

## Datasets

Date published: 2019-11-14

Date modified: 2025-06-11

# 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

<b>Understanding datasets.....</b>	<b>4</b>
<b>Managing Datasets.....</b>	<b>4</b>
Creating Datasets.....	5
Editing Datasets.....	6
Deleting Datasets.....	7
<b>Collaborate with other users.....</b>	<b>7</b>

## Understanding datasets

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

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, etc.

- 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. Advanced search conditions are a subset of attributes for the Apache Atlas type `hive_table`.

- 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](#)

## 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.

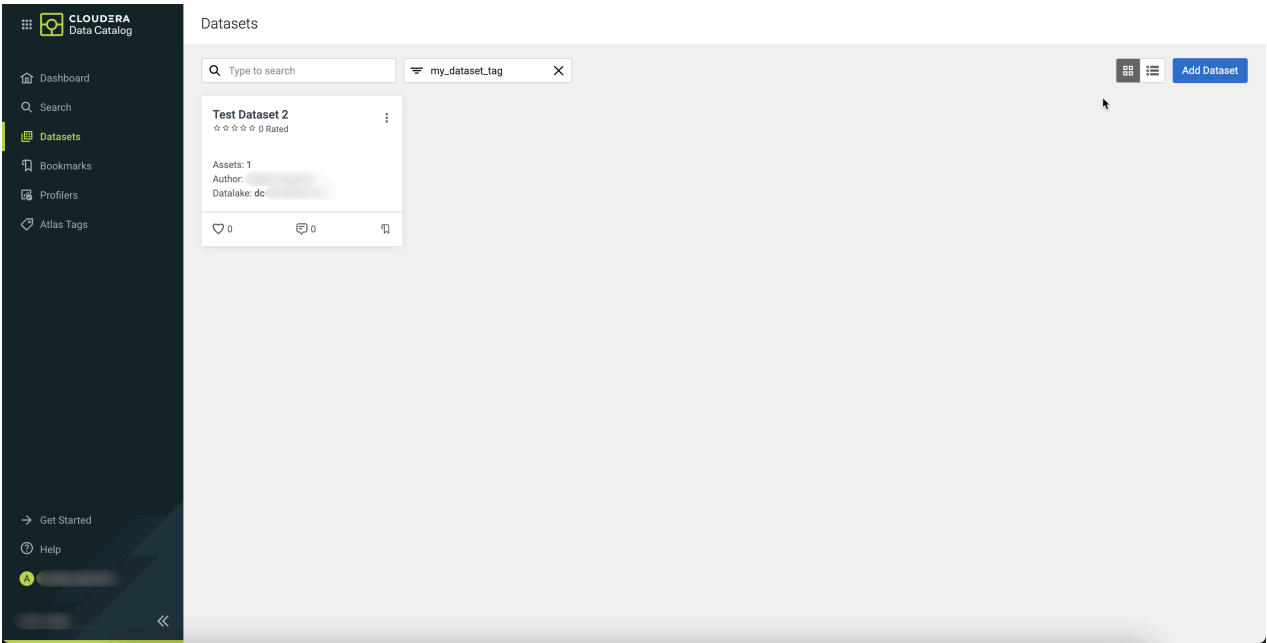
In Cloudera Data Catalog, click Datasets to view all the datasets.


### 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**  
[Understanding datasets](#)


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)	se, pii, geo, finance

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> <p> <b>Note:</b> You can change the status of the asset collection later. Click the lock icon on the <b>Dataset Details</b> page to change the access state of the dataset.</p>	Public/Private

3. Click Next.

The **Dataset Details** page appears for the new dataset.

4. Click Add Assets to add related data assets into your dataset.

The **Asset 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.
- **DB Name:** Enter the name of the database.



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

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

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 **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.

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.



**Note:** Datasets can be deleted only by their creators.

3. Click Confirm.  
You are returned to the Datasets home page.

## Collaborate with other users

You can collaborate and share insights with other users in the enterprise regarding various datasets.

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.

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. 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.

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. allows you to collaborate with other users by adding comments.


Click the  icon to add a comment about this dataset. 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 rate the open dataset. The **Collaborate** tab expands.

Click the stars to provide your own rating. 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 the dataset. You can also click on the tags to search for datasets with similar tags. There are two types of tags. System tags are automatically generated based on the details of the assets in the datasets. You can add more tags that appear in the list of user generated tags.



**Note:** These tags do not synchronize to Atlas.