# Data Catalog Reference

**Date published: 2022-08-21**
**Date modified: 2023-03-10**

# Legal Notice

# Contents

# 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

For more information about the CDP command-line interface and setting up the same, see CDP CLI.

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 stri
ng 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 out
put. 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-imag
e-catalog.json",
```

"catalogName": "cdp-default"

```
},
```

```
        "environmentCrn":
           "crn:cdp:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb
73d:environment:bf795226-b57c-4c4d-8520-82249e57a54f",
```

```
        "credentialCrn":
           "crn:altus:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edb
b73d: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:c
lustertemplate:16a5d8bd-66d3-42ea-8e8d-bd8765873572"
```

```
}
```

```
}
```

# Sample Workflows

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

## Tag Management and Search workflow

In Data Catalog, 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 and Column filters.

### Procedure

1. Data Catalog > Atlas Tags > Create a new tag
2. Enter the name, description for the tag.
3. Inserting 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 in Data Catalog.

   For example, we can assign the tag to the Hive table.
5. Data Catalog > Search for a Hive asset > click on the asset to view the Asset Details page.
6. Under the Manage Classification panel, click + to enable the tag list view.
7. Select the tag that you newly created and assign the tag to the asset.
8. Click Save.

   Next, you can add the tag to a column in the selected asset.
9. Data Catalog > Search for a Hive asset > click on the asset to view the Asset Details page.
10. Select the Schema tab and click Edit Tags.

11. Click + and select the tag from the tag drop-down.
    The tag list contains System Tags and Managed Tags. Managed Tags are selected by default.
12. Select a tag and click Save.
    Once tags are added to the table and column, you can search for the corresponding assets using the tag name.

    To search for the asset using the Entity filter:
13. Data Catalog > Search > Enter the tag name in the Entity filter.
    The asset for which the tag was added is displayed in the search result.

    To search for the asset using the Column filter:
14. Data Catalog > Search > Enter the tag name in the Column filter.
    The asset for whose column the tag was added is displayed in the search result.

## Auto Tagging workflow with Custom Cluster Sensitivity Profiler Rules

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

### About this task

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

### Procedure

1. Data Catalog > Profilers > Select the Tag Rules tab.
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 the 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 the tags value 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 Expression Name 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 tag by selecting the same from the list.

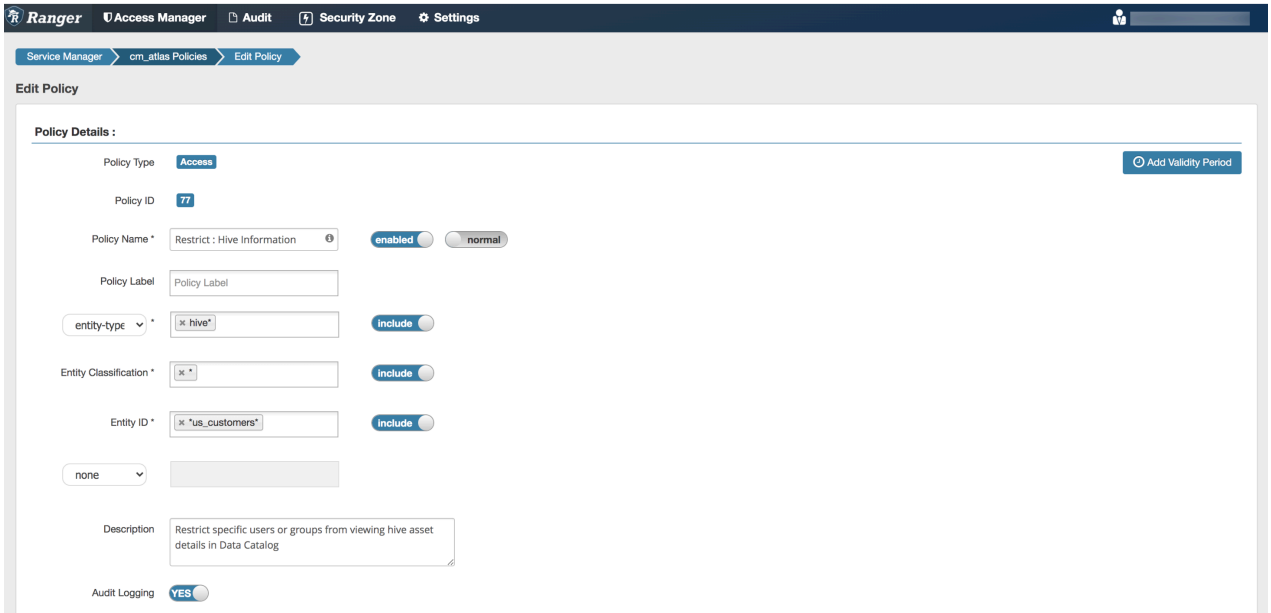
    Once the Cluster Sensitivity Profiler job or On-Demand Profiler picks up the Hive asset for profiling, the newly set-up custom tag must get applied on the Hive column, provided the asset has the column(s) which meet the custom rule criteria.

## Restricting data access for certain users of Data Catalog

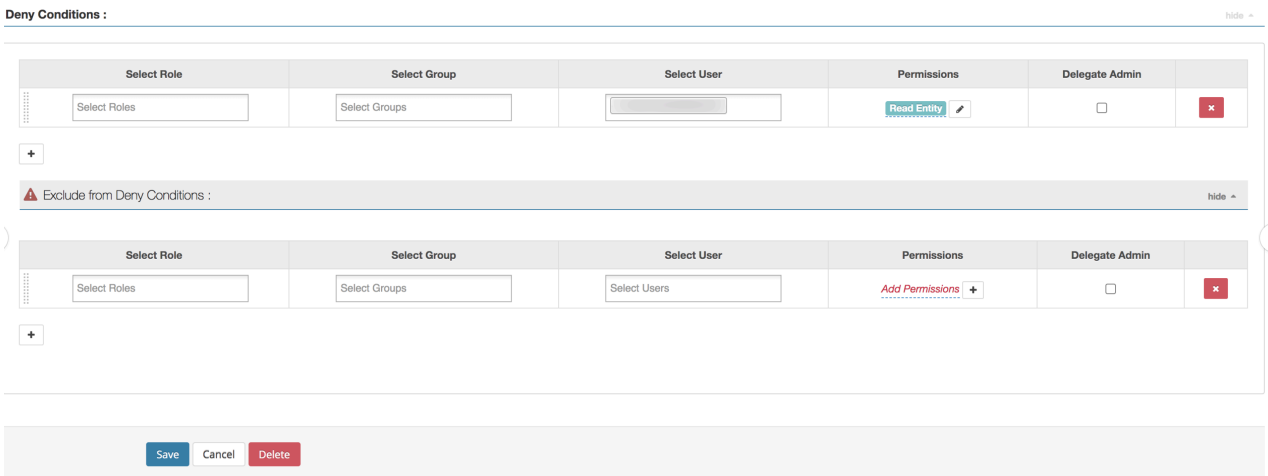
To have a fine-grained access to the user from accessing the assets in Data Catalog, you can perform some additional changes.

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 in Data Catalog.

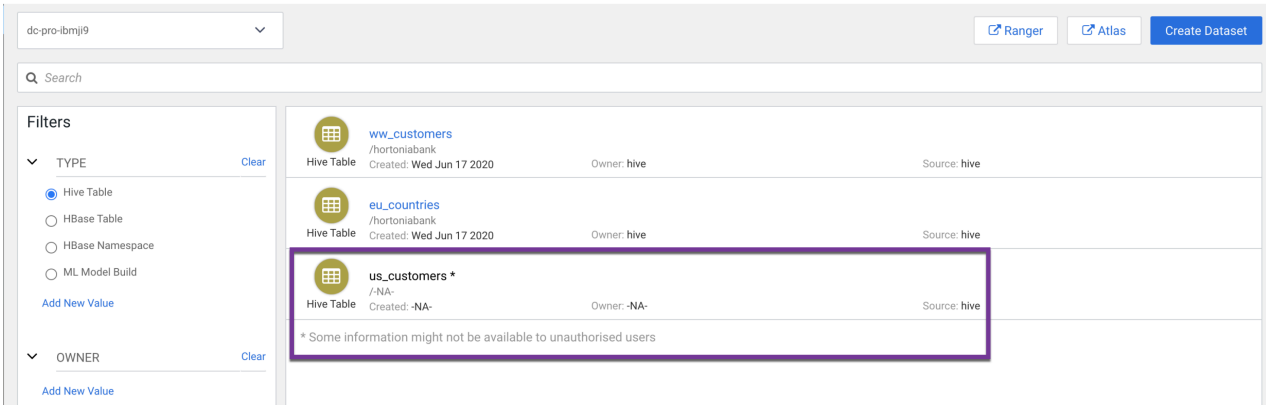
To create the Ranger policy to restrict users from accessing asset details, refer to the following images:




The next image displays the “Deny Conditions” set for the specific user.




The resultant is depicted in the following image, where the user has no permissions to access the specified dataset. In this example, it is us_customers.




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

- Audit summarisation 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 Hive Column profiler respectively.