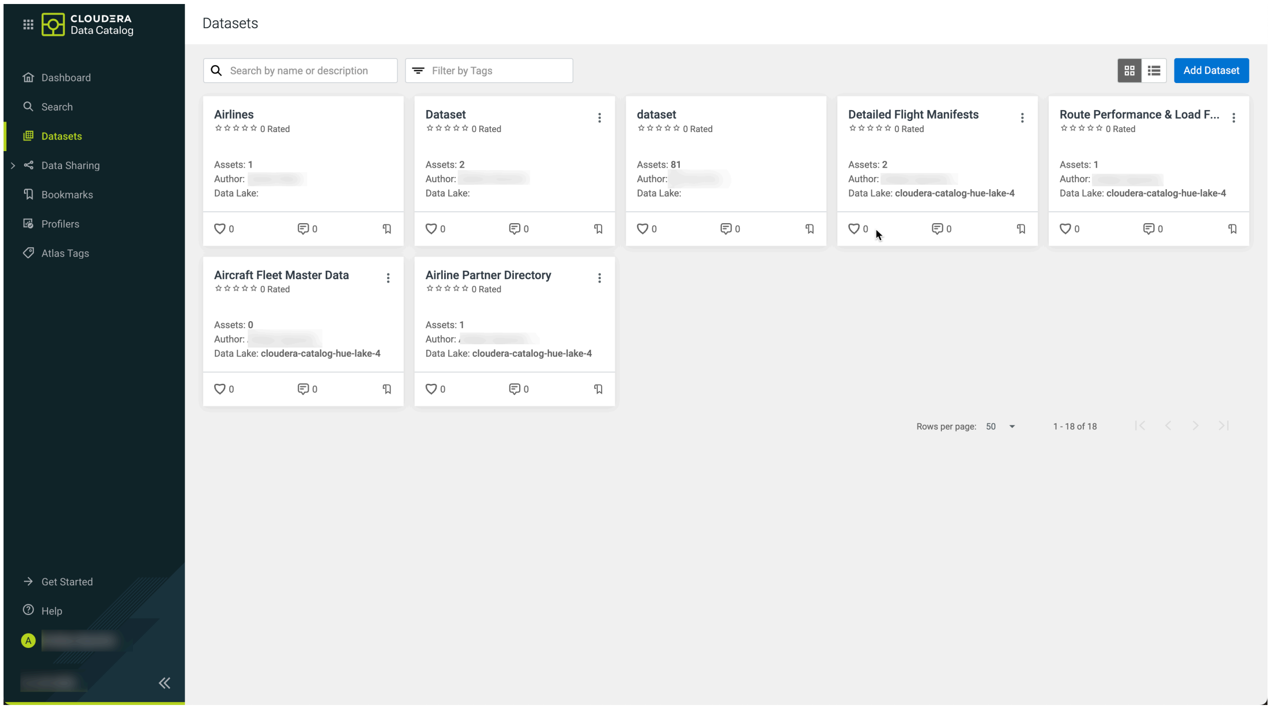
About Cloudera Data Catalog

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

Collaborating with other users

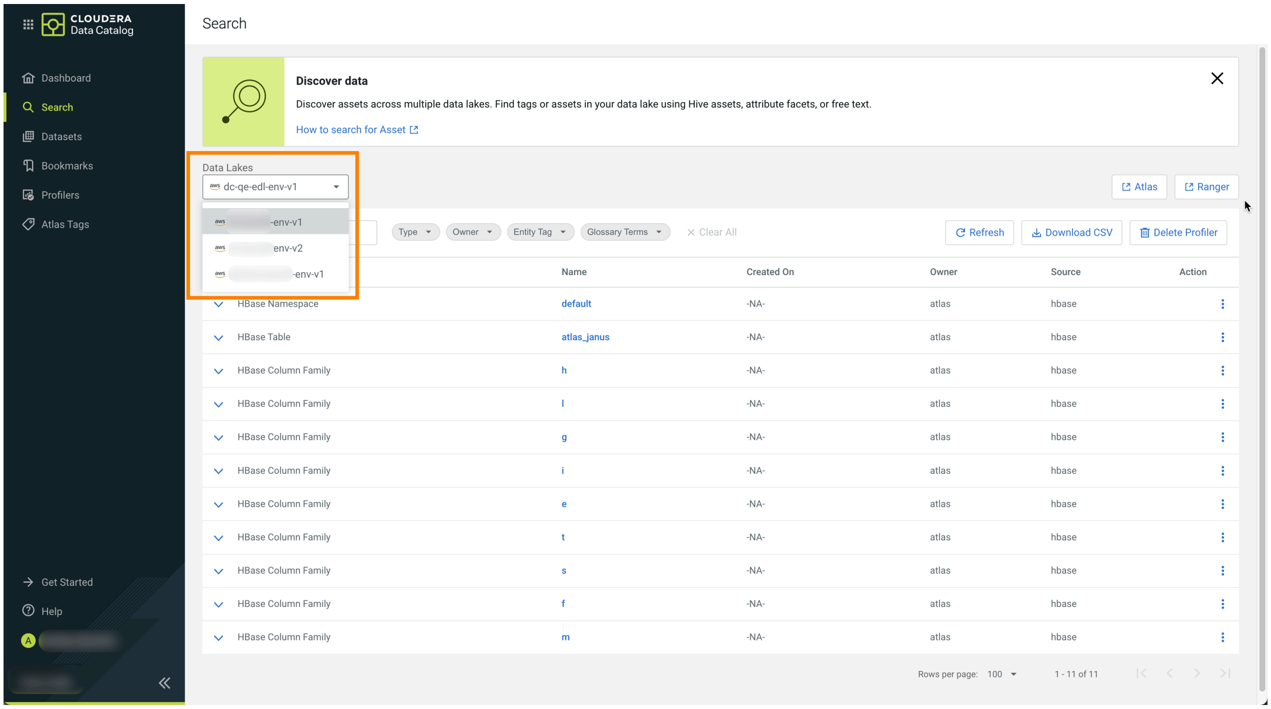
Search overview

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 lakes based on the permissions that are granted. You can choose the available lake by selecting the appropriate radio button.



Discover data

Discover assets across multiple data lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text.

How to search for Asset

Data Lakes: dc-qe-edl-env-v1

- [redacted]-env-v1
- [redacted]env-v2
- [redacted]-env-v1

	Name	Created On	Owner	Source	Action
HBase Namespace	default	-NA-	atlas	hbase	⋮
HBase Table	atlas_janus	-NA-	atlas	hbase	⋮
HBase Column Family	h	-NA-	atlas	hbase	⋮
HBase Column Family	l	-NA-	atlas	hbase	⋮
HBase Column Family	g	-NA-	atlas	hbase	⋮
HBase Column Family	i	-NA-	atlas	hbase	⋮
HBase Column Family	e	-NA-	atlas	hbase	⋮
HBase Column Family	t	-NA-	atlas	hbase	⋮
HBase Column Family	s	-NA-	atlas	hbase	⋮
HBase Column Family	f	-NA-	atlas	hbase	⋮
HBase Column Family	m	-NA-	atlas	hbase	⋮

Rows per page: 100 1 - 11 of 11

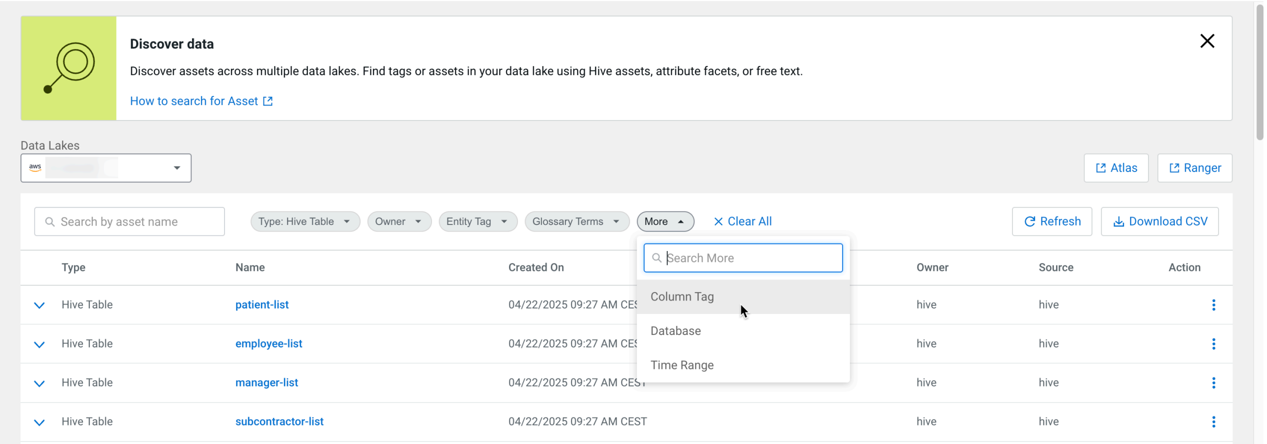
Note:

- 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 show under More.

Search



Discover data

Discover assets across multiple data lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text.

How to search for Asset

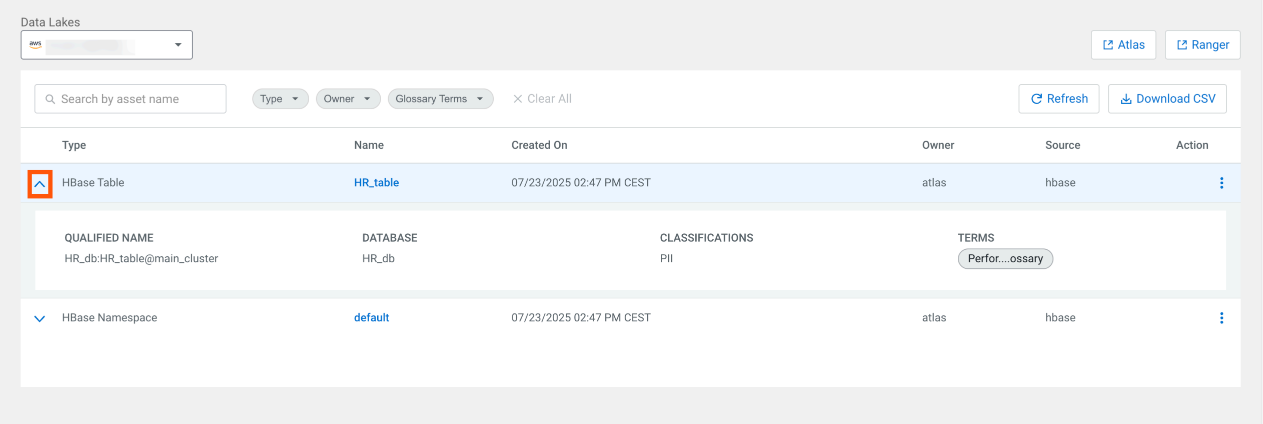
Type	Name	Created On	Owner	Source	Action
Hive Table	patient-list	04/22/2025 09:27 AM CES	hive	hive	⋮
Hive Table	employee-list	04/22/2025 09:27 AM CES	hive	hive	⋮
Hive Table	manager-list	04/22/2025 09:27 AM CEST	hive	hive	⋮
Hive Table	subcontractor-list	04/22/2025 09:27 AM CEST	hive	hive	⋮

Viewing Asset Details

Clicking the ⌄ icon for a search result shows the most important 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**



Type	Name	Created On	Owner	Source	Action
HBase Table	HR_table	07/23/2025 02:47 PM CEST	atlas	hbase	⋮
HBase Namespace	default	07/23/2025 02:47 PM CEST	atlas	hbase	⋮

QUALIFIED NAME	DATABASE	CLASSIFICATIONS	TERMS
HR_db:HR_table@main_cluster	HR_db	PII	Perfor....ossary

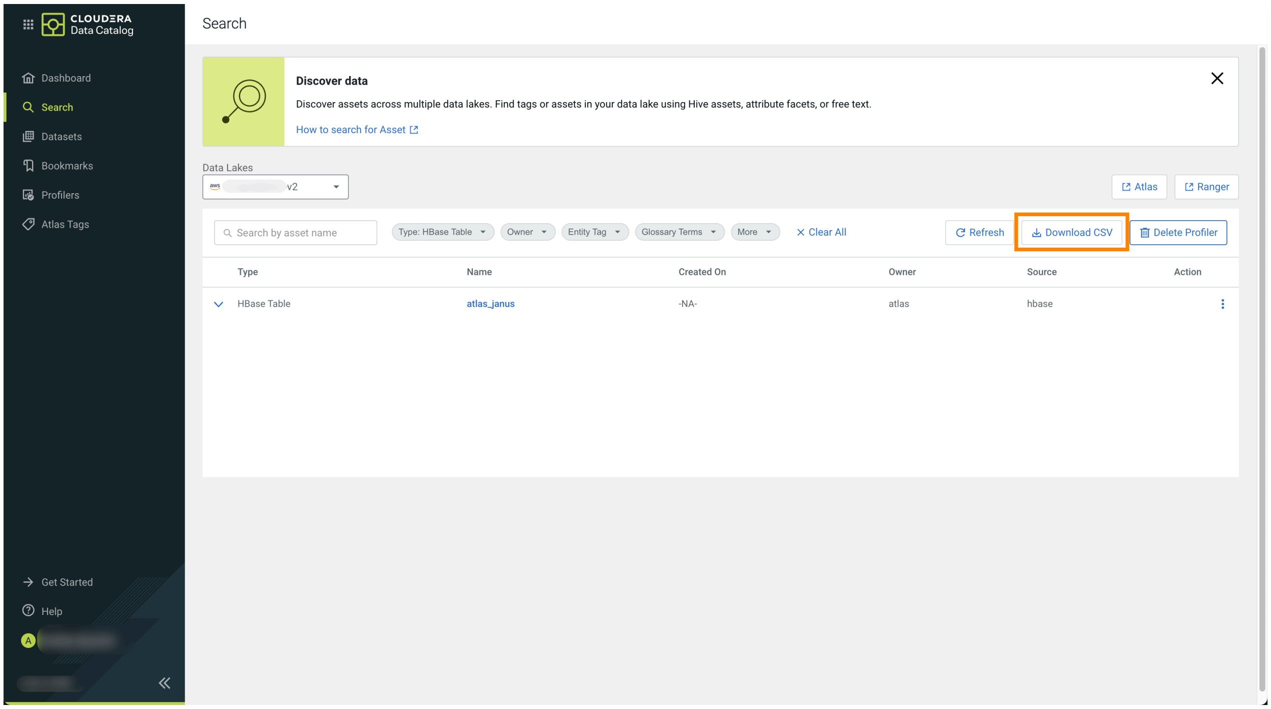
Clicking the Name of the entity will open its **Asset Details**.

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 10000 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, a user can download 10000 assets and later downloads the results for the same query again, then the downloaded CSV files may not contain the search results in the same order as it was downloaded previously.

Click Download CSV to start your download:



Discover data

Discover assets across multiple data lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text.

How to search for Asset

Type	Name	Created On	Owner	Source	Action
HBase Table	atlas_janus	-NA-	atlas	hbase	⋮

Related Information

Search filters

Searching for assets using Atlas glossaries

Additional search options for asset types

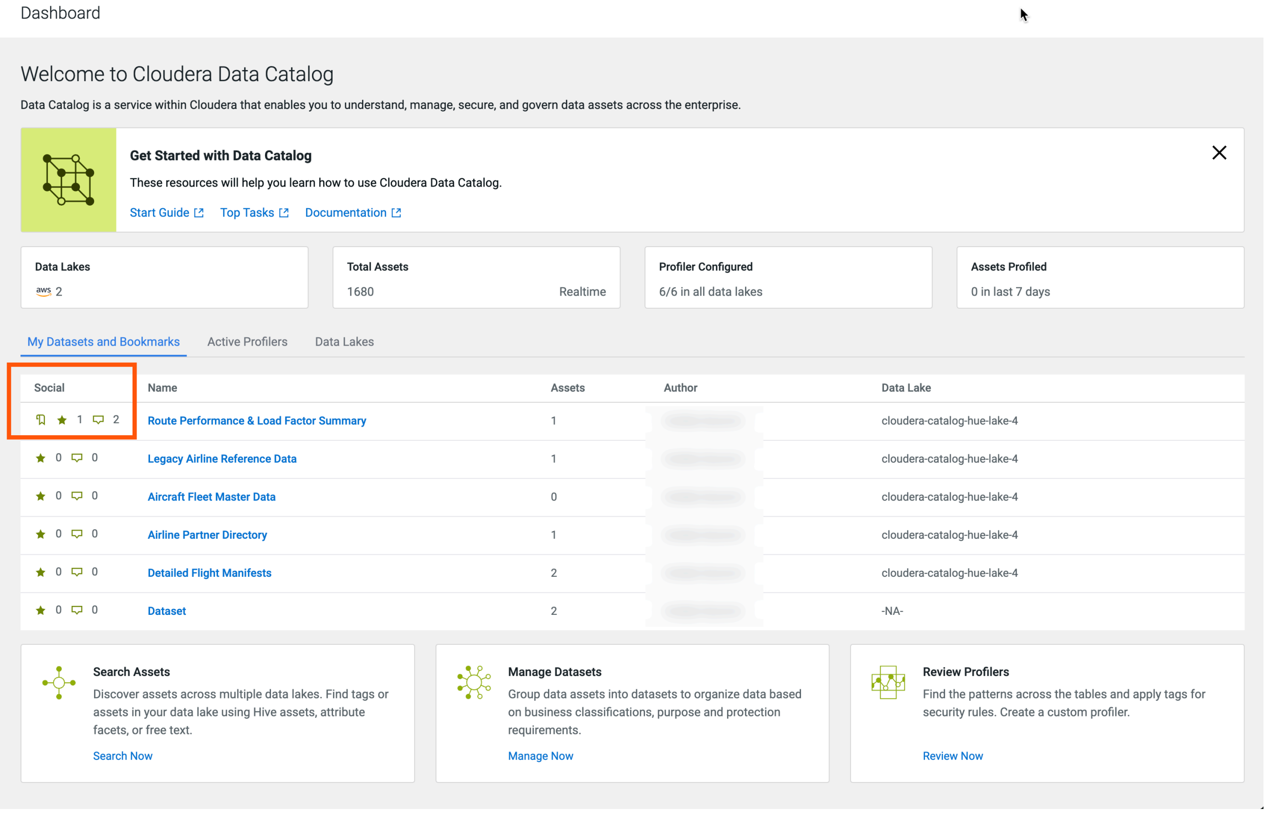
Accessing tables based on Ranger policies

Viewing Data Asset details

Bookmarks overview

Using the tools in Datasets, you can collaborate and share insights with other users in the enterprise.

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.



Dashboard

Welcome to Cloudera Data Catalog

Data Catalog is a service within Cloudera that enables you to understand, manage, secure, and govern data assets across the enterprise.

Get Started with Data Catalog

These resources will help you learn how to use Cloudera Data Catalog.

Start Guide Top Tasks Documentation

Data Lakes	Total Assets	Profiler Configured	Assets Profiled
aws 2	1680 Realtime	6/6 in all data lakes	0 in last 7 days

Social	Name	Assets	Author	Data Lake
★ 1 🗨 2	Route Performance & Load Factor Summary	1	[redacted]	cloudera-catalog-hue-lake-4
★ 0 🗨 0	Legacy Airline Reference Data	1	[redacted]	cloudera-catalog-hue-lake-4
★ 0 🗨 0	Aircraft Fleet Master Data	0	[redacted]	cloudera-catalog-hue-lake-4
★ 0 🗨 0	Airline Partner Directory	1	[redacted]	cloudera-catalog-hue-lake-4
★ 0 🗨 0	Detailed Flight Manifests	2	[redacted]	cloudera-catalog-hue-lake-4
★ 0 🗨 0	Dataset	2	[redacted]	-NA-

Search Assets

Discover assets across multiple data lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text.

Search Now

Manage Datasets

Group data assets into datasets to organize data based on business classifications, purpose and protection requirements.

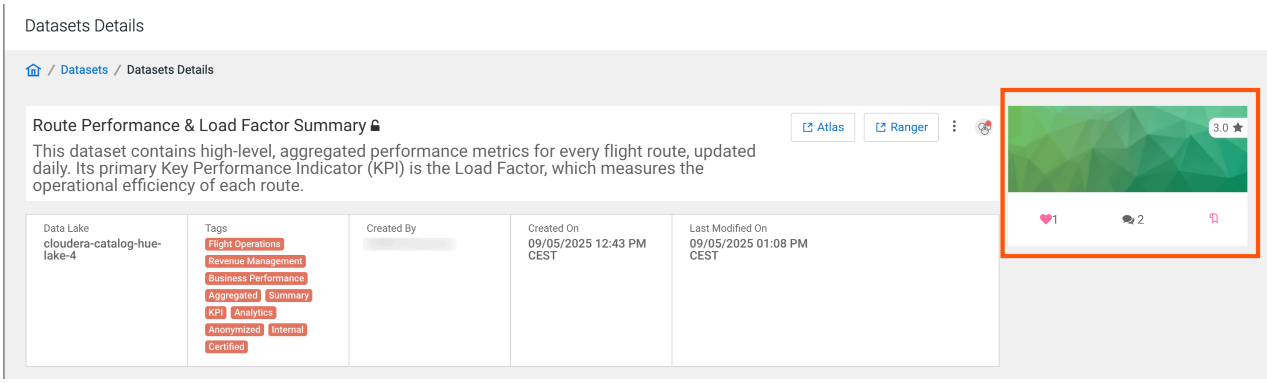
Manage Now

Review Profilers

Find the patterns across the tables and apply tags for security rules. Create a custom profiler.

Review Now

On the right hand side of each dataset **Details** page, you can see additional details about the dataset. The collaboration details are also displayed in this tab.



Datasets Details

Datasets / Datasets Details

Route Performance & Load Factor Summary

This dataset contains high-level, aggregated performance metrics for every flight route, updated daily. Its primary Key Performance Indicator (KPI) is the Load Factor, which measures the operational efficiency of each route.

Data Lake	Tags	Created By	Created On	Last Modified On
cloudera-catalog-hue-lake-4	Flight Operations Revenue Management Business Performance Aggregated Summary KPI Analytics Anonymized Internal Certified	[redacted]	09/05/2025 12:43 PM CEST	09/05/2025 01:08 PM CEST

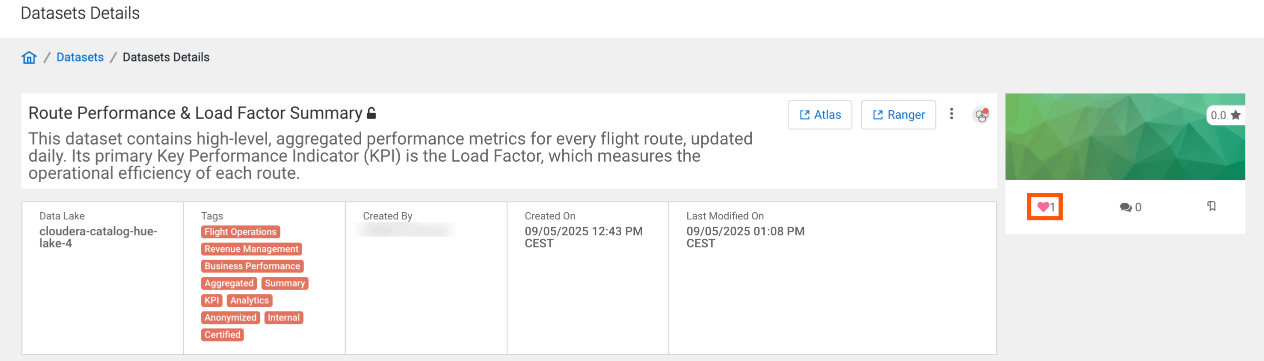
The tab displays the following details:

- The average rating for the asset collection
- The number of likes
- The number of comments
- 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.



Datasets Details

Datasets / Datasets Details

Route Performance & Load Factor Summary

This dataset contains high-level, aggregated performance metrics for every flight route, updated daily. Its primary Key Performance Indicator (KPI) is the Load Factor, which measures the operational efficiency of each route.

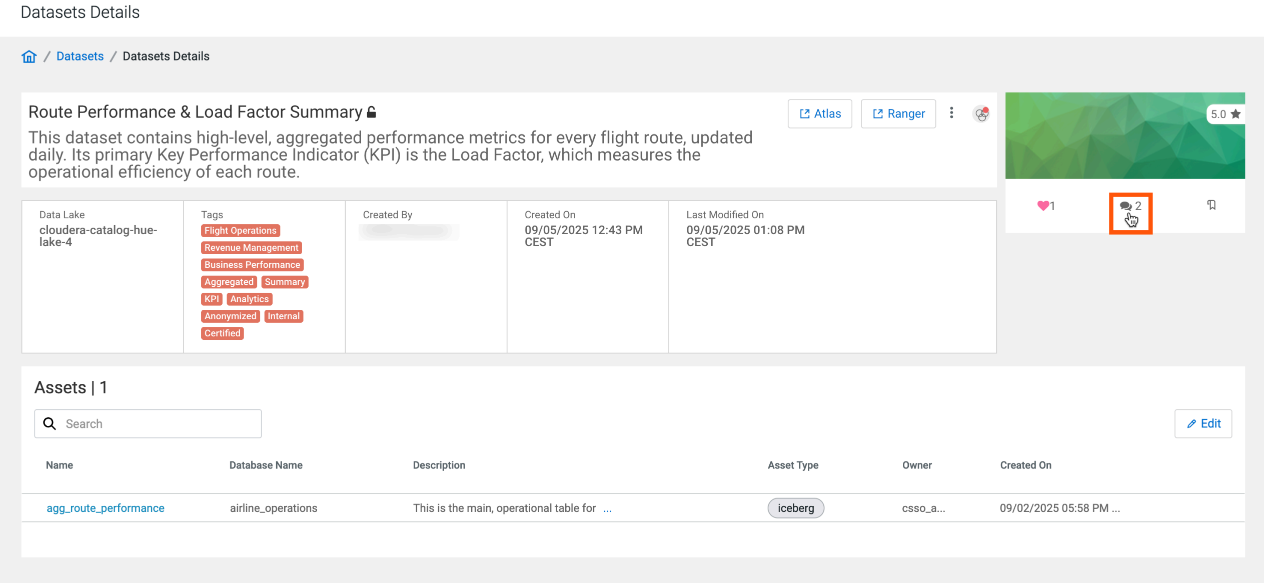
Data Lake	Tags	Created By	Created On	Last Modified On
cloudera-catalog-hue-lake-4	Flight Operations Revenue Management Business Performance Aggregated Summary KPI Analytics Anonymized Internal Certified	[redacted]	09/05/2025 12:43 PM CEST	09/05/2025 01:08 PM CEST

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

Comment and discuss about 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.



Datasets Details

Datasets / Datasets Details

Route Performance & Load Factor Summary

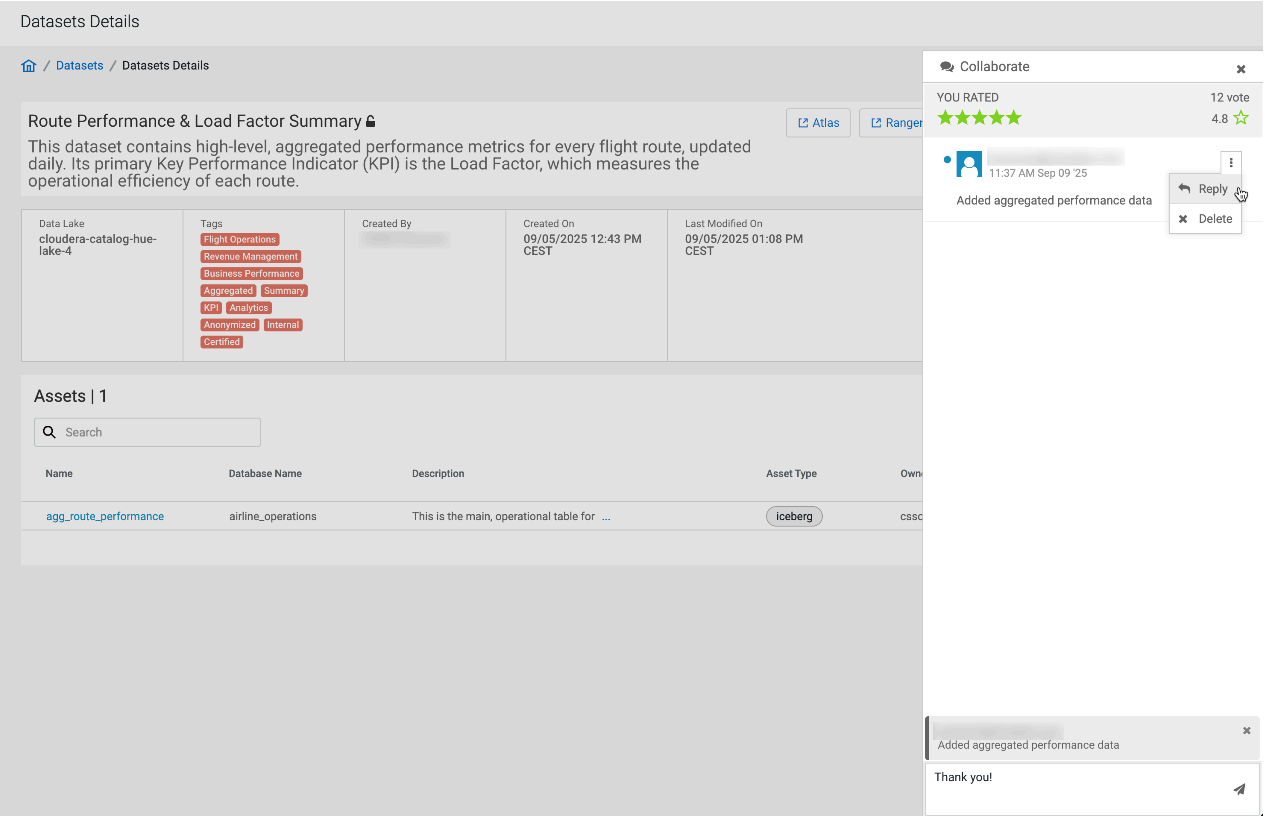
This dataset contains high-level, aggregated performance metrics for every flight route, updated daily. Its primary Key Performance Indicator (KPI) is the Load Factor, which measures the operational efficiency of each route.

Data Lake	Tags	Created By	Created On	Last Modified On
cloudera-catalog-hue-lake-4	Flight Operations Revenue Management Business Performance Aggregated Summary KPI Analytics Anonymized Internal Certified	[redacted]	09/05/2025 12:43 PM CEST	09/05/2025 01:08 PM CEST

Assets | 1

Name	Database Name	Description	Asset Type	Owner	Created On
agg_route_performance	airline_operations	This is the main, operational table for ...	iceberg	csso_a...	09/02/2025 05:58 PM ...

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



Datasets Details

Datasets / Datasets Details

Route Performance & Load Factor Summary

This dataset contains high-level, aggregated performance metrics for every flight route, updated daily. Its primary Key Performance Indicator (KPI) is the Load Factor, which measures the operational efficiency of each route.

Data Lake	Tags	Created By	Created On	Last Modified On
cloudera-catalog-hue-lake-4	Flight Operations Revenue Management Business Performance Aggregated Summary KPI Analytics Anonymized Internal Certified	[redacted]	09/05/2025 12:43 PM CEST	09/05/2025 01:08 PM CEST

Assets | 1

Name	Database Name	Description	Asset Type	Own
agg_route_performance	airline_operations	This is the main, operational table for ...	iceberg	cssc

Collaborate

YOU RATED ★★★★★ 12 vote 4.8 ☆

[redacted] 11:37 AM Sep 09 '25

Added aggregated performance data

[redacted]

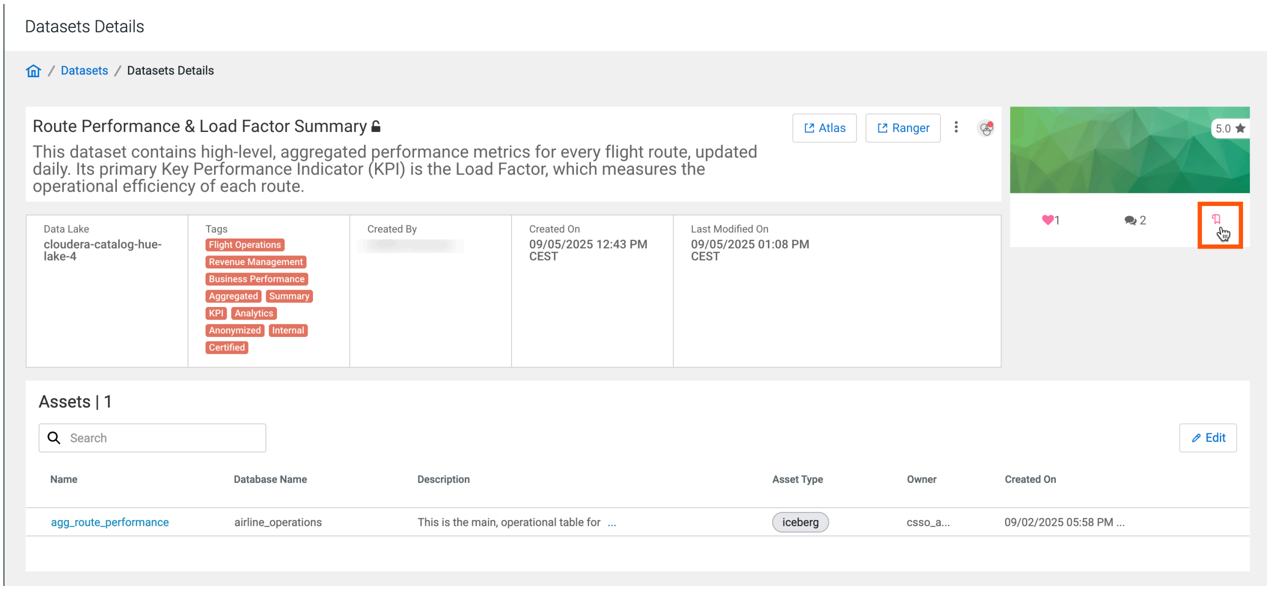
Added aggregated performance data

Thank you!

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.



Datasets Details

Datasets / Datasets Details

Route Performance & Load Factor Summary

This dataset contains high-level, aggregated performance metrics for every flight route, updated daily. Its primary Key Performance Indicator (KPI) is the Load Factor, which measures the operational efficiency of each route.

Data Lake	Tags	Created By	Created On	Last Modified On
cloudera-catalog-hue-lake-4	Flight Operations Revenue Management Business Performance Aggregated Summary KPI Analytics Anonymized Internal Certified	[redacted]	09/05/2025 12:43 PM CEST	09/05/2025 01:08 PM CEST

Assets | 1

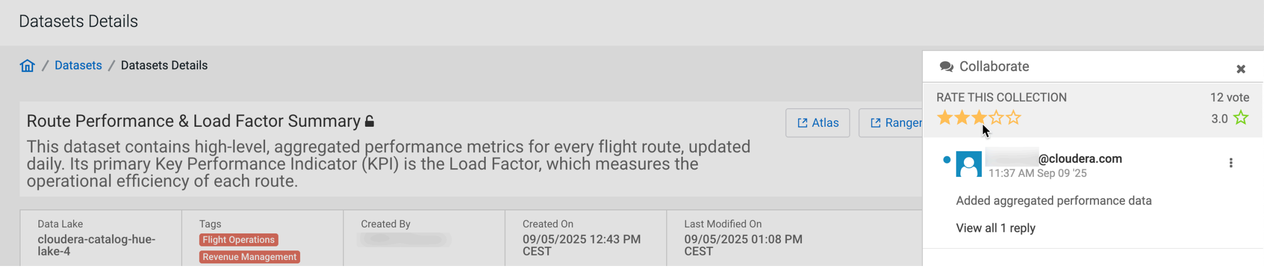
Name	Database Name	Description	Asset Type	Owner	Created On
agg_route_performance	airline_operations	This is the main, operational table for ...	iceberg	csso_a...	09/02/2025 05:58 PM ...

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.



Datasets Details

Datasets / Datasets Details

Route Performance & Load Factor Summary

This dataset contains high-level, aggregated performance metrics for every flight route, updated daily. Its primary Key Performance Indicator (KPI) is the Load Factor, which measures the operational efficiency of each route.

Data Lake	Tags	Created By	Created On	Last Modified On
cloudera-catalog-hue-lake-4	Flight Operations Revenue Management	[redacted]	09/05/2025 12:43 PM CEST	09/05/2025 01:08 PM CEST

Collaborate

RATE THIS COLLECTION ★★★☆☆ 12 vote 3.0 ☆

[redacted]@cloudera.com 11:37 AM Sep 09 '25

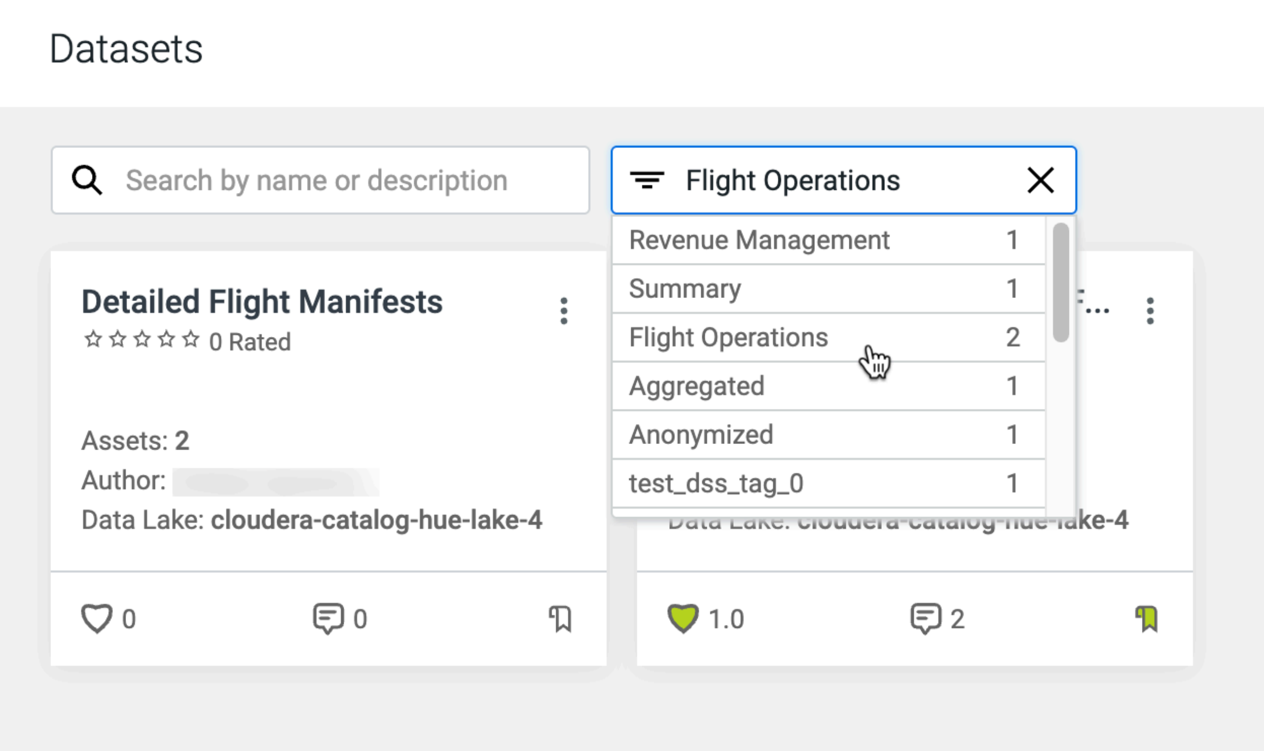
Added aggregated performance data

View all 1 reply

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 of an dataset

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



Datasets

Search by name or description

Flight Operations

Revenue Management	1
Summary	1
Flight Operations	2
Aggregated	1
Anonymized	1
test_dss_tag_0	1

Detailed Flight Manifests

☆☆☆☆☆ 0 Rated

Assets: **2**

Author: [redacted]

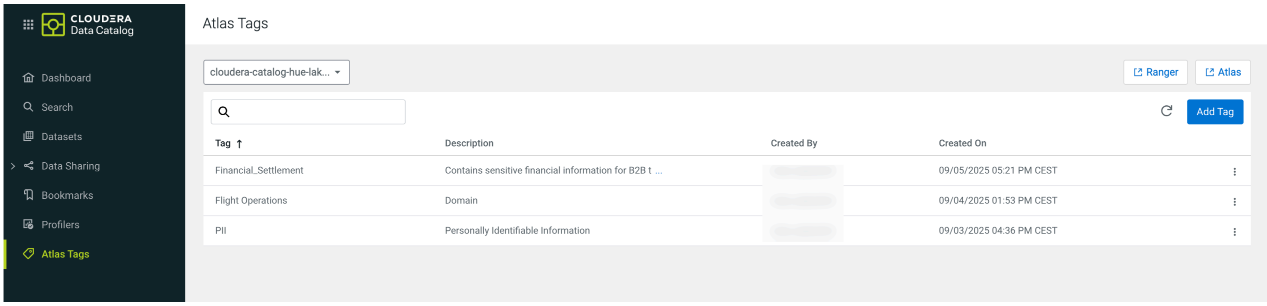
Data Lake: **cloudera-catalog-hue-lake-4**

0
0
Data Lake: cloudera-catalog-hue-lake-4
1.0
2

Note: These tags do not synchronize to Atlas.

Atlas tags overview

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 Data Compliance Profiler or the Cluster Sensitivity Profiler.



Atlas Tags

Tag ↑	Description	Created By	Created On
Financial_Settlement	Contains sensitive financial information for B2B t ...	[redacted]	09/05/2025 05:21 PM CEST
Flight Operations	Domain	[redacted]	09/04/2025 01:53 PM CEST
PII	Personally Identifiable Information	[redacted]	09/03/2025 04:36 PM CEST

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 more information, see the related documents.

Note: You must be a PowerUser to launch and delete profilers. However, users without this rule 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.

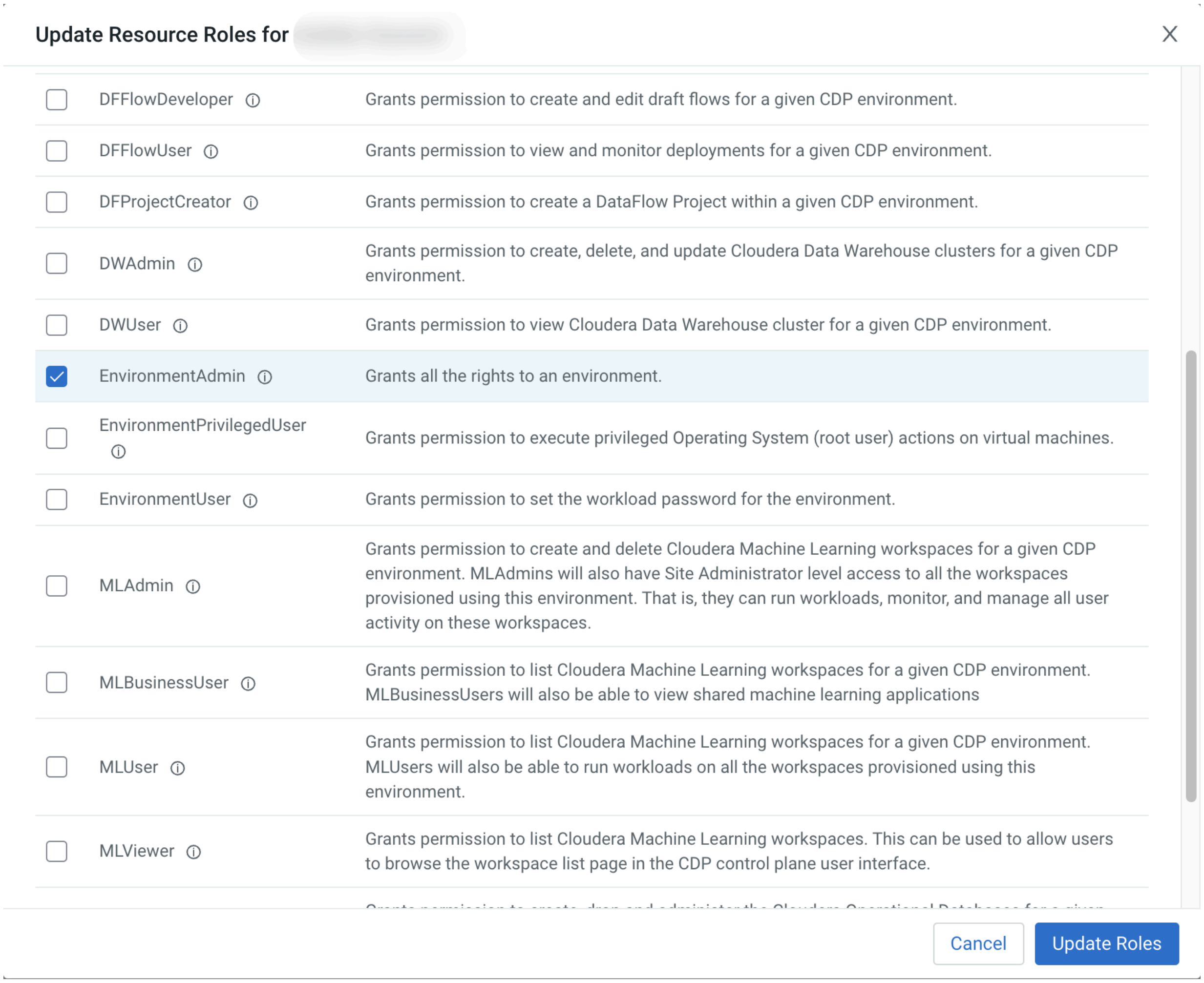
About this task

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

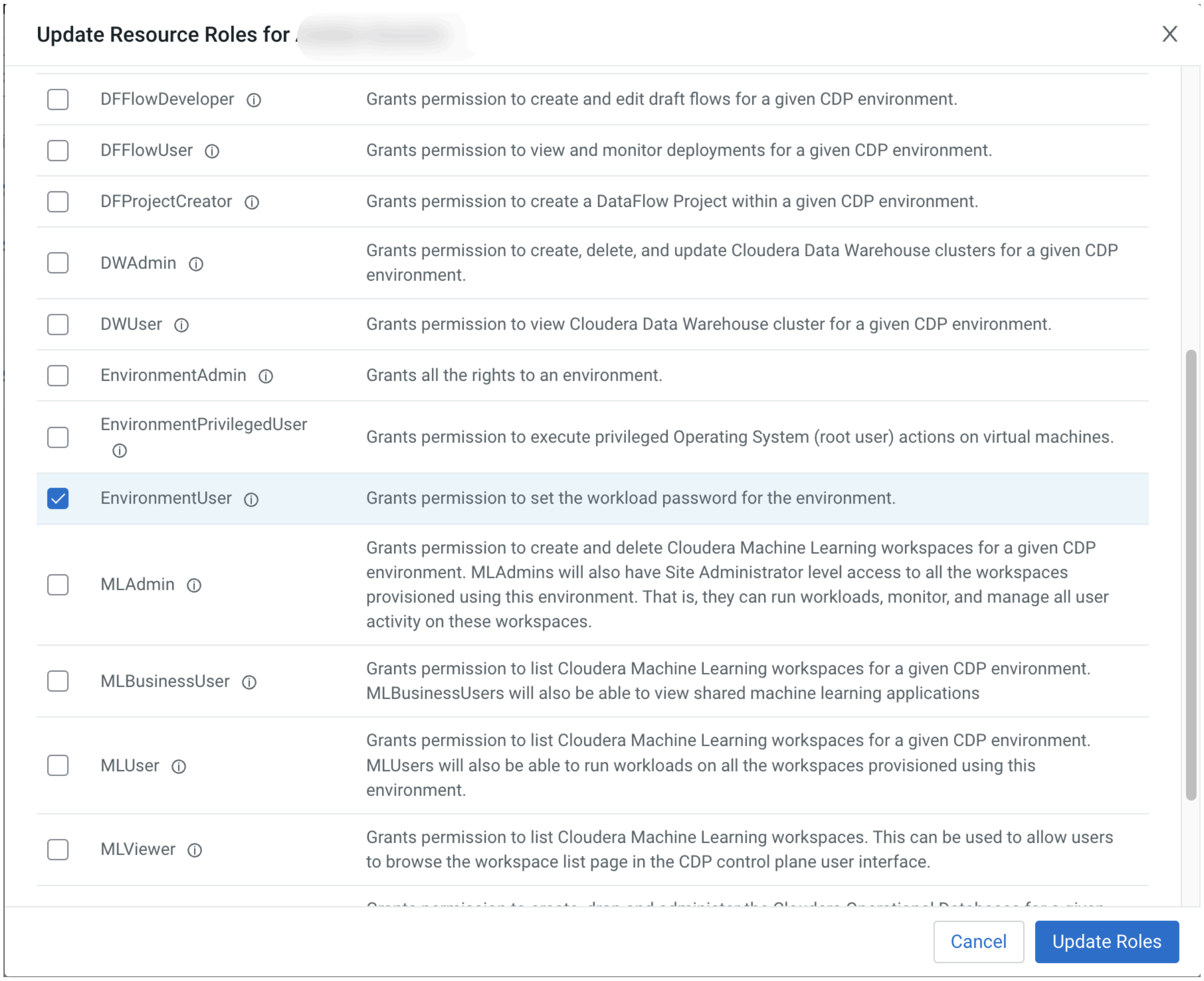
1. From Cloudera Management Console Environments > Select an environment > select the Actions drop-down > Manage Access.

2. Search for the user who requires Cloudera Data Catalog access > Select the check-box, either EnvironmentAdmin or EnvironmentUser > Click Update Roles.



Update Resource Roles for [redacted]

□	DFFlowDeveloper ⓘ	Grants permission to create and edit draft flows for a given CDP environment.
□	DFFlowUser ⓘ	Grants permission to view and monitor deployments for a given CDP environment.
□	DFProjectCreator ⓘ	Grants permission to create a DataFlow Project within a given CDP environment.
□	DWAdmin ⓘ	Grants permission to create, delete, and update Cloudera Data Warehouse clusters for a given CDP environment.
□	DWUser ⓘ	Grants permission to view Cloudera Data Warehouse cluster for a given CDP environment.
☑	EnvironmentAdmin ⓘ	Grants all the rights to an environment.
□	EnvironmentPrivilegedUser ⓘ	Grants permission to execute privileged Operating System (root user) actions on virtual machines.
□	EnvironmentUser ⓘ	Grants permission to set the workload password for the environment.
□	MLAdmin ⓘ	Grants permission to create and delete Cloudera Machine Learning workspaces for a given CDP environment. MLAdmins will also have Site Administrator level access to all the workspaces provisioned using this environment. That is, they can run workloads, monitor, and manage all user activity on these workspaces.
□	MLBusinessUser ⓘ	Grants permission to list Cloudera Machine Learning workspaces for a given CDP environment. MLBusinessUsers will also be able to view shared machine learning applications
□	MLUser ⓘ	Grants permission to list Cloudera Machine Learning workspaces for a given CDP environment. MLUsers will also be able to run workloads on all the workspaces provisioned using this environment.
□	MLViewer ⓘ	Grants permission to list Cloudera Machine Learning workspaces. This can be used to allow users to browse the workspace list page in the CDP control plane user interface.
		Grants permission to create, drop and administer the Cloudera Operational Database for a given



Update Resource Roles for [redacted]

□	DFFlowDeveloper ⓘ	Grants permission to create and edit draft flows for a given CDP environment.
□	DFFlowUser ⓘ	Grants permission to view and monitor deployments for a given CDP environment.
□	DFProjectCreator ⓘ	Grants permission to create a DataFlow Project within a given CDP environment.
□	DWAdmin ⓘ	Grants permission to create, delete, and update Cloudera Data Warehouse clusters for a given CDP environment.
□	DWUser ⓘ	Grants permission to view Cloudera Data Warehouse cluster for a given CDP environment.
□	EnvironmentAdmin ⓘ	Grants all the rights to an environment.
□	EnvironmentPrivilegedUser ⓘ	Grants permission to execute privileged Operating System (root user) actions on virtual machines.
☑	EnvironmentUser ⓘ	Grants permission to set the workload password for the environment.
□	MLAdmin ⓘ	Grants permission to create and delete Cloudera Machine Learning workspaces for a given CDP environment. MLAdmins will also have Site Administrator level access to all the workspaces provisioned using this environment. That is, they can run workloads, monitor, and manage all user activity on these workspaces.
□	MLBusinessUser ⓘ	Grants permission to list Cloudera Machine Learning workspaces for a given CDP environment. MLBusinessUsers will also be able to view shared machine learning applications
□	MLUser ⓘ	Grants permission to list Cloudera Machine Learning workspaces for a given CDP environment. MLUsers will also be able to run workloads on all the workspaces provisioned using this environment.
□	MLViewer ⓘ	Grants permission to list Cloudera Machine Learning workspaces. This can be used to allow users to browse the workspace list page in the CDP control plane user interface.
		Grants permission to create, drop and administer the Cloudera Operational Databases for a given

3. 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