

## Cloudera Data Catalog Top Use Cases

Date published: 2019-11-14

Date modified: 2025-06-11

# CLOUDERA

# Legal Notice

© Cloudera Inc. 2025. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

# Contents

|   |           |
|---|-----------|
| <b>Search for assets.....</b>   | <b>4</b>  |
| <b>Launching profilers in Compute Cluster enabled environments.....</b> | <b>4</b>  |
| <b>Launching profilers in VM based environments.....</b>                | <b>9</b>  |
| <b>Configuring the Activity Profiler.....</b>                           | <b>11</b> |
| <b>Configuring the Ranger Audit Profiler.....</b>                       | <b>14</b> |
| <b>Configuring the Data Compliance profiler.....</b>                    | <b>15</b> |
| <b>Configuring the Cluster Sensitivity Profiler profiler.....</b>       | <b>18</b> |
| <b>Configuring the Statistics Collector profiler.....</b>               | <b>20</b> |
| <b>Configuring the Hive Column Profiler.....</b>                        | <b>23</b> |
| <b>Atlas tag management.....</b>  | <b>26</b> |
| <b>Creating tag rules in compute cluster environments.....</b>          | <b>29</b> |
| <b>Creating tag rules in VM based environments.....</b>                 | <b>34</b> |

## Search for assets

On the Cloudera 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 **Search**, you are looking up names, types, descriptions, and other metadata collected by Cloudera 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 stored assets.

**Note:**

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

### Related Information

[Understanding datasets](#)

[Filters](#)

[Accessing Data Lakes](#)

[Download search results as CSV files](#)

[Integrating Cloudera Data Catalog with AWS Glue Data Catalog](#)

[Prerequisites for accessing Hue tables and databases](#)

[Searching for assets using Atlas glossaries](#)

[Additional search options for asset types](#)

[Accessing tables based on Ranger policies](#)

[Creating classifications for selected assets](#)

[Viewing Data Asset details](#)

## Launching profilers in Compute Cluster enabled environments

In Compute Cluster enabled environments, after you set up the profiler, the Profiler Launcher Services automatically starts the profiler Kubernetes containers.

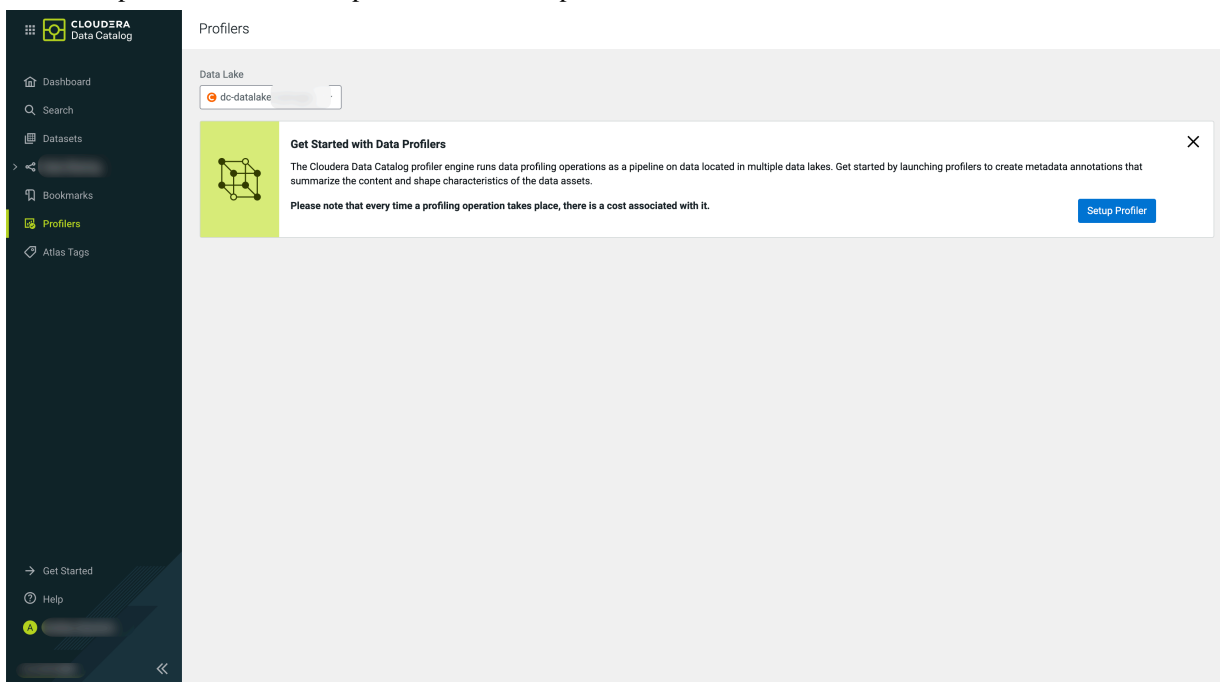


**Note:** You must be a Power User to launch a profiler cluster.

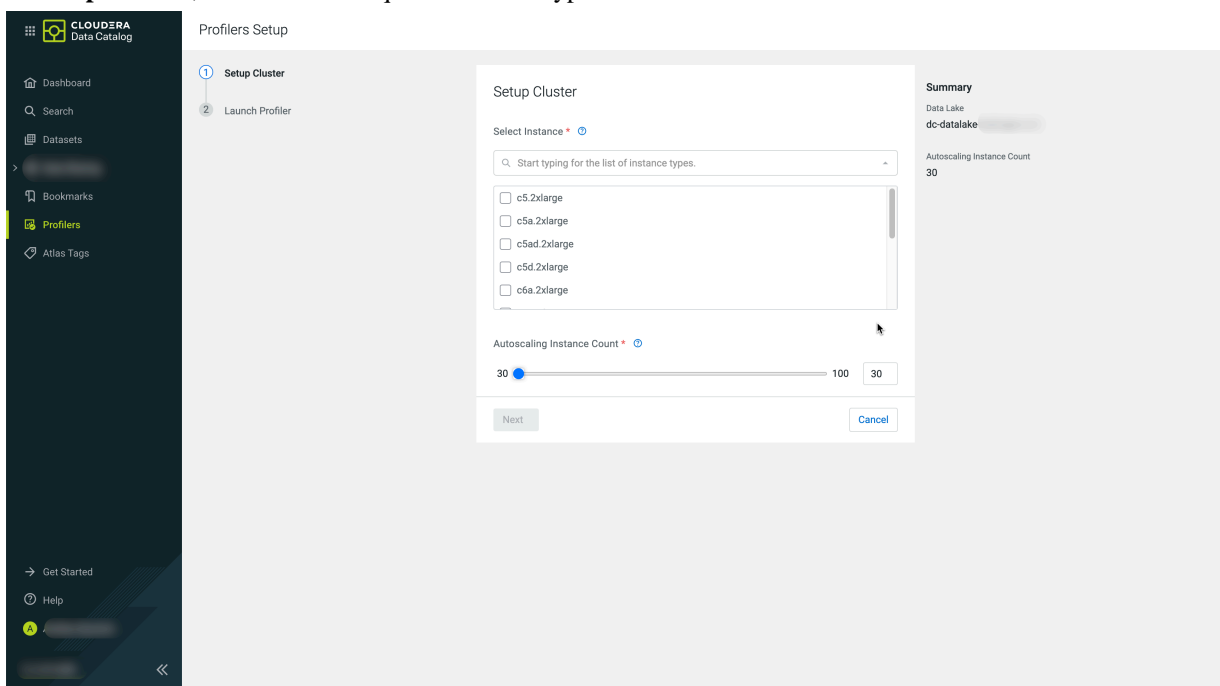
### How to launch the profiler for Compute Cluster enabled environments

1. On the **Profilers** page, select the data lake from which you want to launch the profiler cluster.

2. Click Setup Profiler, to start the profiler cluster setup.



3. In Setup Cluster, search for the required instance types:

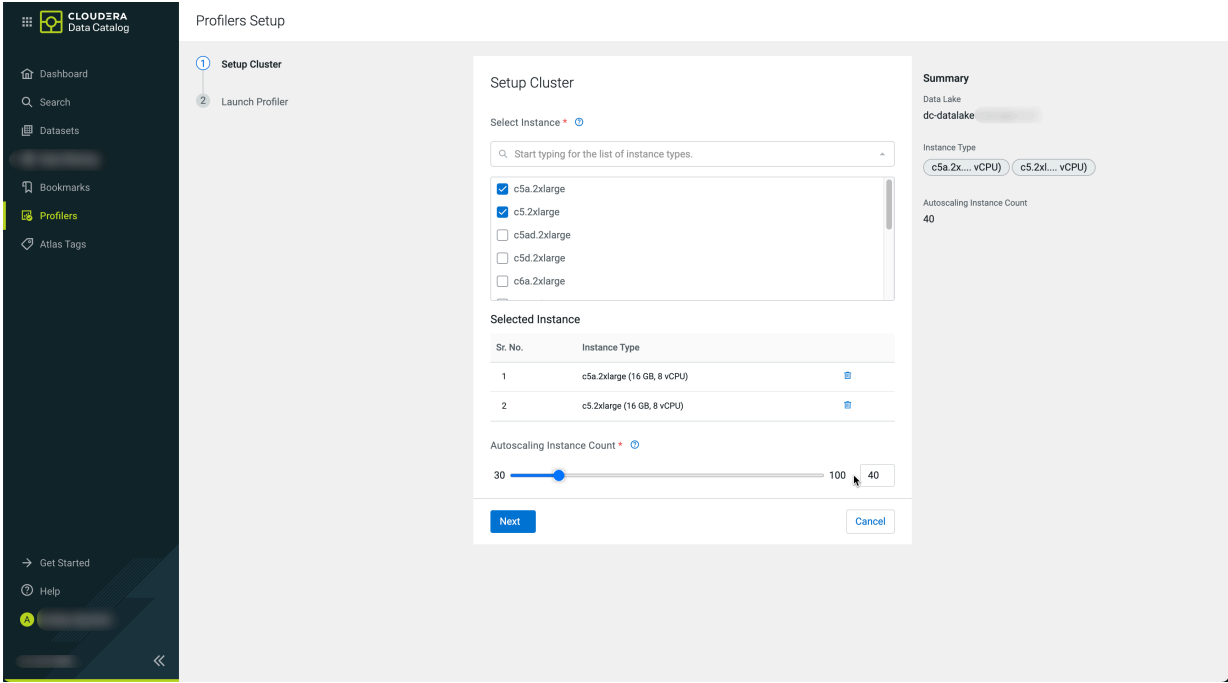


The available instance types depend on the cloud provider of the underlying environment. Choose from them based on your performance and cost requirements.



**Note:** For more information, see [Amazon EC2 Instance types](#) or [Azure Virtual Machine series](#).

4. Select your required instances and set the Autoscaling instance count to define maximum number of workers. The underlying Apache Spark service will manage the actual number of used instances based on workload.



5. Click Next.

6. Select the necessary profilers to be launched.



**Note:** Profilers can be launched later as well. Also, their configuration can be changed after launching them.

Profilers Setup

✓ Setup Cluster

2 Launch Profiler

Launch Profiler

Activity Profiler

Monitor how your data is being used and who it's used by.

Profiler Configuration :

WORKER MEM LIMIT:

4G

NUM WORKERS:

4

THREAD PER WORKER:

3

CRON EXPRESSION:

0 0 \*\*\*

Data Compliance Profiler

Ensure your data is compliant by keeping track of sensitive data types.

Profiler Configuration :

WORKER MEM LIMIT:

11G

NUM WORKERS:

10

THREAD PER WORKER:

3

CRON EXPRESSION:

0 0 \*\*\*

LAST RUN:

Over a period of 2 days

Table Statistics Profiler

Understand the shape of your data with columnar metrics.

Profiler Configuration :

WORKER MEM LIMIT:

11G

NUM WORKERS:

10

THREAD PER WORKER:

3

CRON EXPRESSION:

0 0 \*\*\*

LAST RUN:

Over a period of 2 days

Summary

Data Lake

dc-datalake-l

Instance Type

c5a.2x... vCPU) c5.2xl... vCPU)

Autoscaling Instance Count

40

Profilers

Activity Data C...liance Table ...istics

← Previous


Start Setup

Cancel


## 7. Once the cluster is ready, you can start the individual profilers by clicking Launch.

Profilers


Data Lake  
dc-datalake-hydrogen... [Refresh](#)




**Get Started with Data Profilers**  
The Cloudera Data Catalog profiler engine runs data profiling operations as a pipeline on data located in multiple data lakes. Get started by launching profilers to create metadata annotations that summarize the content and shape characteristics of the data assets. **Please note that every time you start a compute operation, there is a cost associated to it.**



**Activity Profiler**  
Monitor how your data is being used and who it is used by. [Launch](#)



**Data Compliance Profiler**  
Ensure your data is compliant by keeping track of sensitive data types. [Launch](#)




**Statistics Collector Profiler**  
Understand the shape of your data with columnar metrics. [Launch](#)

## Verifying the profiler cluster for Compute Cluster enabled environments

As a final step, you can verify that the node group is ready for the profiler jobs under the Cloudera Management Console Environments Compute Clusters Node Groups pane.

Environments / v2 / Compute Clusters



v2  
cm.cdp.environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:environment:38a40b34-89fb-4f75-a5fa-8a17b090a52e  
US West (Oregon) - us-west-2

[Stop](#) [Actions](#)

**Data Lake Details**

| NAME | NODES | SCALE      | QUICK LINKS   |
|------|-------|------------|---|
| v2   | 2 0 0 | Light Duty | <a href="#">Atlas</a> <a href="#">Ranger</a> <a href="#">Data Catalog</a> |

| STATUS  | STATUS REASON | CRN |
|---------|---------------|-----|
| Running | N/A           | ... |

Data Hubs Data Lake FreeIPA **Compute Clusters** Cluster Definitions Summary

1 Compute Clusters [Add Compute Cluster](#)

| Status  | Name   | CRN |
|---------|--|-----|
| Running | default-compute-cluster <b>Default Cluster</b> | ... |

1 - 1 of 1 |< >| Items per page: 25



default-dc-qe-env-v2-compute-cluster

STATUS: Running | CLUSTER TYPE: Default Cluster | DATE CREATED: 05/08/2024, 05:54:19 | CREATED BY: Deepak Kumar Singh

CRN: [Redacted]

Networking | Encryption | **Node Groups** | Compute Cluster Version | Labels

**Node Groups**

- dcprofiler**
  - Labels: lifite.cloudera.com/instance-group-id: ig-tp04kcyt
  - Root Volume Size (GiB): 50
  - Nodes: 1 (Auto scales between 1 and 10)
- dcprofiler-worker-spot**
  - Labels: lifite.cloudera.com/instance-group-id: ig-q12zn8wn
  - Root Volume Size (GiB): 100
  - Nodes: 0 (Auto scales between 0 and 81)
- lifite-infra**
  - Labels: role.node.kubernetes.io/lifite-infra: true
  - Taints: role.node.kubernetes.io/lifite-infra: true:NoSchedule
  - Root Volume Size (GiB): 40
  - Nodes: 2 (Auto scales between 2 and 4)

## Launching profilers in VM based environments

In VM-based environments, you must first provision the Cloudera Data Hub to launch the profiler cluster to view the profiler results for your assets.



**Note:** You must be a Power User to launch a profiler cluster.

### Profiler cluster in VM based environments

The Profiler Services supports enabling the High Availability (HA) feature.



**Note:** The profiler HA feature is under entitlement. Based on the entitlement, the HA functionality is supported on the Profiler cluster. Contact your Cloudera account representative to activate this feature in your Cloudera environment.



**Attention:** By default when you launch a profiler cluster, the instance type of the Master node will be the following based on the provider:

- AWS - m5.4xlarge
- Azure - Standard\_D16\_v3
- GCP - e2-standard-16

There are two types of Profiler Services:

- Profiler Manager
- Profiler Scheduler

The Profiler Manager service consists of profiler administrators, metrics, and data discovery services. These three entities support HA. The HA feature supports Active-Active mode.



**Important:** The Profiler Scheduler service does not support the HA functionality.

## How to launch the profiler cluster for VM based environments

On the **Search** page, select the data lake from which you want to launch the profiler cluster. Click the Get Started link to proceed.

### Profiler Setup - [redacted]

Setting up the profiler enables the cluster to fetch the data related to the profiled assets. The profiled assets contain summarized information pertaining to Cluster Sensitivity Profiler, Ranger Audit Profiler, and Hive Column Profiler.

☐ **Enable High Availability**

The Profiler High Availability (HA) cluster provides failure resilience for several of the services, including Knox, HDFS, YARN, HMS, and Profiler Manager Service. Services that do not run in HA mode yet include Cloudera Manager, Livy, and Profiler Scheduler Service.

**Setup Profiler**

For setting up the profiler, you have the option to enable or disable the HA.

### Profiler Setup - [redacted]

Setting up the profiler enables the cluster to fetch the data related to the profiled assets. The profiled assets contain summarized information pertaining to Cluster Sensitivity Profiler, Ranger Audit Profiler, and Hive Column Profiler.

☒ **Enable High Availability**

The Profiler High Availability (HA) cluster provides failure resilience for several of the services, including Knox, HDFS, YARN, HMS, and Profiler Manager Service. Services that do not run in HA mode yet include Cloudera Manager, Livy, and Profiler Scheduler Service.

When enabled, the HA Profiler cluster provides greater resiliency and scalability by using more virtual machines that incur additional corresponding cloud provider costs.

**Setup Profiler**

Once you enable HA and click Setup Profiler, Cloudera Data Catalog processes the request and the profiler creation is in progress.

Profiler Cluster is being created

| Type                                      | Name              | Qualified Name                          | Created On      | Owner | Source |
|---|-------------------|---|-----------------|-------|--------|
| <input type="checkbox"/> Azure Container  | container         | abfs://container@sparktestingstorage... | -NA-            | -NA-  | adls   |
| <input type="checkbox"/> AWS S3 V2 Bucket | s3-extractor-test | s3a://s3-extractor-test@cm              | -NA-            | -NA-  | aws    |
| <input type="checkbox"/> Hive Table       | lounge            | airline.lounge@cm                       | Mon Oct 04 2021 | hrt_1 | hive   |

Later, a confirmation message appears that the profiler cluster is created.

Profiler Cluster is provisioned successfully

| Type                                      | Name              | Qualified Name                          | Created On      | Owner | Source |
|---|-------------------|---|-----------------|-------|--------|
| <input type="checkbox"/> Azure Container  | container         | abfs://container@sparktestingstorage... | -NA-            | -NA-  | adls   |
| <input type="checkbox"/> AWS S3 V2 Bucket | s3-extractor-test | s3a://s3-extractor-test@cm              | -NA-            | -NA-  | aws    |
| <input type="checkbox"/> Hive Table       | lounge            | airline.lounge@cm                       | Mon Oct 04 2021 | hrt_1 | hive   |

Next, you can verify the profiler cluster creation under Cloudera Management Console Environments Data Hubs pane.

The newly created profiler cluster looks like the following in Cloudera Management Console:

Environments / v1 / Clusters

aws v1 US West (Oregon) - us-west-2

SDX Data Lake Details

NAME: v1

NODES: 2 (green), 0 (grey), 0 (red)

SCALE: Light Duty

QUICK LINKS: [Atlas](#) [Ranger](#) [Data Catalog](#)

STATUS: Running

STATUS REASON: N/A

CRN: [redacted]

Data Hubs Data Lake FreeIPA Compute Clusters Cluster Definitions Summary

1 Data Hubs

Search

Create Data Hub

| Status  | Name              | Data Hub Type | Runtime | Node Count | Created            |
|---------|-------------------|---------------|---------|------------|--------------------|
| Running | profiler_7_2_18-0 |               | 7.2.18  | 3          | 8/2/2024, 08:36:00 |

1 - 1 of 1 | < > | Items per page: 25

## Configuring the Activity Profiler

Configure the scheduling and the available resources for your profiler.

### Procedure

1. Go to **Profilers** and select your data lake.
2. Go to **Profilers Activity Profiler Profiler Details Configuration All Configurations**

3. Select a schedule to run profiler using either UNIX Cron Expression or the Basic scheduler.



**Note:** Unix (in Compute Cluster enabled environments) cron jobs use the UTC timezone instead of the local timezone of the user.

**Figure 1: Profiler schedule with cron expression**

The screenshot shows the 'Profiler Configuration' window. Under the 'Schedule' section, the 'Cron Expression' radio button is selected. Below it, there is a text input field for the cron expression. A tooltip is visible, stating: 'CRON expression for profiling job which will be run according to UTC. A sample expression is [30 7 \* \* \*] for running jobs at 07:30(am) everyday'.

**Figure 2: Profiler schedule with natural language**

The screenshot shows the 'Profiler Configuration' window. Under the 'Schedule' section, the 'Basic' radio button is selected. The scheduling is configured using natural language: 'At 4 minute of 4 hours on 1 st day of January month on every day of week'. The 'Time Zone' is set to 'UTC'.

**4. Continue with resource settings:****a) Set the Maximum number of executors**

Indicates the number of processes that are used by the distributed computing framework. The recommended value is at least four executors.

**b) Set the Maximum cores per executor**

Indicates the maximum number of cores that can be allocated to an executor.

**c) Set the Executor memory limit in GBs**

**Maximum number of executors \*** 

4

**Maximum cores per Executor \*** 

3

**Executor memory limit in GBs \*** 

4G

**Save**

**Cancel**

**5. Click Save to apply the configuration changes to the selected profiler.**

# Configuring the Ranger Audit Profiler

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

Procedure

- 1. Go to **Profilers** and select your data lake.
- 2. Go to **Profilers Configs**.
- 3. Select **Ranger Audit Profiler**.  
The **Detail** page is displayed.
- 4.



Use the toggle button to enable or disable the profiler.

- 5. Select a schedule to run the profiler using a quartz cron expression.



**Note:** Quartz CRON jobs (in VM-based environments) use the UTC timezone instead of the local timezone of the user.

Detail

Ranger Audit Profiler

Data Lake: dc-env1

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

Active

Schedule\*

0 \*/30 \* ? \* \*

^ Advanced Options

Number of Executors\*

1

Executor Cores\*

1

Executor Memory (in GB)\*

1

Driver Core\*

1

Driver Memory (in GB)\*

1

Save

Cancel

6. Continue with the resource settings.

- In **Advanced Options**, set the following:

- Number of Executors - Enter the number of executors to launch for running this profiler.
- Executor Cores - Enter the number of cores to be used for each executor.
- Executor Memory - Enter the amount of memory in GB to be used per executor process.
- Driver Cores - Enter the number of cores to be used for the driver process.
- Driver Memory - Enter the memory to be used for the driver processes.



**Note:** For more information, see [Configuring SPARK on YARN Applications](#) and [Tuning Resource Allocation](#).

7. Click Save to apply the configuration changes to the selected profiler.

## Configuring the Data Compliance profiler

You can configure the scheduling and the available resources for your profiler.

### Procedure

1. Go to **Profilers** and select your data lake.
2. Go to **Profilers Data Compliance Profiler Details Configuration All Configurations**
3. Select a schedule to run profiler using either UNIX Cron Expression or the Basic scheduler.



**Note:** Unix (in Compute Cluster enabled environments) cron jobs use the UTC timezone instead of the local timezone of the user.

**Figure 3: Profiler schedule with cron expression**

The screenshot shows the 'Profiler Configuration' window. Under the 'Schedule' section, the 'Cron Expression' radio button is selected. Below it, there is a text input field for the cron expression. A tooltip is visible, stating: 'CRON expression for profiling job which will be run according to UTC. A sample expression is [30 7 \* \* \*] for running jobs at 07:30(am) everyday'.

**Figure 4: Profiler schedule with natural language**

The screenshot shows the 'Profiler Configuration' window. Under the 'Schedule' section, the 'Basic' radio button is selected. Below it, the natural language schedule is configured as: 'At 4 minute of 4 hours on 1 st day of January month on every day of week'. The 'Time Zone' is set to 'UTC'.

4. Select Last Run Check and set a period in Day Range if needed.

**Note:**

The Last Run Check enables profilers to avoid profiling the same asset on each scheduled run.

If you have scheduled a cron job, for example set to start in about an hour, and have enabled the Last Run Check configuration for two days, this setup ensures that the job scheduler filters out any asset which was already profiled in the last two days.

If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.



**5. Continue with resource settings:**

## a) Set the Maximum number of executors

Indicates the number of processes that are used by the distributed computing framework. The recommended value is at least 10 executors.

## b) Set the Maximum cores per executor

Indicates the maximum number of cores that can be allocated to an executor.

## c) Set the Executor memory limit in GBs

Maximum number of executors \* 

4

Maximum cores per Executor \* 

3

Executor memory limit in GBs \* 

4G

Save

Cancel



**6. Click Save to apply the configuration changes to the selected profiler.**

## 7. Add **Asset Filtering Rules** as needed to customize the selection of assets to be profiled.



### Note:

- Profiler configurations apply to both scheduled and on-demand profiler jobs.
- In Compute Cluster environments, you cannot enable conflicting Allow and Deny list rules at the same time. Enabling conflicting rules results in an error message.

 Request to create profiler asset filter rule failed. One or more rules with the same condition already exist in your Allow or Deny list. In case it is in the other list, you can disable the rule from that list and retry. 

### a) Set your **Deny List** and **Allow-list**.


The profiler will skip profiling assets that meet any criteria in the **Deny List** and will include assets that meet any criteria in the **Allow List**.

1. Select the **Deny-list** or **Allow List** tab.
2. Click Add New Rule to define new rules.
3. Select the key from the drop-down list and the relevant operator. You can select from the following:

| Key              | Operator   |
|------------------|--|
| Database name    | <ul style="list-style-type: none"> <li>• equals</li> <li>• starts with</li> <li>• ends with</li> </ul>                     |
| Name (of asset)  | <ul style="list-style-type: none"> <li>• equals</li> <li>• contains</li> <li>• starts with</li> <li>• ends with</li> </ul> |
| Owner (of asset) |  |
| Creation date    | <ul style="list-style-type: none"> <li>• greater than</li> <li>• less than</li> </ul>                                      |

4. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
5. Click Add Rule. Once a rule is added (enabled by default), you can toggle the state of the new rule to enable it or disable it as needed.



**Note:** You can check the list of asset impacted by your rule by clicking  > Affected Assets.

## Configuring the Cluster Sensitivity Profiler profiler

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

### Procedure

1. Go to **Profilers** and select your data lake.
2. Go to **Profilers Configs**.

### 3. Select Cluster Sensitivity Profiler.

The **Detail** page is displayed which contains the following sections:

Detail

## Cluster Sensitivity Profiler

Data Lake: **dc-env1**

The Cluster Sensitivity Profiler automatically performs context and content inspection to detect various types of sensitive data. It also suggests suitable classifications or tags based on the type of sensitive content detected or discovered.

☒ Active

Schedule\*  
0 20 \* \* \* ?

Last Run Check\* ☒  
2 Days

Sample Data Size\*  
Number of Rows 100

^ Advanced Options

Number of Executors\* 1

Executor Cores\* 1

Executor Memory (in GB)\* 1

Driver Core\* 1

Driver Memory (in GB)\* 1

### 4.



Use the toggle button to enable or disable the profiler.

### 5. Select a schedule to run the profiler. This is implemented as a quartz cron expression.

For more information, see [Understanding the Cron Expression generator](#).

### 6. Select Last Run Check and set a period if needed.



#### Note:

The Last Run Check enables profilers to avoid profiling the same asset on each scheduled run.

If you have scheduled a cron job, for example set to start in about an hour, and have enabled the Last Run Check configuration for two days, this setup ensures that the job scheduler filters out any asset which was already profiled in the last two days.

If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.

### 7. Set the sample settings for VM-based environments:

#### a. Select the **Sample Data Size**.

1. From the drop down, select the type of sample data size.
2. Enter the value based on the previously selected type.

8. Continue with the resource settings.

a. In **Advanced Options**, set the following:

- Number of Executors - Enter the number of executors to launch for running this profiler.
- Executor Cores - Enter the number of cores to be used for each executor.
- Executor Memory - Enter the amount of memory in GB to be used per executor process.
- Driver Cores - Enter the number of cores to be used for the driver process.
- Driver Memory - Enter the memory to be used for the driver processes.



**Note:** For more information, see [Configuring SPARK on YARN Applications](#) and [Tuning Resource Allocation](#).

9. Click Save to apply the configuration changes to the selected profiler.

10. Add **Asset Filter Rules** as needed to customize the selection of assets to be profiled.



**Note:**

- Profiler configurations apply to both scheduled and on-demand profiler jobs.
- In VM based environments, Deny lists are prioritized over Allow lists.

For example adding a regular expression for a database to the Deny list and adding a regular expression for a table within the first database to the Allow list will result in both entities filtered out. On the other hand, you can include all entities except one from a database by adding the database to the Allow list. Then, add the particular entity from the database to the Deny List.

a) Set your **Deny List** and **Allow-list**.

The profiler will skip profiling assets that meet any criteria in the **Deny List** and will include assets that meet any criteria in the **Allow List**.

1. Select the **Deny-list** or **Allow List** tab.
2. Click Add New to define new rules.
3. Select the key from the drop-down list and the relevant operator. You can select from the following:

| Key              | Operator   |
|------------------|--|
| Database name    | <ul style="list-style-type: none"> <li>• equals</li> <li>• starts with</li> <li>• ends with</li> </ul>                     |
| Name (of asset)  | <ul style="list-style-type: none"> <li>• equals</li> <li>• contains</li> <li>• starts with</li> <li>• ends with</li> </ul> |
| Owner (of asset) |  |
| Creation date    | <ul style="list-style-type: none"> <li>• greater than</li> <li>• less than</li> </ul>                                      |

4. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
5. Click Add Rule. Once a rule is added (enabled by default), you can toggle the state of the new rule to enable it or disable it as needed.

## Configuring the Statistics Collector profiler

You can configure the scheduling and the available resources for your profiler.

### Procedure

1. Go to **Profilers** and select your data lake.

2. Select a schedule to run profiler using either UNIX Cron Expression or the Basic scheduler



**Note:** Unix (in Compute Cluster enabled environments) cron jobs use the UTC timezone instead of the local timezone of the user.

**Figure 5: Profiler schedule with cron expression**

The screenshot shows the 'Profiler Configuration' form. Under the 'Schedule' section, the 'Cron Expression' radio button is selected. Below it, there is a text input field for the cron expression. A tooltip is visible, stating: 'CRON expression for profiling job which will be run according to UTC. A sample expression is [30 7 \* \* \*] for running jobs at 07:30(am) everyday'.

**Figure 6: Profiler schedule with natural language**

The screenshot shows the 'Profiler Configuration' form. Under the 'Schedule' section, the 'Basic' radio button is selected. Below it, the natural language scheduler is configured with the following values: 'At 4 minute of 4 hours on 1 st day of January month on every day of week'. The 'Time Zone' is set to 'UTC'.

3. Select Last Run Check and set a period in Day Range if needed.



**Note:**

The Last Run Check enables profilers to avoid profiling the same asset on each scheduled run.

If you have scheduled a cron job, for example set to start in about an hour, and have enabled the Last Run Check configuration for two days, this setup ensures that the job scheduler filters out any asset which was already profiled in the last two days.

If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.

**4. Continue with resource settings:**

## a) Set the Maximum number of executors

Indicates the number of workers that are used by the distributed computing framework. The recommended value is at least 10 executors.

## b) Set the Maximum cores per executor

Indicates the maximum number of cores that can be allocated to an executor.

## c) Set the Executor memory limit in GBs

Maximum number of executors \* 

4

Maximum cores per Executor \* 

3

Executor memory limit in GBs \* 

4G

Save



Cancel

**5. Click Save to apply the configuration changes to the selected profiler.**

6. Add **Asset Filtering Rules** as needed to customize the selection and deselection of assets which the profiler profiles.

**Note:**

- Profiler configurations apply to both scheduled and on-demand profiler jobs.
- In Compute Cluster environments, you cannot enable conflicting Allow and Deny list rules at the same time. Enabling conflicting rules results in an error message.

 **Request to create profiler asset filter rule failed. One or more rules with the same condition already exist in your Allow or Deny list. In case it is in the other list, you can disable the rule from that list and retry.** 

- a) Set your **Deny List** and **Allow-list**.


The profiler will skip profiling assets that meet any criteria in the **Deny List** and will include assets that meet any criteria in the **Allow List**.

1. Click Add New Rule to define new rules.
2. Use the radio buttons to define your new rule for the Allow or Deny List.
3. Select the key from the drop-down list and the relevant operator. You can select from the following:

| Key              | Operator   |
|------------------|--|
| Database name    | <ul style="list-style-type: none"> <li>• equals</li> <li>• starts with</li> <li>• ends with</li> </ul>                     |
| Name (of asset)  | <ul style="list-style-type: none"> <li>• equals</li> <li>• contains</li> <li>• starts with</li> <li>• ends with</li> </ul> |
| Owner (of asset) |  |
| Creation date    | <ul style="list-style-type: none"> <li>• greater than</li> <li>• less than</li> </ul>                                      |

4. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
5. Click Add Rule. Once a rule is added (enabled by default), you can toggle the state of the new rule to enable it or disable it as needed.



**Note:** You can check the list of assets impacted by your rule by clicking  > Affected Assets.

## Configuring the Hive Column Profiler

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

### Procedure

1. Go to **Profilers** and select your data lake.
2. Go to **Profilers Configs**.

- 3. Select Hive Column Profiler.  
The **Detail** page is displayed.

Detail

## Hive Column Profiler

Data Lake: **dc-env1**

With the Hive Column Profiler, you can view the shape or distribution characteristics of the columnar data within a Hive table.

☒ Active

Schedule\*

0 0 0/6 1/1 \* ? \*

Last Run Check\* ☒

1 Day

Sample Data Size \*

Sample Percentage ▾

100

^ **Advanced Options**

Number of Executors\*

1

Executor Cores\*

1

Executor Memory (in GB)\*

1

Driver Core\*

1

Driver Memory (in GB)\*

1

- 4.

Use the toggle button

☒ Active

to enable or disable the profiler.



5. Select a schedule to run the profiler. This is implemented as a quartz cron expression.



**Note:** Quartz CRON jobs (in VM-based environments) use the UTC timezone instead of the local timezone of the user.

For more information, see [Understanding the Cron Expression generator](#).

6. Select Last Run Check and set a period if needed.



**Note:**

The Last Run Check enables profilers to avoid profiling the same asset on each scheduled run.

If you have scheduled a cron job, for example set to start in about an hour, and have enabled the Last Run Check configuration for two days, this setup ensures that the job scheduler filters out any asset which was already profiled in the last two days.

If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.

7. Set the sample settings:

- a. Select the **Sample Data Size**.

1. From the drop down, select the type of sample data size.
2. Enter the value based on the previously selected type.

8. Continue with the resource settings.

- a. In **Advanced Options**, set the following:

- Number of Executors - Enter the number of executors to launch for running this profiler.
- Executor Cores - Enter the number of cores to be used for each executor.
- Executor Memory - Enter the amount of memory in GB to be used per executor process.
- Driver Cores - Enter the number of cores to be used for the driver process.
- Driver Memory - Enter the memory to be used for the driver processes.



**Note:** For more information, see [Configuring SPARK on YARN Applications](#) and [Tuning Resource Allocation](#).

9. Click Save to apply the configuration changes to the selected profiler.

**10. Add Asset Filter Rules** as needed to customize the selection and deselection of assets which the profiler profiles.



**Note:**

- Profiler configurations apply to both scheduled and on-demand profiler jobs.
- In VM based environments, Deny lists are prioritized over Allow lists.

For example adding a regular expression for a database to the Deny list and adding a regular expression for a table within the first database to the Allow list will result in both entities filtered out. On the other hand, you can include all entities except one from a database by adding the database to the Allow list. Then, add the particular entity from the database to the Deny List.

a) Set your **Deny List** and **Allow-list**.

The profiler will skip profiling assets that meet any criteria in the **Deny List** and will include assets that meet any criteria in the **Allow List**.

1. Select the **Deny-list** or **Allow List** tab.
2. Click Add New to define new rules.
3. Select the key from the drop-down list and the relevant operator. You can select from the following:

| Key              | Operator   |
|------------------|--|
| Database name    | <ul style="list-style-type: none"> <li>• equals</li> <li>• starts with</li> <li>• ends with</li> </ul>                     |
| Name (of asset)  | <ul style="list-style-type: none"> <li>• equals</li> <li>• contains</li> <li>• starts with</li> <li>• ends with</li> </ul> |
| Owner (of asset) |  |
| Creation date    | <ul style="list-style-type: none"> <li>• greater than</li> <li>• less than</li> </ul>                                      |

4. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
5. Click Add Rule. Once a rule is added (enabled by default), you can toggle the state of the new rule to enable it or disable it as needed.

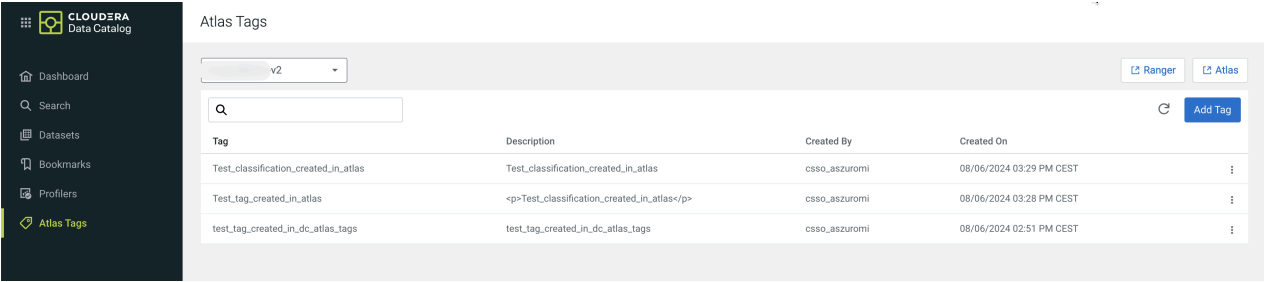
## Atlas tag management

From the Atlas Tags menu, you can create, modify, and delete any of the Apache Atlas classifications.

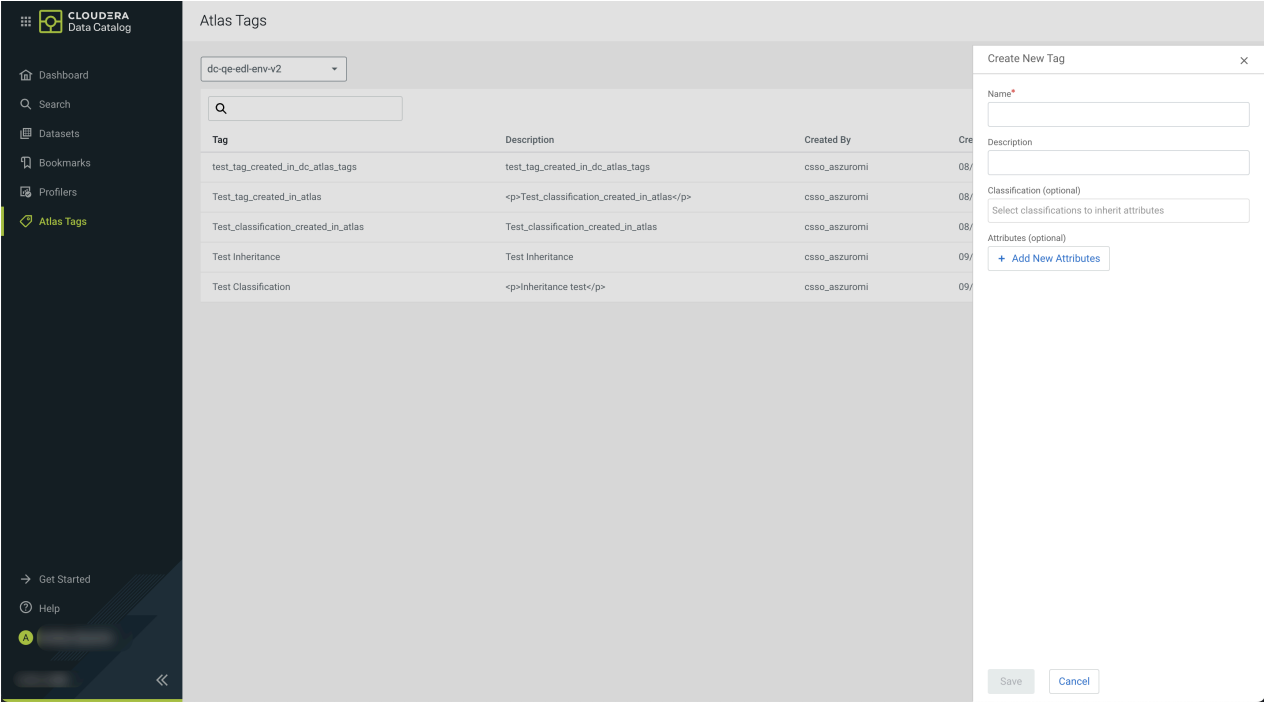
Atlas Tags allows the user to perform the following activities with a selected data lake for tag management:

- Selecting a data lake
- Searching for a tag
- Adding a tag
- Editing a tag
- Deleting a tag

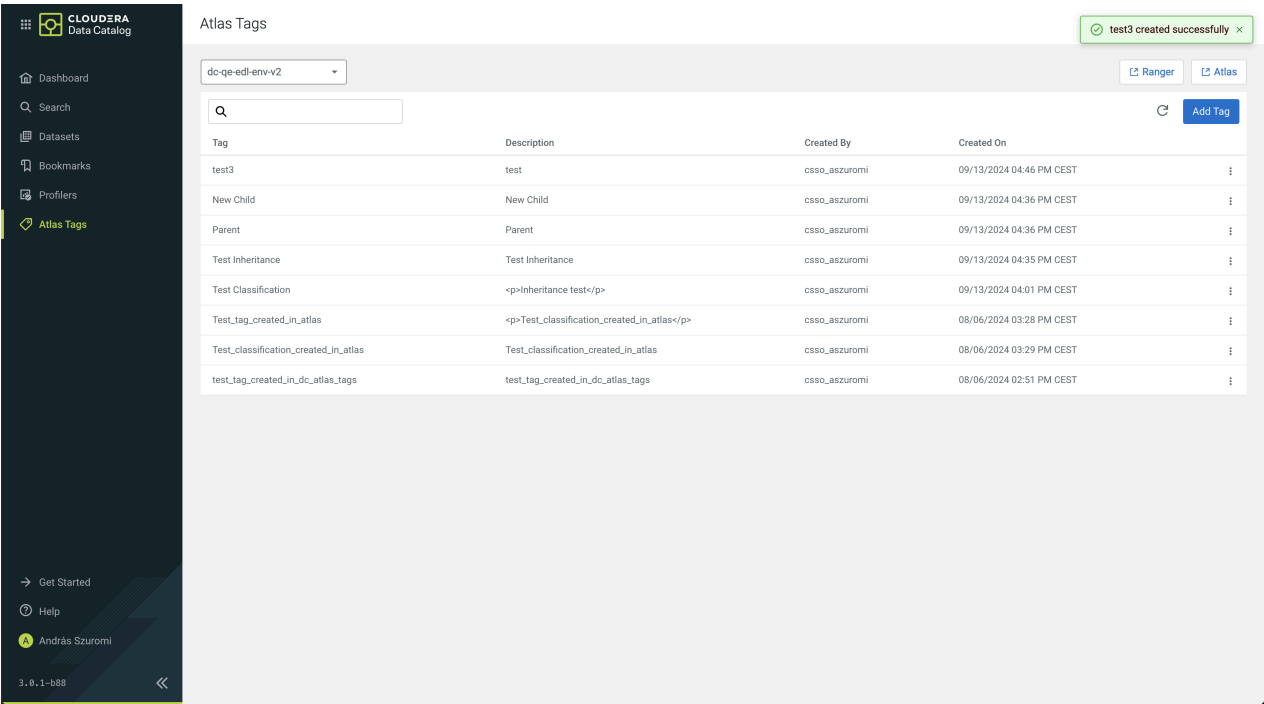
You can create a new Cloudera Data Catalog tag in the **Atlas Tags**, which are synced to Atlas. Click Add Tag to open the **Create a new tag** page.



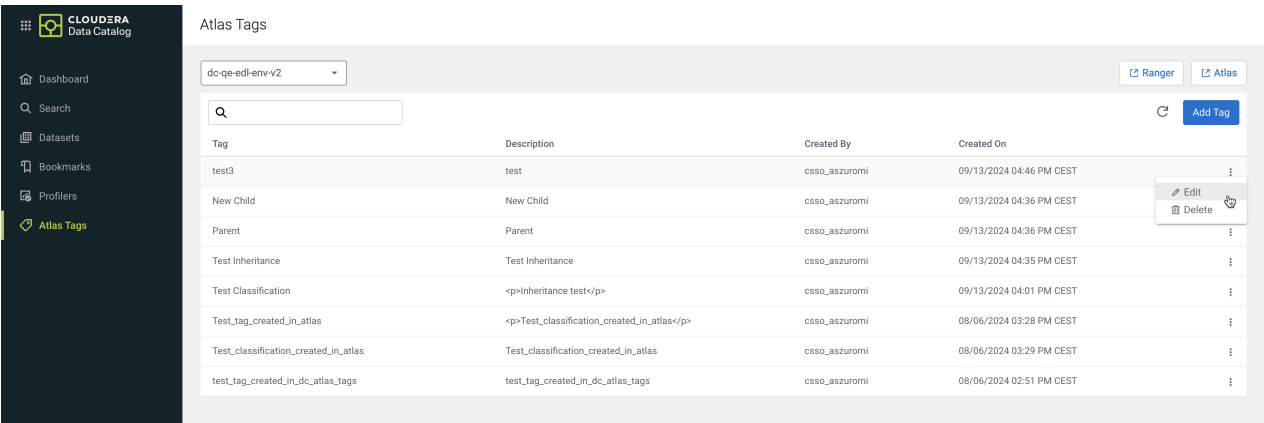
In **Create New Tag**, you can define the tag name, description and the "super-classification" from which the attributes are inherited for the sub-classification (or tag in Cloudera Data Catalog)



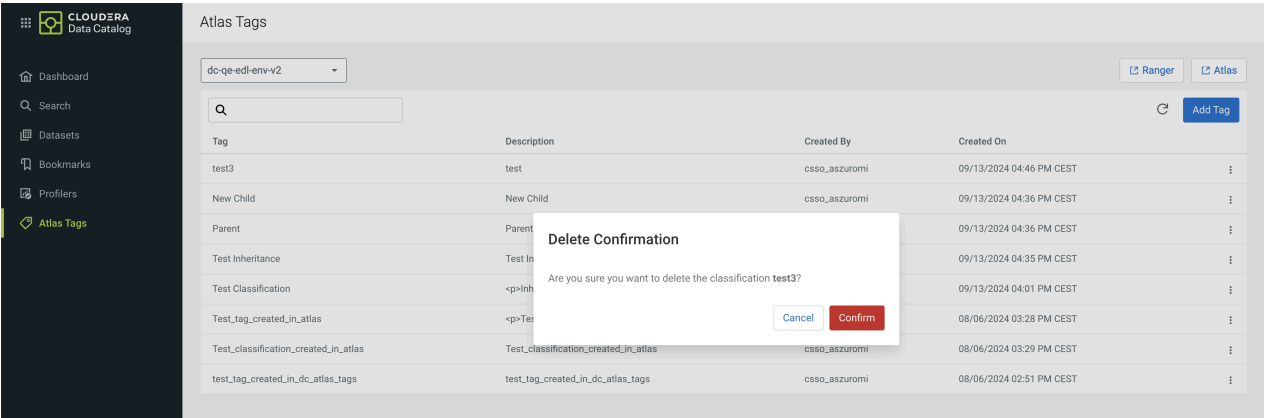
You can add or update Atlas tags. The created or updated tag is highlighted in the tag list as seen in the following diagram.



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.



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



Related Information  
Propagated asset tagging

## Creating tag rules in compute cluster environments in VM based environments

## Creating tag rules in compute cluster environments

With tag rules, you can apply Apache Atlas classifications to your assets based on regex expressions or similarity to a set of values in a table.

### About this task

### Procedure

1. To start applying tags, go to **Profilers** and select your data lake.
2. Go to **Profilers Data Compliance Tag Rules**.
3. Click **+ Create Tag Rule**.
4. Name your tag rule and add a description to it in **General Information**.

Create Tag Rule

- 1 General Information
- 2 Configure Tag Rule
- 3 Test Tag Rule
- 4 Review

### General Information

**About**

Tag Rule Name \*

Description \*

**Tags**

In Atlas, your tags appear as classifications. Atlas classifications / Data Catalogs tags are synchronized between both services.

[Create New Atlas Tag](#)

**SELECT TAGS**

Select tags to add them to your rule.

[Refresh Atlas Tag](#)

**Selected Parent Tags**

| Parent Tags                         | Children Tags   |
|-------------------------------------|---|
| <input checked="" type="radio"/> dp | <input type="radio"/> dp_HRV...ection <input type="radio"/> dp_ukp...number +74 |

**Selected Child Tags**

| Children Tags                                    | Parent Tags              |
|--|--------------------------|
| <input checked="" type="radio"/> dp_HRV...ection | <input type="radio"/> dp |

**Data Pattern Type**

☒ **Regular Expression**  
 Generate an expression manually or by file upload to create a data pattern.

☐ **Single Column File Upload**  
 Upload a file that contains all potential values for classification in a single column.

[Next →](#) [Cancel](#)

**General Information**

TAG RULE NAME  
**Test**

DESCRIPTION  
**test**

PARENT TAG  
☒ dp

Child TAG  
☒ dp\_HRV...ection

5. Select the tags to be applied from the list of available tags synchronized from the list of Atlas classifications. If you select a child tag, its parent tag is also automatically selected. By default, if the child tag is applied to a column, the table receives the parent tag.

**6. Select your Data Pattern Type:****Option****Regular Expression**

You can upload a text file containing your regex expression or directly type it in the **Configure Tag Rule** page. The required format of the CSV file can be seen by clicking [Download Sample Tag Rule](#).

Continue in step [7](#) on page 30.

**Single Column File Upload**

Upload a CSV file with values to be matched against the actual values in your tables. After uploading your file, continue with step [11](#) on page 31.

Creating regular expression based tag rule:

**7. Define your regular expression for the table name.**

**Note:** Cloudera recommends using PCRE2 compatible regular expressions. Non-compliant regular expressions may show reduced performance.

For more information, see [PCRE - Perl Compatible Regular Expressions](#).

8. When using **Column Level** regex expressions, you can define multiple expression for both of the following:

- Column Name
- Column Values

Create Tag Rule



**Note:** Regular expressions matching the same type of entity (column name or value) have the OR logical relationship between them. When using multiple regular expressions of the same type (table name, column name or value), even if one of the regular expressions match, it is considered as a match.

9. Define the Column Value Weightage in percentage with the slider.

The remainder percentage is the column name weightage percentage. The results of the individual regex matches are weighted according to this setting before determining the final result confidence for applying the tag.



**Note:** A correctly formatted file is automatically processed by Cloudera Data Catalog. All details will be filled in this case.

Tag rule testing:

10. You can make a sanity check of your tag rule in **Test Tag Rule** by uploading a sample dataset in CSV format.




**Note:** A final test called "Dry Run" is still needed to be passed to enable your tag rule.

11. Review all your input before clicking Create Tag Rule.

a) Click Confirm to finalize your tag rule.

Your tag rule is created with **Status Disabled** (🔒) and the **Test Status** will be Test Pending.

- Click  > Dry Run.

### Profilers Details

/ [Profilers](#) / [Profilers Details](#)

Data Compliance Profiler
 

Disable Profiler

RECENT JOB ID

WAUKUBUB

TOTAL JOBS

75

TOTAL PROFILED ASSETS

1219

LAST RUN

04/01/2025 10:15 PM CEST

NEXT RUN

04/01/2025 11:15 PM CEST

SCHEDULE FREQUENCY (UTC)

every hour at minute 15

Job History

Configuration

Tag Rules

Status

Associated Tag

Rule Type

Last Modified By

✕ Clear All

Create Tag Rule

| Status | Name                   | Parent Tags      | Child Tags       | Rule Type | Last Modified By | Modified On              | Validation Status | Action                    |
|--------|------------------------|------------------|------------------|-----------|------------------|--------------------------|-------------------|---------------------------|
|        | test_aadhar_rule       | teuhb.....r_card |                  | Custom    |                  | 04/01/2025 01:47 PM CEST | Dry Run Pending   | ⋮                         |
|        | AUT_Passport_Detection | dp               | dp_AUT.....ction | System    | NA               | 01/13/2025 08:09 AM CEST | Validated         | Edit<br>Dry Run<br>Delete |
|        | LVA_IBAN_Detection     | dp               | dp_LVA.....ction | System    | NA               | 01/13/2025 08:09 AM CEST | Validated         | ⋮                         |

The **Dry Run Test** pane opens.



13. Click Run to start an on-demand dry run profiling job on up to 10 tables from your data.

>>

Dry Run Test

Test Connection with Catalog Data

customer

☒ test123.customer\_iceberg

☐ test123.customer\_parquet


Selected Assets

| Sr. No. | Asset Name               |  |
|---------|--------------------------|--|
| 1       | test123.customer_iceberg |  |

Start Run

Close

Your tag rule becomes VALIDATED after a successful dry run.

14. After the "Dry run" test was passed, click  > Enable to start your using your tag rule on your live data.

## Creating tag rules in VM based environments

With tag rules, you can apply Apache Atlas classifications to your assets based on regex expressions.

### About this task

#### Procedure

1. To start applying tags, go to **Profilers** and select your data lake.
2. Go to Profilers Tag Rules .
3. Click + New.
4. Name your tag rule and add a description to it.
5. Select the tags to be applied from the list of available tags synchronized from the list of Atlas classifications.  
Multiple tags can be selected.

6. In **Column Name Expression**, select at least one regular expression to use a match it against for column names.  
Select from the same regular expression you had created under the **Resources** pane.

## Resources

▼ Regex



DeployRegex1669236475651

SampleRegex\_1586378290804

DeployRegex1670015816812

SampleRegex\_1.6183997393e+1

SampleRegex\_1618318507327

DeployRegex1670618720012

SampleRegex\_1.61840620023e+1



**Note:** You can select multiple expressions connected by AND, OR, NOT logical operators.

#### Tag Rules

#### Custom Rule

**Name \***

**Description**

**Tags \***

this\_is\_test\_tag x

**Column Name Expression**

**Column Value Expression**

#### Resources

Regex

- DeployRegex1669236475651
- SampleRegex\_1586378290804
- DeployRegex1670015816812
- SampleRegex\_1.6183997393e+1
- SampleRegex\_1618318507327
- DeployRegex1670618720012
- SampleRegex\_1.61849620033e+
- SampleRegex\_1.58583859178e+

- In **Column Value Expression**, select at least one regular expression to use a match it against for column names. The **Column Name Expression** matches are considered with a 15% weightage in the final score when calculating if the tag needs to be applied. The **Column Value Expression** matches receive the remaining 85% weightage. The column name expression results are binary (TRUE, FALSE), while by column value a certain ratio of all values can be matched.
- Click Save & Validate.

9. Enter some sample data manually to check the validity of your regular expression, then click Submit Validation.

## Data For Validation

Sample to test column name expression

sales\_property

Sample to test column value expression

sales\_property

Datalake where the validation will run

dc-profiler ▼

Close

Submit Validation

The status for the newly created regular expression validation is displayed on the **Tags Rules** tab. Once the validation is successful, you can deploy the rule.