

Data Catalog 1.5.3

Data Catalog Operations

Date published: 2023-10-10

Date modified: 2024-03-05

CLOUDERA

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2024. 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.....	5
Create Datasets.....	5
Edit Datasets.....	6
Delete Datasets.....	7
Collaborate with other users.....	7
Search for Assets.....	8
Filters.....	8
Prepopulating Asset Owners.....	8
Viewing Ranger and Atlas applications.....	9
Accessing Data Lakes.....	10
Download CSV option.....	10
Searching for assets using Glossary.....	11
Using Terms in Data Catalog.....	12
Mapping glossary terms.....	12
Searching for assets using glossary terms.....	15
Additional search options for asset types.....	17
Searching for assets in Data Catalog using additional search options.....	18
Accessing Tables based on Ranger policies.....	19
Creating Classification for selected assets.....	21
Adding Classifications / Terms for selected assets.....	21
Additional Entity type selection for searching Assets.....	21
Viewing Data Asset Details.....	23
Viewing Data Assets.....	24
View Data Asset Schema.....	25
Navigating from the container asset to the parent asset from Asset Details page.....	26
View Authorization Policies on a Data Asset.....	26
View Data Asset Audit Logs.....	27
Navigation Support for Hive entity within Lineage.....	27
Adding Hive asset to one or more datasets on Asset Details screen.....	28
Viewing Atlas Entity Audits.....	30
Managing Profilers.....	32
Scheduling Profiler jobs on your Private Cloud cluster.....	33
Launching profilers using Command-line.....	35
Deleting profilers.....	38
On-Demand Profilers.....	39
Tracking Profiler Jobs.....	40
Viewing Profiler Jobs.....	41
Viewing Profiler Configurations.....	42
Additional Configuration for Ranger Audit Profiler.....	42
Additional Configuration for Hive Column Profiler.....	43
Additional Configuration for Cluster Sensitivity Profiler.....	45
Understanding Cron Expression generator.....	46
Setting Asset filter rules.....	46
Enable or Disable Profilers.....	49
Profiler Tag Rules.....	50
Tag Management.....	51
Tagging Multiple Assets.....	53
Creating Custom Profiler Rules.....	56
Adding Custom Regular Expressions.....	56

Adding Lookup Files.....	57
Using Behaviors.....	57
Regular expressions.....	58
Using DSL Grammar.....	58

Managing Datasets

You can view, create, edit, and delete Datasets.

On the Data Catalog menu, 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 that contain the search string.

Filter Datasets by Tags

You can filter Datasets and view Dataset with the tags. 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 filter results.

Related Information


[Understanding asset collections](#)

Create 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 one 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 your to quickly catalog, search and retrieve asset collections as well as share such information with others in the future. (Optional)	se, pii, geo, finance
Public/Private	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.  Note: You can later change the status of the asset collection. Click the lock icon on the Dataset Details page to change the access state of the dataset.	

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 Basic Search.

a) Search using the name of the asset by entering the name in the search bar.

b) Use filters to search for specific assets based on the attributes of assets. Click Filter to display the filters available.

- Created Time: From the dropdown list, 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.
- Tag: Enter the names of the tags.

c) Select one more than one filter if needed.

d) Click Search to view the assets. The Results appear.

e) Click Reset to reset the filters and search again.

f) From the list, click to select the assets that you like to add to your dataset.

6. Search for assets using Advanced Search, 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 Done.**

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

8. Close the Search tab.

The Datasets Details page appears.

9. Click Save.

Edit 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.

Delete Datasets

You might want to delete an 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 an Datasets does not delete the assets contained therein, it only disassembles the Datasets. You can re-create Datasets or reassign assets to new Datasets.

Procedure

1. From Data Catalog Datasets page, click the More Options icon beside the name of the Dataset you want to delete.
2. Click Delete.
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 asset collection page, you can see additional details about the dataset. The collaboration details are also displayed in this tab. The tab displays the following details - average rating for the asset collection, the number of likes, the number of comments, and 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 like icon on the Dataset page displays the total number of likes received by this Dataset.

Click the like 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. Data Catalog allows you to collaborate with other users by adding comments.

Click the comment icon to add a comment about this Dataset. The Collaborate tab expands. Click Actions menu 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 bookmark 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 on the left navigation menu.

Rate the Dataset

You can also rate the datasets on a scale of one to five. Click the star 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.

Search for Assets

On the 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 Data Catalog Search, you are looking up names, types, descriptions, and other metadata collected by 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 assets that Data Catalog stores.

Related Information

[Understanding data assets](#)

Filters

When you select a property value, a filter breadcrumb shows above the search results.

You can further refine your search results using filters as follows:

- Owner - From all the owner names that appear, you can select the owner to further refine the results and display those search results with the selected owner.
- Database - Select the database to view all the assets stored in that database. This filter is applicable to Hive and HBase tables only.



Note: For information purposes, Database filter is displayed as Namespace in case of HBase tables.

- Entity Tag - Use entity tags to refine your search results. You can add business metadata as entity tags in Atlas and use these tags to refine your search results and view the details of the required data asset.
- Created Within - You can choose to refine your search results of assets within the data lake to view the data assets created within the last 7 days, 15 days, or 30 days. You can also add custom values such as 5 days or 10 days to view specific information.
- Created Before - Depending on the time when the assets were created, you can choose to refine the search results and view data assets created before 1 day, 7 days, or 15 days. You can add custom values to view data assets created before the days of your preference such as 8 days or 12 days.



Note: These two filters (Created Within and Created Before) are applicable only when Atlas provides the created time for the assets.

- Column Tag - You can search for Hive and HBase table assets by tags that have been applied on their children entities, that is, columns or column families using the column tags filter.
- Glossary - You can filter assets based on business glossary terms. You can search for any asset without any entity type restrictions.

Click Clear for any filter to clear the selection. You can use a combination of filters to view the required data assets.

Prepopulating Asset Owners

In Data Catalog, under the search page, you can filter for assets based on the owners.

Rather than having to type in the owners manually, the available asset owners are listed in drop down. Select the record from the list and add it as a filter criteria

For example, in the following diagram, the selected asset TYPE is “Hive”.

For the selected TYPE the owner “hive” is available in the drop-down and based on this condition, the assets can be filtered in the search page.

Data Catalog / Search

Type	Name	Location	Created On	Owner	Source
Hive Table	scheduled_queries	/information_schema	Mon Sep 07 2020	hive	hive
Hive Table	home_stay	/travel	Mon Sep 07 2020	hrt_1	hive
Hive Table	day_resort	/resort	Mon Sep 07 2020	hrt_qa	hive
Hive Table	weather	/wonders	Mon Sep 07 2020	hrt_qa	hive
Hive Table	lounge_classic	/airline	Mon Sep 07 2020	hrt_1	hive
Hive Table	call_center	/tpcds_bin_partitioned_parquet_50	Mon May 11 2020	csso_mhussain	hive
Hive Table	date_dim	/tpcds_bin_partitioned_parquet_50	Mon May 11 2020	csso_mhussain	hive
Hive Table	compactions	/sys	Mon Sep 07 2020	hive	hive
Hive Table	tables	/information_schema	Mon Sep 07 2020	hive	hive
Hive Table	column_privileges	/information_schema	Mon Sep 07 2020	hive	hive
Hive Table	table_privileges	/information_schema	Mon Sep 07 2020	hive	hive
Hive Table	lounge_premium	/airline	Mon Sep 07 2020	hrt_1	hive
Hive Table	lounge	/airline	Mon Sep 07 2020	hrt_1	hive
Hive Table	version	/sys	Mon Sep 07 2020	hive	hive
Hive Table	flight	/airline	Mon Sep 07 2020	hrt_1	hive
Hive Table	world	/wonders	Mon Sep 07 2020	hrt_qa	hive
Hive Table	schemata	/information_schema	Mon Sep 07 2020	hive	hive
Hive Table	partition_stats_view	/sys	Mon Sep 07 2020	hive	hive
Hive Table	scheduled_executions	/information_schema	Mon Sep 07 2020	hive	hive
Hive Table	cdh_version	/sys	Mon Sep 07 2020	hive	hive

Type	Name	Qualified Name	Created On	Owner	Source
Hive Table	inventory	test_dss_db.inventory@cm	Wed Nov 01 2023	hive	hive
Hive Table	catalog_returns	test_dss_db.catalog_returns@cm	Wed Nov 01 2023	hive	hive
Hive Table	store_sales	test_dss_db.store_sales@cm	Wed Nov 01 2023	hive	hive
Hive Table	web_site_one	test_dss_db.web_site_one@cm	Wed Nov 01 2023	hive	hive
Hive Table	sales	marketing.sales@cm	Wed Nov 01 2023	hive	hive
Hive Table	persons	personal_data.persons@cm	Wed Nov 01 2023	hive	hive
Hive Table	web_site_five	test_dss_db.web_site_five@cm	Wed Nov 01 2023	hive	hive
Hive Table	ww_customers	hortoniabank.ww_customers@cm	Wed Nov 01 2023	hive	hive
Hive Table	tax_2015	finance.tax_2015@cm	Wed Nov 01 2023	hive	hive
Hive Table	provider_summary	claim.provider_summary@cm	Wed Nov 01 2023	hive	hive
Hive Table	claim_savings	cost_savings.claim_savings@cm	Wed Nov 01 2023	hive	hive
Hive Table	web_site_four	test_dss_db.web_site_four@cm	Wed Nov 01 2023	hive	hive
Hive Table	catalog_page	test_dss_db.catalog_page@cm	Wed Nov 01 2023	hive	hive
Hive Table	store_returns	test_dss_db.store_returns@cm	Wed Nov 01 2023	hive	hive
Hive Table	household_demographics	test_dss_db.household_demographics@cm	Wed Nov 01 2023	hive	hive

Viewing Ranger and Atlas applications

For the selected data lake, click Atlas and Ranger links to navigate to the respective services in a new browser tab.

The Atlas and Ranger buttons seen on the search page of Data Catalog allows you to navigate to the specific Base cluster component.

Clicking on Atlas and Ranger links enables you to sign into the respective services and proceed further.



Note: When you click on Atlas and Ranger buttons, you must separately sign into these services and proceed further.

Accessing Data Lakes

In the Data Catalog search dashboard, the accessible data lakes are displayed under the search panel.

For more information about Data Lakes, see [Data Lake Security](#).

For example, in the following diagram, the logged in user has access to all the listed data lakes.

Download CSV option

The download CSV feature in Data Catalog allows you to download the search result for the current / specified query with the selected data lake. The feature allows you to download upto 10000 rows for the current search query.

You can access the [data lake](#) (OCP / ECS cluster) to query.

The CSV format does not confirm with any specific order or continuation in the downloaded results. For example, a user downloads 10000 assets once and later downloads the CSV again with about 10000 assets. The CSV may not contain the search results in the same order as it was downloaded earlier or previously.

The following example images provides a sample download flow.

Search

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

Data Lakes

- cod-7213 43
- glue.cod-7213 NA
- cod-7212 31
- glue.cod-7212 NA
- cod-7213-gcp 9
- spark-p5zj3y 16
- Cluster 1, rachit NA
- Cluster 1, dc_rid NA

Filters

TYPE

- ☐ Hive Table
- ☐ HBase Table
- [+ Add New Value](#)

OWNERS

- ☒ atlas
- ☒ hbase
- ☐ hive
- ☐ public

Setup the Profiler for spark-p5zj3y

Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started >](#)

spark-p5zj3y | 16

Type	Name	Qualified Name	Created On	Owner	Action
<input type="checkbox"/> HBase Table	hbase-acl	hbase.acl@cm	Wed Dec 01 2021	hbase	
<input type="checkbox"/> HBase Namespace	hbase	hbase@cm	-NA-	hbase	
<input type="checkbox"/> HBase Column Family	l	hbase.acl.l@cm	Wed Dec 01 2021	hbase	
<input type="checkbox"/> HBase Column Family	dt	default:ATLAS_ENTITY_AUDIT_EVENTS.dt@cm	Wed Dec 01 2021	hbase	
<input type="checkbox"/> HBase Table	ATLAS_ENTITY_AUDIT_EVENTS	default:ATLAS_ENTITY_AUDIT_EVENTS@cm	Wed Dec 01 2021	hbase	
<input type="checkbox"/> HBase Column Family	m	default:atlas_janus.m@cm	Wed Dec 01 2021	atlas	
<input type="checkbox"/> HBase Column Family	g	default:atlas_janus.g@cm	Wed Dec 01 2021	atlas	
<input type="checkbox"/> HBase Column Family	l	default:atlas_janus.l@cm	Wed Dec 01 2021	atlas	
<input type="checkbox"/> HBase Column Family	h	default:atlas_janus.h@cm	-NA-	atlas	
<input type="checkbox"/> HBase Column Family	f	default:atlas_janus.f@cm	-NA-	atlas	
<input type="checkbox"/> HBase Column Family	t	default:atlas_janus.t@cm	-NA-	atlas	
<input type="checkbox"/> HBase Table	atlas_janus	default:atlas_janus@cm	Wed Dec 01 2021	atlas	
<input type="checkbox"/> HBase Column Family	s	default:atlas_janus.s@cm	Wed Dec 01 2021	atlas	
<input type="checkbox"/> HBase Column Family	i	default:atlas_janus.i@cm	Wed Dec 01 2021	atlas	
<input type="checkbox"/> HBase Namespace	default	default@cm	-NA-	atlas	

Search

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

Data Lakes

- ☐ aws cod-7213 0
- ☐ glue:cod-7213 NA
- ☐ cod-7212 0
- ☐ glue:cod-7212 NA
- ☐ cod-7213-gcp 0
- ☒ spark-p5zj3y 663
- ☐ Cluster 1, rachit NA
- ☐ Cluster 1, dc_rid NA

Filters

TYPE

- ☐ Hive Table
- ☐ HBase Table
- [+ Add New Value](#) [Clear](#)

OWNERS

- ☐ atlas
- ☐ hbase
- ☐ hive
- ☐ public
- [Clear](#)

Setup the Profiler for spark-p5zj3y

Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started >](#)

spark-p5zj3y | 663

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	global_privs	sys.global_privs@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	database_params	sys.database_params@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	columns_v2	sys.columns_v2@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	bucketing_cols	sys.bucketing_cols@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	db_privs	sys.db_privs@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	tbls	sys.tbls@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	partitions	sys.partitions@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	version	sys.version@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	mv_creation_metadata	sys.mv_creation_metadata@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	db_version	sys.db_version@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	partition_params	sys.partition_params@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	part_privs	sys.part_privs@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	part_col_privs	sys.part_col_privs@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	roles	sys.roles@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	serdes	sys.serdes@cm	Wed Dec 01 2021	hive	hive

[Download CSV File](#)

Your file is getting downloaded

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

Data Lakes

- ☐ aws cod-7213 0
- ☐ glue:cod-7213 NA
- ☐ cod-7212 0
- ☐ glue:cod-7212 NA
- ☐ cod-7213-gcp 0
- ☒ spark-p5zj3y 663
- ☐ Cluster 1, rachit NA
- ☐ Cluster 1, dc_rid NA

Filters

TYPE

- ☐ Hive Table
- ☐ HBase Table
- [+ Add New Value](#) [Clear](#)

OWNERS

- ☐ atlas
- ☐ hbase
- ☐ hive
- [Clear](#)

Setup the Profiler for spark-p5zj3y

Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started >](#)

spark-p5zj3y | 663

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	global_privs	sys.global_privs@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	database_params	sys.database_params@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	columns_v2	sys.columns_v2@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	bucketing_cols	sys.bucketing_cols@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	db_privs	sys.db_privs@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	tbls	sys.tbls@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	partitions	sys.partitions@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	version	sys.version@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	mv_creation_metadata	sys.mv_creation_metadata@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	db_version	sys.db_version@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	partition_params	sys.partition_params@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	part_privs	sys.part_privs@cm	Wed Dec 01 2021	hive	hive
<input type="checkbox"/> Hive Table	part_col_privs	sys.part_col_privs@cm	Wed Dec 01 2021	hive	hive

[Download CSV File](#)

Your file has been downloaded

[Show all](#)

Searching for assets using Glossary

Use glossaries to define a common set of search terms that data users across your organization use to describe their data.

Data can describe a wide variety of content: lists of names or text or columns full of numbers. You can use algorithms to describe data as having a specific pattern, of being within a range or having wide variation, but what's missing from these descriptions is what does the data mean in a given business context and what is it used for? Is this column of integers the count of pallets that entered a warehouse on a given day or number of visitors for each room in a conference center?

The glossary is a way to organize the context information that your business uses to make sense of your data beyond what can be figured out just by looking at the content. The glossary holds the terms you've agreed upon across your organization so business users can use familiar terms to find what they are looking for.

Glossaries enable you to define a hierarchical set of business terms that represents your business domain.

Glossary terms can be thought of as of a flat (but searchable) list of business terms organized by glossaries. Unlike classifications, terms are not propagated through lineage relationships: the context of the term is what's important, so propagation may or may not make sense.

Using Terms in Data Catalog

You can use the Asset Details page in Data Catalog to add or modify “terms” for your selected assets.

A new widget called “Terms” is available in the Asset Details page. You can define rich glossary vocabularies using the natural terminology (technical terms and/or business terms). To semantically relate the term(s) to each other. And finally to map assets to glossary terms(s).

You can assign terms with entities, search for entities, filter entities by glossary term(s), and also search for entities by using associated term(s).



Note: When you work with terms in Data Catalog and map them to your assets, you can search for the same datasets in Atlas by using the corresponding terms.

Asset Details

The screenshot shows the 'world' Asset Details page. The 'Properties' section lists details like Type (HIVE TABLE), # of Columns (4), Data Lake, Datasets (1), Owner (hrt_qa), Created On, Last Access Time, Table Type (MANAGED_TABLE), Database (wonders), DB Catalog (cm), and Parent (wonders). The 'Classifiers' section shows a single classifier 'test_dss_atlas_tag_169ef_val25,...'. The 'Terms' widget is highlighted with a red box and contains an 'Add Terms' button. The 'Profilers' section shows two profilers: 'Cluster Sensitivity Profiler' and 'Hive Column Profiler', both with 'Run' buttons.

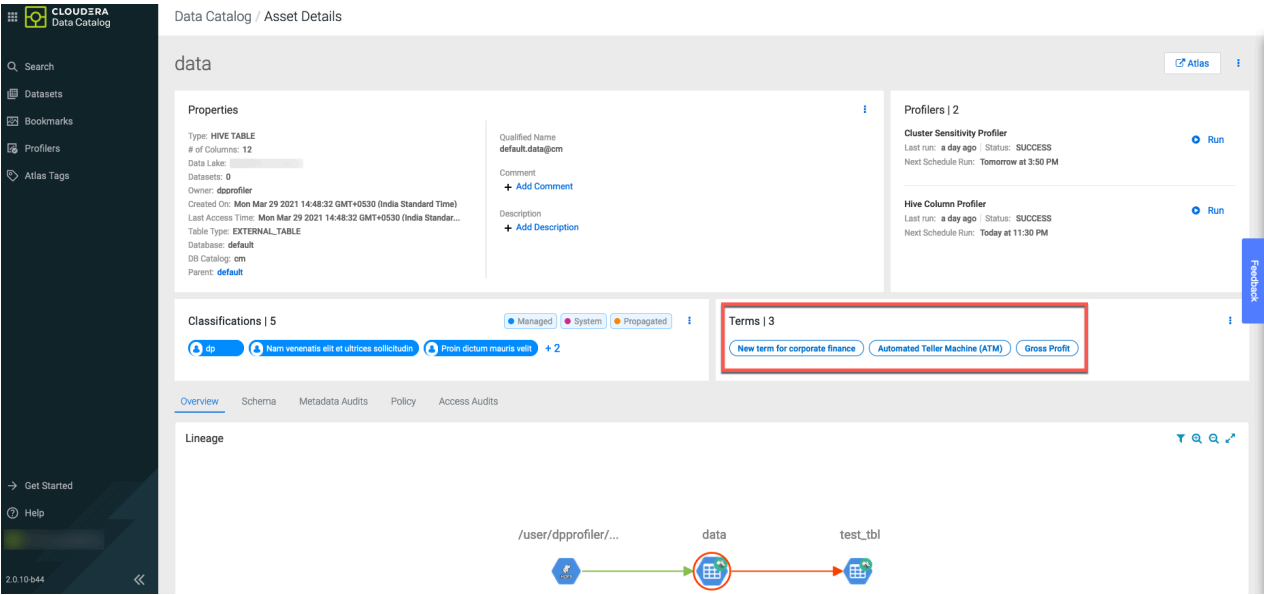
Mapping glossary terms

Data Catalog contains the glossary terms that are created in Atlas.

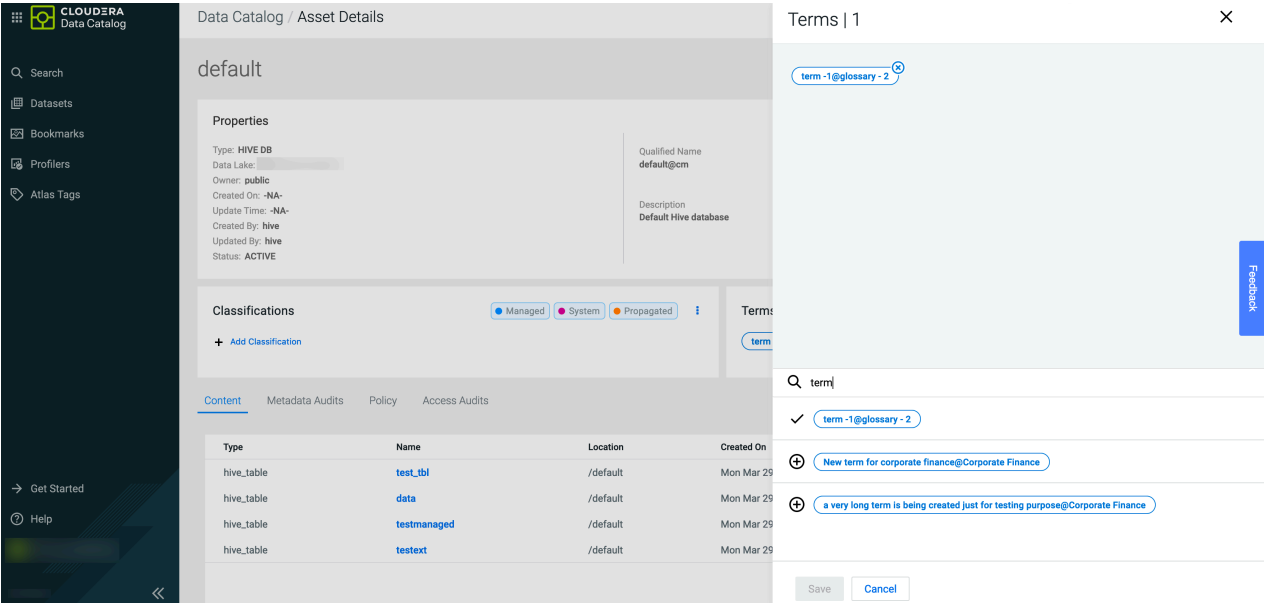
You can search for those terms in Data Catalog and map specific terms with Data assets. You can search for terms in Data Catalog to either add and delete them from the selected data asset. The selected asset displays the total number of terms associated or mapped accordingly.

When you map a specific term for your dataset, the term is displayed in the following format:

```
<termname>@glossaryname>
```

You can use the icon in the Terms widget on the Asset Details page to add new terms for your data asset. Click Save to save the changes.



CloudERA Data Catalog / Asset Details

test

Successfully updated the glossary terms for the asset

[Atlas](#)

Properties

Type: HIVE DB
 Data Lake:
 Owner: hive
 Created On: -NA-
 Update Time: -NA-
 Created By: hive
 Updated By: csso_ssudhindra
 Status: ACTIVE

Qualified Name: test@cm
 Description: [+ Add Description](#)

Classifications | 1

Managed System Propagated

dp_email

Terms | 1

a very long term is being created just fo...

Content Metadata Audits Policy Access Audits

Type	Name	Location	Created On	Owner	Source
hive_table	test100	/test	Mon Mar 29 2021	hive	hive
hive_table	testmanaged100	/test	Mon Mar 29 2021	hive	hive

You can search for the same asset in the corresponding Atlas environment as shown in the example image.

Apache Atlas

SEARCH CLASSIFICATION GLOSSARY

Basic Advanced

Search By Type: _ALL_ENTITY_TYPES

Search By Classification: Select Classification

data (hive_table)

Classifications: Nam venenatis elit et ultrices sollicitudin Proin dictum mauris velit dp dp_email dp_ukpassportnumber

Terms: New term for corporate finance Automated Teller Machine (ATM) Gross Profit

Properties Lineage Relationships Classifications Audits Schema

Additionally, you can also associate terms to your datasets by selecting one or more assets on the Data Catalog search page. You can associate terms with multiple datasets at a time.

CloudERA Data Catalog / Search

Search

Data Lakes

Type	Name	Location	Created On	Owner	Source
<input checked="" type="checkbox"/>	Hive Table	testmanaged	/default	Mon Mar 29 2021	hive
<input checked="" type="checkbox"/>	Hive Table	test100	/test	Mon Mar 29 2021	hive
<input checked="" type="checkbox"/>	Hive Table	test_tbl	/default	Mon Mar 29 2021	dp profiler
<input checked="" type="checkbox"/>	Hive Table	data	/default	Mon Mar 29 2021	dp profiler
<input checked="" type="checkbox"/>	Hive Table	testmanaged100	/test	Mon Mar 29 2021	hive
<input checked="" type="checkbox"/>	Hive Table	testtext	/default	Mon Mar 29 2021	hive

Action Ranger Atlas Create Dataset

Add Tag Add Term

CloudERA Data Catalog / Search

Search

Data Lakes

Type	Name	Location	Created On
<input checked="" type="checkbox"/>	Hive Table	testmanaged	Mon Mar 29 2021
<input checked="" type="checkbox"/>	Hive Table	test100	Mon Mar 29 2021
<input checked="" type="checkbox"/>	Hive Table	test_tbl	Mon Mar 29 2021
<input checked="" type="checkbox"/>	Hive Table	data	Mon Mar 29 2021
<input checked="" type="checkbox"/>	Hive Table	testmanaged100	Mon Mar 29 2021
<input checked="" type="checkbox"/>	Hive Table	testtext	Mon Mar 29 2021

Add Terms to 6 selected entities

Term	Glossary	
<input type="checkbox"/>	term -1	glossary - 2
<input type="checkbox"/>	Accounting Rate of Re...	Corporate Finance
<input type="checkbox"/>	New term for corporat...	Corporate Finance
<input type="checkbox"/>	Automated Teller Mac...	Corporate Finance
<input type="checkbox"/>	Compound Annual Gr...	Corporate Finance
<input type="checkbox"/>	Gross Profit	Corporate Finance
<input type="checkbox"/>	a very long term is bei...	Corporate Finance
<input type="checkbox"/>	Employee Stock Owne...	Corporate Finance

When you select a Hive table asset and navigate to the Asset Details page, under the Schema tab, you can view the list of terms associated with the asset.

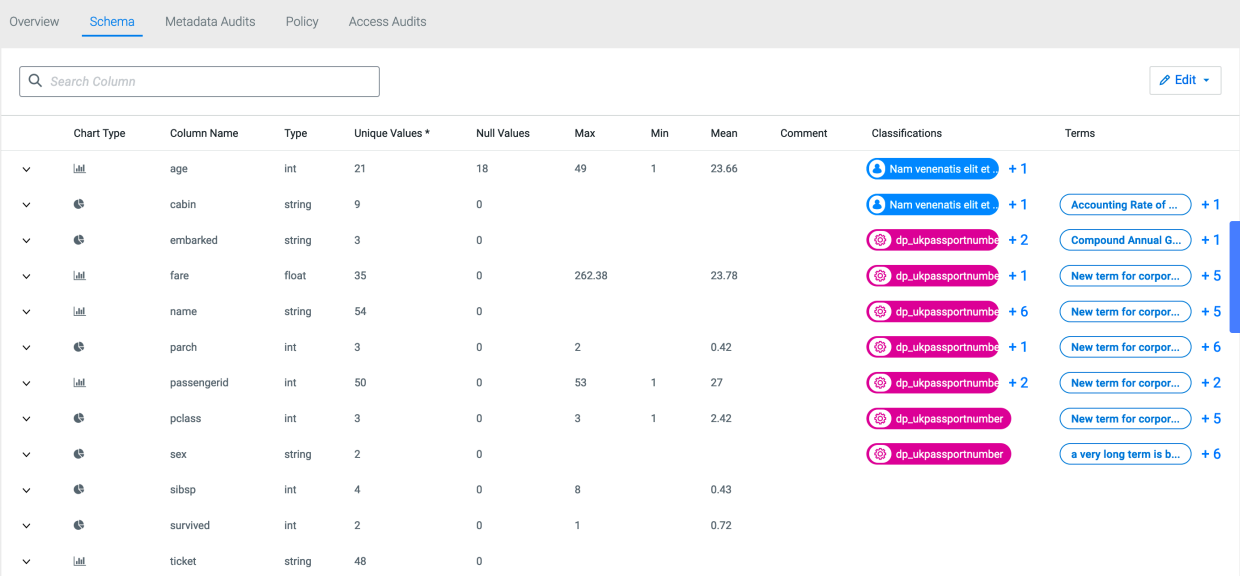


Chart Type	Column Name	Type	Unique Values *	Null Values	Max	Min	Mean	Comment	Classifications	Terms
▼	age	int	21	18	49	1	23.66		Nam venenatis elit et +1	
▼	cabin	string	9	0					Nam venenatis elit et +1	Accounting Rate of ... +1
▼	embarked	string	3	0					dp.ukpassportnumber +2	Compound Annual G... +1
▼	fare	float	35	0	262.38		23.78		dp.ukpassportnumber +1	New term for corpor... +5
▼	name	string	54	0					dp.ukpassportnumber +6	New term for corpor... +5
▼	parch	int	3	0	2		0.42		dp.ukpassportnumber +1	New term for corpor... +6
▼	passengerid	int	50	0	53	1	27		dp.ukpassportnumber +2	New term for corpor... +2
▼	pclass	int	3	0	3	1	2.42		dp.ukpassportnumber	New term for corpor... +5
▼	sex	string	2	0					dp.ukpassportnumber	a very long term is b... +6
▼	sibsp	int	4	0	8		0.43			
▼	survived	int	2	0	1		0.72			
▼	ticket	string	48	0						

Rows per page: 20 1 - 12 of 12

You can add or update the terms for the associated datasets by clicking the Edit button.

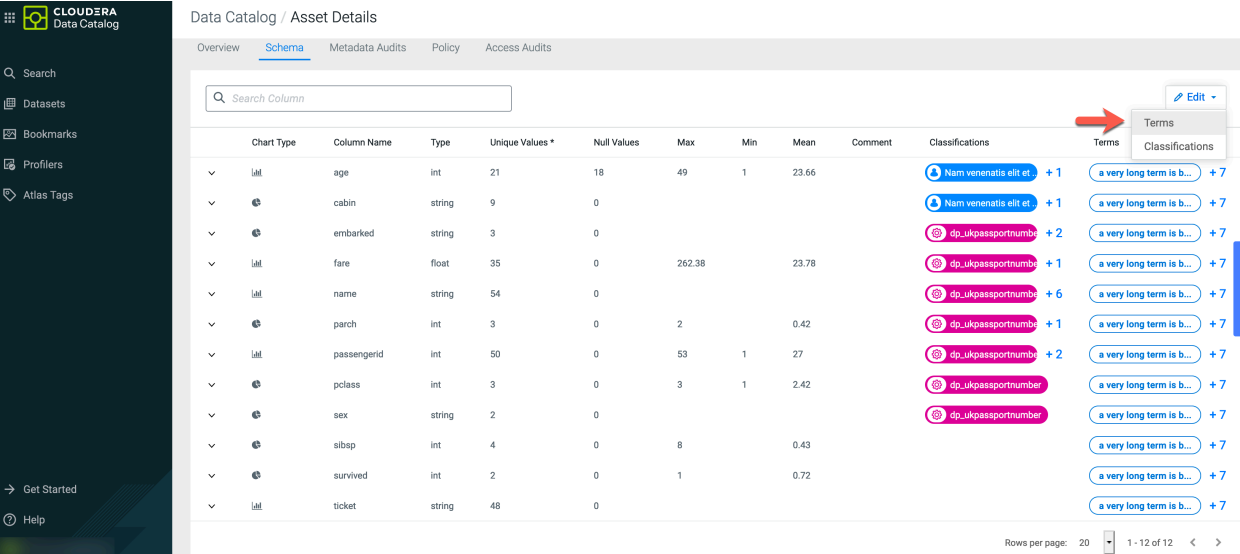



Chart Type	Column Name	Type	Unique Values *	Null Values	Max	Min	Mean	Comment	Classifications	Terms
▼	age	int	21	18	49	1	23.66		Nam venenatis elit et +1	a very long term is b... +7
▼	cabin	string	9	0					Nam venenatis elit et +1	a very long term is b... +7
▼	embarked	string	3	0					dp.ukpassportnumber +2	a very long term is b... +7
▼	fare	float	35	0	262.38		23.78		dp.ukpassportnumber +1	a very long term is b... +7
▼	name	string	54	0					dp.ukpassportnumber +6	a very long term is b... +7
▼	parch	int	3	0	2		0.42		dp.ukpassportnumber +1	a very long term is b... +7
▼	passengerid	int	50	0	53	1	27		dp.ukpassportnumber +2	a very long term is b... +7
▼	pclass	int	3	0	3	1	2.42		dp.ukpassportnumber	a very long term is b... +7
▼	sex	string	2	0					dp.ukpassportnumber	a very long term is b... +7
▼	sibsp	int	4	0	8		0.43			a very long term is b... +7
▼	survived	int	2	0	1		0.72			a very long term is b... +7
▼	ticket	string	48	0						a very long term is b... +7

Rows per page: 20 1 - 12 of 12

Searching for assets using glossary terms

You can search for the datasets using the Glossary terms filter available on the Data Catalog search page.






CloudERA


Data Catalog

 **Search**


 Datasets


 Bookmarks

 Profilers

 Atlas Tags

 Get Started

 Help



Data Catalog / Search

<input type="radio"/>	<div></div>	NA
<input type="radio"/>	<div></div>	NA
<input type="radio"/>	<div></div>	NA

Filters

TYPE

Clear ^

- ☐ Hive Table
- ☐ HBase Table
- [+ Add New Value](#)

OWNERS

Clear ^

- ☐ atlas
- ☐ dpprofiler
- ☐ hive
- ☐ public

ENTITY TAG

Clear ^

- [+ Add New Value](#)

GLOSSARY TERMS

Clear ^

- [+ Add New Value](#)

Additional search options for asset types

Using Data Catalog, you can add or edit asset description values to search for data assets across both Data Catalog and Atlas services by using the asset content.

In the Asset Details page for each asset type that you select, you can add or edit comment and description fields. For each asset type in Data Catalog, you can add or edit comments or include a description. Including these values for the selected asset helps you to identify your chosen asset when you perform the search operation.

Later, using the same set of values (comment or description), you can search for the asset types in Atlas.



Note: The comment and description options are supported only for Hive table and Hive Column assets. For other asset types, only the description option is supported.

Data Catalog / Asset Details

ww_customers

Atlas

Properties

Type: HIVE TABLE

of Columns: 40

Data Lake:

Datasets: 0

Owner: hive

Created On: Tue Mar 09 2021 10:48:45 GMT+0530 (India Stand...)

Last Access Time: Tue Mar 09 2021 10:48:45 GMT+0530 (Indi...

Table Type: EXTERNAL_TABLE

Database: hortoniabank

DB Catalog:

Parent: hortoniabank

Qualified Name

hortoniabank.ww_customers@cm

Comment

+ Add Comment

Description

+ Add Description

Profilers | 2

Cluster Sensitivity Profiler

Last run: 8 hours ago | Status: SUCCESS

Next Schedule Run: Thursday at 11:50 AM

Run

Hive Column Profiler

Last run: 8 hours ago | Status: SUCCESS

Next Schedule Run: Tomorrow at 5:30 PM

Run

Click + besides Comment and Description to include the respective values.

Data Catalog / Asset Details

ww_customers

Atlas

Properties

Cancel Save

Type: HIVE TABLE

of Columns: 40

Data Lake:

Datasets: 0

Owner: hive

Created On: Tue Mar 09 2021 10:48:45 GMT+0530 (India Stand...)

Last Access Time: Tue Mar 09 2021 10:48:45 GMT+0530 (Indi...

Table Type: EXTERNAL_TABLE

Database: hortoniabank

DB Catalog:

Parent: hortoniabank

Qualified Name

hortoniabank.ww_customers@cm

Comment

passport_number

Description

visa_number

Profilers | 2

Cluster Sensitivity Profiler

Last run: 9 hours ago | Status: SUCCESS

Next Schedule Run: Thursday at 11:50 AM

Run

Hive Column Profiler

Last run: 8 hours ago | Status: SUCCESS

Next Schedule Run: Tomorrow at 5:30 PM

Run

Click Save to save the changes.

Data Catalog / Asset Details

ww_customers

Atlas

Asset details were updated successfully.

Properties

Type: HIVE TABLE

of Columns: 40

Data Lake:

Datasets: 0

Owner: hive

Created On: Tue Mar 09 2021 10:48:45 GMT+0530 (India Stand...)

Last Access Time: Tue Mar 09 2021 10:48:45 GMT+0530 (Indi...

Table Type: EXTERNAL_TABLE

Database: hortoniabank

DB Catalog:

Parent: hortoniabank

Qualified Name

hortoniabank.ww_customers@cm

Comment

passport_number

Description

visa_number

Profilers | 2

Cluster Sensitivity Profiler

Last run: 9 hours ago | Status: SUCCESS

Next Schedule Run: Thursday at 11:50 AM

Run


Hive Column Profiler

Last run: 8 hours ago | Status: SUCCESS

Next Schedule Run: Tomorrow at 5:30 PM

Run




Note: You can also edit the already saved value by clicking the  icon.

Clicking on the Atlas button will navigate to the corresponding Atlas asset page as displayed.



ww_customers (hive_table)

Classifications: 

Terms: 

Properties Lineage Relationships Classifications Audits Schema

Technical properties

columns (40)

title
givenname
middleinitial

comment passport_number

createTime 03/09/2021 10:48:45 AM (IST)

db

hortoniabank

dcProfiledData

```
{
  samplePercent: "100.0",
  rowCount: 50000,
}
```

description visa_number

User-defined properties

Add

Labels

Add

Business Metadata

Add

[Switch to Beta UI](#)

Searching for assets in Data Catalog using additional search options

Consider a scenario in Data Catalog, where you select a data asset type and under the Asset Details page, you insert a comment and provide the description for the selected asset.

Navigate to the Data Catalog search query pane and enter the Comment and Description value(s) that you saved for the selected asset type in Data Catalog. The result page displays the asset type that you added for the Comment and Description fields in Data Catalog.

When you query for the entered Comment value for the selected asset type in Data Catalog, the relevant asset type is displayed in the search result page.

Data Catalog / Search

Visa Number

Atlas Ranger

Data Lakes

Asset Name	Type	Owner	Source
cloudera-kc3fw4 1	Hive Table	sys.role_map@cm	hive

Filters

TYPE

- ☒ Hive Table
- ☐ HBase Table

[Add New Value](#) [Clear](#)

OWNERS

- ☒ hive

[Clear](#)

DATABASE

- ☒ information_schema
- ☐ sys

Setup the Profiler for cloudera-kc3fw4 | 1

Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started](#)

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	role_map	sys.role_map@cm	Fri Aug 27 2021	hive	hive

[Action](#)

Clicking on the asset type in Data Catalog displays the comment and description values as it was assigned in Data Catalog.

Data Catalog / Asset Details

role_map

Atlas

Properties

Type: HIVE TABLE
 # of Columns: 8
 Data Lake:
 Datasets: 0
 Owner: hive
 Created On: Fri Aug 27 2021 10:51:33 GMT+0530 (India Standard Time)
 Last Access Time: Fri Aug 27 2021 10:51:33 GMT+0530 (India Standard Time)
 Table Type: EXTERNAL_TABLE
 Database: sys
 DB Catalog: cm
 Parent: sys

Qualified Name: sys.role_map@cm
 Comment: Visa Number
 Description: Passport Number

Classifications Managed System Propagated
 Add Classification

Terms
 Add Terms

Overview Schema Metadata Audits Policy Access Audits

Lineage Filter By: Depth: 3 Process Node: Hide

When you query for the entered Description value for the selected asset type in Data Catalog, the relevant asset type is displayed in the search result page.

Data Catalog / Search

Search: Passport Number

Atlas Ranger

Data Lakes

Filters

TYPE

☒ Hive Table
☐ HBase Table
 Add New Value Clear

OWNERS

☐ hive Clear

DATABASE

☐ information_schema
☐ sys

Setup the Profiler for
 Profiler runs data profiling operations as a pipeline on data located in the data lake which helps to view the profiled results for assets. Setting up a Profiler launches a cluster with various services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. [Get Started](#)

cloudera-kc3fw4 | 1 Action

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	role_map	sys.role_map@cm	Fri Aug 27 2021	hive	hive

Your search query displays the results.

Accessing Tables based on Ranger policies

In Data Catalog service, when a table (in blue color link) is clicked, the Asset Details view page is displayed.

If a user is not authorized to click or view table details, it implies that the user permissions have not been set-up in the Ranger.



Note: The user permissions to view table details are configured in Ranger.

Edit Policy

Last Response Time
10/31/2023 05:00:08 PM

Service Manager > cm_atlas Policies > Edit Policy

Policy Details

Policy Type

Access

ⓘ Add Validity Period

Policy ID*

81

Policy Name*

rs_test

ⓘ

Enabled

Normal

Policy Label

Select...

Entity Type

hive_table x hbase_table x

Include

Entity Classification*

rs_test x

Include

Entity ID*

x

Include

None

Select...

Description

Audit Logging*

Yes

Allow Conditions:

hide

Select Roles

Select...

Select Groups

Select...

Select Users

Select...

Permissions

Add Permissions

+

Delegate Admin

x

CLOUDERA
Data Catalog

Dashboard

Search

Datasets

Bookmarks

Profilers

Atlas Tags

Help

hrt_e2e_admin

1.5.2-b126

Asset Details

Authorization Unsuccessful. Contact Admin(Error Code - 403)

List properties for entity request failed Error from Atlas. Forbidden : Received 403. Body {"errorCode":"ATLAS-403-00-001","errorMessage":"hrt_e2e_admin is not authorized to perform read entity: guid=ce70cd53-891e-46d0-a43e-caf7f87927a9"}.

CREATED BEFORE









Clear

Last 1 day

Last 7 days

Last 15 days

Add New Value

Hive Table	/sys	Created	Tue Apr 07 2020	Owner	hive	Source	hive	
	scheduled_queries	/information_schema	Created	Tue Apr 07 2020	Owner	hive	Source	hive
Hive Table	schemata	/information_schema	Created	Tue Apr 07 2020	Owner	hive	Source	hive
	table_stats_view	/sys	Created	Tue Apr 07 2020	Owner	hive	Source	hive
	scheduled_executions	/information_schema	Created	Tue Apr 07 2020	Owner	hive	Source	hive
	andromeda	/-	Created	-	Owner	-	Source	hive
	milky	/-	Created	-	Owner	-	Source	hive
	bear	/-	Created	-	Owner	-	Source	hive
	n170	/-	Created	-	Owner	-	Source	hive
	umajor5	/-	Created	-	Owner	-	Source	hive

Creating Classification for selected assets

You can create a classification that can be associated with an asset.

1. From Data Catalog > navigate to the search page.
2. You can perform one or more of the following:
 - Select Add Classifications on action button in search page
 - Select Add classification in classification widget on Asset Details page.
3. On the Add Classification slider, click Create button.
4. Enter the necessary values in the fields and click the Create button.

Adding Classifications / Terms for selected assets

You can add classification or terms that can be associated with an asset.

Procedure

1. From Data Catalog > navigate to the search page.
2. You can perform one or more of the following:
 - a) Select Add Classifications / Terms on action button in the search page.
 - b) Select Add Classifications / Terms in classification widget on Asset Details page.
3. On the Add Classifications / terms slider, click on the Add icon against classification / term.
4. Enter other values in the fields, if required and click Save.

Additional Entity type selection for searching Assets

Using the Data Catalog service, you can search for assets by using the entity types.

Data Catalog users can search and discover assets of more types. Users can search assets of types just like they do for Hive Table with some restrictions.

Supported entity types include:

- HBase Table
- HBase Column Family
- HBase Namespace
- HDFS Path
- Hive DB
- Hive Table
- Hive Column
- ML Project
- ML Model Build
- ML Model Deployment
- NiFi Flow
- NiFi Data
- Impala Process
- Impala Column Lineage
- Impala Process Execution
- Kafka Topic
- RDBMS DB
- RDBMS Column
- RDBMS Foreign Key
- RDBMS Index
- RDBMS Instance
- RDBMS Table
- Spark Process
- Spark Application
- Spark Column
- Spark Column Lineage
- Spark DB
- Spark ML Directory
- Spark ML Model
- Spark ML Pipeline
- Spark Process Execution
- Spark Table

Selecting a type triggers a search query for that type. Currently two types of entities are supported but totally about twelve types of generic entities can be selected to search for assets depending on the data lake.

Owners data is derived from the response received from type based queries.

The following example diagrams depict the entity type selection search results.

Search

Search

Atlas Ranger

Data Lakes

157
66
NA

Filters

TYPE

☒ Hive Table

☐ HBase Table

Add New Value Clear

OWNERS

☐ csso_mhussain

☐ csso_santhosh

☐ hive

☐ hrt_1

☐ hrt_0a

Clear

Profiler Cluster is provisioned successfully

Action

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	global_privs	sys.global_privs@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	partition_key_vals	sys.partition_key_vals@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	partition_keys	sys.partition_keys@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	tbls	sys.tbls@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	sort_cols	sys.sort_cols@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	skewed_string_list_values	sys.skewed_string_list_values@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	skewed_values	sys.skewed_values@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	compaction_queue	sys.compaction_queue@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	key_constraints	sys.key_constraints@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	wm_mappings	sys.wm_mappings@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	wm_resourceplans	sys.wm_resourceplans@cm	Mon Oct 04 2021	hive	hive
<input type="checkbox"/> Hive Table	wm_triggers	sys.wm_triggers@cm	Mon Oct 04 2021	hive	hive

Search

Search

Atlas Ranger

Data Lakes

5
672
NA

Filters

TYPE

☐ Hive Table

☒ HBase Table

Add New Value Clear

OWNERS

☐ atlas

☐ hbase

Clear

NAMESPACE

☐ hbase

Clear

Profiler Cluster is provisioned successfully

Action

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> HBase Table	atlas_janus	default:atlas_janus@cm	Mon Oct 04 2021	atlas	hbase
<input type="checkbox"/> HBase Table	t1	default:t1@cm	Tue Jun 09 2020	hbase	hbase
<input type="checkbox"/> HBase Table	t2	default:t2@cm	Tue Jun 09 2020	hbase	hbase
<input type="checkbox"/> HBase Table	ATLAS_ENTITY_AUDIT_EVENTS	default:ATLAS_ENTITY_AUDIT_EVENTS@cm	Fri Jun 05 2020	atlas	hbase
<input type="checkbox"/> HBase Table	hbase:acl	hbase:acl@cm	Tue Jun 09 2020	hbase	hbase

Viewing Data Asset Details

The Asset Details page comprises four to five tabs (Overview, Schema, Metadata Audit, Policy, and Access Audits).

To access the Asset Details page, click an asset in the Data Catalog Search page. This brings you to the Overview tab, the first of the four tabs that form the Asset Details page.

- Asset properties: Number of rows, number of columns, number of partitions, and owner.
- Overview: Displays an overview for the data asset.
 - Lineage: Shows the chain of custody for the data from relevant metadata repositories such as Apache Atlas. Lineage shows both upstream paths (lineage) into and downstream paths (impact) out of a given asset.

- Schema: Displays the schema of the data asset for structured data (such as Hive tables) from the relevant metadata repositories (such as Atlas).
- Metadata Audits: Displays the change logs per asset fetched from Apache Atlas.
- Content: Only visible for container types like databases. When the tab is visible, the overview tab is not applicable.
- Policy: The policy view shows security (authorization) policies defined on assets such as those present in Apache Ranger. It includes both resource (physical asset based) as well as classification based policies
- Access Audits: The data asset audit logs page shows the most recent access audits from Apache Ranger.

Viewing Data Assets

The Data Asset Overview page displays all the Apache Atlas metadata associated with a particular data asset.

About this task

The Data Asset Overview page displays:

Asset properties: Displays properties information relevant to asset type, like in case of Hive table - Number of rows, number of columns, number of partitions, and the owner.

From the Data Catalog search page, click to select a data asset.

The Asset Overview window opens.

The following matrix captures the supported fields for different asset types:

Asset Type	Lineage	Tagging	Access Metrics	Schema	Policy	Audit	Atlas Punch out
Hive DB	Not Supported	Yes	Not Supported	Not Supported	Yes	Yes	Yes
Hive Table	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hive Column	Yes	Yes	Not Supported	Not Supported	Yes	Yes	Yes
Hbase Namespace	Yes	Yes	Not Supported	Not Supported	Yes	Yes	Yes
Hbase Table	Yes	Yes	Not Supported	Yes	Yes	Yes	Yes
Hbase Column Family	Yes	Yes	Not Supported	Not Supported	Yes	Yes	Yes
impala_process	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
impala_column_lineage	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
impala_process_execution	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
ML_Project	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
ML_Model_Build	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
ML_Model_Deploy	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_db	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes

Asset Type	Lineage	Tagging	Access Metrics	Schema	Policy	Audit	Atlas Punch out
rdbms_column	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_foreign_key	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_index	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_instance	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_table	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_process	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_application	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_column	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_column_lineage	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_db	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_ml_directory	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_ml_model	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_ml_pipeline	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_process_execution	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_table	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes

View Data Asset Schema

From the Asset Details Schema page, you can view the schema of the data asset for structured data (such as Hive tables) from the relevant metadata repositories (such as Atlas).

Procedure

1. From the Data Catalog search page, select an asset.
The Asset Overview window opens.
2. Click Schema.
The Schema table shows the data asset schema as retrieved from Apache Atlas.
3. (Optional) To edit tags:
 - a) Click Edit Tags.
 - b) Click the (+) icon.
 - c) Select or deselect the tags you choose, then click Save.

You can now manage and edit tags at the table level.

Navigating from the container asset to the parent asset from Asset Details page

A generic Assets Details page is available for container data types like buckets and databases.

A Contents tab (similar to the Schema tab) lists all the contents of the selected entity. Clicking on any element available in the selected entity list navigates you to the Asset Details page.

Data Catalog / Asset Details

The screenshot shows the 'default' asset details page. The 'Properties' section includes: Type: HBASE NAMESPACE, Data Lake: [redacted], Owner: atlas, Created On: -NA-, Update Time: -NA-, Created By: atlas, Updated By: [redacted], Status: ACTIVE. The 'Qualified Name' is default@cm and the 'Description' is default. The 'Classifications' section shows 'Managed' as selected. The 'Terms' section shows 'new_term1@new_glossary'. The 'Contents' tab is active, showing a table with columns: Type, Name, Location, Created On, Owner, Source. The table contains one entry: Type: hbase_table, Name: ATLAS_ENTITY_AUDIT_EVENTS, Location: /default, Created On: Wed Apr 07 2021, Owner: atlas, Source: hbase.

For example, for a database entity having a list of tables, clicking on any listed table navigates to the Asset Details page of the same table. Helps you understand the parent-child relationship as far as asset management is concerned. The Contents tab displays entities that are contained within assets of container entity types. The entities in the table of Contents tab are clickable, which will allow you to navigate to the Asset Details page of these contained assets.

The following table lists the entity types, their parent, and contents.

Type	Parent	Content
Hive DB	-	Hive Table
HBase Namespace	-	HBase Table
ML Project	-	ML Model Build
ML Model Build	ML Project	ML Model Deployment
RDBMS Instance	-	RDBMS DB
RDBMS DB	RDBMS Instance	RDBMS Table

View Authorization Policies on a Data Asset

The Asset Details Policy page displays all the Apache Ranger policy details associated with a particular data asset. This helps you understand how data access is secured and protected: what users can see what data (or metadata) under what conditions (security policies, data protection, and anonymization).

Procedure

1. From the Data Catalog search page, select a data asset.
The Asset Overview window opens.

2. Click the Policy tab.

The Policy table shows the data asset policies as retrieved from Apache Ranger.

View Data Asset Audit Logs

The Asset Details Audit page displays all the Apache Ranger audit events associated with a particular data asset. This helps you to view who has accessed what data from a forensic audit or compliance perspective, and to visualize access patterns and identify anomalies.

Procedure

1. From the Data Catalog search page, select a data asset.

The Asset Overview window opens.

2. Click the Audit tab.

The Audit table shows the most recent raw audit event data by type of access and access outcome (authorized/unauthorized).

3. (Optional) You can filter the audit results by Access Type or Result.

Access type: SELECT, UPDATE, CREATE, DROP, ALTER, INDEX, READ, WRITE.

Result: ALLOWED, DENIED.


Navigation Support for Hive entity within Lineage

When you click a Hive entity within lineage, the Asset Details page of the selected Hive entity is displayed.

Previously, when you clicked on any entity for which slider information was available, a slider would display the entity details. As of now, as seen in the corresponding images, the Asset Details page of the Hive entity is displayed. The option selected in Depth drop-down and Show Process nodes are now displayed on the upper-left corner of the Lineage module.

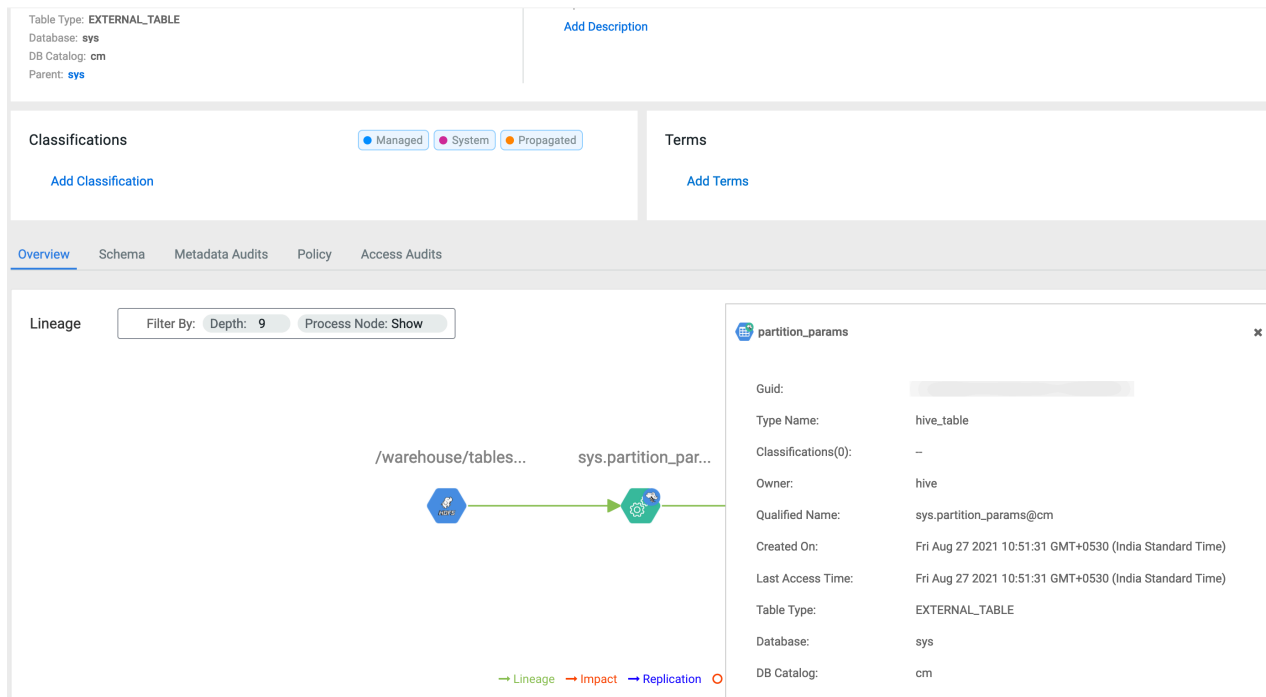
Data Catalog / Asset Details

The screenshot displays the 'Asset Details' page for a Hive entity. The top section contains metadata: Data Lake: cloudera-kc3fw4, Datasets: 0, Owner: hive, Created On: Fri Aug 27 2021 10:51:33 GMT+0530 (India Standard Time), Last Access Time: Fri Aug 27 2021 10:51:33 GMT+0530 (India Standard Time), Table Type: EXTERNAL_TABLE, Database: sys, DB Catalog: cm, Parent: sys. There are links for 'Add Comment' and 'Add Description'. Below this, there are tabs for 'Classifications' (Managed, System, Propagated) and 'Terms' (Add Terms). A navigation bar shows 'Overview', 'Schema', 'Metadata Audits', 'Policy', and 'Access Audits'. The 'Lineage' section has filters for 'Filter By: Depth: 3' and 'Process Node: Hide'. The lineage diagram shows a flow from '/warehouse/tables...' to 'role_map'.

Alternatively, if you do not want to navigate away from the current page and want to view the information with respect to any entity, hover on the entity and click the information icon  to view the details.

The screenshot depicts the slider information for the clicked entity:

Data Catalog / Asset Details

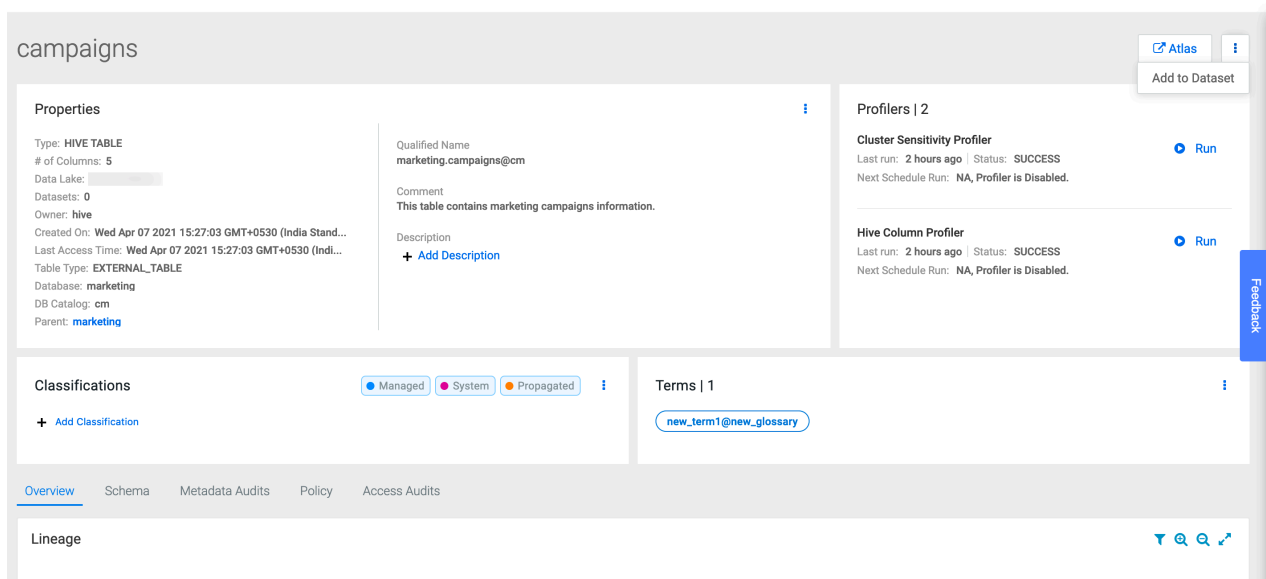


The screenshot shows the 'Asset Details' page for the entity 'sys.partition_params'. The top section includes metadata: Table Type: EXTERNAL_TABLE, Database: sys, DB Catalog: cm, and Parent: sys. Below this are tabs for 'Classifications' (Managed, System, Propagated) and 'Terms'. The main content area is divided into 'Overview' (selected), 'Schema', 'Metadata Audits', 'Policy', and 'Access Audits'. The 'Overview' tab displays a lineage diagram showing data flow from '/warehouse/tables...' to 'sys.partition_params'. A right-hand panel provides detailed information for 'partition_params', including Guid, Type Name (hive_table), Owner (hive), Qualified Name (sys.partition_params@cm), Created On, Last Access Time, Table Type (EXTERNAL_TABLE), Database (sys), and DB Catalog (cm).

Adding Hive asset to one or more datasets on Asset Details screen

On the Asset Details screen, users are provided with an option to add the asset to the dataset as shown in the diagram.

Data Catalog / Asset Details



The screenshot shows the 'Asset Details' page for the entity 'marketing.campaigns@cm'. The top section includes metadata: Type: HIVE TABLE, # of Columns: 5, Data Lake: , Datasets: 0, Owner: hive, Created On, Last Access Time, Table Type: EXTERNAL_TABLE, Database: marketing, DB Catalog: cm, and Parent: marketing. Below this are tabs for 'Classifications' (Managed, System, Propagated) and 'Terms'. The main content area is divided into 'Overview' (selected), 'Schema', 'Metadata Audits', 'Policy', and 'Access Audits'. The 'Overview' tab displays a 'Properties' section with a 'Qualified Name' and a 'Description'. A right-hand panel shows 'Profiles | 2', including 'Cluster Sensitivity Profiler' and 'Hive Column Profiler'. A 'Feedback' button is visible on the right side. The bottom section shows a 'Lineage' tab.

The Add to Dataset window provides an option to add the asset into one or more existing datasets or even create a new one.

Datasets that already contain the asset are disabled and marked as checked. Datasets which are currently in edit state are disabled and marked with a *.

The screenshot shows the 'campaigns' asset details page. The 'Add to Dataset' modal is open, displaying a table of existing datasets. The modal has search filters for 'Name' and 'Tag', and a '+ New Dataset' button. The table lists datasets with columns for Name, Owner, and # of Assets. Some datasets are checked, indicating they already contain the asset.

<input type="checkbox"/>	Name	Owner	# of Assets
<input checked="" type="checkbox"/>	Credit card	Srinivas Sudhindra	1
<input type="checkbox"/>	Town1234	Srinivas Sudhindra	1
<input checked="" type="checkbox"/>	City	Srinivas Sudhindra	1
<input type="checkbox"/>	Visa	Srinivas Sudhindra	1
<input checked="" type="checkbox"/>	SSN1	Srinivas Sudhindra	1
<input type="checkbox"/>	Passport123	Srinivas Sudhindra	1

Buttons at the bottom of the modal: Add Asset, Cancel.

Users can search for an existing dataset by name or by tags applied on the dataset. Users can select one or more datasets from the list and then click on the Add Asset button which adds the asset to these dataset(s).

The screenshot shows the 'warehouse' asset details page. The 'Add to Dataset' modal is open, displaying the 'New Dataset' form. The form has fields for Name, Description, and Tags. There is a 'Public' toggle switch and 'Create' and 'Cancel' buttons.

Buttons at the bottom of the modal: Create, Cancel.

There are instances, where there are no datasets present or the user just wants to create a new dataset to add the asset. In that case, the user can click on the New Dataset button which opens up a new dataset form. Once the user fills in the form and clicks on the Create button, a new dataset with the given properties is created and the asset is added to it automatically. This is reflected in the datasets list where the newly added dataset is highlighted.

Data Catalog / Asset Details

Add asset to datasets completed

warehouse

Properties

Type: HIVE TABLE
 # of Columns: 14
 Data Lake:
 Datasets: 6
 Owner:
 Created On: Mon May 11 2020 13:02:22 GMT+0530 (India Stan...
 Last Access Time: Mon May 11 2020 13:02:22 GMT+0530 (Indi...
 Table Type: EXTERNAL_TABLE
 Database:
 DB Catalog: cm
 Parent:
 Qualified Name:
 Comment:
 + Add Comment
 Description:
 + Add Description

Classifications

+ Add Classification

Overview Schema Metadata Audits Policy Access Audits

Lineage

Add to Dataset

Search by Name Search by Tag + New Dataset

<input type="checkbox"/>	Name↑	Owner↑	# of Assets↑
<input checked="" type="checkbox"/>	Link test	Srinivas Sudhindra	1
<input checked="" type="checkbox"/>	Demotest	Srinivas Sudhindra	1
<input checked="" type="checkbox"/>	Social Security Number	Srinivas Sudhindra	1
<input checked="" type="checkbox"/>	Visa number	Srinivas Sudhindra	1
<input type="checkbox"/>	passport12	Srinivas Sudhindra	0
<input type="checkbox"/>	Demo1234	Srinivas Sudhindra	0
<input checked="" type="checkbox"/>	Test1357	Srinivas Sudhindra	2
<input type="checkbox"/>	Demo	Srinivas Sudhindra	1
<input checked="" type="checkbox"/>	Hive123	Srinivas Sudhindra	2

Feedback

Viewing Atlas Entity Audits

In Data Catalog, Atlas audits help Data Stewards to identify and track the entity changes or modifications that are performed over a period of time.

Information about the Atlas entity audit events are displayed for each entity in the Asset Details page in Data Catalog. Using this information, Data Stewards can distinguish between entity audits and data audits that emanate from Ranger.

On the Asset Details page, a new tab called Metadata Audits displays information related to the selected entity type and about the events that occurred based on the user activities.

Overview Schema **Metadata Audits** Policy Access Audits

Clicking on Metadata Audits, tab, you can view manage information about:

- The user who made the changes to the specific entity
- The time when the entity was changed
- The kind of change that was made to the entity
- Any other relevant changes pertaining to the audit entries

The changes that can be identified for:

- Created entities and related updates
- Tagged entities
- Labeled entities
- Export and Import operations

For example, the following image displays information about the Atlas audit events that are performed by each Atlas user that is displayed in the Asset Details page in Data Catalog.

ATLAS_ENTITY_AUDIT_EVENTS

Atlas

Properties

Type: HBASE TABLE

Data Lake:

Owner: atlas

Created On: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)

Modified Time: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)

Namespace GUID: f4658406-3a4b-4076-afb0-6e74745934b6

URI: ATLAS_ENTITY_AUDIT_EVENTS

Parent: default

Qualified Name

default:ATLAS_ENTITY_AUDIT_EVE ...

Description

ATLAS_ENTITY_AUDIT_EVENTS

Classifications

Managed

System

Propagated

+ Add Classification

Terms

+ Add Terms

Overview

Schema

Metadata Audits

Policy

Access Audits

User

Timestamp

Actions

atlas

Wed Apr 07 2021 10:50:48 GMT+0530 (India Standard Time)

Entity created

Clicking on any line item displays the JSON format, which is directly derived from Atlas, in other words the source of data available in Atlas.

ATLAS_ENTITY_AUDIT_EVENTS

Atlas

Properties

Type: HBASE TABLE

Data Lake:

Owner: atlas

Created On: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)

Modified Time: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)

Namespace GUID: f4658406-3a4b-4076-afb0-6e74745934b6

URI: ATLAS_ENTITY_AUDIT_EVENTS

Parent: default

Qualified Name

default:ATLAS_ENTITY_AUDIT_EVE ...

Description

ATLAS_ENTITY_AUDIT_EVENTS

Classifications

Managed

System

Propagated

+ Add Classification

Terms

+ Add Terms

Overview

Schema

Metadata Audits

Policy

Access Audits

User

Timestamp

Actions

atlas

Wed Apr 07 2021 10:50:48 GMT+0530 (India Standard Time)

Entity created

Created:

```
{
  "typeName": "hbase_table",
  "attributes": {
    "owner": "atlas",
    "isNormalizationEnabled": false,

```

Use the toggle icon (on the top-right corner) for viewing Atlas Audits in different formats. By default, you can view Metadata Audits in tabular format in the Asset Details page and when you toggle the view icon, you can view the Timeline format. The events are listed as timelines in this format.

Data Catalog / Asset Details

ATLAS_ENTITY_AUDIT_EVENTS [Atlas](#)

Properties

Type: HBASE TABLE
 Data Lake:
 Owner: atlas
 Created On: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)
 Modified Time: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)
 Namespace GUID: f4658406-3a4b-4076-afb0-6e74745934b6
 URI: ATLAS_ENTITY_AUDIT_EVENTS
 Parent: default

Qualified Name
 default:ATLAS_ENTITY_AUDIT_EVE ...

Description
 ATLAS_ENTITY_AUDIT_EVENTS

Classifications Managed System Propagated
 + Add Classification

Terms
 + Add Terms

Overview Schema **Metadata Audits** Policy Access Audits

Timeline

Wed Apr 07 2021 10:50:48 GMT+0530 (India Standard Time) **atlas**
 Entity created

Feedback

Clicking on a user in the Timeline format displays the JSON data, which is again derived from Atlas.

Data Catalog / Asset Details

ATLAS_ENTITY_AUDIT_EVENTS [Atlas](#)

Properties

Type: HBASE TABLE
 Data Lake:
 Owner: atlas
 Created On: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)
 Modified Time: Wed Apr 07 2021 10:50:42 GMT+0530 (India Standard Time)
 Namespace GUID: f4658406-3a4b-4076-afb0-6e74745934b6
 URI: ATLAS_ENTITY_AUDIT_EVENTS
 Parent: default

Qualified Name
 default:ATLAS_ENTITY_AUDIT_EVE ...

Description
 ATLAS_ENTITY_AUDIT_EVENTS

Classifications Managed System Propagated
 + Add Classification

Terms
 + Add Terms

Overview Schema **Metadata Audits** Policy Access Audits

Timeline

Wed Apr 07 2021 10:50:48 GMT+0530 (India Standard Time) **atlas**
 Entity created

Details

atlas

Snapshot

Created:

```
{
  "typeName": "hbase_table",
  "attributes": {
    "owner": "atlas",
    ...
  }
}
```

Feedback

Managing Profilers

Kubernetes enables profiler job scheduling and runs profiler jobs on-demand and on schedule.

A service called Profiler Launcher Service (PLS) is made available to launch the Data Catalog profiler. The PLS is deployed in the Control Plane during the stack installation and the Management Console application (DC-API) makes an HTTP call to schedule the jobs. PLS is authorized to schedule and run Kubernetes jobs in the targeted cluster. You must install a PLS service in each Kubernetes / OCP cluster and a single control plane application to manage all the profiler jobs.



Note: You must be a **PowerUser** to launch the Profiler.

Table 1: List of built-in profilers

Profiler Name	Description
Cluster Sensitivity Profiler	A sensitive data profiler- PII, PCI, HIPAA, etc.
Ranger Audit Profiler	A Ranger audit log summarizer.
Hive Column Profiler	Provides summary statistics like Maximum, Minimum, Mean, Unique, and Null values at the Hive column level.

Scheduling Profiler jobs on your Private Cloud cluster

You must schedule the profilers to view the profiler results for your assets and datasets. You must be a **PowerUser** to schedule these jobs.

Profiler launcher service is installed at the setting up of the cluster. Later, you can schedule or run jobs on demand from the Data Catalog UI.

You must first note the following scenarios when working with profilers:

- Your profiler is not launched for the selected data lake
- Your profiler is already launched for the selected data lake.

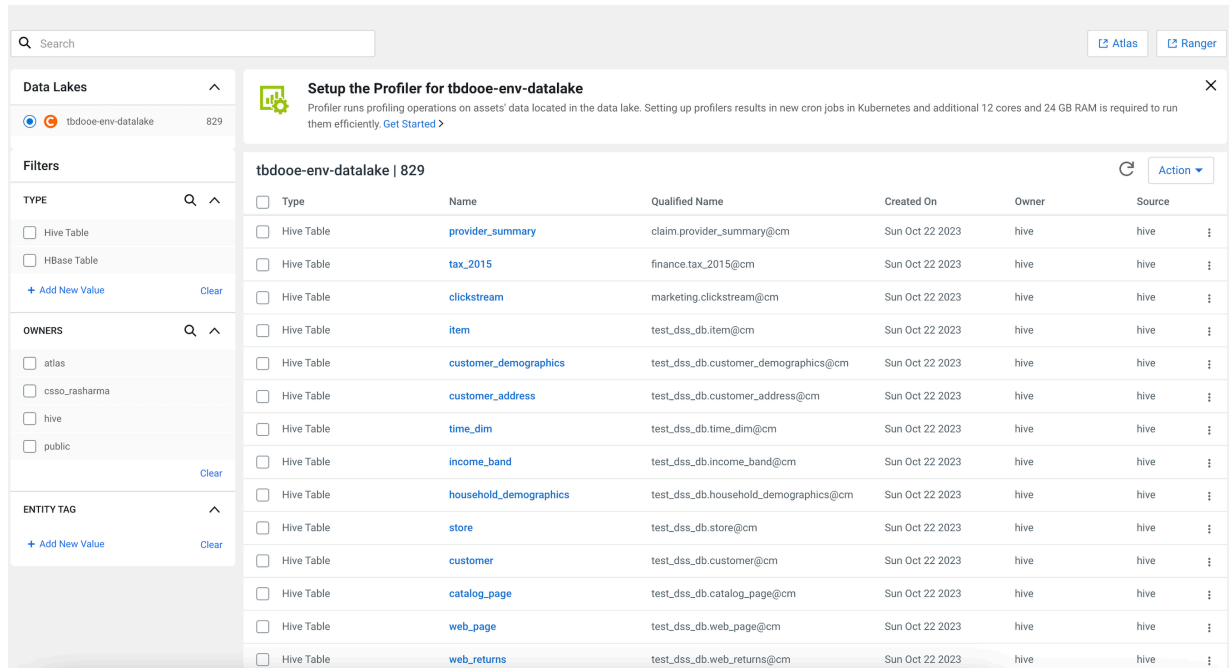
If your profiler is not launched for the selected data lake, follow these steps:

1. From the Search menu > select the data lake for which you want to profile the data.

The setup profiler page is displayed for the data lake.

2. Click the Get Started link.

Search

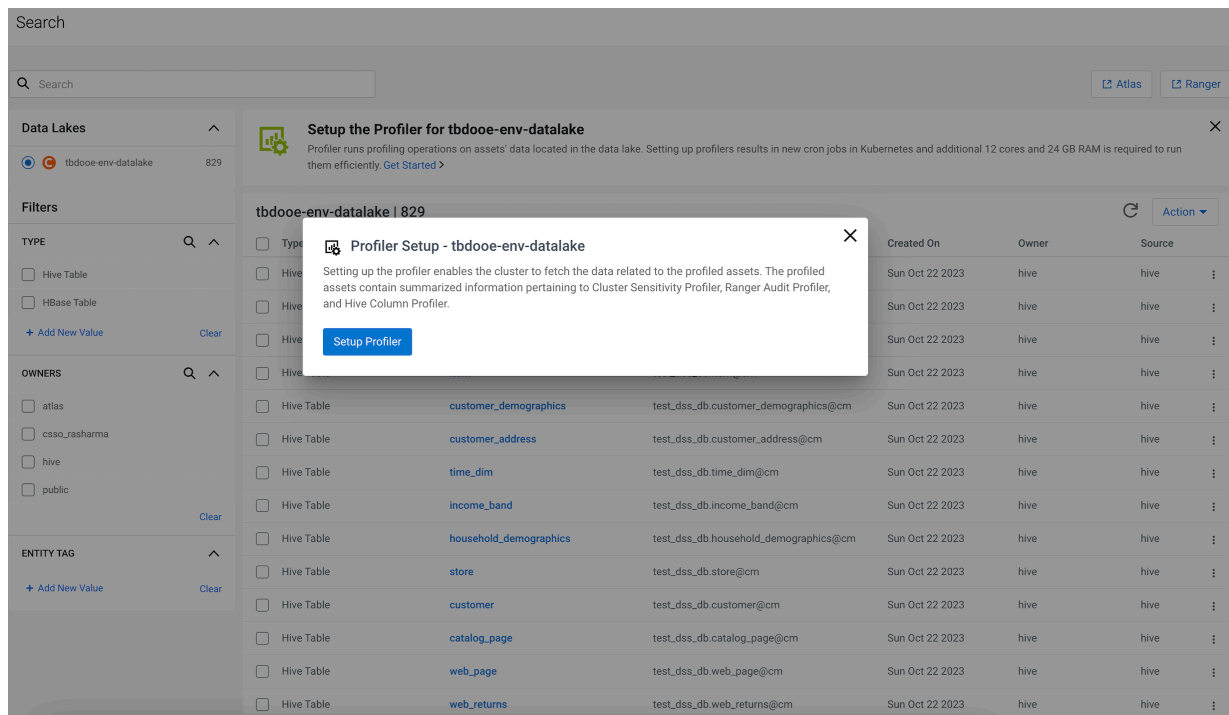


Setup the Profiler for tbdooe-env-datalake

Profiler runs profiling operations on assets' data located in the data lake. Setting up profilers results in new cron jobs in Kubernetes and additional 12 cores and 24 GB RAM is required to run them efficiently. [Get Started >](#)

Type	Name	Qualified Name	Created On	Owner	Source
Hive Table	provider_summary	claim.provider_summary@cm	Sun Oct 22 2023	hive	hive
Hive Table	tax_2015	finance.tax_2015@cm	Sun Oct 22 2023	hive	hive
Hive Table	clickstream	marketing.clickstream@cm	Sun Oct 22 2023	hive	hive
Hive Table	item	test_dss_db.item@cm	Sun Oct 22 2023	hive	hive
Hive Table	customer_demographics	test_dss_db.customer_demographics@cm	Sun Oct 22 2023	hive	hive
Hive Table	customer_address	test_dss_db.customer_address@cm	Sun Oct 22 2023	hive	hive
Hive Table	time_dim	test_dss_db.time_dim@cm	Sun Oct 22 2023	hive	hive
Hive Table	income_band	test_dss_db.income_band@cm	Sun Oct 22 2023	hive	hive
Hive Table	household_demographics	test_dss_db.household_demographics@cm	Sun Oct 22 2023	hive	hive
Hive Table	store	test_dss_db.store@cm	Sun Oct 22 2023	hive	hive
Hive Table	customer	test_dss_db.customer@cm	Sun Oct 22 2023	hive	hive
Hive Table	catalog_page	test_dss_db.catalog_page@cm	Sun Oct 22 2023	hive	hive
Hive Table	web_page	test_dss_db.web_page@cm	Sun Oct 22 2023	hive	hive
Hive Table	web_returns	test_dss_db.web_returns@cm	Sun Oct 22 2023	hive	hive

The Profiler setup for the data lake window appears.



Setup the Profiler for tbdooe-env-datalake

Profiler runs profiling operations on assets' data located in the data lake. Setting up profilers results in new cron jobs in Kubernetes and additional 12 cores and 24 GB RAM is required to run them efficiently. [Get Started >](#)

Type	Name	Qualified Name	Created On	Owner	Source
Hive Table	customer_demographics	test_dss_db.customer_demographics@cm	Sun Oct 22 2023	hive	hive
Hive Table	customer_address	test_dss_db.customer_address@cm	Sun Oct 22 2023	hive	hive
Hive Table	time_dim	test_dss_db.time_dim@cm	Sun Oct 22 2023	hive	hive
Hive Table	income_band	test_dss_db.income_band@cm	Sun Oct 22 2023	hive	hive
Hive Table	household_demographics	test_dss_db.household_demographics@cm	Sun Oct 22 2023	hive	hive
Hive Table	store	test_dss_db.store@cm	Sun Oct 22 2023	hive	hive
Hive Table	customer	test_dss_db.customer@cm	Sun Oct 22 2023	hive	hive
Hive Table	catalog_page	test_dss_db.catalog_page@cm	Sun Oct 22 2023	hive	hive
Hive Table	web_page	test_dss_db.web_page@cm	Sun Oct 22 2023	hive	hive
Hive Table	web_returns	test_dss_db.web_returns@cm	Sun Oct 22 2023	hive	hive

3. Click Setup Profiler button to launch the profiler

The High Availability (HA) feature for profilers, including launching and managing jobs are supported by default. No separate action is required to enable the HA functionality or its components.



Note: Once you schedule the profiler jobs, navigate to the Profilers page to view the status of the respective profiler jobs.

Launching profilers using Command-line

Data Catalog now supports launching Data profilers using the Command-Line Interface (CLI) option.

This, apart from launching the profilers using the Data Catalog UI. The CLI will be one executable and will not have any external dependencies. You can execute some operations in the Data Catalog service using the CDP CLI commands.

Users must have valid permission(s) to launch profilers on a data lake.

For more information about the access details, see [Prerequisites to access Data Catalog service](#).

You must have the following entitlement granted to use this feature:

DATA_CATALOG_ENABLE_API_SERVICE

In your CDP CLI environment, enter the following command to get started in the CLI mode.

```
cdp datacatalog --help
```

This command provides information about the available commands in Data Catalog.

The output is displayed as:

NAME

datacatalog

DESCRIPTION

Cloudera Data Catalog Service is a web service, using this service user can execute operations like launching profilers in Data Catalog.

AVAILABLE SUBCOMMANDS

launch-profilers

You get additional information about this command by using:

```
cdp datacatalog launch-profilers --help
```

NAME

launch-profilers -

DESCRIPTION

Launches DataCatalog profilers in a given datalake.

SYNOPSIS

launch-profilers

--datalake <value>

[--cli-input-json <value>]

[--generate-cli-skeleton]

OPTIONS

--datalake (string) The Name or CRN of the Datalake.

```
--cli-input-json (string) Performs service operation based on the JSON string provided. The JSON string follows the format provided by --generate-cli-s
```

keleton. If other arguments are provided on the command line, the CLI values will override the JSON-provided values.

`--generate-cli-skeleton` (boolean) Prints a sample input JSON to standard output. Note the specified operation is not run if this argument is specified. The sample input can be used as an argument for `--cli-input-json`.

OUTPUT

`datahubCluster` -> (object)

Information about a cluster.

`clusterName` -> (string)

The name of the cluster.

`crn` -> (string)

The CRN of the cluster.

`creationDate` -> (datetime)

The date when the cluster was created.

`clusterStatus` -> (string)

The status of the cluster.

`nodeCount` -> (integer)

The cluster node count.

`workloadType` -> (string)

The workload type for the cluster.

`cloudPlatform` -> (string)

The cloud platform.

`imageDetails` -> (object)

The details of the image used for cluster instances.

`name` -> (string)

The name of the image used for cluster instances.

`id` -> (string)

The ID of the image used for cluster instances.

This is internally generated by the cloud provider to Uniquely identify the image.

`catalogUrl` -> (string)

The image catalog URL.

catalogName -> (string)

The image catalog name.

environmentCrn -> (string)

The CRN of the environment.

credentialCrn -> (string)

The CRN of the credential.

datalakeCrn -> (string)

The CRN of the attached datalake.

clusterTemplateCrn -> (string)

The CRN of the cluster template used for the cluster

creation.

You can use the following CLI command to launch the Data profiler:

```
cdp datacatalog launch-profilers --datalake <datalake name or datalake CRN>
```

Example

```
cdp datacatalog launch-profilers --datalake test-env-ycloud
```

```
{
```

```
"datahubCluster": {
```

```
"clusterName": "cdp-dc-profilers-24835599",
```

```
  "crn":
```

```
    "crn:cdp:datahub:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:cluster:dfaa7646-d77f-4099-a3ac-6628e1576160",
```

```
"creationDate": "2021-06-04T11:31:23.735000+00:00",
```

```
"clusterStatus": "REQUESTED",
```

```
"nodeCount": 3,
```

```
"workloadType": "v6-cdp-datacatalog-profiler_7_2_8-1",
```

```
"cloudPlatform": "YARN",
```

```
"imageDetails": {
```

```
  "name":
```

```
    "docker-sandbox.infra.cloudera.com/cloudbreak/centos-76:2020-05-18-17-16-16",
```

```
"id": "d558405b-b8ba-4425-94cc-a8baff9ffb2c",
```

```
  "catalogUrl":
```

```
    "https://cloudbreak-imagecatalog.s3.amazonaws.com/v3-test-cb-image-catalog.json",
```

```
"catalogName": "cdp-default"
```

```
},
```

```
    "environmentCrn":
      "crn:cdp:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:environment:bf795226-b57c-4c4d-8520-82249e57a54f",
```

```
    "credentialCrn":
      "crn:altus:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:credential:3adc8ddf-9ff9-44c9-bc47-1587db19f539",
```

```
    "datalakeCrn":
      "crn:cdp:datalake:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:datalake:5e6471cf-7cb8-42cf-bda4-61d419cfbc53",
```

```
    "clusterTemplateCrn":
      "crn:cdp:datahub:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:cluster-template:16a5d8bd-66d3-42ea-8e8d-bd8765873572"
```

```
}
```

```
}
```

Deleting profilers

Deleting profiler container (pod) jobs removes all the Custom Sensitivity Profiler rules and other updates to the specified profiler.

About this task

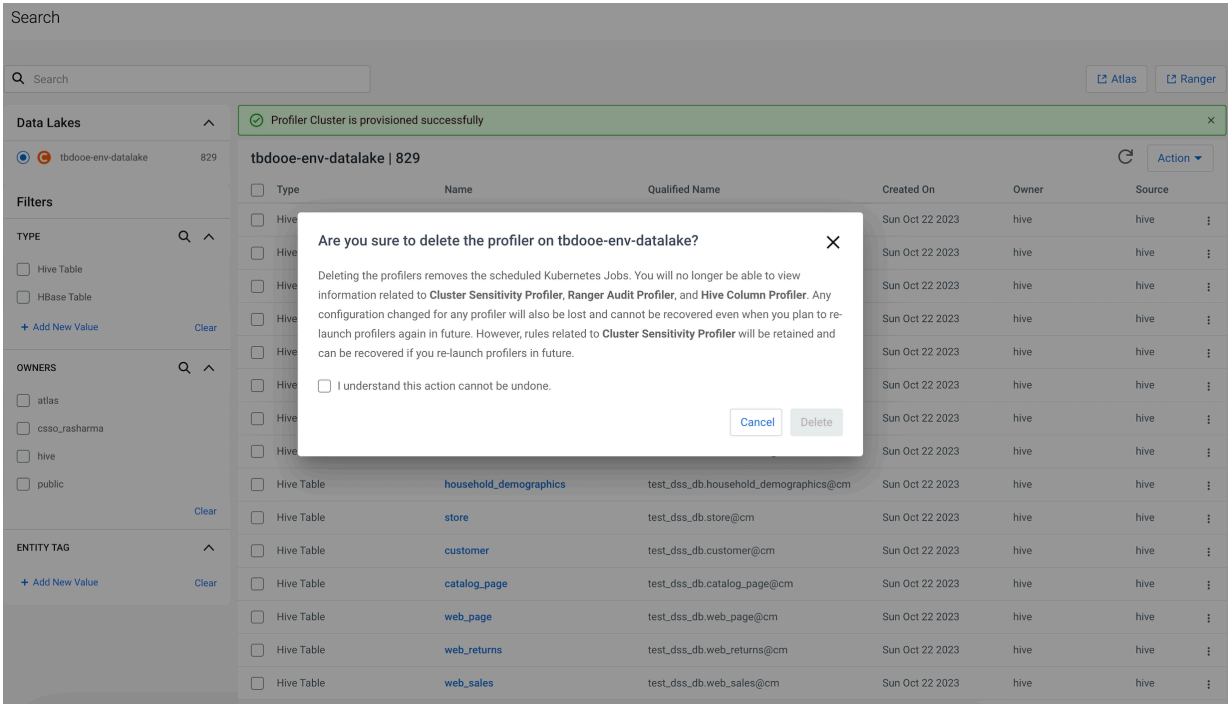
To overcome this situation, when you decide to delete the profiler cluster, there is a provision to retain the status of the Custom Sensitivity Profiler rules. If your profiler cluster has rules that are not changed or updated, you can directly delete the profiler cluster. If the rules were modified or updated, you have an option to download the modified rules along with deletion. The modified rules consist of the suspended System rules and the deployed Custom rules. Using the downloaded rules, you can manually add or modify them to your newly added profiler cluster.

- When you delete the scheduled jobs, the associated Kubernetes cron job object is deleted from the Kubernetes cluster.
- The associated data of the profilers from the Management Console database is also deleted for the specified data lake.

Procedure

1. On the search page, select the Data Lake from the list.
2. Click the Actions drop-down menu and select Delete Cluster.
3. Click Yes to proceed.

4. If you agree, select the warning message I understand this action cannot be undone.




5. Click Delete.
The application displays the following message.

The profiler cluster is deleted successfully.

On-Demand Profilers

You can use on-demand profilers to profile specific assets without depending on the cron-based scheduling of profilers jobs. On-demand profiler option is available on the asset details page of the selected asset.

For example, the diagram displays the Asset Details page of an asset. Run On-Demand profiler for Hive Column Statistics and Custom Sensitivity Profiler by clicking on the appropriate Run button. The next scheduled run provides details about the next scheduled profiling for the respective profilers.

 **Note:** You can use the On-Demand Profiler feature to profile both External and Managed tables.

Profilers | 2

Hive Column Profiler

Last run: 10 mins ago
Status: SUCCESS

Next Schedule Run: Today at 11:30 PM

Run

Cluster Sensitivity Profiler

Last run: 12 mins ago
Status: SUCCESS

Next Schedule Run: NA, Profiler is Disabled.

Run

Tracking Profiler Jobs

The Data Catalog profiler page is updated to provide a better user experience for tracking respective profiler jobs.

A new placeholder named “Schedule” is introduced under the Profilers section to provide tracking information of each profiler job. Under Schedule, you can find the type of profiler job that has run or in progress or has completed profiling data assets.

Jobs

Configs

Tag Rules

Filters

Clear All

Job Status

Finished

11

Running

1

Failed

0

Profilers

Cluster Sensitivity Profiler

0

Hive Column Profiler

0

Ranger Audit Profiler

12

Schedule : 5 | Running | Today

Schedule : 3 | Finished | Today

Schedule : 2 | Finished | Today

Status

Job ID ↑

Start Time

Stage

Queue

Assets Profiled

Ranger Audit

Finished

4

Dec 10 2020 09:05:51

Metrics Service

–

NA

Finished

3

Dec 10 2020 09:01:47

Metrics Service

–

NA

Finished

2

Dec 10 2020 09:00:04

Livy

default

NA

Finished

1

Dec 10 2020 09:00:01

Scheduler Service

–

NA

D

W

M

For each profiler job, you can view the details about:

- Job Status
- Type
- Job ID
- Start Time
- Stage
- Job Queue
- Total assets profiled

Data Catalog / Profilers

The screenshot shows the 'Data Catalog / Profilers' interface. At the top, there's a dropdown menu set to 'dc-pro-bxgaxu'. Below it are tabs for 'Jobs', 'Configs', and 'Tag Rules'. On the left, there's a 'Filters' sidebar with 'Clear All' and sections for 'Job Status' (Finished: 4, Running: 2, Failed: 2) and 'Profilers' (Cluster Sensitivity Profiler: 0, Ranger Audit Profiler: 8, Hive Column Profiler: 0). The main table has a summary bar at the top showing 'Schedule : 100 | Running | Today' and a table with columns: Status, Job ID, Start Time, Stage, Queue, and Assets Profiled. The table lists jobs for 'Ranger Audit', including one running (Job ID 280) and one finished (Job ID 279). Below the table, there are more summary bars for other profilers.

Status	Job ID	Start Time	Stage	Queue	Assets Profiled
Ranger Audit					
Running	280	Dec 09 2020 09:27:17	Livy	default	NA
Finished	279	Dec 09 2020 09:27:16	Scheduler Service	-	NA

Using this data can help you to troubleshoot failed jobs or even understand how the jobs were profiled and other pertinent information that can help you to manage your profiled assets. Whenever the Schedule status appears in green, it indicates that the profiler job has run successfully. When the color appears in blue and red, it indicates that the profiler job is running or has failed.

Profiler job runs in three phases:

- Scheduler Service - Part of Profiler Admin which queues the profiler requests.
- Livy - This service is managed by YARN and where the actual asset profiling takes place.
- Metrics Service - Reads the profiled data files and publishes them.



Note: More than one occurrence of Scheduler Service or Livy indicates that there could be more assets to be profiled. For example, if a HBase schedule has about 80 assets to be profiled, the first 50 assets would be profiled in the first Livy batch and the other assets get profiled in the next batch.

Clicking on each profiled asset would navigate to the profiled asset details page. The asset profiled page provides information about the profiled asset, profiled status, the profiled job id, and other relevant details.

In case of Ranger Audit profiling, there could be a “NA” status for the total number of assets profiled. It indicates that the auditing that happens is dependent on the Ranger policies. In other words, the Ranger policies are actually profiled and not the assets.

Important: Currently, the On-Demand schedule is not supported for this version of the profiler. The job schedule is either grayed out or disabled in such a scenario.

Viewing Profiler Jobs

You can monitor the overall health of your profiler jobs by viewing their status on the Profiler Jobs .

Each profiler runs a Spark job on a user-defined schedule defined via the profiler configuration. You can view the status of each of those jobs for all your clusters.

Monitoring the profiler jobs has the following uses:

- By seeing long-term trends in job execution, you can determine the overall health of your profilers.
- Knowing when jobs first failed can help when troubleshooting problems with profilers.

You can take the following actions:

1. Filter by job status or profiler.
2. Sort by start time.
3. Click to show a day, week, or month of jobs.
4. Sort by jobs ID, status, start time, cluster, queue, or profilers.

5. Expand or narrow to show a day, week, or month of jobs.

Related Information

[Understanding the sensitive data profiler](#)

[Understanding the ranger audit profiler](#)

Viewing Profiler Configurations

You can monitor the overall health of individual profilers by viewing their status on Profiler Configs .

Profilers / Configs

tbdoee-env-datalake ▾					
Jobs <u>Configs</u> Tag Rules					
Profiler Configuration					
Name	Last Run Time	Last Run Status	Next Scheduled Run	Config Version	Status
Ranger Audit Profiler	14 hours ago	SUCCESS	Tomorrow at 12:00 AM (UTC)	1	Active
Cluster Sensitivity Profiler	14 hours ago	FAILED	Tomorrow at 12:00 AM (UTC)	1	Active
Hive Column Profiler	14 hours ago	FAILED	Tomorrow at 12:00 AM (UTC)	1	Active

Monitoring the profiler configurations has the following uses:

- Verify which profilers are active and inactive.
- Verify the status of the profiler runs.
- View the last run time and status and the next scheduled run.

Related Information

[Understanding the sensitive data profiler](#)

[Understanding the ranger audit profiler](#)

Additional Configuration for Ranger Audit Profiler

In addition to the generic configuration, there are additional parameters for the Ranger Audit Profiler that can optionally be edited.

Procedure

1. Click Profilers in the main navigation menu on the left..
2. Click Configs to view all of the configured profilers.

3. Select Ranger Audit Profiler for which you need to edit the profiler configuration.



You can use the toggle button to enable / disable the Ranger Audit Profiler.

The Ranger Audit Profiler detail page is displayed which contains the following sections:

- Profiler Configurations
- Executor Configurations

Profiler Configurations

- Sampling configurations enables you to regulate sampling behaviour of the profilers. When an asset/table is profiled, instead of scanning the whole table, the profiler sample selects records as it finds them.
- Sample Count: Indicates the number of times a table must be sampled for profiling. A value less than 3 and higher than 30 is not recommended.
- Sample Factor: Controls the randomisation of records. Less value promote better random samples and higher values results in poor samples. A value 0.001 indicates that the data that is retrieved from Hive and a new random number is generated. If the value is less than or equal to the provided proportion (0.001), it will be chosen in the result set. If the value is greater, it is ignored.
- Sample Records: Indicates the number of records to be retrieved in a given sample. Consider this as LIMIT clause of the SQL query.

Executor Configurations

Executor Configurations are the runtime configuration. These configuration must be changed if you are changing the Pod configurations and when there is a requirement for additional compute power.

- Number of workers: Indicates the number of processes that are used by the distributed computing framework.
- Number of threads per worker: Indicates the number of threads used by each worker to complete the job.
- Worker Memory limit in GB: To avoid over utilization of memory, this parameter forces an upper threshold memory usage for a given worker. For example, if you have a 8 GB Pod and 4 threads, the value of this parameter must be 2 GB.

Additional Configuration for Hive Column Profiler

In addition to the generic configuration, there are additional parameters for the Hive Column Profiler that can optionally be edited.

Procedure

1. Click Profilers in the main navigation menu on the left.
2. Click Configs to view all of the configured profilers.
3. Select the cluster for which you need to edit profiler configuration.

The list of profilers for the selected clusters is displayed.

4. Select the cluster for which you need to edit profiler configuration.



You can use the toggle button to enable / disable the Hive Column Profiler.

The Hive Column Profiler detail page is displayed which contains the following sections:

- Profiler Configurations
- Pod Configurations
- Executor Configurations
- Asset Filter Rules

Profiler Configurations

- Sampling or Profiler configurations enables you to regulate sampling behaviour of the profilers. When an asset/table is profiled, instead of scanning the whole table, the profiler sample selects records as it finds them.
- Sample Count: Indicates the number of times a table must be sampled for profiling. A value less than 3 and higher than 30 is not recommended.
- Sample Factor: Controls the randomisation of records. Less value promote better random samples and higher values results in poor samples. A value 0.001 indicates that the data that is retrieved from Hive and a new random number is generated. If the value is less than or equal to the provided proportion (0.001), it will be chosen in the result set. If the value is greater, it is ignored.
- Sample Records: Indicates the number of records to be retrieved in a given sample. Consider this as LIMIT clause of the SQL query.

Profiler Configurations

As all profilers are submitted as Kubernetes jobs, you must decide if you want to add or reduce resources to handle workload of various sizes.

Pod configurations specify the resources that would be allocated to a pod when the profiler job starts to run.

- Pod CPU limit: Indicates the maximum number of cores that can be allocated to a Pod. The accepted values examples are 0.5, 1, 2, 500m, and 250m.
- Pod CPU Requirements: This is the minimum number of CPUs that will be allocated to a Pod when its provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The accepted values examples are 0.5, 1, 2, 500m, and 250m.
- Pod Memory limit: Maximum amount of memory can be allocated to a Pod. The accepted values examples are: 128974848, 129e6, 129M, 128974848000m, and 123Mi.
- Pod Memory Requirements: This is the minimum amount of RAM that will be allocated to a Pod when it is provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit.

Executor Configurations

Executor Configurations are the runtime configuration. These configuration must be changed if you are changing the Pod configurations and when there is a requirement for additional compute power.

- Number of workers: Indicates the number of processes that are used by the distributed computing framework.
- Number of threads per worker: Indicates the number of threads used by each worker to complete the job.
- Worker Memory limit in GB: To avoid over utilization of memory, this parameter forces an upper threshold memory usage for a given worker. For example, if you have a 8 GB Pod and 4 threads, the value of this parameter must be 2 GB.

Additional Configuration for Cluster Sensitivity Profiler

In addition to the generic configuration, there are additional parameters for the Cluster Sensitivity Profiler that can optionally be edited.

Procedure

1. Click Profilers in the main navigation menu on the left.
2. Click Configs to view all of the configured profilers.
3. Select Cluster Sensitivity Profiler for which you need to edit the profiler configuration.



You can use the toggle button to enable / disable the Cluster Sensitivity Profiler.

The Cluster Sensitivity Profiler detail page is displayed which contains the following sections:

- Profiler Configurations
- Pod Configurations
- Executor Configurations
- Asset Filter Rules

Profiler Configurations

- Sampling configurations enables you to regulate sampling behaviour of the profilers. When an asset/table is profiled, instead of scanning the whole table, the profiler sample selects records as it finds them.
- Sample Count: Indicates the number of times a table must be sampled for profiling. A value less than 3 and higher than 30 is not recommended.
- Sample Factor: Controls the randomisation of records. Less value promote better random samples and higher values results in poor samples. A value 0.001 indicates that the data that is retrieved from Hive and a new random number is generated. If the value is less than or equal to the provided proportion (0.001), it will be chosen in the result set. If the value is greater, it is ignored.
- Sample Records: Indicates the number of records to be retrieved in a given sample. Consider this as LIMIT clause of the SQL query.

Pod Configurations

As all profilers are submitted as Kubernetes jobs, you must decide if you want to add or reduce resources to handle workload of various sizes.

Pod configurations specify the resources that would be allocated to a pod when the profiler job starts to run.

- Pod CPU limit: Indicates the maximum number of cores that can be allocated to a Pod. The accepted values examples are 0.5, 1, 2, 500m, and 250m.
- Pod CPU Requirements: This is the minimum number of CPUs that will be allocated to a Pod when its provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The accepted values examples are 0.5, 1, 2, 500m, and 250m.
- Pod Memory limit: Maximum amount of memory can be allocated to a Pod. The accepted values examples are: 128974848, 129e6, 129M, 128974848000m, and 123Mi.
- Pod Memory Requirements: This is the minimum amount of RAM that will be allocated to a Pod when it is provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed)

for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit.

Executor Configurations

Executor Configurations are the runtime configuration. These configuration must be changed if you are changing the Pod configurations and when there is a requirement for additional compute power.

- Number of workers: Indicates the number of processes that are used by the distributed computing framework.
- Number of threads per worker: Indicates the number of threads used by each worker to complete the job.
- Worker Memory limit in GB: To avoid over utilization of memory, this parameter forces an upper threshold memory usage for a given worker. For example, if you have a 8 GB Pod and 4 threads, the value of this parameter must be 2 GB.

Related Information

[Understanding the sensitive data profiler](#)

Understanding Cron Expression generator

A cron expression details about when the schedule executes and visualizes the next execution dates of your cron expression. The cron expression utilizes the quartz engine.

The cron expression uses a typical format:

Each * in the cron represents a unique value.

Cron Expression: 0 18 * * *
Represented by Minute hour day(month) month day(week) As an example, "At 10:30 on day-of-month 15 in May." 30 10 15 5 *
Consider another use case example:
"At 10:30 on Sunday in May." 30 10 * 5 7

You can change the value of cron as and when it is required depending on how you want to schedule your profiler job.

Setting Asset filter rules

Add Asset filter rules as needed to customize the selection and deselection of assets which the profiler profiles.



Note: You can configure the Deny-list and Allow-list for both Cluster Sensitivity Profiler and Hive Column Profiler. The same filter rules do not apply to Ranger Audit Profiler.

Data Catalog / Profilers / Configs / Detail

Hive Column Profiler

Data Lake:

You can view who has accessed which data from a forensic audit or compliance perspective, visualize access patterns, and identify anomalies in access patterns using the Ranger Audit Profiler.

Active

Schedule*

0 0 0/6-1/1 *?*

Last Run Check*

1 Day

Sample Data Size *

Sample Percentage

100

Advance Options

Asset Filter Rules

Deny List

Allow List

Profiler will skip profiling assets which meet any of deny list rules

Search Deny List

Add New

Status	Key ↑	Operator	Value
--------	-------	----------	-------

Data Catalog / Profilers / Configs / Detail

Ranger Audit Profiler

Data Lake:

You can view the shape or distribution characteristics of the columnar data within a Hive table based on the Hive Column Profiler.

☒ Active

Schedule*

0 */30 * ? * *

^ Advance Options

Number of Executors*

1 ⓘ

Executor Cores*

1 ⓘ

Executor Memory (in GB)*

1 ⓘ

Driver Core*

1 ⓘ

Driver Memory (in GB)*

1 ⓘ

SAVE Cancel

- Deny-list - The profiler will skip profiling assets that meet any defined Deny-list criteria.
 - Select the Deny-list tab.
 - Click Add New to include rules for Deny-list.
 - Select the key from the drop down list. You can select a database name, name of the asset, name of the owner of the asset, path to the assets, or created date.
 - Select the operator from the drop down list. Depending on the keys selected, you can select an operator such as equals, contains. For example, you can select the name of assets that contain a particular string.
 - Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example here.
 - Click Done. Once it is added, you can toggle the state of the new rule to enable it or disable it as needed.

Asset Filter Rules

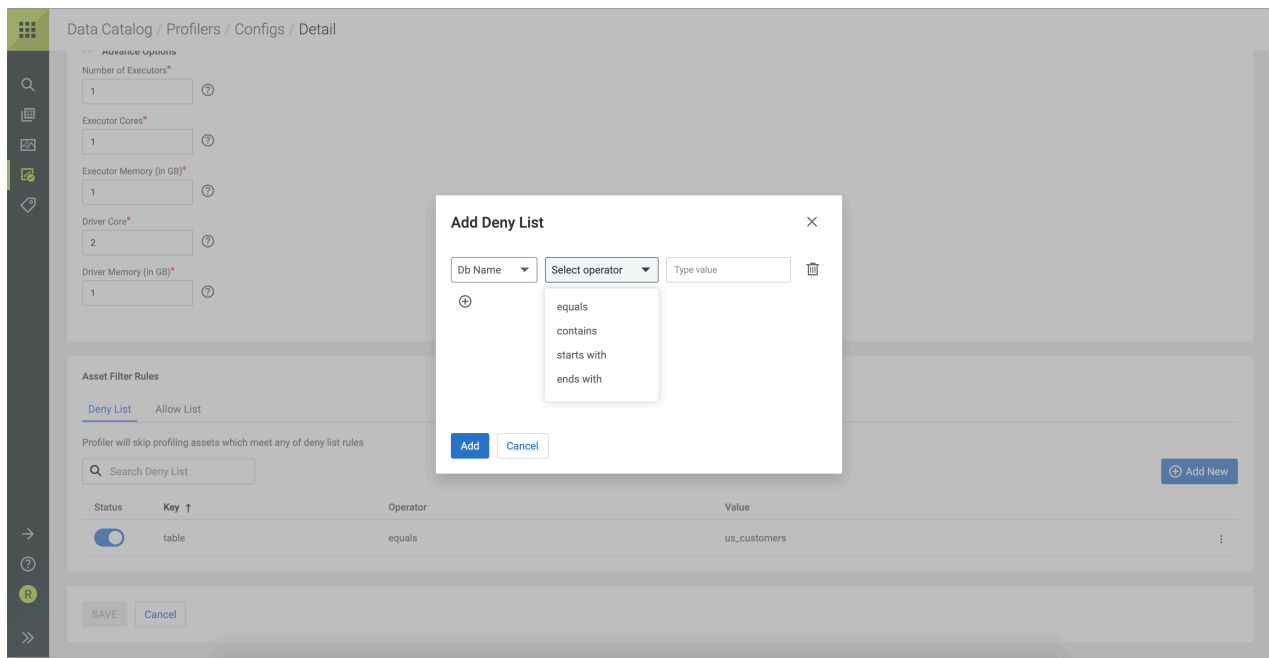
[Deny List](#) [Allow List](#)

Profiler will skip profiling assets which meet any of deny list rules

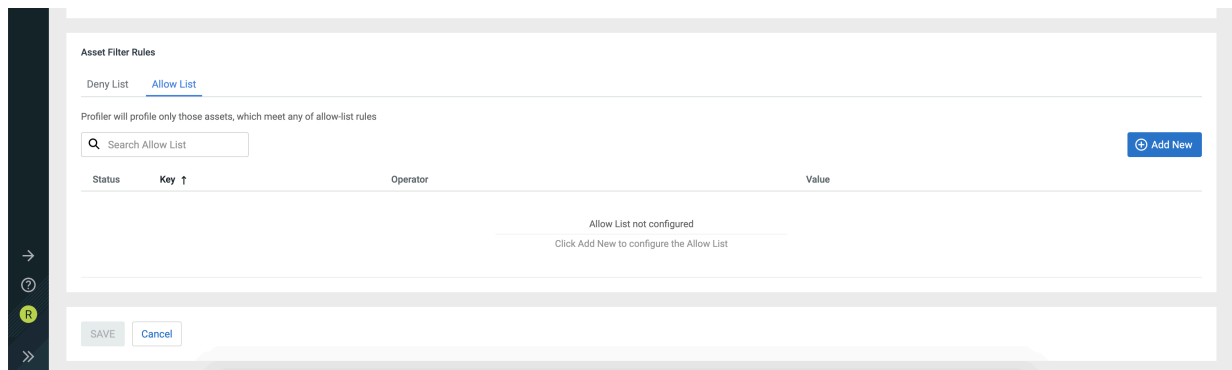
Q Search Deny List Add New

Status	Key ↑	Operator	Value
<input checked="" type="checkbox"/>	table	equals	us_customers

SAVE Cancel



- Allow-list - The profiler will include only assets that satisfy any defined Allow-list criteria. If no Allow-list is defined, the profiler will profile all the assets.
 - Select the Allow-list tab.
 - Click Add New to include rules for the Allow-list.
 - Select the key from the drop down list. You can select a database name, name of the asset, name of the owner of the asset, path to the assets, or created date.
 - Select the operator from the drop down list. Depending on the keys selected, you can select an operator such as equals, contains. For example, you can select the name of assets that contain a particular string.
 - Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
 - Click Done. Once it is added, you can toggle the state of the new rule to enable or disable it as needed.



Note: If an asset meets both Allow-list and Deny-list rules, the Deny-list rule overrides the Allow-list.

Enable or Disable Profilers

By default profilers are scheduled to run at every 24 hours at midnight UTC timezone.

Procedure

1. From Profiler Configs
2. Select the profiler to proceed further.



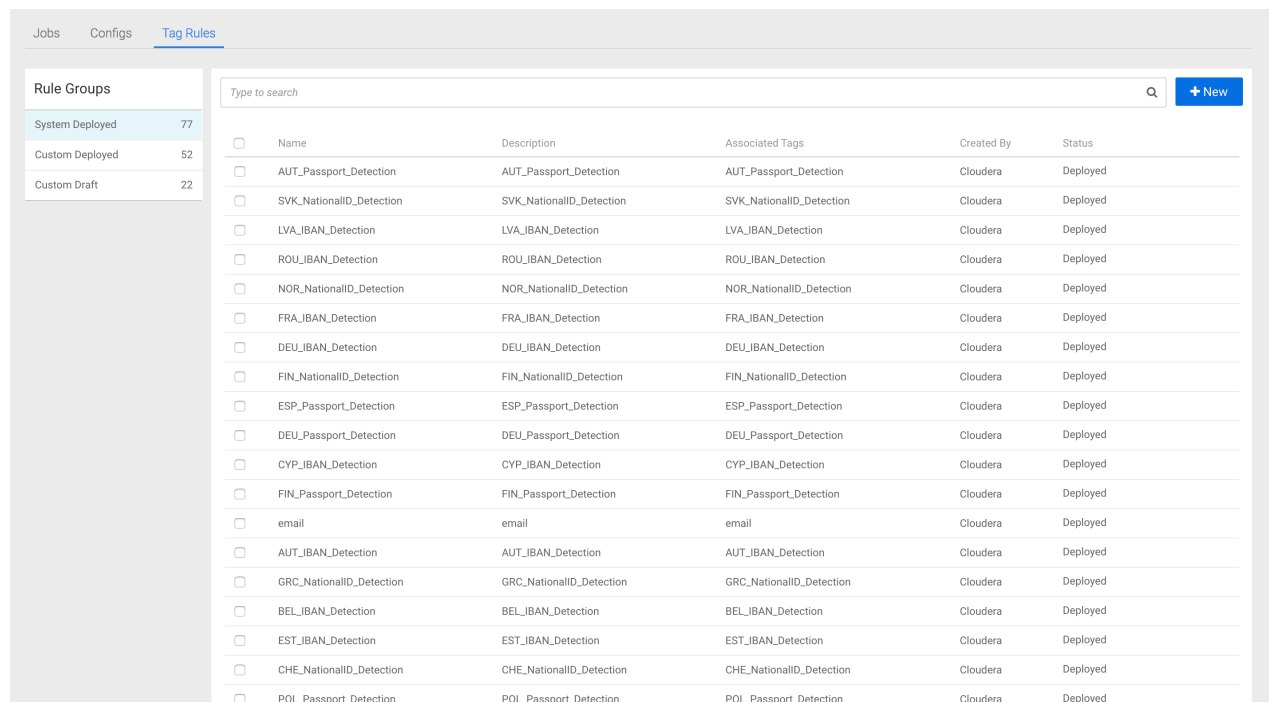
Name	Config Version	Status
Cluster Sensitivity Profiler	5	Active
Hive Column Profiler	1	Active
Ranger Audit Profiler	1	Active

Profiler Tag Rules

You can use preconfigured tag rules or create new rules based on regular expressions and allow or deny files on specific columns in your tables.

Rules are categorized into three groups:

- **System Deployed** : These are in-built rules that cannot be edited.
- **Custom Deployed**: Tag rules that you create and deploy on clusters after validation will appear under this category. Hover your mouse over the tag rules to deploy or suspend them as needed. You can also edit these tag rules.
- **Custom Draft** : You can create new tag rules and save them for later validation and deployment on clusters. Such rules appear under this category.



Name	Description	Associated Tags	Created By	Status
AUT_Passport_Detection	AUT_Passport_Detection	AUT_Passport_Detection	Cloudera	Deployed
SVK_NationalID_Detection	SVK_NationalID_Detection	SVK_NationalID_Detection	Cloudera	Deployed
LVA_IBAN_Detection	LVA_IBAN_Detection	LVA_IBAN_Detection	Cloudera	Deployed
ROU_IBAN_Detection	ROU_IBAN_Detection	ROU_IBAN_Detection	Cloudera	Deployed
NOR_NationalID_Detection	NOR_NationalID_Detection	NOR_NationalID_Detection	Cloudera	Deployed
FRA_IBAN_Detection	FRA_IBAN_Detection	FRA_IBAN_Detection	Cloudera	Deployed
DEU_IBAN_Detection	DEU_IBAN_Detection	DEU_IBAN_Detection	Cloudera	Deployed
FIN_NationalID_Detection	FIN_NationalID_Detection	FIN_NationalID_Detection	Cloudera	Deployed
ESP_Passport_Detection	ESP_Passport_Detection	ESP_Passport_Detection	Cloudera	Deployed
DEU_Passport_Detection	DEU_Passport_Detection	DEU_Passport_Detection	Cloudera	Deployed
CYP_IBAN_Detection	CYP_IBAN_Detection	CYP_IBAN_Detection	Cloudera	Deployed
FIN_Passport_Detection	FIN_Passport_Detection	FIN_Passport_Detection	Cloudera	Deployed
email	email	email	Cloudera	Deployed
AUT_IBAN_Detection	AUT_IBAN_Detection	AUT_IBAN_Detection	Cloudera	Deployed
GRC_NationalID_Detection	GRC_NationalID_Detection	GRC_NationalID_Detection	Cloudera	Deployed
BEL_IBAN_Detection	BEL_IBAN_Detection	BEL_IBAN_Detection	Cloudera	Deployed
EST_IBAN_Detection	EST_IBAN_Detection	EST_IBAN_Detection	Cloudera	Deployed
CHE_NationalID_Detection	CHE_NationalID_Detection	CHE_NationalID_Detection	Cloudera	Deployed
POL_Passport_Detection	POL_Passport_Detection	POL_Passport_Detection	Cloudera	Deployed

Tag Management

From Atlas tags UI in Data Catalog, you can create, modify, and delete any of the Atlas tags in a Data Catalog instance.

You can access the Atlas link by logging into Data Catalog > Atlas Tags .

Atlas Tags allows the user to perform the following activities with a selected Data Lake for tag management:

- Selecting a Data Lake
- Searching for a tag
- Adding a tag
- Editing a tag
- Deleting a tag

The new Atlas tags UI is displayed as seen in the diagram.

Data Catalog / Atlas Tags

Tags (Atlas Tags)

Tag ↑	Description	Created By	Created On
dp	dp	dp profiler	Mar 12 2021 12:40:11
dp_bankaccount	dp_bankaccount	dp profiler	Mar 12 2021 15:53:23
dp_BEL_Passport_Detection	dp_BEL_Passport_Detection	dp profiler	Mar 12 2021 15:53:23
dp_BGR_IBAN_Detection	dp_BGR_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23
dp_BGR_NationalID_Detection	dp_BGR_NationalID_Detection	dp profiler	Mar 12 2021 15:53:19
dp_CHE_IBAN_Detection	dp_CHE_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23
dp_creditcard	dp_creditcard	dp profiler	Mar 12 2021 15:53:19
dp_CZE_NationalID_Detection	dp_CZE_NationalID_Detection	dp profiler	Mar 12 2021 15:53:19
dp_DEU_IBAN_Detection	dp_DEU_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23
dp_DEU_Passport_Detection	dp_DEU_Passport_Detection	dp profiler	Mar 12 2021 15:53:19
dp_DNK_NationalID_Detection	dp_DNK_NationalID_Detection	dp profiler	Mar 12 2021 15:53:19
dp_email	dp_email	dp profiler	Mar 12 2021 15:53:19
dp_ESP_Passport_Detection	dp_ESP_Passport_Detection	dp profiler	Mar 12 2021 15:53:23
dp_EST_IBAN_Detection	dp_EST_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23
dp_EST_NationalID_Detection	dp_EST_NationalID_Detection	dp profiler	Mar 12 2021 15:53:19

You can create a new tag in the Atlas tags UI. The following diagram provides an overview about the Create a new tag page.

The screenshot shows the Cloudera Data Catalog interface. On the left is a sidebar with navigation links: Search, Datasets, Bookmarks, Profilers, and Atlas Tags (highlighted). The main content area is titled 'Data Catalog / Atlas Tags'. It features a search bar and a table of existing tags. A modal window titled 'Create a new tag' is open on the right, allowing users to add a new tag with fields for Name, Description, and classification, along with an 'Add new attributes' button.

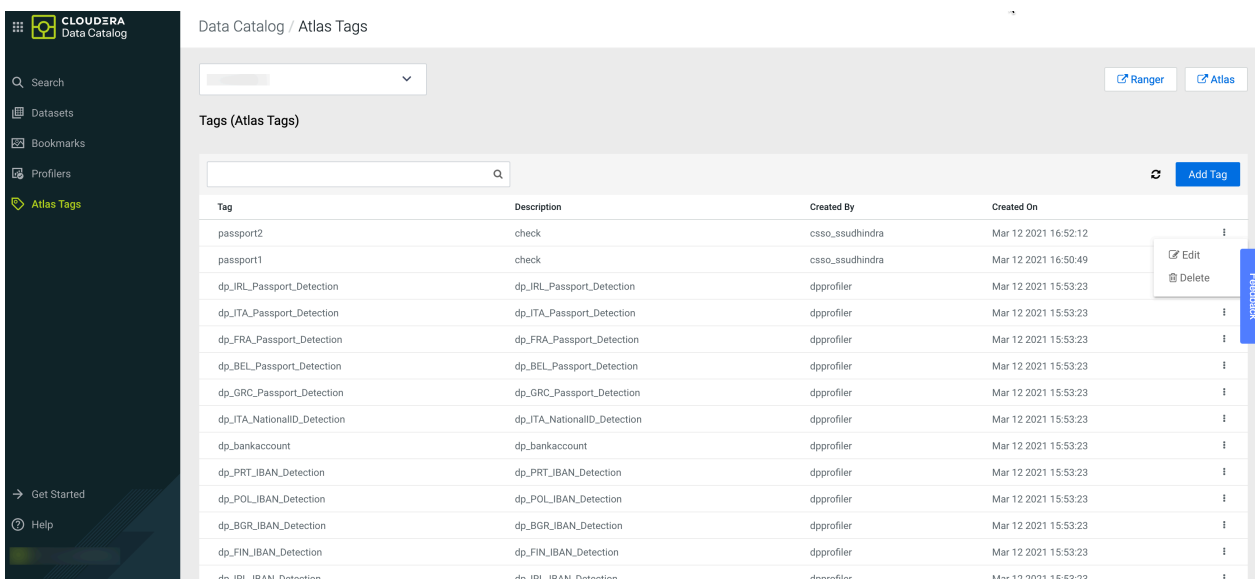
Tag ↑	Description	Created By
dp	dp	dp profiler
dp_bankaccount	dp_bankaccount	dp profiler
dp_BEL_Passport_Detection	dp_BEL_Passport1_Detection	dp profiler
dp_BGR_IBAN_Detection	dp_BGR_IBAN_Detection	dp profiler
dp_BGR_NationalID_Detection	dp_BGR_NationalID_Detection	dp profiler
dp_CHE_IBAN_Detection	dp_CHE_IBAN_Detection	dp profiler
dp_creditcard	dp_creditcard	dp profiler
dp_CZE_NationalID_Detection	dp_CZE_NationalID_Detection	dp profiler
dp_DEU_IBAN_Detection	dp_DEU_IBAN_Detection	dp profiler
dp_DEU_Passport_Detection	dp_DEU_Passport1_Detection	dp profiler
dp_DNK_NationalID_Detection	dp_DNK_NationalID_Detection	dp profiler
dp_email	dp_email	dp profiler
dp_ESP_Passport_Detection	dp_ESP_Passport1_Detection	dp profiler
dp_EST_IBAN_Detection	dp_EST_IBAN_Detection	dp profiler
dp_EST_NationalID_Detection	dp_EST_NationalID_Detection	dp profiler

You can add or update Atlas tags. The created or updated tag is highlighted in the tag list as seen in the following diagram.

This screenshot shows the same Cloudera Data Catalog interface after a new tag has been added. A green notification banner at the top states 'passport2 created successfully'. The tag list now includes 'passport2' and 'passport1'. The 'passport2' tag is highlighted. The table also includes a 'Created On' column. A blue 'Add Tag' button is visible in the top right of the tag list area.

Tag	Description	Created By	Created On
passport2	check	csso_ssudhindra	Mar 12 2021 16:52:12
passport1	check	csso_ssudhindra	Mar 12 2021 16:50:49
dp_IRL_Passport_Detection	dp_IRL_Passport_Detection	dp profiler	Mar 12 2021 15:53:23
dp_ITA_Passport_Detection	dp_ITA_Passport_Detection	dp profiler	Mar 12 2021 15:53:23
dp_FRA_Passport_Detection	dp_FRA_Passport_Detection	dp profiler	Mar 12 2021 15:53:23
dp_BEL_Passport_Detection	dp_BEL_Passport_Detection	dp profiler	Mar 12 2021 15:53:23
dp_GRC_Passport_Detection	dp_GRC_Passport_Detection	dp profiler	Mar 12 2021 15:53:23
dp_ITA_NationalID_Detection	dp_ITA_NationalID_Detection	dp profiler	Mar 12 2021 15:53:23
dp_bankaccount	dp_bankaccount	dp profiler	Mar 12 2021 15:53:23
dp_PRT_IBAN_Detection	dp_PRT_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23
dp_POL_IBAN_Detection	dp_POL_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23
dp_BGR_IBAN_Detection	dp_BGR_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23
dp_FIN_IBAN_Detection	dp_FIN_IBAN_Detection	dp profiler	Mar 12 2021 15:53:23

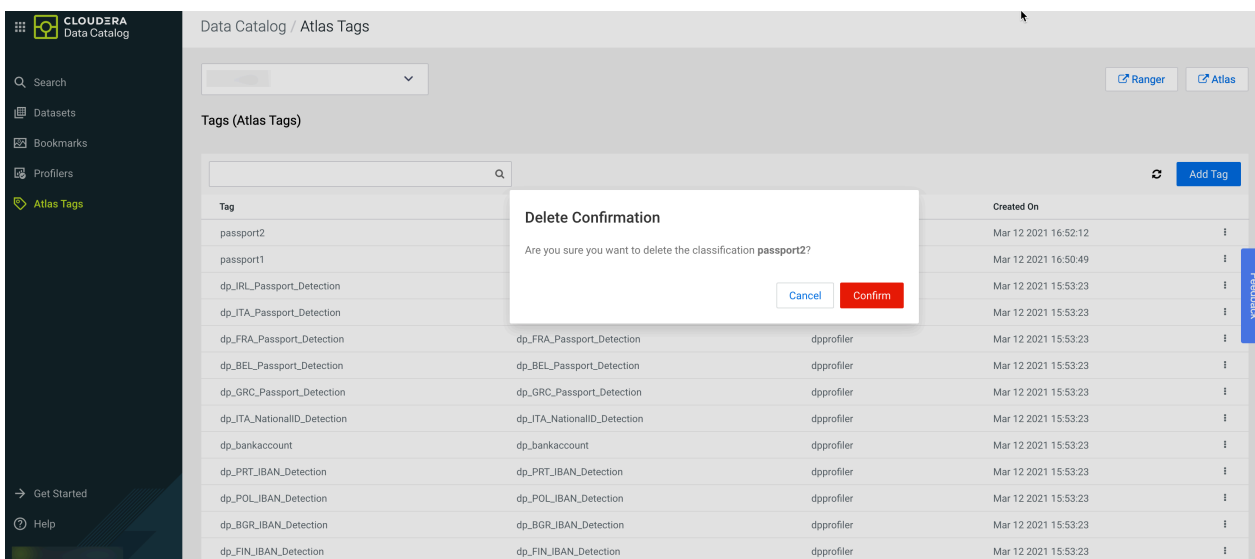
You can also edit or delete the Atlas tag as shown in the image. When you are editing the tag, you can only change the description or add new attributes.



The screenshot shows the Cloudera Data Catalog interface with the 'Atlas Tags' section selected in the left sidebar. The main content area displays a table of tags. A search bar is at the top, and there are 'Ranger' and 'Atlas' filters. A 'Delete' button is visible in the top right corner of the table area.

Tag	Description	Created By	Created On
passport2	check	csso_ssudhinda	Mar 12 2021 16:52:12
passport1	check	csso_ssudhinda	Mar 12 2021 16:50:49
dp_JRL_Passport_Detection	dp_JRL_Passport_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_ITA_Passport_Detection	dp_ITA_Passport_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_FRA_Passport_Detection	dp_FRA_Passport_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_BEL_Passport_Detection	dp_BEL_Passport_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_GRC_Passport_Detection	dp_GRC_Passport_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_ITA_NationalID_Detection	dp_ITA_NationalID_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_bankaccount	dp_bankaccount	dpprofiler	Mar 12 2021 15:53:23
dp_PRT_IBAN_Detection	dp_PRT_IBAN_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_POL_IBAN_Detection	dp_POL_IBAN_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_BGR_IBAN_Detection	dp_BGR_IBAN_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_FIN_IBAN_Detection	dp_FIN_IBAN_Detection	dpprofiler	Mar 12 2021 15:53:23
dp_JRL_IBAN_Detection	dp_JRL_IBAN_Detection	dpprofiler	Mar 12 2021 15:53:23

You can delete one Atlas tag at a time. A separate confirmation message appears.



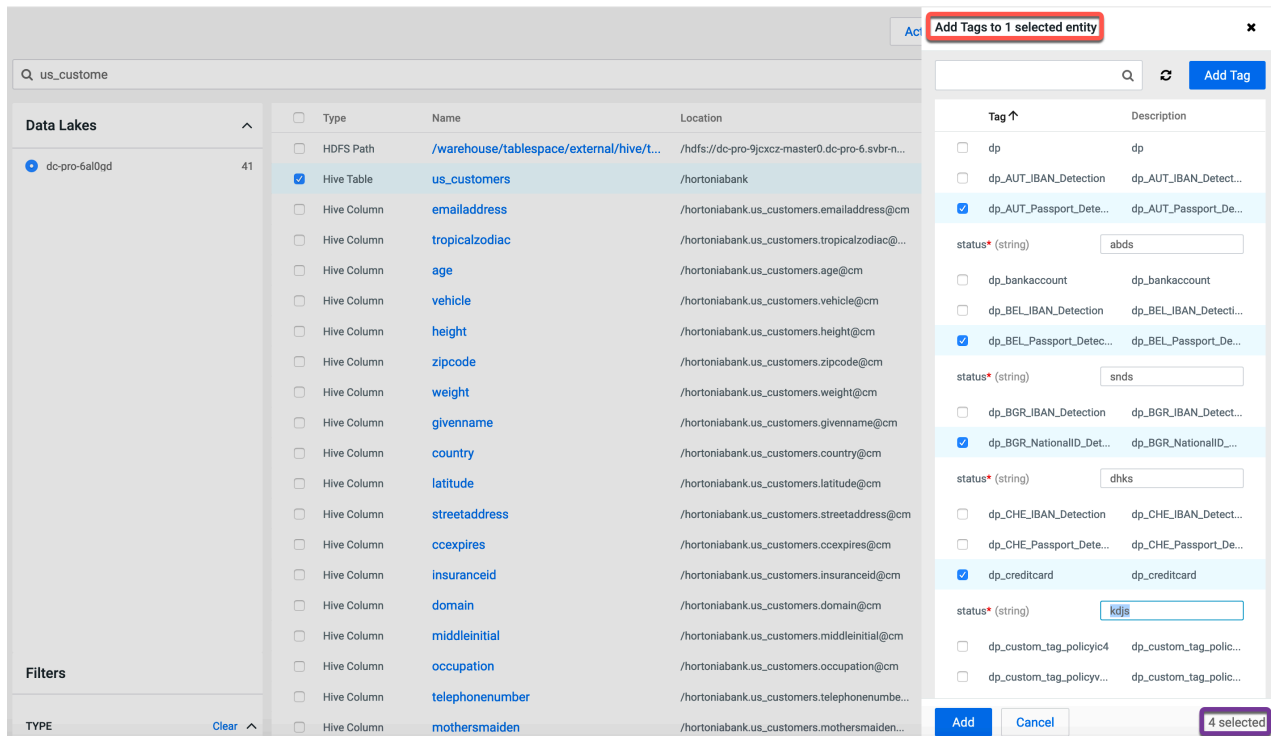
The screenshot shows the same Cloudera Data Catalog interface, but with a 'Delete Confirmation' dialog box open in the center. The dialog asks 'Are you sure you want to delete the classification passport2?' and has 'Cancel' and 'Confirm' buttons. The background table of tags is still visible.

Tagging Multiple Assets

On the Data Catalog search page, you can add tags to multiple assets based on the asset type that you select based on the search result.

When you select an asset, you can add one or more available tags to the selected asset. You can also create one or more new tags and associate the newly created tags to the selected asset. The number of selected assets that you plan to tag is displayed. As you add the number of tags to one or more selected assets, The Add Tag panel displays the number of tags assigned.

Data Catalog / Search



Add Tags to 1 selected entity

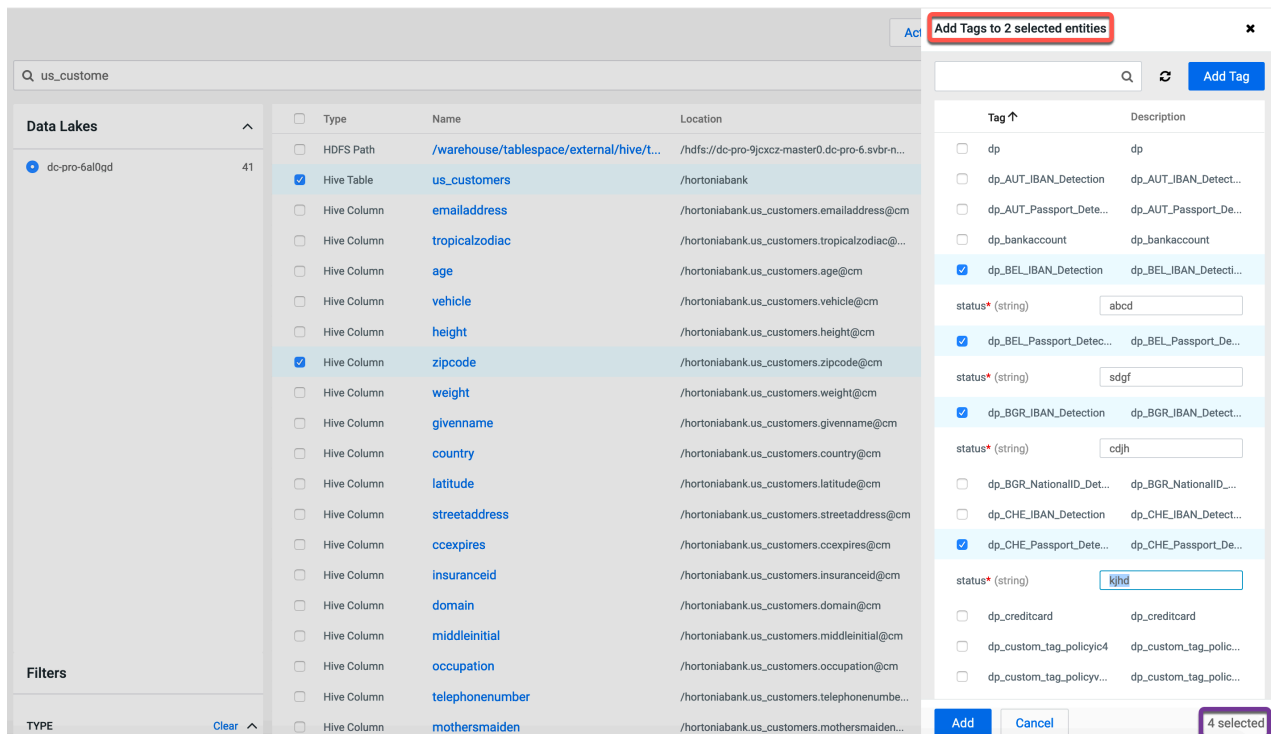
Tag	Description
<input type="checkbox"/> dp	dp
<input type="checkbox"/> dp_AUT_IBAN_Detection	dp_AUT_IBAN_Detect...
<input checked="" type="checkbox"/> dp_AUT_Passport_Dete...	dp_AUT_Passport_De...
status* (string) abds	
<input type="checkbox"/> dp_bankaccount	dp_bankaccount
<input type="checkbox"/> dp_BEL_IBAN_Detection	dp_BEL_IBAN_Detecti...
<input checked="" type="checkbox"/> dp_BEL_Passport_Detec...	dp_BEL_Passport_De...
status* (string) snds	
<input type="checkbox"/> dp_BGR_IBAN_Detection	dp_BGR_IBAN_Detect...
<input checked="" type="checkbox"/> dp_BGR_NationalID_Det...	dp_BGR_NationalID_D...
status* (string) dhks	
<input type="checkbox"/> dp_CHE_IBAN_Detection	dp_CHE_IBAN_Detect...
<input type="checkbox"/> dp_CHE_Passport_Dete...	dp_CHE_Passport_De...
<input checked="" type="checkbox"/> dp_creditcard	dp_creditcard
status* (string) kjds	
<input type="checkbox"/> dp_custom_tag_policyic4	dp_custom_tag_polic...
<input type="checkbox"/> dp_custom_tag_policyv...	dp_custom_tag_polic...

Add **Cancel** **4 selected**



Note: If you do not save your changes without clicking the Add button in Add Tags panel, the changes are not retained in the Data Catalog instance. You have to retag the assets and later click the Add button.

Data Catalog / Search



Add Tags to 2 selected entities

Tag	Description
<input type="checkbox"/> dp	dp
<input type="checkbox"/> dp_AUT_IBAN_Detection	dp_AUT_IBAN_Detect...
<input type="checkbox"/> dp_AUT_Passport_Dete...	dp_AUT_Passport_De...
<input type="checkbox"/> dp_bankaccount	dp_bankaccount
<input checked="" type="checkbox"/> dp_BEL_IBAN_Detection	dp_BEL_IBAN_Detecti...
status* (string) abcd	
<input checked="" type="checkbox"/> dp_BEL_Passport_Detec...	dp_BEL_Passport_De...
status* (string) sdfg	
<input checked="" type="checkbox"/> dp_BGR_IBAN_Detection	dp_BGR_IBAN_Detect...
status* (string) cdjh	
<input type="checkbox"/> dp_BGR_NationalID_Det...	dp_BGR_NationalID_...
<input type="checkbox"/> dp_CHE_IBAN_Detection	dp_CHE_IBAN_Detect...
<input checked="" type="checkbox"/> dp_CHE_Passport_Dete...	dp_CHE_Passport_De...
status* (string) kjhd	
<input type="checkbox"/> dp_creditcard	dp_creditcard
<input type="checkbox"/> dp_custom_tag_policyic4	dp_custom_tag_polic...
<input type="checkbox"/> dp_custom_tag_policyv...	dp_custom_tag_polic...

Add **Cancel** **4 selected**

When you add one or more tags to the selected entities, the assigned tags are displayed having been tagged to the number of selected entities. Another scenario could throw a message saying that the selected asset is already tagged.

Propagated Asset tagging

Data Catalog supports the concept of propagated tags. This feature is derived from Apache Atlas.

Whenever you add a new tag, you can mark them as propagated and use those tags accordingly while tagging assets.

Data Catalog / Asset Details

The screenshot shows the 'employee1' asset details page. The 'Properties' section on the left lists: Type: HIVE TABLE, # of Columns: 4, Data Lake: [redacted], Datasets: 0, Owner: [redacted], Created On: Mon Mar 15 2021 14:21:43 GMT+0530 (India Stan...), Last Access Time: Mon Mar 15 2021 14:21:43 GMT+0530 (Indi...), Table Type: EXTERNAL_TABLE, Database: default, DB Catalog: [redacted], Parent: default. The 'Classifications' section shows four tags: test_tag1, test_tag, dp_email, and dp. The 'Lineage' section shows a flow from /warehouse/tables... to employee1 to employee2. The 'Profilers' section on the right shows two profilers: Cluster Sensitivity Profiler (Last run: 21 mins ago, Status: SUCCESS, Next Schedule Run: Wednesday at 3:50 PM) and Hive Column Profiler (Last run: -, Status: NA, Next Schedule Run: Today at 5:30 PM). A 'Feedback' button is visible on the right side.

For example, consider table1 as a parent asset and table2 as a child asset. Create a tag and mark that tag as propagated, and later apply the same tag to table1. The tag gets applied to table2 as well. Propagated tag works on the basis of parent -> child tagging relationship.

Data Catalog / Asset Details

The screenshot shows the 'employee2' asset details page. The 'Properties' section on the left lists: Type: HIVE TABLE, # of Columns: 4, Data Lake: [redacted], Datasets: 0, Owner: [redacted], Created On: Mon Mar 15 2021 15:01:11 GMT+0530 (India Stan...), Last Access Time: Mon Mar 15 2021 15:01:11 GMT+0530 (Indi...), Table Type: MANAGED_TABLE, Database: default, DB Catalog: [redacted], Parent: default. The 'Classifications' section shows four tags: test_tag1, test_tag, dp_email, and dp. The 'Lineage' section shows a flow from /warehouse/tables... to employee1 to employee2. The 'Profilers' section on the right shows two profilers: Cluster Sensitivity Profiler (Last run: -, Status: NA, Next Schedule Run: Today at 3:50 PM) and Hive Column Profiler (Last run: -, Status: NA, Next Schedule Run: Today at 5:30 PM). A 'Feedback' button is visible on the right side.



Note: When you delete or remove the propagated tag from the parent asset, the same tag is removed from all the child assets.



Attention: The propagated tag concept is not supported with child -> parent relationships.

Creating Custom Profiler Rules

You can create a custom profiler by adding the required tags, regex entries, and attaching whitelist or blacklist files to specific columns within your tables.

Procedure

1. On the Profilers page, click Tag Rules.
2. On the Tag Rules tab, click New to create a new profiler tag rule.
3. Enter the name of the new custom profiler tag rule.
4. Enter the description for the custom tag rule.
5. Select the Tags. You can select tags from the drop down list and or enter a new value to create a new tag.
New tags that you create here are added with a dp_ prefix in the list of Atlas tags. For example, if you add a new tag called credit_card, this tag will be added as dp_credit_card in Atlas.
6. Enter the rule for the column name. As you enter the values, regex name and resource names are auto populated. Select the column that is needed for your custom profiler.
7. Enter the column value for the DSL.

Based on your entry, Data Catalog auto populates values from the entries already available in the Resources tab. You can use a combination of regex entries and whitelist or blacklist files and other behaviors. For more information about behaviors, see DSL Grammar.

8. Click Save and Validate.

Data Catalog / Profilers

Custom Rule

Name *

Description

Tags *

Column Name Expression

Column Value Expression *

Resources

~ Regex 🔍 +

SampleRegex_1580209003967
SampleRegex_1.58020939186e+12
DeployRegex1580209681238
SampleRegex_1.58020999412e+12
SampleRegex_1.58021014275e+12
SampleRegex_1.58021014308e+12
DeployRegex1580210288950
SampleRegex_1580276618318
SampleRegex_1580277217453

In the validation pop up window that appears, enter data to validate your custom profiler tag rule. Make sure you separate each data entry with a new line.

9. Click Save to create a tag rule and validate and deploy it later.

Adding Custom Regular Expressions

To use custom regex entries within your new custom profiler tag rules, you can also add new regex values.

Procedure

1. Click Resources in the right panel on the New Custom Profiler Rules page.

2. Click + icon on the Regex tab. The Regular Expression Editor page appears.
3. Enter the name of the new regular expression.
4. Enter a valid regular expression.

For example:

```
\b((([a-zA-Z0-9_-\.\ ]+))@((\[[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.))|((([a-zA-Z0-9_-\ ]+\.\.)))([a-zA-Z]{2,4}|[0-9]{1,3})(\|)?))\b
```

5. Enter the list of test strings to evaluate the new regular expression.
If the test string is valid, then the match information gets auto populated in the Match Information box.
6. Click Save to add the new regular expression to the list of Regex Resources.

Adding Lookup Files

When you have too many allowed and denied entries and cannot add them inline, you can create allowed or denied files with one value in each line and add them to your DSL.

Procedure

1. Click Resources in the right panel on the New Custom Profiler Rules page.
2. Click + icon on the Lookups tab. The New Lookup File page appears.
3. Enter the name of the new Lookup file.
4. Click Choose File to upload the file.
5. Click Save.

Using Behaviors

You can use various behaviors to take single inputs of type text and evaluate them to a Boolean value.

The profiler can take column values of any type and pass the values to each behaviour as text. Behaviors include the following:

1. Regular expressions
2. File based allowlist and denylist checks

Regular expressions

You can include one or more regular expressions and evaluate to True if one of these matches the provided value.

Keyword: `regex`

A regex that matches everything can be defined as follows:

```
regex(\"[\\\\\\\\s\\\\\\\\S]+\")
```

A regex that includes multiple expressions can be defined as follows:

```
regex(\"[\\\\\\\\s\\\\\\\\S]+\", \"^[0-9]*$\")
```

Using DSL Grammar

Using DSL grammar, you can combine different behaviours in intuitive ways to bring out functionality while creating custom profiler rules.

The two behaviors available in this framework are as follows:

1. `falseIdentity` - Always evaluates to false, regardless of the input.
2. `trueIdentity` - Always evaluates to true, regardless of the input.

These two behaviors are used in the following examples and descriptions.

Binary AND operator

Keyword: `and`

And works the same way it does other languages. Hence following observations.

```
falseIdentity and trueIdentity == falseIdentity
```

```
falseIdentity and falseIdentity == falseIdentity
```

```
trueIdentity and trueIdentity == trueIdentity
```

```
trueIdentity and falseIdentity == falseIdentity
```

Here we are using `==` to show their equality.

Binary OR operator

The or operator works the same way it does in other languages.

```
falseIdentity or trueIdentity == trueIdentity
```

```
falseIdentity or falseIdentity == falseIdentity
```

```
trueIdentity or trueIdentity == trueIdentity
```

```
trueIdentity or falseIdentity == trueIdentity
```

Expand DSL to use as follows.

```
val rule1= falseIdentity and trueIdentity and trueIdentity
```

```
val rule2= trueIdentity and trueIdentity and trueIdentity
```

```
val rule3=rule1 and rule2
```

```
rule3 or trueIdentity
```

The above expression evaluates to true.