

Datasets

Date published: 2023-12-16

Date modified: 2025-11-10

CLOUDERA

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

Managing datasets.....	4
Creating datasets.....	4
Editing datasets.....	9
Deleting datasets.....	10
 Bookmarks overview.....	 11

Managing datasets

You can view, create, edit, and delete datasets to manage and govern various kinds of data objects as a single unit through a unified interface.

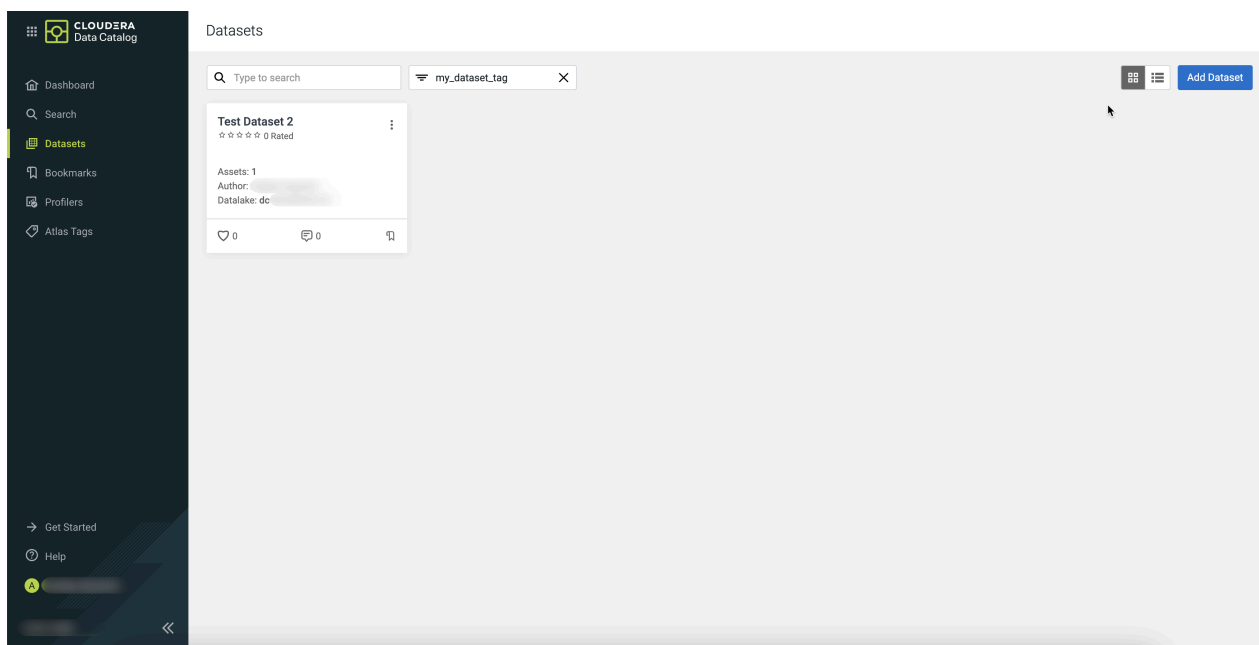
In Cloudera Data Catalog, click Datasets to view all the datasets.

Search for datasets

On the **Datasets** page, enter a search string in the search box to view all asset collections with names or descriptions that contain the search string.

Filter datasets by tags

You can view and filter datasets with tags added during dataset creation. Select the tag from the drop down list or enter the tag in the filter box. Any dataset with the filter tag assigned to a column will appear in the results.



Note: The tags added during dataset creation are not synchronized to Atlas. They can be used for organization only in .

Related Information

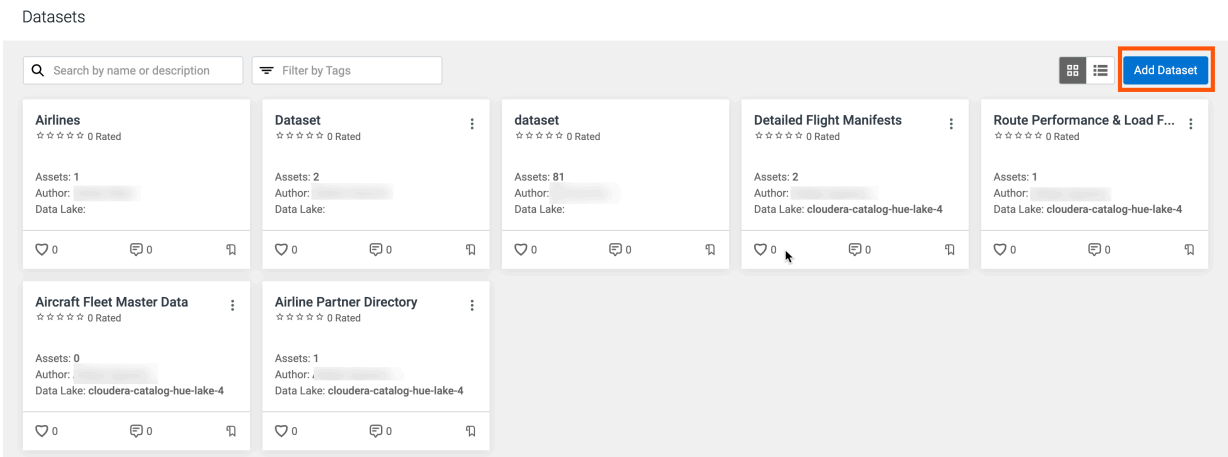
[Datasets overview](#)

Creating datasets

You can group data assets into datasets. This enables you to organize data based on business classifications, purpose, protection requirements, or more. Examples of datasets are: customer profiles, sales assets, financials, PII, and HR data.

Procedure

- 1. From the **Datasets** page, click Add Datasets.
The **Add** page appears.




- 2. Enter the following information.

Field Name	Description	Example Values
Name	Enter an appropriate dataset name. This name cannot be duplicated across the system. (Mandatory)	Customer Profiles, Sales Assets, Financials
Description	Describe the purpose or intent of the dataset. (Mandatory)	Contains customer profiles: data assets for US and WW.
Data Lake	Assign the dataset to a data lake. Choose from a list of available data lakes. (Mandatory)	dss_bbsh_clust3
Tags	Add tags to your dataset for context and subsequent lookup. Tags enable you to quickly catalog, search and retrieve asset collections in Cloudera Data Catalog, as well as, share such information with others in the future. (Optional) ¹	se, pii, geo, finance

¹



Note: These tags are not synchronized with Atlas.



Field Name	Description	Example Values
Public/Private	<p>Select Public if you want other users to have access to this dataset. Select Private if only you want to have access to this dataset.</p> <div><p>Note: You can change the status of the asset collection later. Click the lock icon on the Dataset Details page to change the access state of the dataset.</p></div>	Public/Private

Add

Name*

Legacy Airline Reference Data

Description*

B *I* U ~~S~~      

This dataset contains historical or superseded airline reference tables. The information in these assets may be out of date or incomplete and has been replaced by the `Airline Partner Directory` dataset. This data is preserved for historical analysis or archival purposes only.

Data Lake*

cloudera-catalog-hue-lake-4

Tags

Add tags to your dataset for context and subsequent lookup.

Legacy 

Archive 

Internal 

Partnerships 



Public

Next

Cancel

3. Click Next.
The **Dataset Details** page appears for the new dataset.
4. Click Add Assets to add related data assets into your dataset.

Details

Home

Datasets

Datasets Details

Legacy Airline Reference Data

This dataset contains historical or superseded airline reference tables. The information in these assets may be out of date or incomplete and has been replaced by the Airline Partner Directory dataset. This data is preserved for historical analysis or archival purposes only.

Data Lake cloudera-catalog-hue-lake-4	Tags Informal Legacy Archive Partnerships	Created By [Redacted]	Created On 09/05/2025 06:25 PM CEST	Last Modified On 09/05/2025 06:25 PM CEST
--	---	--------------------------	--	--

Assets

Add related data assets into your dataset, by searching this data lake.

Add Assets

Save

Cancel

The **Search** page appears.

5. Search for assets using the search bar.
 - a) Use filters to search for specific assets based on the attributes of assets. Click Filter to display the filters available.

- **Created:** Select the time to refine the search on the basis of when the asset has been created.
- **Owner:** Enter the name of the owner to refine the search on the basis of the owners of the assets.
- **Database name:** Enter the name of the database.



Note: The data base name filter uses the "begins with" logic.

- **Tags:** Enter the names of the tags after selecting its type (Table/Column).

Datasets Details

Search

Q airlines

Created OnWheneverOwnerEnter OwnerDatabase nameEnter Database nameTagsTable TagEnter Table Tag

SearchReset

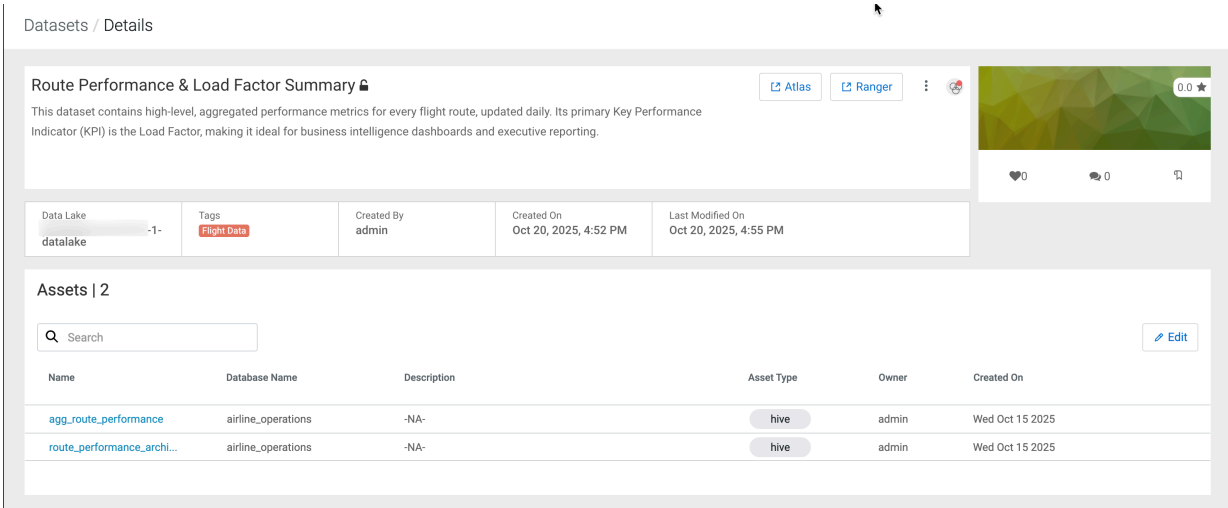
Results

<input type="checkbox"/>	Name	Database Name	Description	Asset Type	Owner	Created On
<input type="checkbox"/>	airlines_new	airline_operations	This is a master reference table contain ...	iceberg		09/05/2025 03:09 PM ...
<input type="checkbox"/>	airlines	airline_operations	This table has been superseded by the mo ...	iceberg		09/05/2025 06:28 PM ...

- b) Select one more than one filter if needed.
 - c) Click Search to view the assets.
 - d) Click Reset to reset the filters and search again.
 - e) From the list, click to select the assets that you like to add to your dataset.
6. Search for assets using the **Advanced** tab, if needed. Advanced 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. Advanced search conditions are a subset of attributes for the Apache Atlas type `hive_table`.
7. Click Add.

The assets are added to the dataset and the **Search** page is refreshed.

- 8. Close the **Search** tab by clicking Done.
The Datasets Details page appears.
- 9. Click Save to finish editing your dataset.

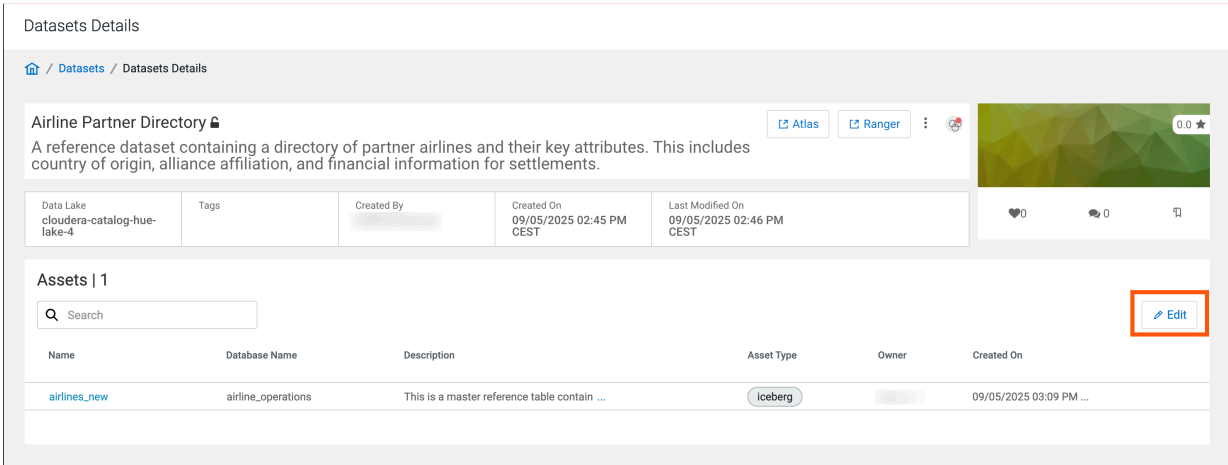


Editing datasets

You can edit datasets by adding or removing assets and changing the access state of the datasets.

Procedure

- 1. Click a dataset in the list to edit it. The **Dataset Details** page of that dataset appears.



- 2. On the **Assets** tab, click Edit to edit the content of this dataset. The dataset appears in edit mode.
If another user is editing this dataset, an error message will appear saying that this dataset is being edited by another user and you cannot edit it.

3. Add or remove assets in the dataset.
 - a) Click Add to add new assets to this dataset.

Datasets Details

[Home](#) / [Datasets](#) / Datasets Details

Airline Partner Directory

A reference dataset containing a directory of partner airlines and their key attributes. This includes country of origin, alliance affiliation, and financial information for settlements.

[Atlas](#) [Ranger](#)

Data Lake cloudera-catalog-hue-lake-4	Tags	Created By	Created On 09/05/2025 02:45 PM CEST	Last Modified On 09/05/2025 02:46 PM CEST	0 0
--	------	------------	--	--	------

Assets | 1

<input type="checkbox"/>	Name	Database Name	Description	Asset Type	Owner	Created On
<input type="checkbox"/>	airlines_new	airline_operations	This is a master reference table contain ...	iceberg		09/05/2025 03:09 PM ...

* Some information might not be available to unauthorized users.

[Save](#) [Cancel](#) [+ Add](#)

- b) Select one or more assets and click Remove to remove assets from this dataset.
4. Click Save to save the changes that you made to the dataset.
5. Click Cancel to undo any changes that you made to this dataset.



Note: You also can edit the metadata (name, description, and tags) of the datasets. Being an owner of specific datasets, and making them private, you can update the name, description, and tags.

Deleting datasets

You might want to delete a datasets if you no longer need to track those datasets, or if you want to reassign those assets to another dataset. You can delete datasets at any time. Deleting datasets does not delete the assets contained therein, it only disassembles the datasets. You can recreate datasets or reassign assets to new datasets.

Procedure

1. From **Datasets** page, click the icon beside the name of the dataset you want to delete.
2. Click Delete.

Datasets Details

[Home](#) / [Datasets](#) / Datasets Details

Legacy Airline Reference Data

This dataset contains historical or superseded airline reference tables. The information in these assets may be out of date or incomplete and has been replaced by the Airline Partner Directory dataset. This data is preserved for historical analysis or archival purposes only.

[Atlas](#) [Ranger](#)

Data Lake cloudera-catalog-hue-lake-4	Tags Internal Legacy Archive Partnerships	Created By	Created On 09/05/2025 06:40 PM CEST	Last Modified On 09/05/2025 06:47 PM CEST	0 0
--	--	------------	--	--	------

Assets | 1

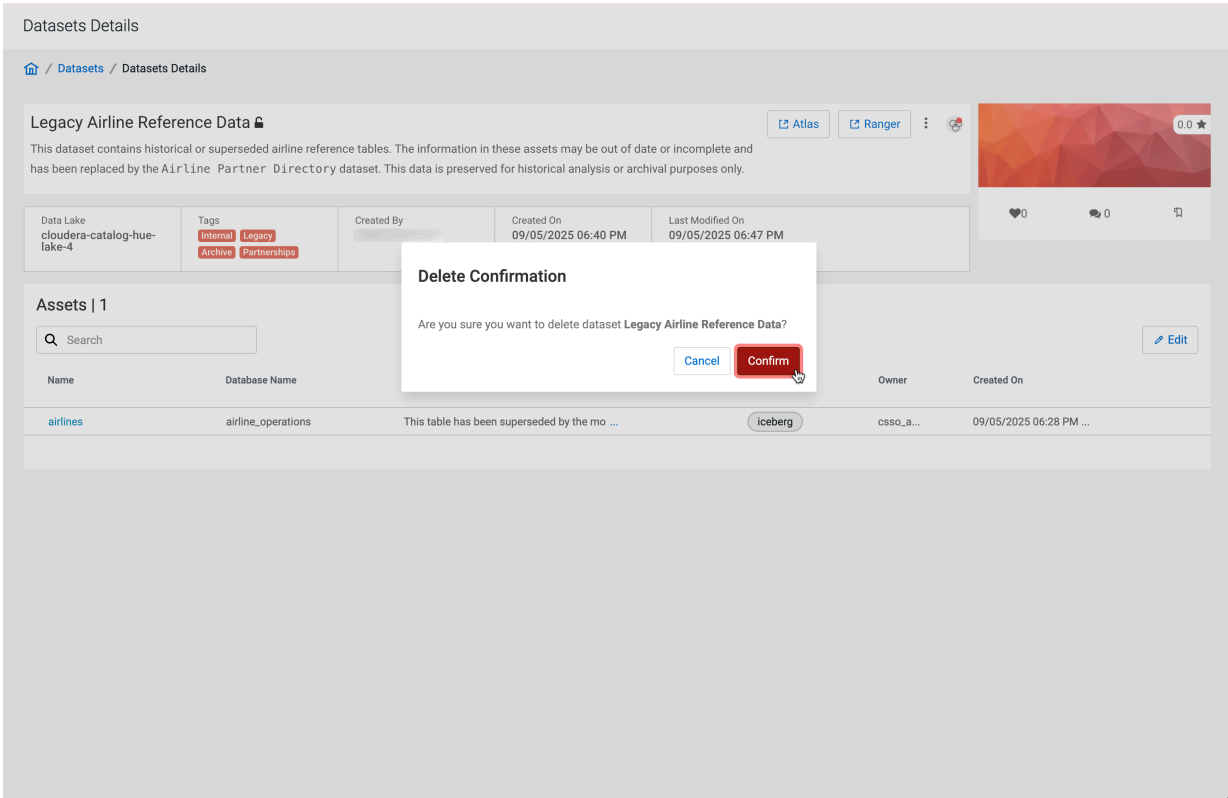
<input type="checkbox"/>	Name	Database Name	Description	Asset Type	Owner	Created On
<input type="checkbox"/>	airlines	airline_operations	This table has been superseded by the mo ...	iceberg	csso_a...	09/05/2025 06:28 PM ...

[Delete](#) [Edit](#)



Note: Datasets can be deleted only by their creators.

3. Click Confirm.



You are returned to the Datasets home page.


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
aws 2


Total Assets
1680
Realtime

Profiler Configured
6/6 in all data lakes

Assets Profiled
0 in last 7 days


[My Datasets and Bookmarks](#) [Active Profilers](#) [Data Lakes](#)

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

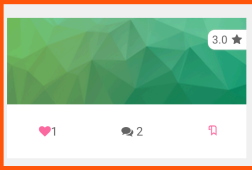
[/ Datasets / Datasets Details](#)

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.

[Atlas](#) [Ranger](#)

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



3.0 ★


♥ 1  2 

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

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.

Atlas

Ranger

0.0

★

Data Lake

cloudera-catalog-hue-lake-4

Tags

Flight Operations

Revenue Management

Business Performance

Aggregated

Summary

KPI

Analytics

Anonymized

Internal

Certified

Created By

Created On


09/05/2025 12:43 PM CEST

Last Modified On

09/05/2025 01:08 PM CEST


♥1

0

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

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.

Atlas

Ranger

5.0

★

♥1


2

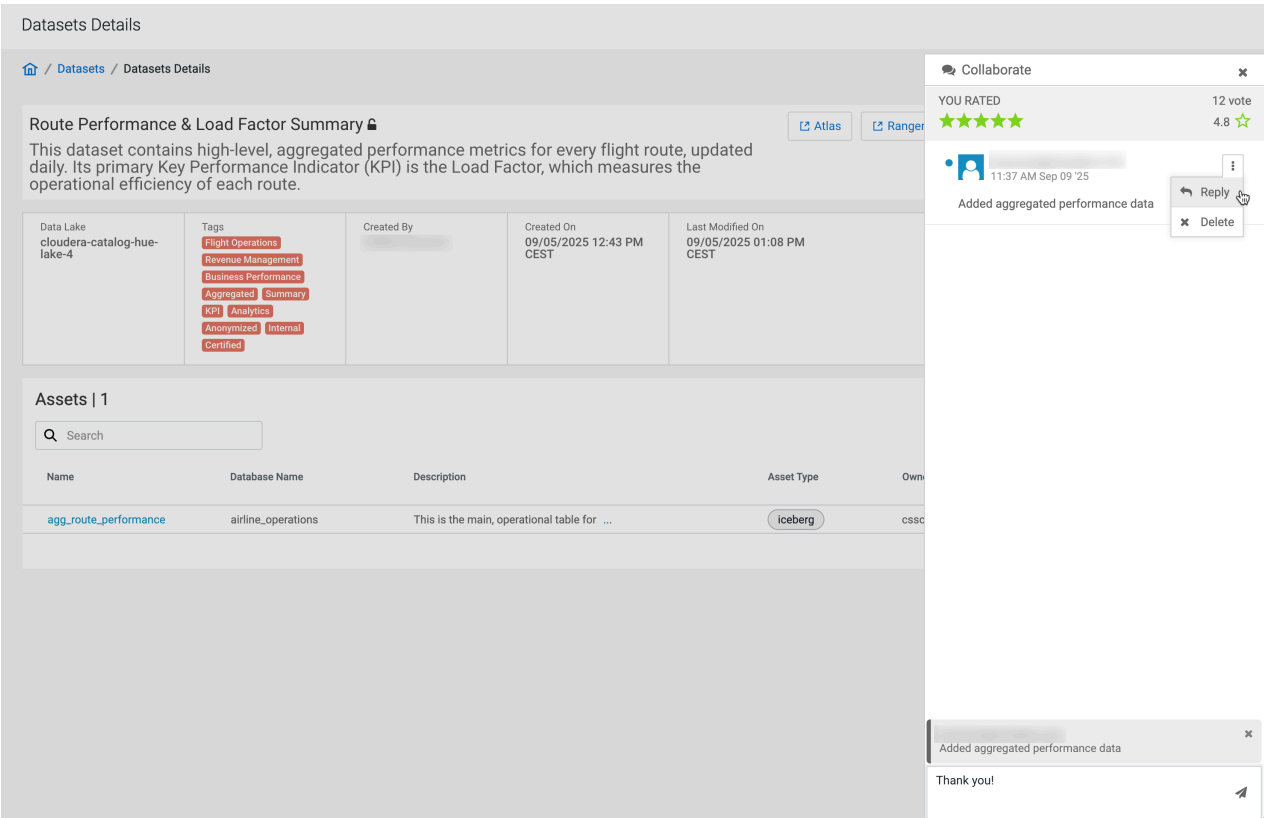
Assets | 1

Q Search

Edit


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 ...

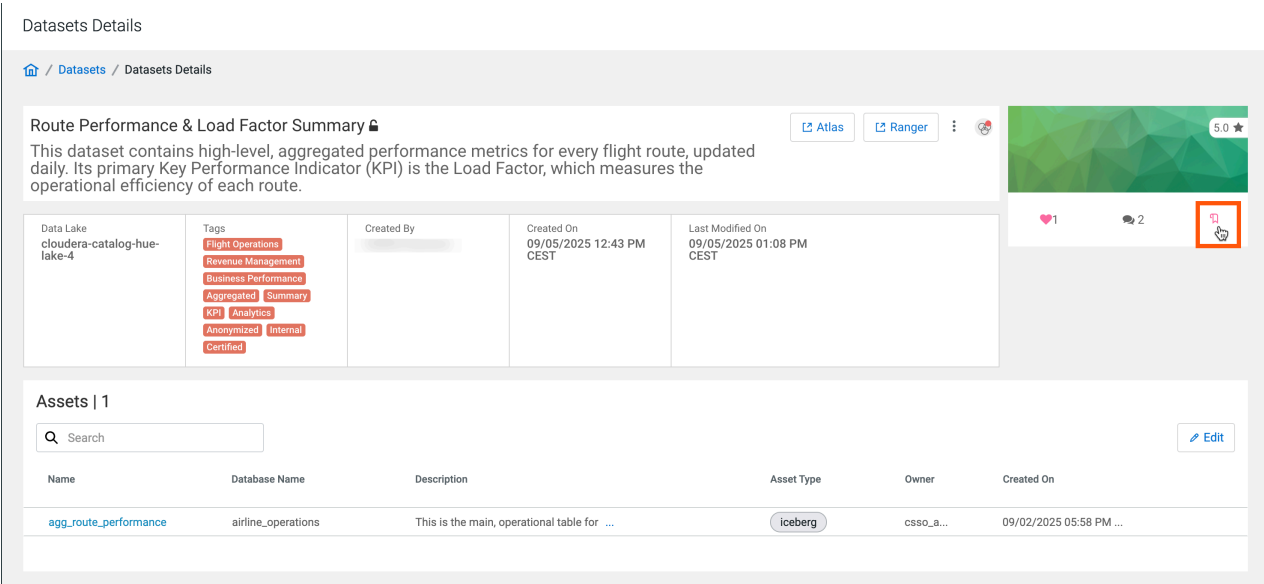
The **Collaborate** tab expands. Click  icon to reply to an existing comment. You can continue to add comments for each dataset.



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.



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.

Home

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	<div>Flight Operations</div> <div>Revenue Management</div>		09/05/2025 12:43 PM CEST	09/05/2025 01:08 PM CEST

Collaborate

RATE THIS COLLECTION

★

★

★

★

★

12 vote

3.0

Profile Icon

@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

The screenshot shows the Databricks interface. At the top, there is a search bar with the placeholder text "Search by name or description". Below the search bar, there is a card titled "Detailed Flight Manifests" with a rating of 0 stars. To the right of this card, a dropdown menu is open, showing a list of items: "Revenue Management" (1), "Summary" (1), "Flight Operations" (2), "Aggregated" (1), "Anonymized" (1), and "test_dss_tag_0" (1). A hand cursor is pointing at the "Flight Operations" item. The background is a light gray.



Note: These tags do not synchronize to Atlas.