CDP Private Cloud Data Services / Data Catalog 1.5.5

Cloudera Data Catalog Reference

Date published: 2023-12-16
Date modified: 2025-11-10

https://docs.cloudera.com/

Legal Notice

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

Launching profilers using the command-line

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

The CLI is one executable and does not have any external dependencies. You can execute some operations in the Cloudera Data Catalog service using the Cloudera CLI commands.

Users must have valid permissions to launch profilers on a data lake.

For more information about the access details, see Prerequisites to access Cloudera Data Catalog.

Prerequisites

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

DATA_CATALOG_ENABLE_API_SERVICE

The Cloudera Data Catalog CLI

In your Cloudera 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 Cloudera 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
```

Parameters for profiler launch command

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.
```

```
NAME
       launch-profilers - Launches DataCatalog profilers in a given datalak
e.

DESCRIPTION
       Launches DataCatalog profilers in a given datalake.

SYNOPSIS

            launch-profilers
          --datalake <value>
          [--enable-ha | --no-enable-ha]
          [--profilers <value>]
          [--instance-types <value>]
          [--max-nodes <value>]
          [--cli-input-json <value>]
          [--generate-cli-skeleton]

OPTIONS
       --datalake (string)
          The CRN of the Datalake.

       --enable-ha | --no-enable-ha (boolean)
          Enables High Availability (HA) for datacatalog profilers (default
          value is false). The High Availability (HA) Profiler cluster
          provides failure resilience and scalability but incurs additional
          cost.
       --profilers (array)
          List of profiler names that need to be launched. (Applicable only
          for compute cluster enabled environments).

       Syntax:

          "string" "string" ...

       --instance-types (array)
          List of instance types to be used for the auto-scaling node group
          setup (Applicable only for compute cluster enabled environments).

       Syntax:

          "string" "string" ...
       --max-nodes (integer)
          Maximum number of nodes that can be spawned inside the auto-scal
ing
          node group, in the range of 30 to 100 (both inclusive). (Applicabl
e
          only for compute cluster enabled environments).

       --cli-input-json (string)
          Performs service operation based on the JSON string provided. The
          JSON string follows the format provided by --generate-cli-skelet
on.
          If other arguments are provided on the command line, the CLI value
s
          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 i
nput
          can be used as an argument for --cli-input-json.
OUTPUT
       success -> (boolean)
          Status of the profiler launch operation.

       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.

FORM FACTORS
       public
```

Note:

- The following parameters are only applicable to Compute Cluster environments (they are ignored in VM-based environments):
  - --profilers ****VALUE****
  - --instance-types ****VALUE****
  - --max-nodes ****VALUE****

Parameters for profiler delete command

You get additional information about this command by using:

cdp datacatalog delete-profiler --help

```
NAME
       delete-profiler - Deletes DataCatalog profiler in a given datalake.
DESCRIPTION
       Deletes DataCatalog profiler in a given datalake.

SYNOPSIS
```

```
             delete-profiler
           --datalake <value>
           [--cli-input-json <value>]
           [--generate-cli-skeleton]

OPTIONS
       --datalake (string)
          The 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-skeleton
.
          If other arguments are provided on the command line, the CLI val
ues
          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 inp
ut
          can be used as an argument for --cli-input-json.

OUTPUT
FORM FACTORS
       public
```

Launching the profiler

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

```
cdp datacatalog launch-profilers --datalake [***DATALAKE CRN***]
```

Example:

```
cdp datacatalog launch-profilers --datalake crn:cdp:data
lake:DATACENTERNAME:c*****b-ccce-4**d-a**1-8********8:datalake:4*****5e-c**
1-4**2-8**e-1*******2
{
    "success": true
}
```

Related Information

Cloudera CLI 0.9.124 documentation / datacatalog command

Sample Workflows

When you work with Cloudera Data Catalog, you can use the workflows described in this section.

Tag management and search workflow

You can create a tag, assign the tag to an asset, and search for an asset using various filters.

About this task

You can create a tag and search for any asset using the tag. You can also apply a tag to a column. Later, you can search assets based on the Entity Tag and Column Tag filters.

Procedure

1. Go to **Atlas Tags** and click Add Tag
2. Enter the name, description for the tag.
3. Inserting a classification to inherit attributes and adding new attributes are optional.
4. Click Save.
   The new tag is created.

   Next, you must assign the newly created tag to one of the existing assets.

   For example, we can assign the tag to the Hive table.
5. Go to **Search** and search for an asset with columns, such as a Hive table, then click on the asset to view the **Asset Details** page.
6. Click +Add Classification, to display the **Classifications** view.
7. Select the tag that you newly created by typing in its name and assign the tag to the asset.
8. Click Save.

   Next, you can add the tag to a column in the selected asset.
9. Return to the previously selected asset: Go to **Search** and search for an asset with columns, such as a Hive table, then click on the asset to view the **Asset Details** page.
10. Select the **Schema** tab and click Edit, then select Classifications.
11. Click + and select the tag by entering its name.

    Note: The tag list contains System Tags and Managed Tags. Managed Tags are selected by default.

    - System Tags are applied by the tag rules in the **System Deployed** section. For example, IBAN, passport number detection rules.
    - Managed Tags are applied by the tag rules in the **Custom Deployed** section. These are user created tag rules. However, custom tag rules can also apply system tags.
12. Select a tag with the + icon and click Save.

Once tags are added to the table and column, you can search for the corresponding assets using the tag name.

13. Go to **Search** and enter the tag name in the Entity Tag or the Column tag filter.
    The asset for which the tag was added is displayed as the search result.

Auto-tagging workflow with custom Cluster Sensitivity Profiler rules

You can auto-tag workflows while working with Cluster Sensitivity Profiler.

About this task

Use the following information to create a custom tag and assign the same to the Cluster Sensitivity Profiler.

Procedure

1. Go to Profilers Tag Rules .
2. Click + New to open the **Custom Rule** window.
3. Under the **Resources** pane, click + to open the **Regular Expression Editor** window.
4. Enter the name and input the regular expression value in such a way that it matches your test string.
   If your regular expression value is [a-z][a-z][a-z][a-z] and the test string is “baby”, there is a match.
5. Click Save.
6. On the Custom Rule window, enter the name and description.
7. Enter the tags value and select the Column Expression Name from the drop-down.

   You must select the same regular expression you had created under the Resources pane.

8. Enter a value for the **Tags** field and select the **Column Value Expression** from the drop-down.

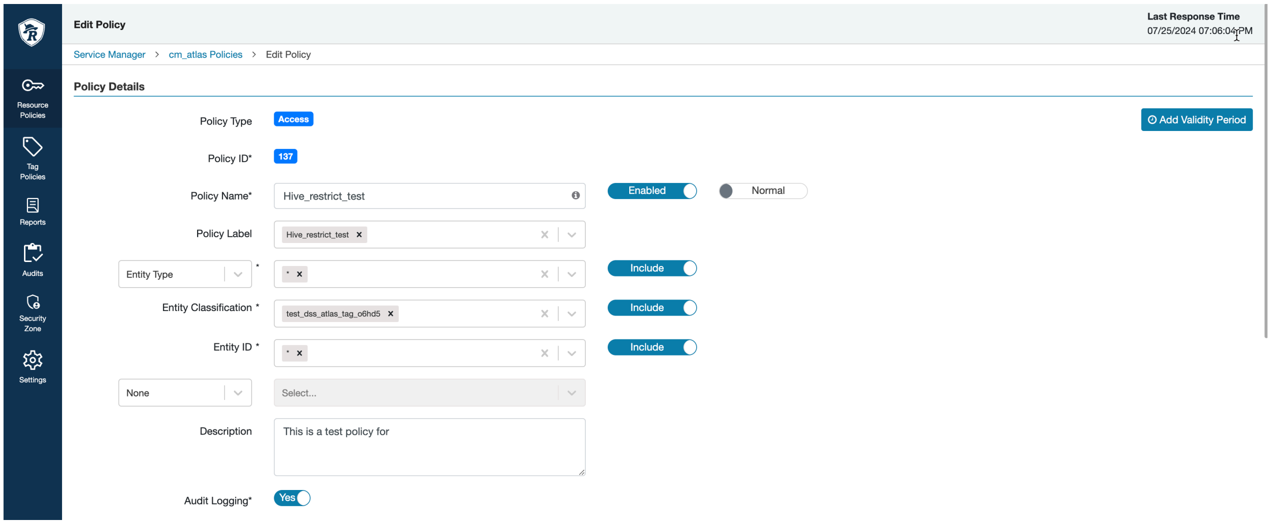
   You must select the same regular expression you had created under the Resources pane.
9. Click Save & Validate.
   The Data For Validation window appears.
10. Enter the sample values to validate if **Column Name Expression** and **Column Value Expression** entities match. Make sure that the correct data lake is selected to validate the entries.
11. Click Submit Validation.
    The status for the newly created regular expression validation is displayed on the Tags Rule tab. Once the validation is successful, you can deploy the rule.
12. Click Done.
    On the Rule Groups pane, verify if the rule is available under the Custom Deployed list. You can also suspend the rule by selecting the it from the list.

    Once the Cluster Sensitivity Profiler job or On-Demand Profiler picks up the Hive asset for profiling, the newly set up custom tag gets applied on the Hive column, provided the asset has the column(s) which meet the custom rule criteria.

Restricting access for certain users of Cloudera Data Catalog

To have a fine-grained access to the same user from accessing the assets in Cloudera Data Catalog, you can perform some additional changes. For example, if you want to restrict some users from accessing specific table information, you must set-up a Ranger policy such that these users will not have access to the asset details.

To create the Ranger policy to restrict users from accessing asset details, for example, with a specific classification, refer to the following images:



Edit Policy

Last Response Time
07/25/2024 07:06:04 PM

Service Manager > cm_atlas Policies > Edit Policy

Policy Details

Add Validity Period

Policy Type: Access

Policy ID*: 137

Policy Name*: Hive_restrict_test — Enabled — Normal

Policy Label: Hive_restrict_test

Entity Type *: * — Include

Entity Classification *: test_dss_atlas_tag_o6hd5 — Include

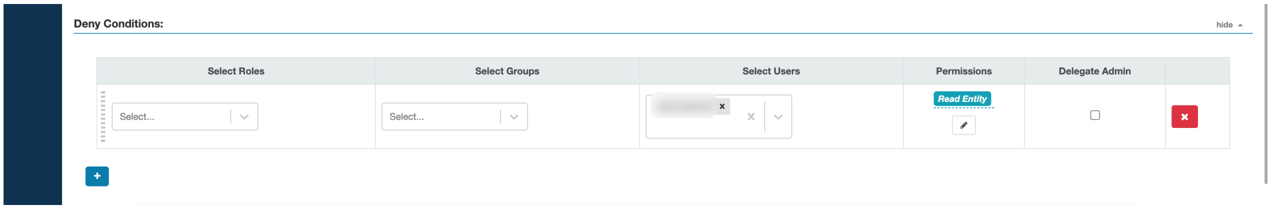
Entity ID *: * — Include

None: Select...

Description: This is a test policy for

Audit Logging*: Yes

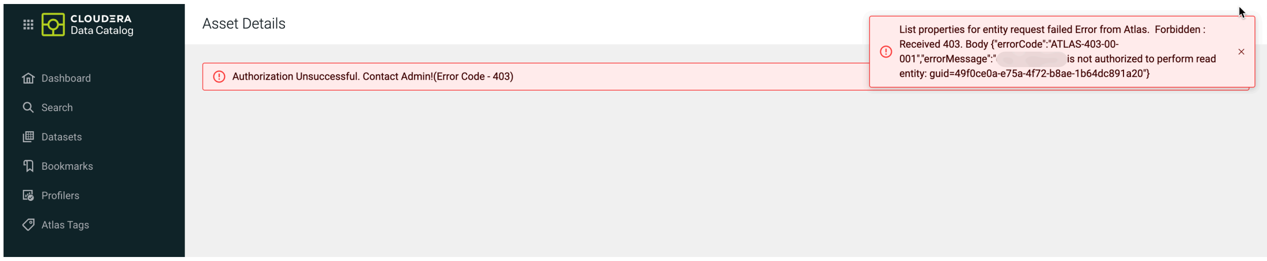
The following image displays the **Deny Conditions** set for the specific user.



Deny Conditions:

Select Roles	Select Groups	Select Users	Permissions	Delegate Admin	
Select...	Select...	[redacted]	Read Entity	☐	✖

The result is depicted in the following image, where the user has no permissions to access the specified dataset.



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":"[redacted] is not authorized to perform read entity: guid=49f0ce0a-e75a-4f72-b8ae-1b64dc891a20"}

Reducing resource consumption with restricted users

Additionally, when you plan to restrict data access, please note the following:

- Audit summarization for the asset evolves from the Ranger Audit Profiler and Metrics service.
- Various Hive Column Statistical metrics for columns of the asset evolves from Atlas as part of the profile_data of a column.

To ensure that the data related to audit summary and Hive Column Statistics are not visible to the subscribers, you must make sure to turn off the audit profiler and the Hive Column Profiler respectively.