CDP Public Cloud / Data Catalog

Profilers in VM Based Environments

Date published: 2019-11-14
Date modified: 2025-10-17

https://docs.cloudera.com/

Legal Notice

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.

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 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:

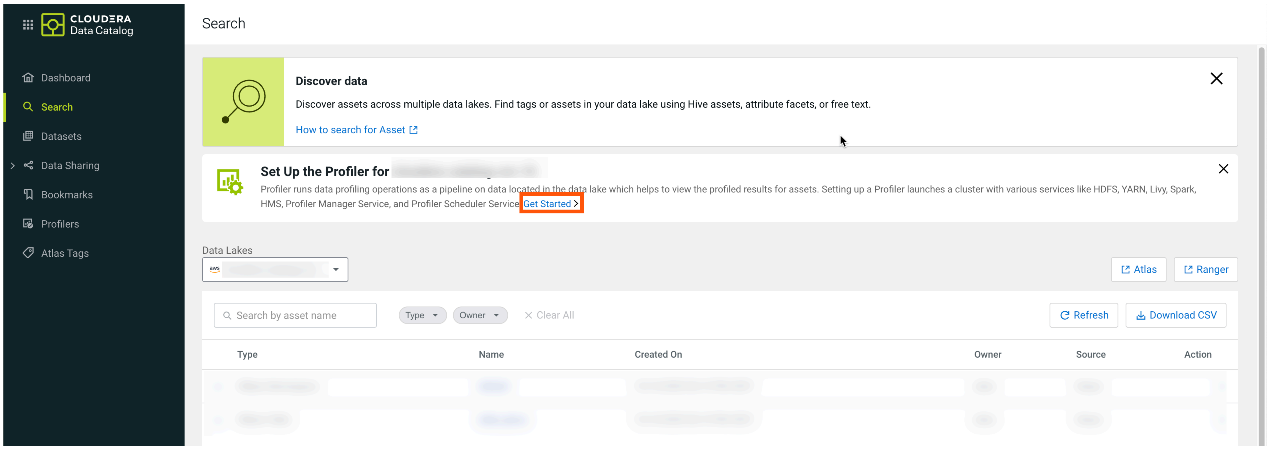
- Profiler Manager
- Profiler Scheduler

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.



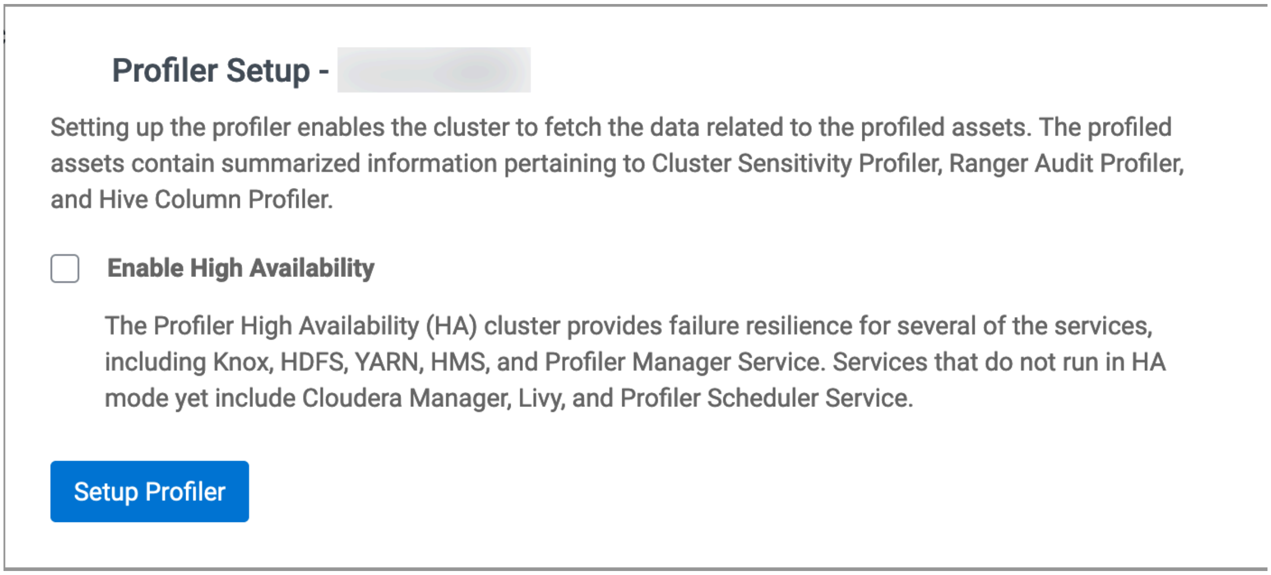
Discover data

Discover assets across multiple data lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text.

How to search for Asset

Set Up the Profiler for [redacted]

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 >



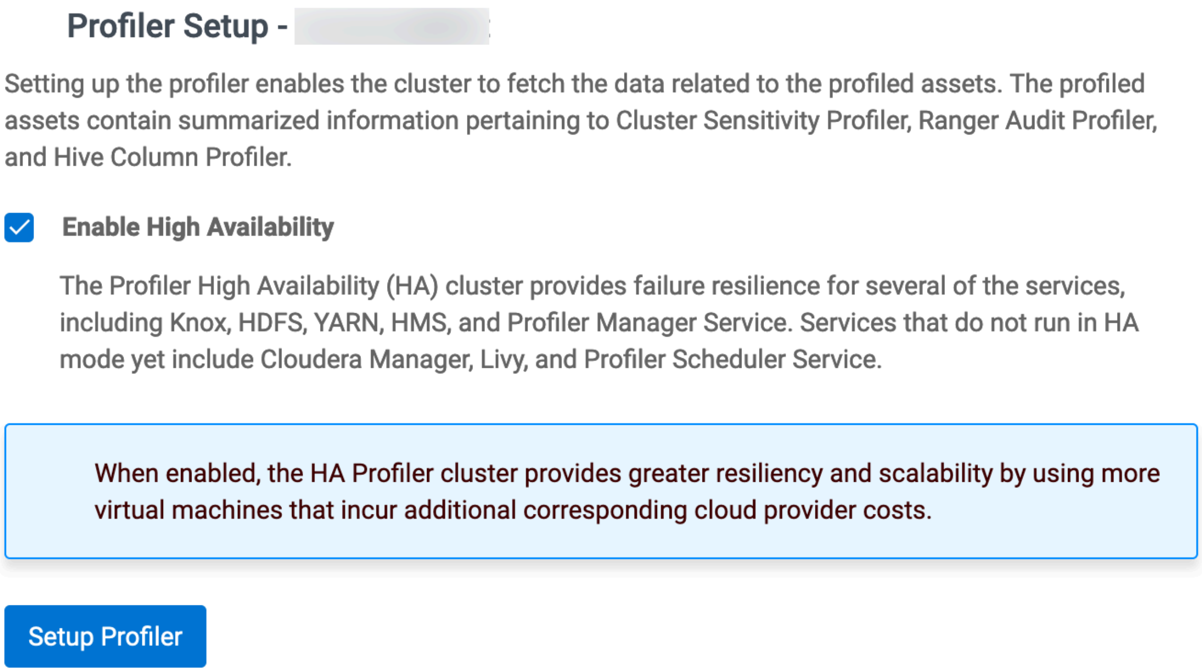
Profiler Setup - [redacted]

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.

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



Profiler Setup - [redacted]

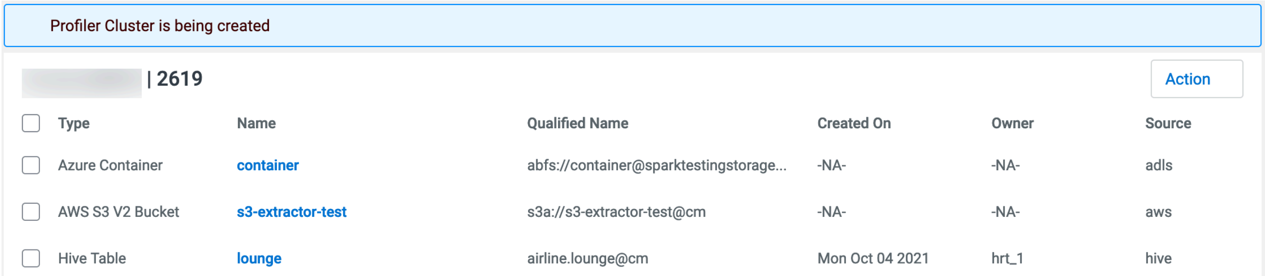
Setting up the profiler enables the cluster to fetch the data related to the profiled assets. The profiled assets contain summarized information pertaining to Cluster Sensitivity Profiler, Ranger Audit Profiler, and Hive Column Profiler.

☑ **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.

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

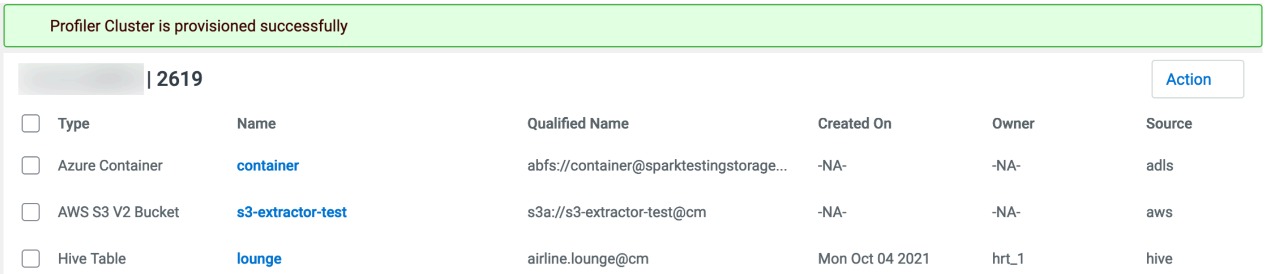


Profiler Cluster is being created

[redacted] | 2619

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

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



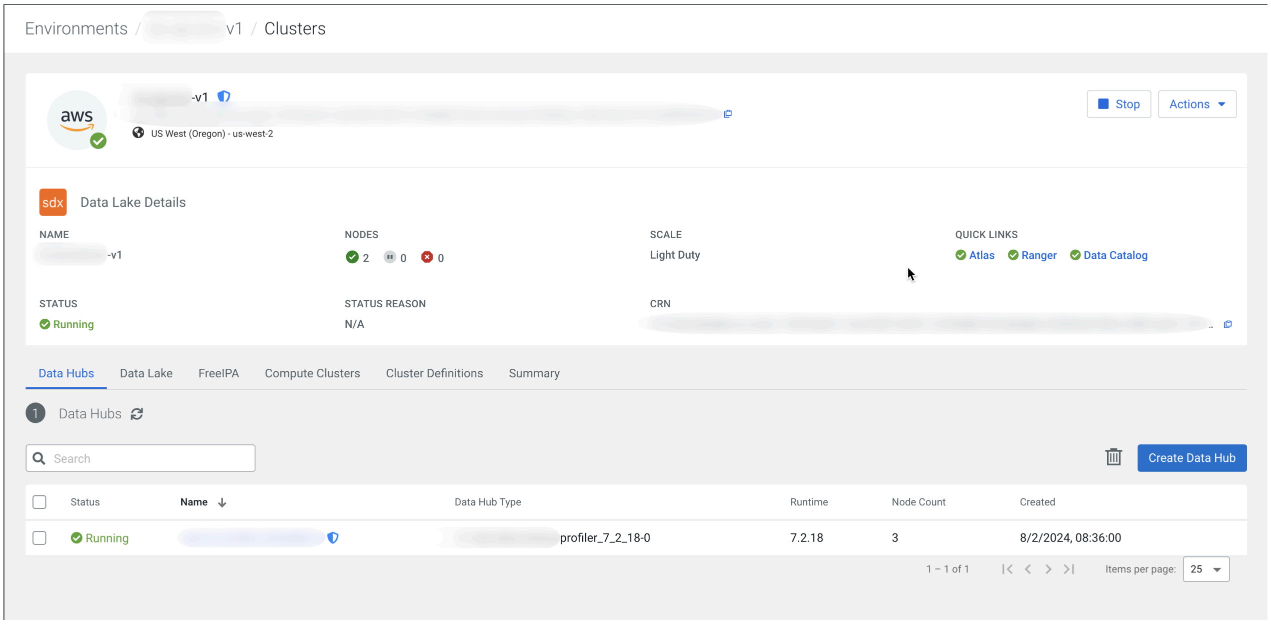
Profiler Cluster is provisioned successfully

[redacted] | 2619

☐	Type	Name	Qualified Name	Created On	Owner	Source
☐	Azure Container	container	abfs://container@sparktestingstorage...	-NA-	-NA-	adls
☐	AWS S3 V2 Bucket	s3-extractor-test	s3a://s3-extractor-test@cm	-NA-	-NA-	aws
☐	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 / [redacted] v1 / Clusters

[redacted]-v1

[redacted]

US West (Oregon) - us-west-2

Data Lake Details

NAME: [redacted]-v1

NODES: ✓ 2 ⏸ 0 ✗ 0

SCALE: Light Duty

QUICK LINKS: ✓ Atlas ✓ Ranger ✓ Data Catalog

STATUS: ✓ Running

STATUS REASON: N/A

CRN: [redacted]

☐	Status	Name ↓	Data Hub Type	Runtime	Node Count	Created
☐	✓ Running	[redacted]	[redacted]profiler_7_2_18-0	7.2.18	3	8/2/2024, 08:36:00

1 – 1 of 1 Items per page: 25

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

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

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 for Cloudera on cloud 7.2.18. and earlier versions.

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:

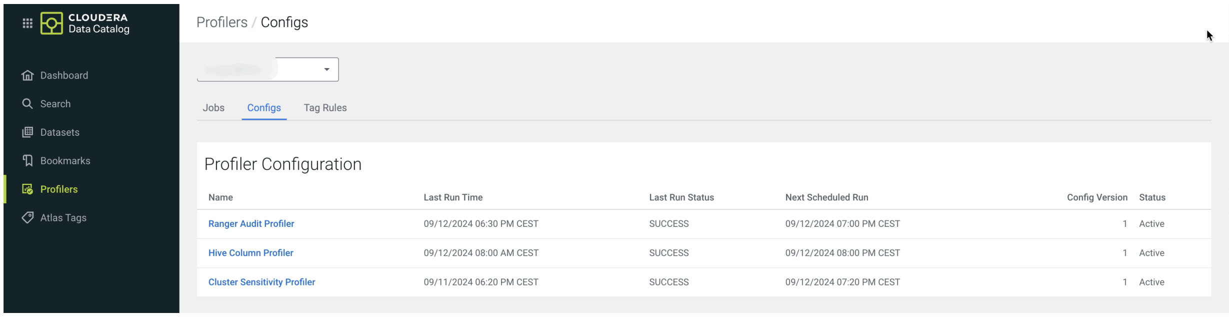
```
cdp datacatalog launch-profilers --datalake crn:cdp:data
lake:DATACENTERNAME:c*****b-ccce-4**d-a**1-8********8:datalake:4*****5e-c**
1-4**2-8**e-1*******2
{
    "success": true
}
```

Enabling or disabling profilers in VM-based environments

By default, profilers are enabled and run every 30 minutes. If you want to disable (or re-enable) a profiler, you can do this by selecting the appropriate profiler from the Configs tab.

Procedure

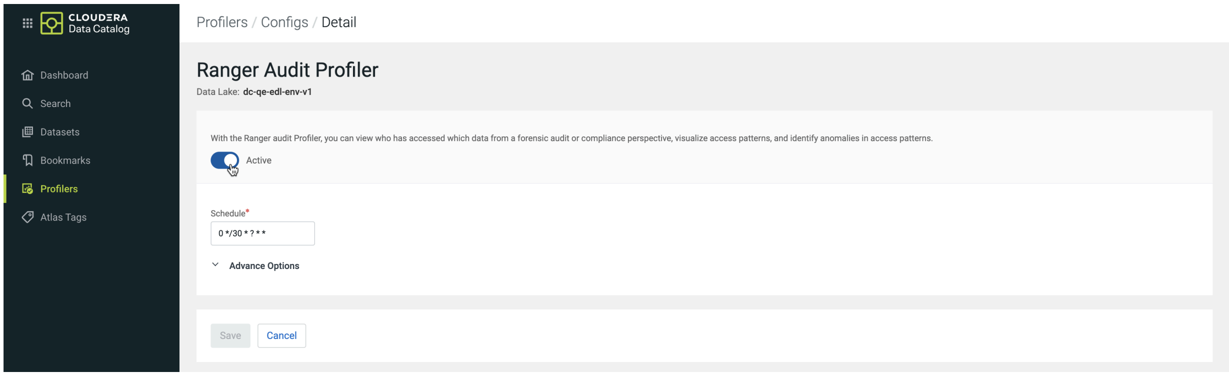
1. Go to Profilers Configs .
2. Select the profiler to proceed further.



Profiler Configuration

Name	Last Run Time	Last Run Status	Next Scheduled Run	Config Version	Status
Ranger Audit Profiler	09/12/2024 06:30 PM CEST	SUCCESS	09/12/2024 07:00 PM CEST	1	Active
Hive Column Profiler	09/12/2024 08:00 AM CEST	SUCCESS	09/12/2024 08:00 PM CEST	1	Active
Cluster Sensitivity Profiler	09/11/2024 06:20 PM CEST	SUCCESS	09/12/2024 07:20 PM CEST	1	Active

3. Switch the toggle to the desired state.



Ranger Audit Profiler

Data Lake: dc-qe-edl-env-v1

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 * ? * *

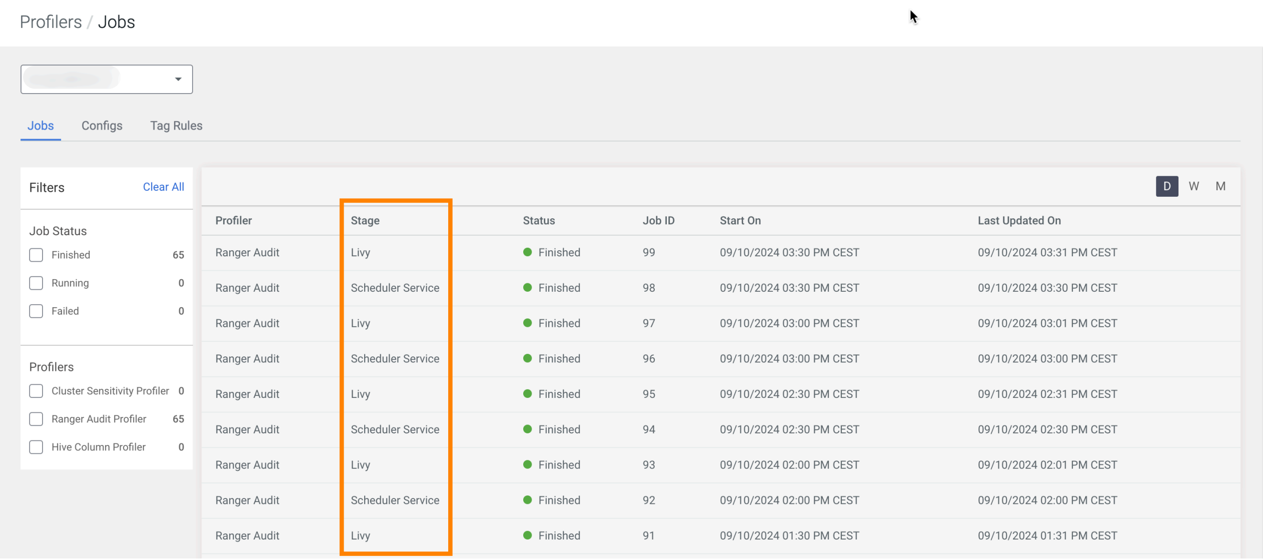
Tracking profiler jobs in VM-based environments

Use the Profilers > Jobs page for tracking the respective profiler jobs.

Under Profilers Jobs , you can have an overview of your started profiler jobs. By using the D, W, M filters, you can go back up to a day, week or a month, to see your previous jobs. Use this page to quickly check if your profiler jobs are failing.

In VM-based environments, Profilers Jobs can show you the current profiling **Stage** based on the relevant service used:

Figure 1: Profiling jobs in a VM-based environment



Filters

Job Status

□ Finished	65
□ Running	0
□ Failed	0

Profilers

□ Cluster Sensitivity Profiler	0
□ Ranger Audit Profiler	65
□ Hive Column Profiler	0

Profiler	Stage	Status	Job ID	Start On	Last Updated On
Ranger Audit	Livy	● Finished	99	09/10/2024 03:30 PM CEST	09/10/2024 03:31 PM CEST
Ranger Audit	Scheduler Service	● Finished	98	09/10/2024 03:30 PM CEST	09/10/2024 03:30 PM CEST
Ranger Audit	Livy	● Finished	97	09/10/2024 03:00 PM CEST	09/10/2024 03:01 PM CEST
Ranger Audit	Scheduler Service	● Finished	96	09/10/2024 03:00 PM CEST	09/10/2024 03:00 PM CEST
Ranger Audit	Livy	● Finished	95	09/10/2024 02:30 PM CEST	09/10/2024 02:31 PM CEST
Ranger Audit	Scheduler Service	● Finished	94	09/10/2024 02:30 PM CEST	09/10/2024 02:30 PM CEST
Ranger Audit	Livy	● Finished	93	09/10/2024 02:00 PM CEST	09/10/2024 02:01 PM CEST
Ranger Audit	Scheduler Service	● Finished	92	09/10/2024 02:00 PM CEST	09/10/2024 02:00 PM CEST
Ranger Audit	Livy	● Finished	91	09/10/2024 01:30 PM CEST	09/10/2024 01:31 PM CEST

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

- **Profiler** type
- Profiler **Status**
- **Stage** (for VM-based environments)
- Job **Status**
- **Job ID**
- **Start Time**
- **Last Updated On**

Using this data can help you to troubleshoot failed jobs or even understand how the assets were profiled and other pertinent information that can help you to manage your profiled assets.

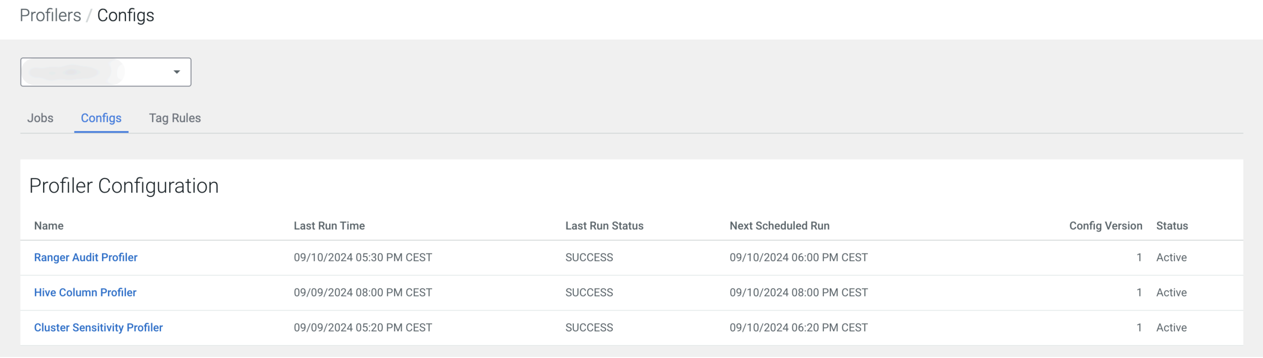
In VM-based environments, profiler job runs in the following phases:

1. Scheduler Service - The part of Profiler Admin which queues the profiler requests. This shows that profiling job is being created and queued for execution based on resource availability.
2. Livy - This service is managed by YARN and is used to submit the Apache Spark jobs after which the actual asset profiling takes place. Livy submits the profiler jobs to Spark for execution.
3. Metrics Service - Reads the profiled data files and publishes them in the correct format. For Ranger Audit Profiler results are synchronized only to the Cloudera Data Catalog database. For Hive Column Profiler and Cluster Sensitivity Profiler the results are also synced to Apache Atlas as these profilers affect the Atlas classifications and also, the information has to be available during the next run.

Note: More than one occurrence of Scheduler Service or Livy indicates that there could be more assets to be profiled. For example, if an 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.

Viewing profiler configurations in VM-based environments

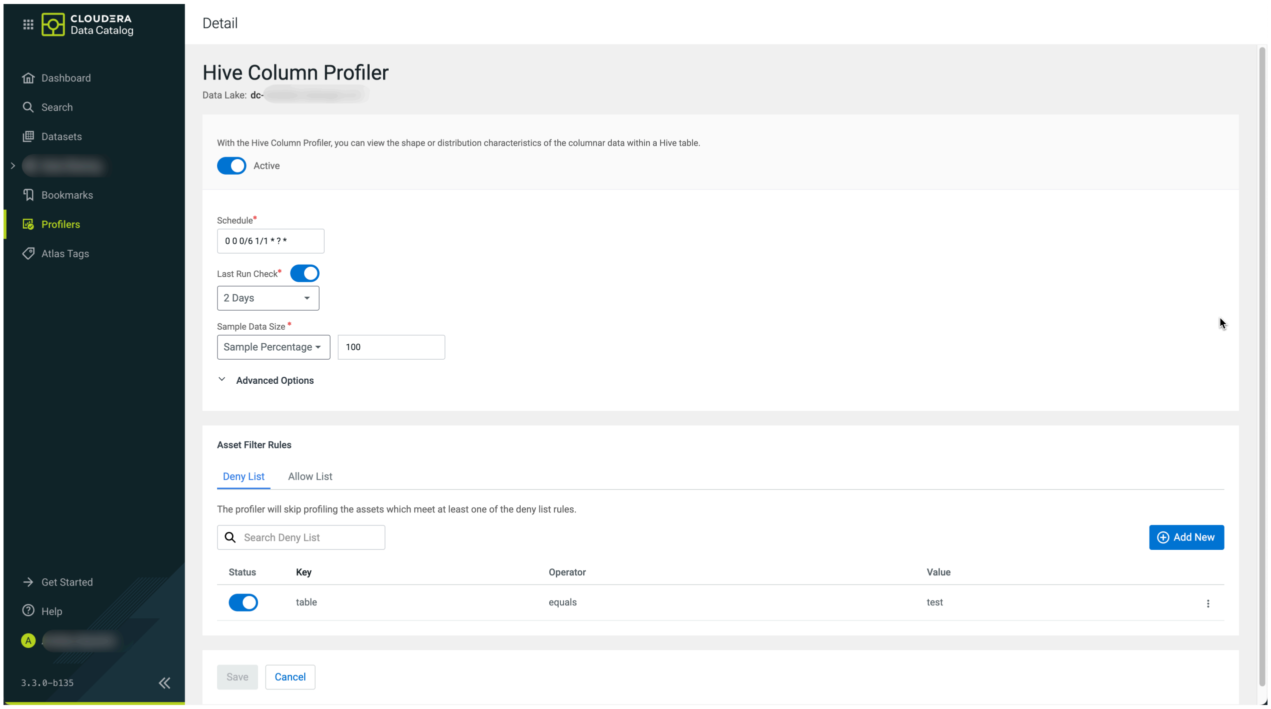
You can monitor the last status of individual profilers by viewing them in Profiler > Configs. Also, you can change their resources, sensitivity and scheduling.



Profiler Configuration

Name	Last Run Time	Last Run Status	Next Scheduled Run	Config Version	Status
Ranger Audit Profiler	09/10/2024 05:30 PM CEST	SUCCESS	09/10/2024 06:00 PM CEST	1	Active
Hive Column Profiler	09/09/2024 08:00 PM CEST	SUCCESS	09/10/2024 08:00 PM CEST	1	Active
Cluster Sensitivity Profiler	09/09/2024 05:20 PM CEST	SUCCESS	09/10/2024 06:20 PM CEST	1	Active

Select one of the profilers to open the **Detail** menu.



Hive Column Profiler

Data Lake: dc-[redacted]

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*: 2 Days

Sample Data Size*: Sample Percentage 100

Advanced Options

Asset Filter Rules

Deny List | Allow List

The profiler will skip profiling the assets which meet at least one of the deny list rules.

Search Deny List

Add New

Status	Key	Operator	Value
●	table	equals	test

Save Cancel

Monitoring the profiler configurations has the following uses:

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

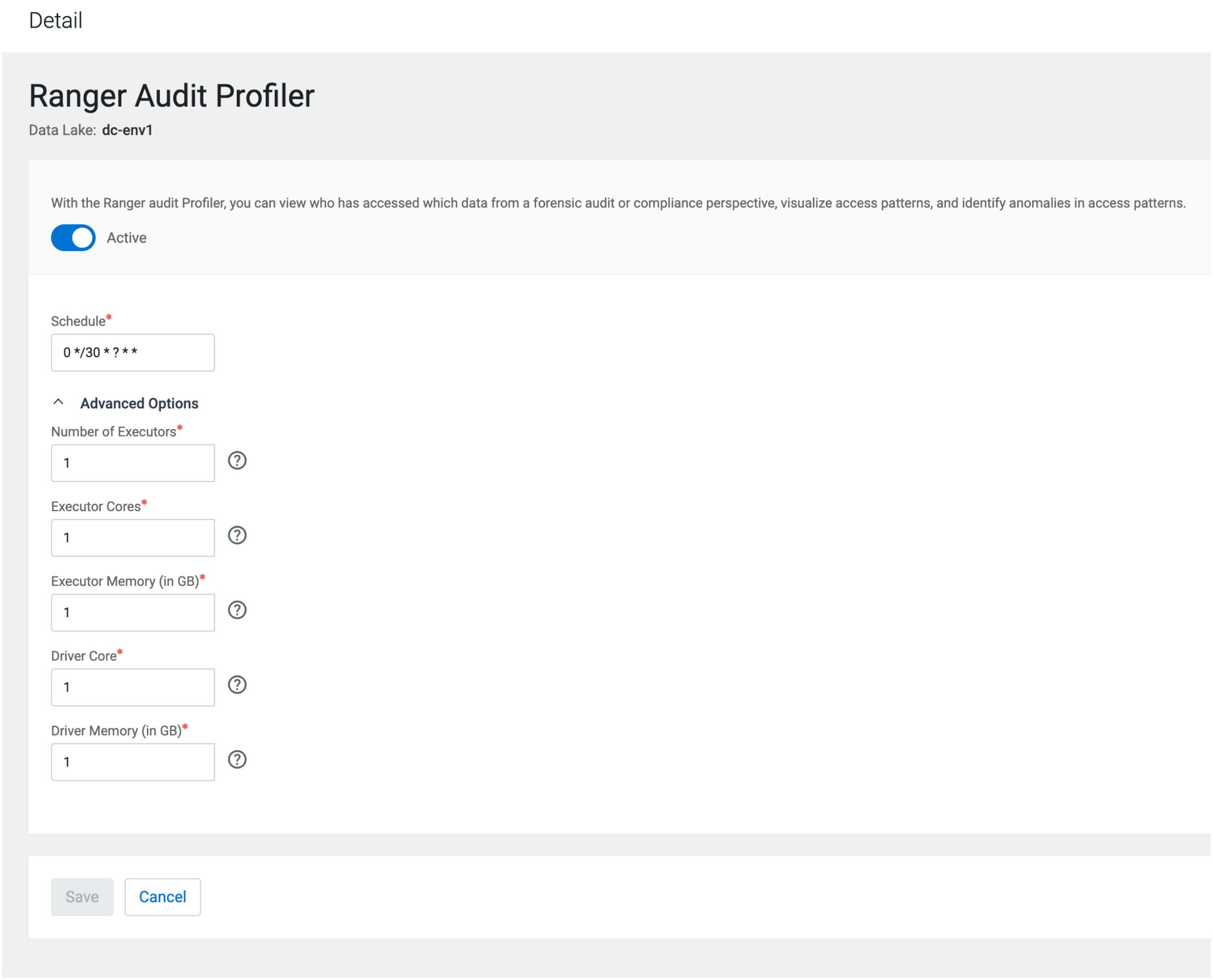
Configuring the Ranger Audit Profiler

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

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 Cluster Sensitivity Profiler

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

Before you begin

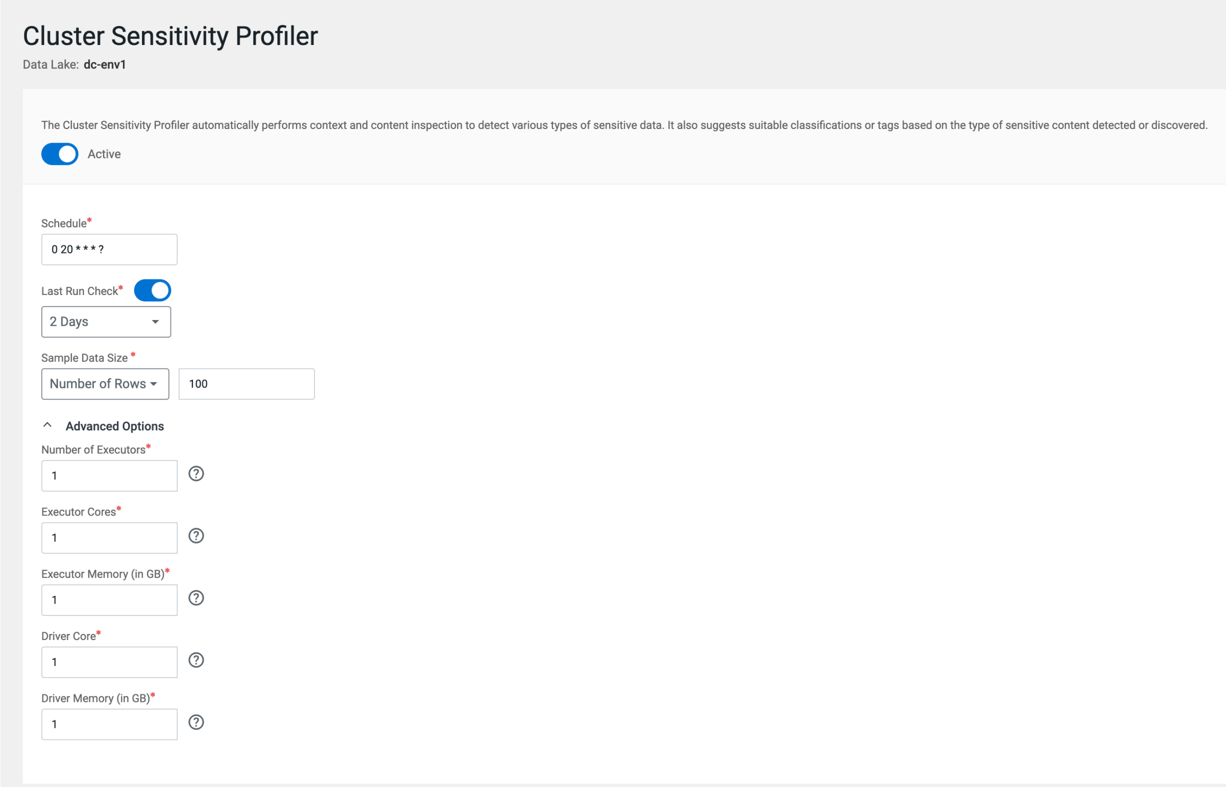
You need the DataCatalogCspRuleManager role, to create, to deploy new Custom Sensitivity Profiler rules, to create new regex expressions, and to run validations on newly created rules.

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.

   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 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 and deselection of assets which the profiler profiles.

Note:

- Profiler configurations apply to both scheduled and on-demand profiler jobs.
- Asset filtering rules apply to assets, such as tables, and not to complete databases.
- Multiple asset filtering rules are evaluated together as if connected by the OR operator.
- 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.

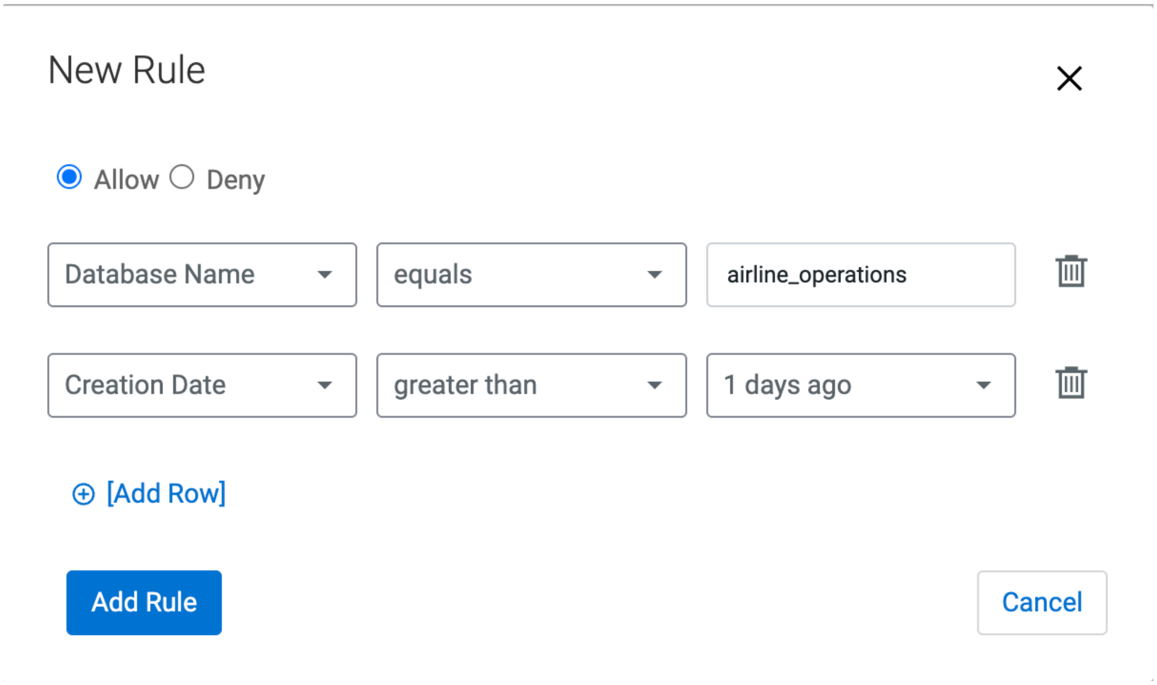
3. Select the key from the drop-down list and the relevant operator. You can select from the following:

Key	Operator
Database name	• equals • starts with • ends with
Name (of asset)	• equals • contains • starts with • ends with
Owner (of asset)	

Key	Operator
Creation date[1]	• greater than • less than

Note: **Name** refers here to the actual name of the asset and not to its **Qualified Name**.

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.



New Rule

● Allow ○ Deny

Database Name | equals | airline_operations

Creation Date | greater than | 1 days ago

Setting up column name based tagging

In VM-based environments with Cloudera on cloud 7.2.18.500 or later, you can use column name based tagging to ensure profiling columns whose data quality might not trigger the column value based checks of the Cluster Sensitivity Profiler. Typically, this can be used for tables where a large ratio of rows contain a different type of data or no data at all compared to the targeted data type that needs to be profiled.

Before you begin

A new classification must be created in Apache Atlas in advance. This classification (called tag in Cloudera Data Catalog) will be matched with tag rules to trigger the profiling. For more information, see Creating classifications.

Procedure

1. Create a tag rule for the tag previously created in Atlas to be applied to the column to be profiled.
   a) Go to Profilers Tag Rules .
   b) Click + New.
   c) In the **Tags** field, enter the name of the previously created Atlas classification.

[1] By Creation Date, Greater than 7 days means an asset older than seven days. Less than 7 days means an asset younger than seven days.

2. Click + in the **Resources** tab of **Tag Rules** to create your regular expression matching your column name.

   Note: The regular expression must be a full match to the column name that you created this rule for.

3. Select the regular expression matching the column name in **Column Name Expression**.

   Note: As this rule is exclusively created to allow columns to get tagged based on their name, skip the **Column Value Expression** field.

4. Go to the profiler's Cloudera Data Hub with the following path: Cloudera Manager Clusters profiler_scheduler Configuration .

   a) Search for "spark" and edit Profiler Scheduler Spark conf.

5. Add the following configuration snippet to set the level of confidence for the profiler to apply a tag:

```
spark.sensitive.tagRule.<***TAG RULE NAME***>.<***TAG NAME***>.<***COLUMN
 NAME***>.confidence = value=100
```

   Note:

   Although the range of 0 to 100 (both inclusive) is supported, it is recommended to set the value to 100, since this rule is exclusively to be used for column name matching.

   The column name and the column value tests both add up to a total of 100% weightage. If the confidence assigned to the column name matching is x then the confidence assigned to the column value is (100-x) by default. The profiler will suggest a tag for the column if the combined match score from testing both the column name and the column values add up to 70% or more.

   Multiple configuration snippets can be used, each with a different tag name for different Cluster Sensitivity Profilers.

6. Click Save Changes.
7. Wait until the changes are saved and the Restart button appears. Restart the scheduler service.

Profiler tag rules in VM-based environments

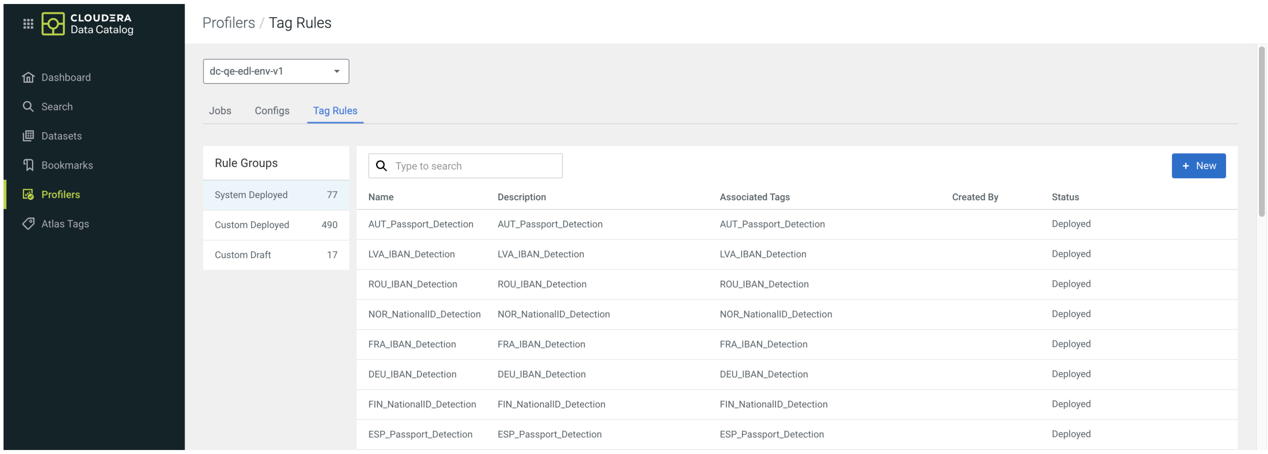
You can use preconfigured tag rules or create new rules based on regular expressions and values in your data to be profiled by the Cluster Sensitivity Profiler. When a tag rule is matching your data, the selected Apache Atlas classification (also known as a Cloudera Data Catalog tag) is applied.

Note: The improved tag rules are available for Compute Cluster enabled environments. In VM-based environments, tag rules are valid for all data lakes, while tag rules in Compute Cluster enabled environments are data lake specific.

Tag rule types

Tag Rules are categorized based on their type into the following groups:

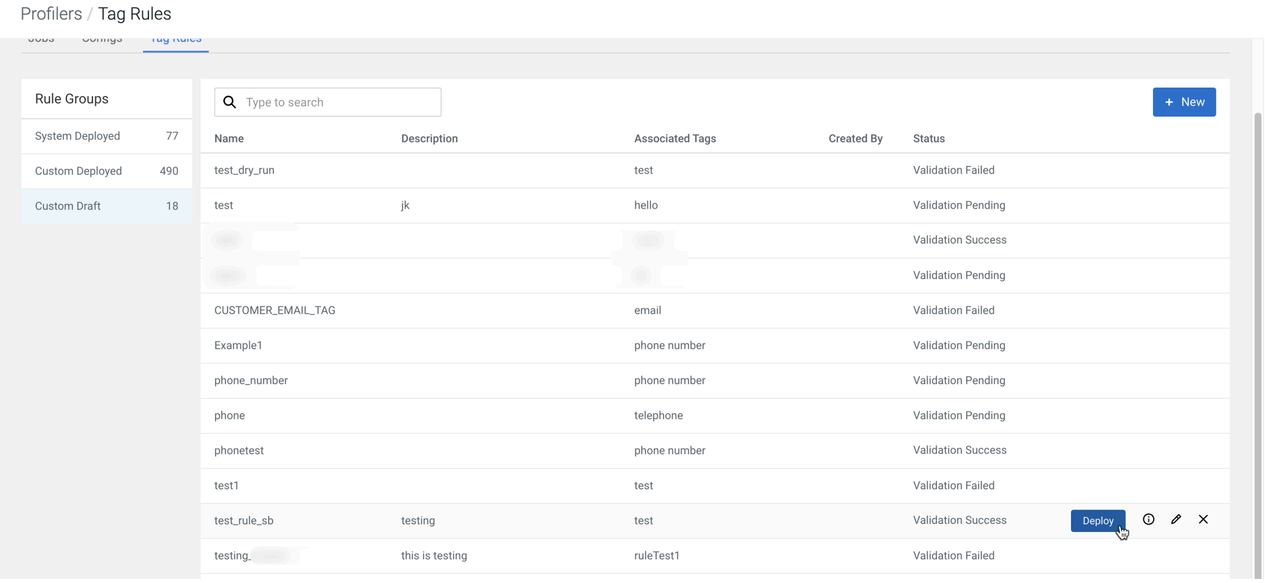
- **System Deployed**: These are built-in rules that cannot be edited. You can only enable or disable them for your data.
- **Custom Deployed**: Tag rules that you create, edit and deploy on clusters after validation will appear under this category. Click the ⋮ icon in the **Action** column to enable your custom tag rules. 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.



Rule Groups	
System Deployed	77
Custom Deployed	490
Custom Draft	17

Name	Description	Associated Tags	Created By	Status
AUT_Passport_Detection	AUT_Passport_Detection	AUT_Passport_Detection		Deployed
LVA_IBAN_Detection	LVA_IBAN_Detection	LVA_IBAN_Detection		Deployed
ROU_IBAN_Detection	ROU_IBAN_Detection	ROU_IBAN_Detection		Deployed
NOR_NationalID_Detection	NOR_NationalID_Detection	NOR_NationalID_Detection		Deployed
FRA_IBAN_Detection	FRA_IBAN_Detection	FRA_IBAN_Detection		Deployed
DEU_IBAN_Detection	DEU_IBAN_Detection	DEU_IBAN_Detection		Deployed
FIN_NationalID_Detection	FIN_NationalID_Detection	FIN_NationalID_Detection		Deployed
ESP_Passport_Detection	ESP_Passport_Detection	ESP_Passport_Detection		Deployed

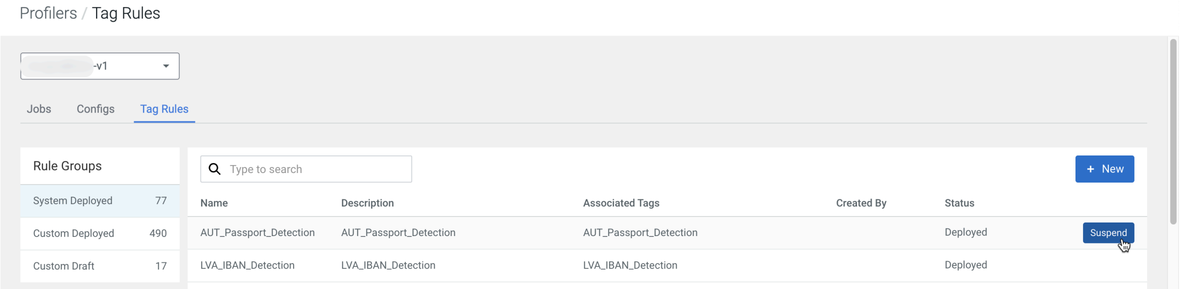
After creating your rule, you have to validate them. Only then you can click Enable.



Rule Groups	
System Deployed	77
Custom Deployed	490
Custom Draft	18

Name	Description	Associated Tags	Created By	Status
test_dry_run		test		Validation Failed
test	jk	hello		Validation Pending
[redacted]		[redacted]		Validation Success
[redacted]		[redacted]		Validation Pending
CUSTOMER_EMAIL_TAG		email		Validation Failed
Example1		phone number		Validation Pending
phone_number		phone number		Validation Pending
phone		telephone		Validation Pending
phonetest		phone number		Validation Success
test1		test		Validation Failed
test_rule_sb	testing	test		Validation Success
testing_[redacted]	this is testing	ruleTest1		Validation Failed

Note: Tag Rules can be temporarily suspended.



Rule Groups	
System Deployed	77
Custom Deployed	490
Custom Draft	17

Name	Description	Associated Tags	Created By	Status
AUT_Passport_Detection	AUT_Passport_Detection	AUT_Passport_Detection		Deployed
LVA_IBAN_Detection	LVA_IBAN_Detection	LVA_IBAN_Detection		Deployed

Tag rule inputs

Tag Rules can be applied based on the following inputs:

Input type	VM based environments	Compute Cluster enabled environments
Column name value	Manually entered regex pattern	• Manually entered regex pattern • Uploaded regex pattern

Input type	VM based environments	Compute Cluster enabled environments
Column value	Manually entered regex pattern	• Manually entered regex pattern • Uploaded regex pattern • CSV files with data which will be matched against column values for your tables in your data lake.
Table name		• Manually entered regex pattern • Uploaded regex pattern

Match thresholds and weightage

The System Deployed rules have a preset match threshold: A matching column name means a 15% confidence value. This is increased by 85% by a matching column value.

Tag rule testing

After creating your tag rule, you have to test it:

By VM-based environments validate them with manually entered test data and, then deploy them from the Custom Draft status.

Creating tag rules in VM based environments

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

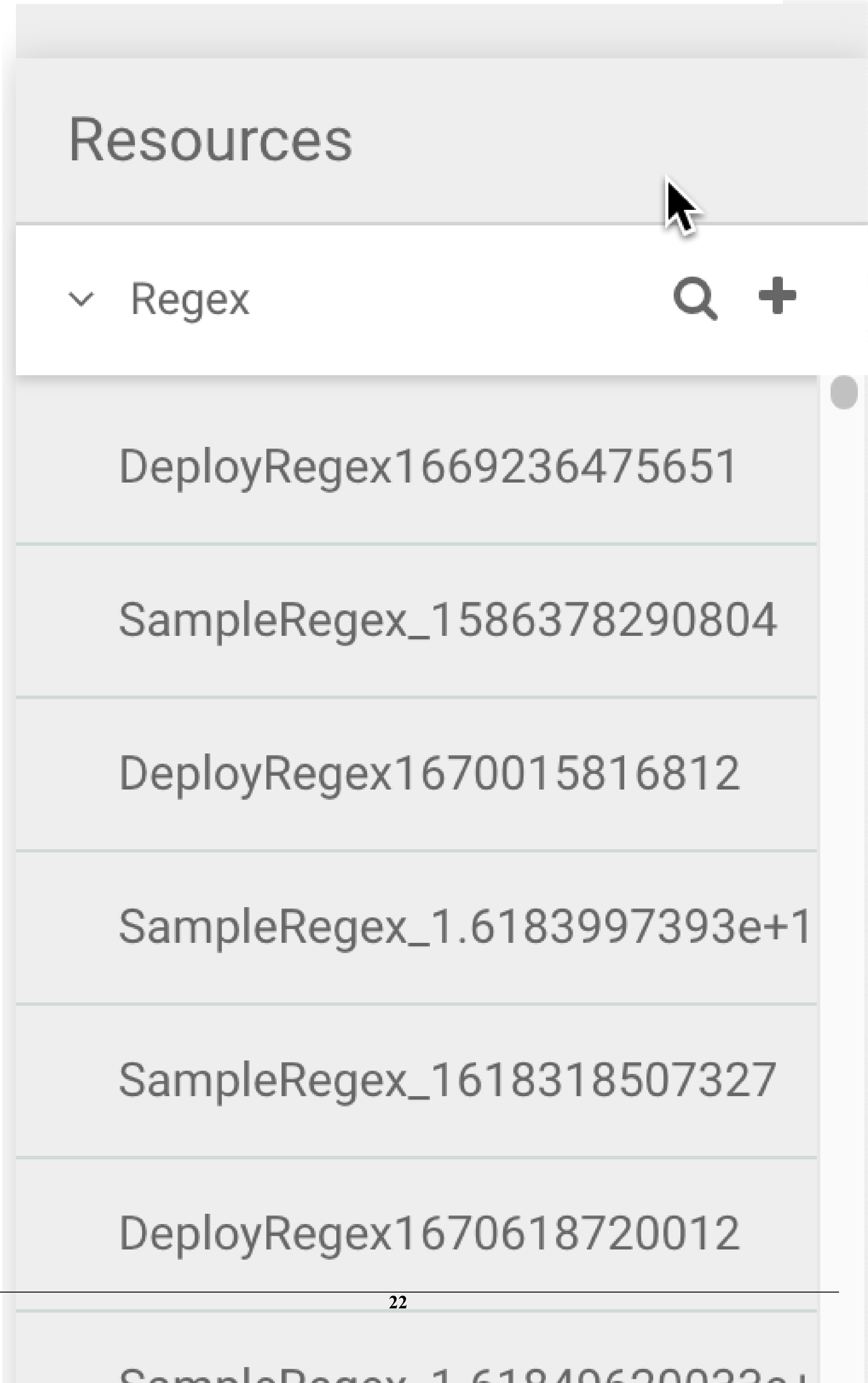
About this task

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

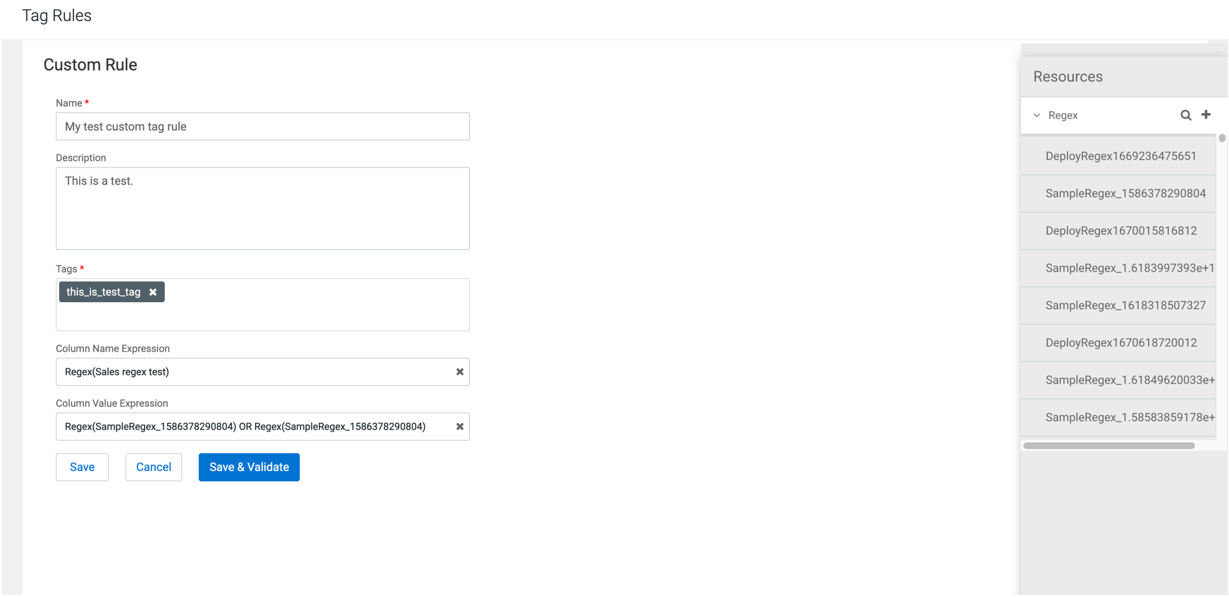
DeployRegex1670015816812

SampleRegex_1.6183997393e+1

SampleRegex_1618318507327

DeployRegex1670618720012

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



Tag Rules

Custom Rule

Name *: My test custom tag rule

Description: This is a test.

Tags *: this_is_test_tag

Column Name Expression: Regex(Sales regex test)

Column Value Expression: Regex(SampleRegex_1586378290804) OR Regex(SampleRegex_1586378290804)

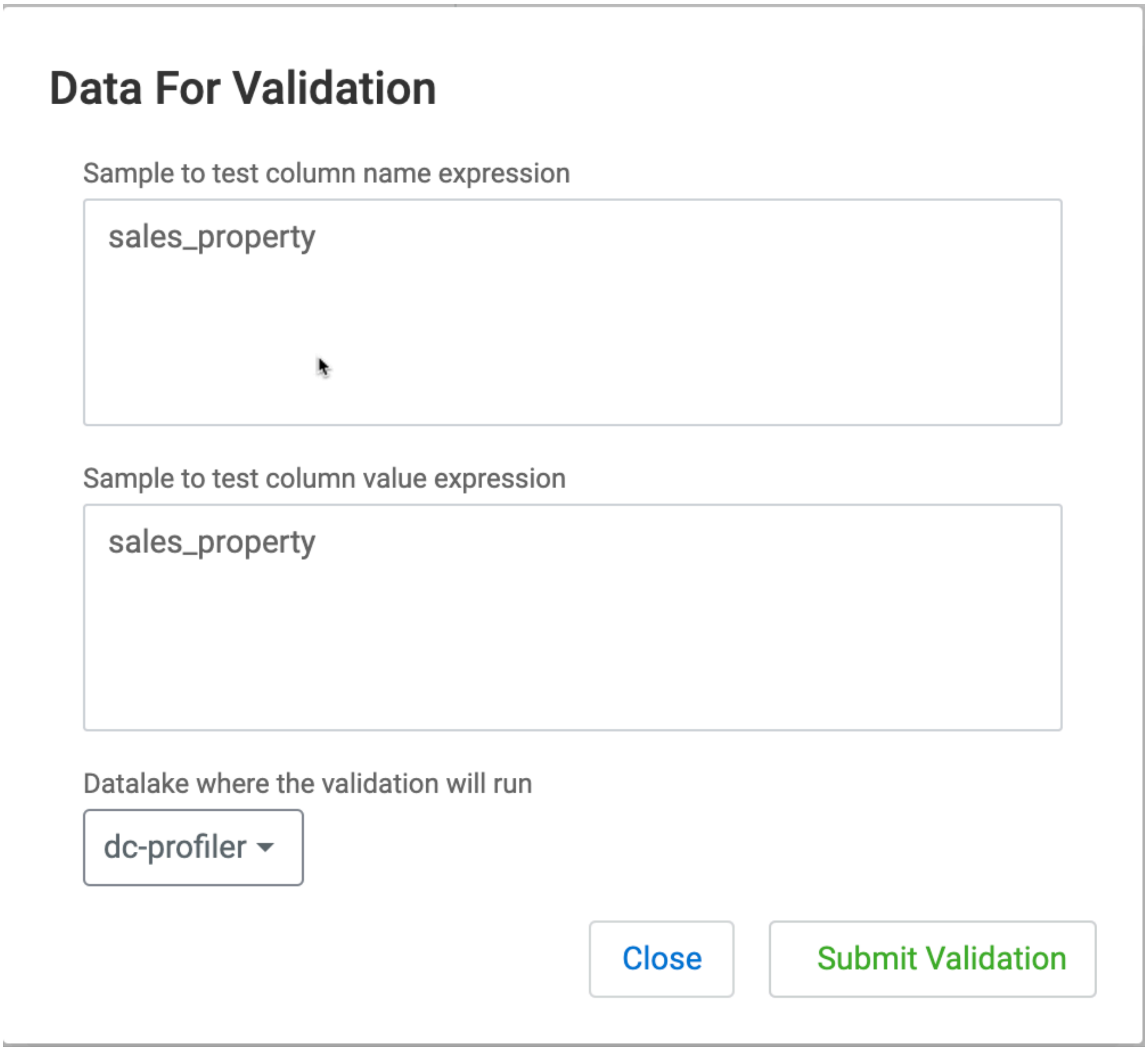
Save
Cancel
Save & Validate
Resources
Regex
DeployRegex1669236475651
SampleRegex_1586378290804
DeployRegex1670015816812
SampleRegex_1.6183997393e+1
SampleRegex_1618318507327
DeployRegex1670618720012
SampleRegex_1.61849620033e+
SampleRegex_1.58583859178e+

7. 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.

8. 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

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.

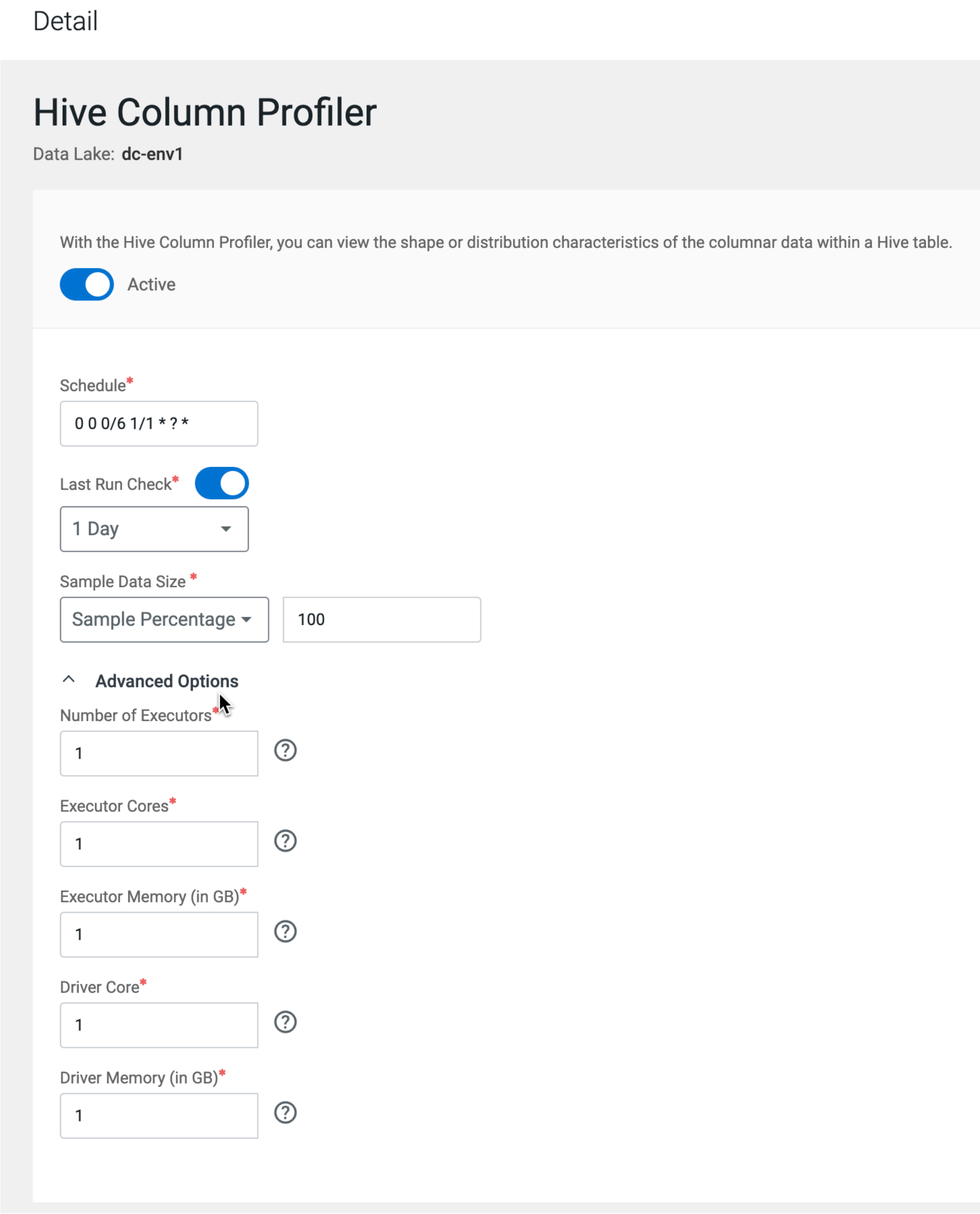
Configuring the Hive Column Profiler

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 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.
    - Asset filtering rules apply to assets, such as tables, and not to complete databases.
    - Multiple asset filtering rules are evaluated together as if connected by the OR operator.
    - 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**.

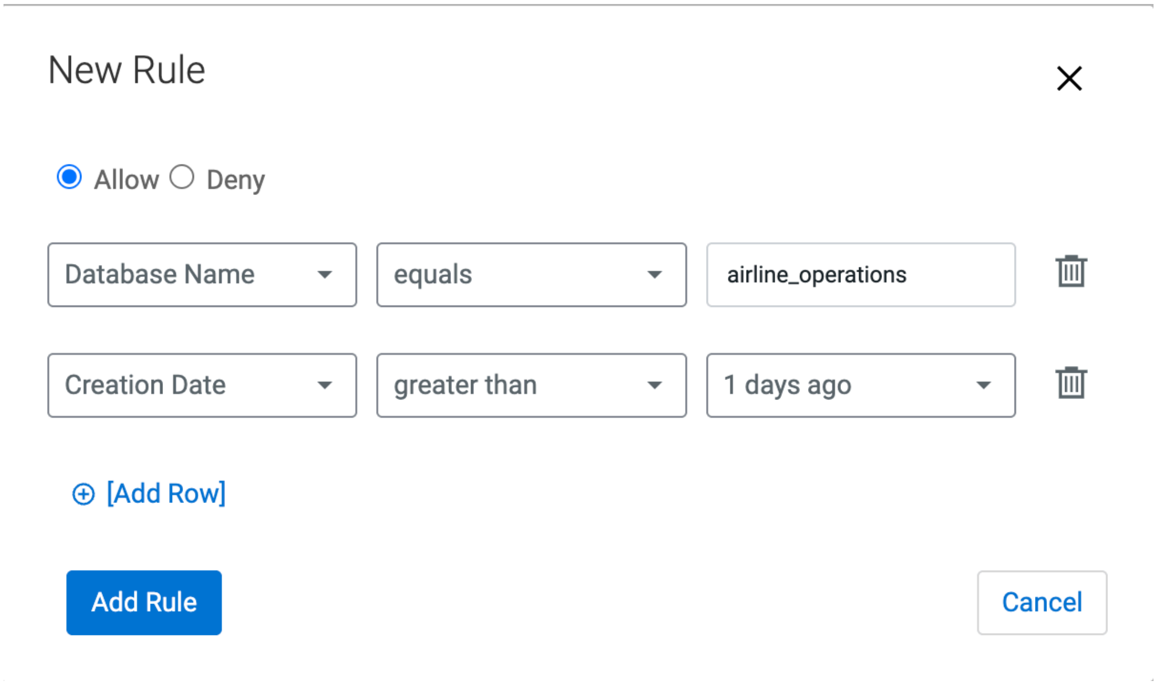
    1. Select the **Deny-list** or **Allow List** tab.
    2. Click Add New to define new rules.
    3. Select the key from the drop-down list and the relevant operator. You can select from the following:

Key	Operator
Database name	• equals • starts with • ends with
Name (of asset)	• equals • contains • starts with • ends with
Owner (of asset)	

Key	Operator
Creation date[2]	• greater than • less than

Note: Name refers here to the actual name of the asset and not to its **Qualified Name**.

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.



New Rule

● Allow ○ Deny

Database Name	equals	airline_operations
Creation Date	greater than	1 days ago

Understanding the Cron Expression generator

In the Profiler > Configs > Detail page, a cron expression defines when the profiler schedule executes and visualizes the next execution dates of your profiling jobs.

The Unix (in Compute Cluster enabled environments) and quartz (in VM-based environments) cron expressions use the following typical format:

Each * in the cron represents a unique value.

For VM-based environments

Schedule: * * * * * ? *

In this format the * characters represent the following units:

1. seconds (0-59)
2. minutes (0-59)
3. hours (0-23)

[2] By Creation Date, Greater than 7 days means an asset older than seven days. Less than 7 days means an asset younger than seven days.

4. day of the month (1-31)
5. month (1-12 or JAN-DEC)
6. day of the week (1-7 or SUN-SAT)
7. year (optional, 1970-2099)

Consider the following examples:

1 2 3 2 5 ? 2021

This cron expression is scheduled to run the profiler job at: 03:02:01am, on the 2nd day, in May, in 2021.

Note: The ? character is a replacement for the day of the week (or a day of the month). It is not specified on which day (or month) of the week the job has to run.

*** * * ? * ***

Every second

0 * * ? * *

Every minute (every 60th second)

0 0 * ? * *

Every hour (every 60th minute)

0 0 13 * * ?

At 13:00:00 every day

0 0 13 ? * WED

At 13:00:00, on every Wednesday, every month

0 0 13 ? * MON-FRI

At 13:00:00, on every weekday, every month

0 0 12 2 * ?

Every month on the 2nd, at noon

For Compute Cluster enabled environments

Cron Expression: 0 18 * * *

In this format the * characters represent the following units:

1. minute (0-59)
2. hour (0-23)
3. day of the month (1-31)
4. month (1-12)
5. day of the week (0-6)

Consider the following examples:

30 10 15 5 *

This cron expression is scheduled to run the profiler job at: “At 10:30 on 15th day-of-month in Ma y.”

Note: The * character is a replacement for the day of the week. It is not specified on which day of the week the job has to run.

30 10 * 5 7

This cron expression is scheduled to run the profiler job at: “At 10:30 on Sunday in May”.

Note: The * character is a replacement for the day of the month. It is not specified on which day of the month the job has to run.

5 * * * *

Every fifth minute of the hour. (18:05, 19:05, 20:05, etc.)

5 5 * * *

At 05:05 every day.

5 5 5 * *

At 05:05 on every fifth day of the month. (07-05 05:05:00, 08-05 05:05:00, 09-05 05:05:00, etc.)

5 5 5 5 *

At 05:05 on every fifth day of May (fifth month of the year); (2026-05-05 05:05:00, 2027-05-05 05:05:00, 2028-05-05 05:05:00, etc.)

5 5 5 5 5

At 05:05 on fifth day of the month OR if its a Friday (fifth day of the week) in May (fifth month of the year); 2026-05-01 05:05:00, 2026-05-05 05:05:00, 2026-05-08 05:05:00, etc.

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

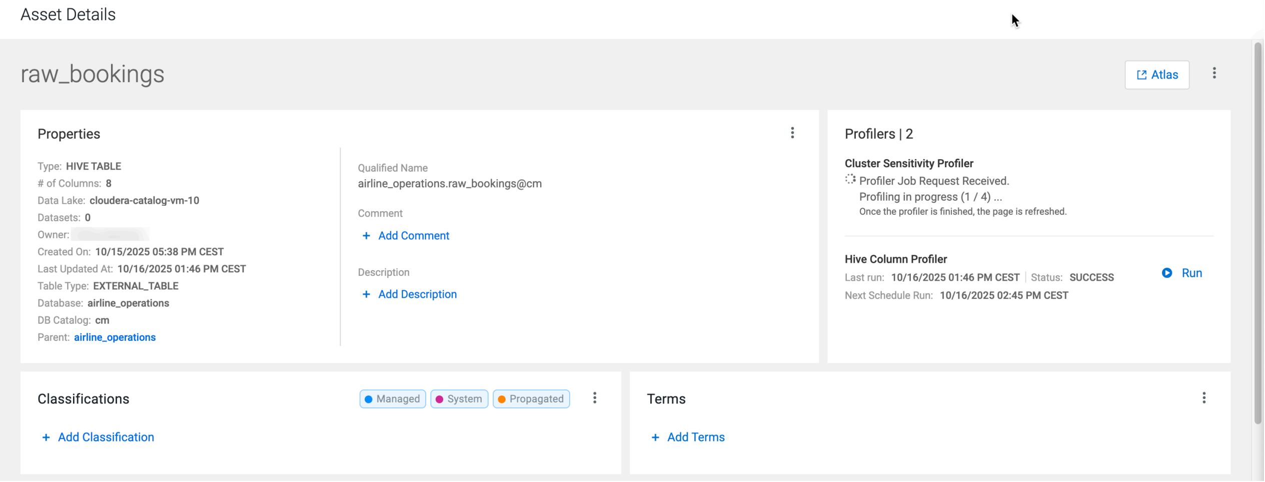
Note: Both Unix (in Compute Cluster enabled environments) and quartz CRON jobs (in VM-based environments) use the UTC timezone instead of the local timezone of the user.

On-Demand Profilers in VM based environments

You can use On-Demand Profilers to profile specific assets without depending on the cron-based scheduling of profilers jobs. The On-Demand Profiler option is available in the Asset Details of the selected asset.

The following image shows the **Asset Details** page of an asset. You can run an On-Demand Profiler for Hive Column Profiler and Cluster 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.



Asset Details

raw_bookings

Properties

Type: HIVE TABLE
# of Columns: 8
Data Lake: cloudera-catalog-vm-10
Datasets: 0
Owner: [redacted]
Created On: 10/15/2025 05:38 PM CEST
Last Updated At: 10/16/2025 01:46 PM CEST
Table Type: EXTERNAL_TABLE
Database: airline_operations
DB Catalog: cm
Parent: airline_operations

Qualified Name
airline_operations.raw_bookings@cm

Comment

Description

Profilers | 2

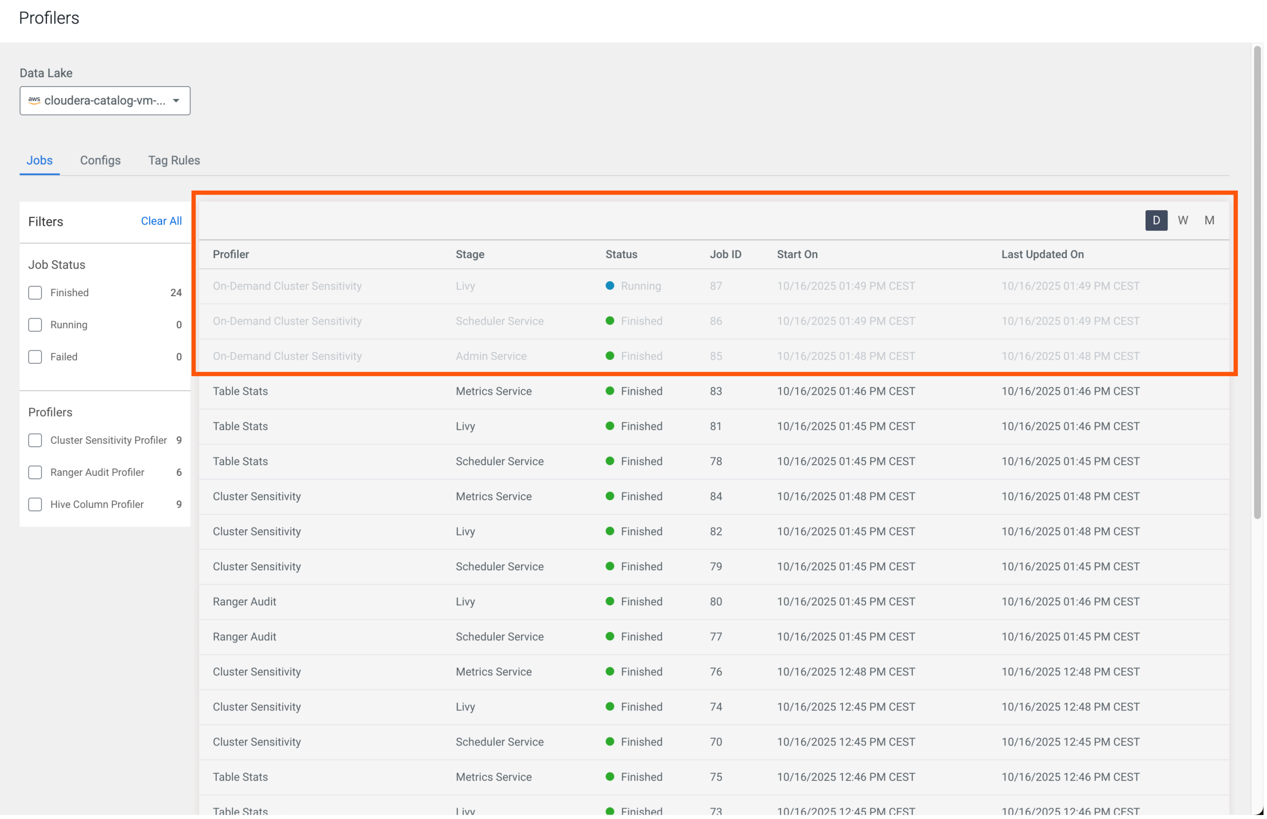
Cluster Sensitivity Profiler
Profiler Job Request Received.
Profiling in progress (1 / 4) ...
Once the profiler is finished, the page is refreshed.

Hive Column Profiler
Last run: 10/16/2025 01:46 PM CEST | Status: SUCCESS
Next Schedule Run: 10/16/2025 02:45 PM CEST

Classifications

Terms

Figure 2: Tracking on-demand profilers



Profilers

Data Lake

cloudera-catalog-vm-...

Filters Clear All

Job Status

□ Finished	24
□ Running	0
□ Failed	0

Profilers

□ Cluster Sensitivity Profiler	9
□ Ranger Audit Profiler	6
□ Hive Column Profiler	9

Profiler	Stage	Status	Job ID	Start On	Last Updated On
On-Demand Cluster Sensitivity	Livy	● Running	87	10/16/2025 01:49 PM CEST	10/16/2025 01:49 PM CEST
On-Demand Cluster Sensitivity	Scheduler Service	● Finished	86	10/16/2025 01:49 PM CEST	10/16/2025 01:49 PM CEST
On-Demand Cluster Sensitivity	Admin Service	● Finished	85	10/16/2025 01:48 PM CEST	10/16/2025 01:48 PM CEST
Table Stats	Metrics Service	● Finished	83	10/16/2025 01:46 PM CEST	10/16/2025 01:46 PM CEST
Table Stats	Livy	● Finished	81	10/16/2025 01:45 PM CEST	10/16/2025 01:46 PM CEST
Table Stats	Scheduler Service	● Finished	78	10/16/2025 01:45 PM CEST	10/16/2025 01:45 PM CEST
Cluster Sensitivity	Metrics Service	● Finished	84	10/16/2025 01:48 PM CEST	10/16/2025 01:48 PM CEST
Cluster Sensitivity	Livy	● Finished	82	10/16/2025 01:45 PM CEST	10/16/2025 01:48 PM CEST
Cluster Sensitivity	Scheduler Service	● Finished	79	10/16/2025 01:45 PM CEST	10/16/2025 01:45 PM CEST
Ranger Audit	Livy	● Finished	80	10/16/2025 01:45 PM CEST	10/16/2025 01:46 PM CEST
Ranger Audit	Scheduler Service	● Finished	77	10/16/2025 01:45 PM CEST	10/16/2025 01:45 PM CEST
Cluster Sensitivity	Metrics Service	● Finished	76	10/16/2025 12:48 PM CEST	10/16/2025 12:48 PM CEST
Cluster Sensitivity	Livy	● Finished	74	10/16/2025 12:45 PM CEST	10/16/2025 12:48 PM CEST
Cluster Sensitivity	Scheduler Service	● Finished	70	10/16/2025 12:45 PM CEST	10/16/2025 12:45 PM CEST
Table Stats	Metrics Service	● Finished	75	10/16/2025 12:46 PM CEST	10/16/2025 12:46 PM CEST
Table Stats	Livy	● Finished	73	10/16/2025 12:45 PM CEST	10/16/2025 12:46 PM CEST

Note:

- Profiler configurations apply to both scheduled and on-demand profiler jobs.

Backing up and restoring the profiler database

Using certain scripts that can be executed by the root users, you can back up of the profiler databases. Later, if you want to delete the existing Cloudera Data Hub cluster and launch a new cluster, you will have an option to restore the old data.

Important: Backing up and restoring the profiler database is only available in VM-based environments.

Cloudera Data Catalog includes profiler services that run data profiling operations on data that is located in multiple data lakes. In VM-based environments, the profiler services run on a Cloudera Data Hub cluster. When you delete the Cloudera Data Hub cluster, the profiled data and the user configuration information stored in the local databases are lost.

Profiler clusters run on the Cloudera Data Hub cluster using embedded databases:

- profiler_agent
- profiler_metrics

Note: If you download the modified Cluster Sensitivity Profiler rules before deleting the profiler cluster, and later when you create a new profiler cluster, you can restore the state of the rules manually. If the system rules are part of the downloaded files, you must Suspend those rules. If custom rules are part of the downloaded files, you must deploy those rules. This is applicable if the profiler cluster has Cloudera Runtime below 7.2.14 version.

About the back up script

The Backup and Restore script can be used only on Amazon Web Services, Microsoft Azure, and Google Cloud Platform clusters where they support cloud storage.

Scenarios for using the script

- When you upgrade the data lake cluster and want to preserve profiler data in the Cloudera Data Hub cluster.
- When you want to delete the Cloudera Data Hub cluster but preserve the profiler data.
- When you want to relaunch the profiler and access the older processed data.
- **Note:** For users using Cloudera Data Catalog on Cloudera Runtime 7.2.14 version, note the following:
  - No user action or manual intervention needed after the upgrading Cloudera Data Hub cluster to the 7.2.14 version.
  - Also, as an example use case scenario, in case a new profiler cluster is launched that contains Custom Sensitivity Profiler tags and which is deleted and relaunched later, the changes are retained and no further action is required.
  - No user action is required to backup and restore the profiler data. The changes are automatically restored.

When upgrading a Cloudera Runtime version earlier than 7.2.11 to version 7.2.11:

Go to the following locations to pick up your scripts:

Back up

```
bash    /opt/cloudera/parcels/PROFILER_MANAGER/profileradmin/scripts/users/backup_db.sh
```

Restore

```
bash    /opt/cloudera/parcels/PROFILER_MANAGER/profileradmin/scripts/users/restore_db.sh
```

When upgrading a version below or equal to Cloudera Runtime version 7.2.11 to 7.2.12:

Go to the following locations to pick up your scripts:

Back up

```
bash    /opt/cloudera/parcels/PROFILER_MANAGER/profileradmin/scripts/users/backup_db.sh
```

Restore

```
bash    /opt/cloudera/parcels/CDH/lib/profiler_manager/profileradmin/scripts/users/restore_db.sh
```

When backing up and restoring for a cluster having the Cloudera Runtime version 7.2.12 and onwards:

Navigate to the following location to pick up your scrips:

Back up

```
bash    /opt/cloudera/parcels/CDH/lib/profiler_manager/profileradmin/scripts/users/backup_db.sh
```

Restore

```
bash    /opt/cloudera/parcels/CDH/lib/profiler_manager/profileradmin/scripts/users/restore_db.sh
```

Running the back up script

Running the profiler Backup and Restore script has multiple phases.

About this task

First, you need to back up your profiler database and next you can restore it.

Backing up the profiler database

1. Stop the Profiler Manager and Profiler Scheduler services from the Cloudera Manager instance of the Cloudera Data Hub cluster.

2. Use SSH to connect to the node where the Profiler Manager is installed as a root user.
3. Execute the backup_db.sh script:

   Attention: Users of Cloudera Runtime below 7.2.8 version should contact Cloudera Support.

   Note:

   - If the profiler cluster has Cloudera Runtime version 7.2.11 or earlier, you run the following command:

     ```
     bash /opt/cloudera/parcels/PROFILER_MANAGER/profileradmin/scripts/
     users/backup_db.sh
     ```

   - If the profiler cluster has the Cloudera Runtime version 7.2.12 or higher you must run the following command:

     ```
     bash /opt/cloudera/parcels/CDH/lib/profiler_manager/profileradmin/
     scripts/users/backup_db.sh
     ```

4. Delete the Profiler cluster.
5. Install a new version of Profiler cluster:

- [Scenario-1] When the data lake upgrade is successfully completed.
- [Scenario-2] When the user decides to launch a new version of the Profiler cluster.

Restoring the profiler database

1. Stop the Profiler Manager and Profiler Scheduler services from the Cloudera Manager instance of the Cloudera Data Hub cluster.
2. Use SSH to connect to the node where Profiler Manager is installed as a root user.
3. Execute the restore_db.sh script.

   Attention: Users of Cloudera Runtime below 7.2.8 version should contact Cloudera Support.

   Note:

   - If the profiler cluster has the Cloudera Runtime version 7.2.11 or earlier, you must run the following command:

     ```
     bash /opt/cloudera/parcels/PROFILER_MANAGER/profileradmin/scripts/
     users/restore_db.sh
     ```

   - If the profiler cluster having the Cloudera Runtime version 7.2.12 or higher, you must run the following command:

     ```
     bash /opt/cloudera/parcels/CDH/lib/profiler_manager/profileradmin/
     scripts/users/restore_db.sh
     ```

4. Start the Profiler Manager and Profiler Scheduler services from Cloudera Manager.

Note: When you upgrade the data lake cluster and a new version of profiler cluster is installed, the profiler configurations that have been modified by users in the older version is replaced with new values as the following:

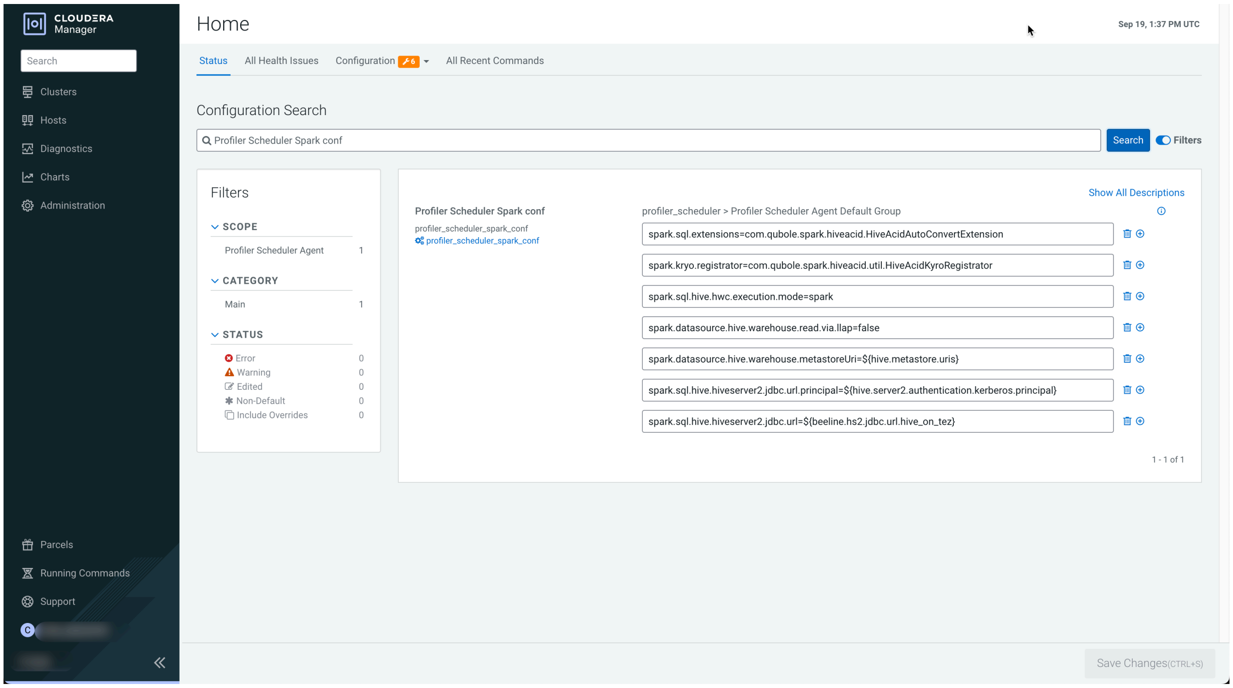
- Schedule
- Last Run Check
- Number of Executors
- Executor Cores
- Executor Memory (in GB)
- Driver Core
- Driver Memory (in GB)

Profiling table data in non-default buckets

In VM-based environments, you must configure a parameter in Profiler Scheduler in your instance to profile table data in non-default buckets.

Procedure

1. In Cloudera Data Catalog, make a note of your environment's name in the **Search** menu.
2. Go Cloudera Management Console Environments
3. Search for your environment, then switch to the **Data hubs** tab.
4. Open you Cloudera Data Hub by clicking its name.
5. Open the CM URL under **Cloudera Manager Info**.
6. In Cloudera Manager go to Configuration Configuration Search .
7. Search for the term Profiler Scheduler Spark conf.
   The **Profiler Scheduler Spark conf** configuration snippet appears.
8. Add spark.yarn.access.hadoopFileSystems=s3a://default-bucket,s3a://bucket-1,s3a://bucket-2 to **Profiler Scheduler Spark conf** to enable profiling for bucket-1 and bucket-2 non-default buckets.



Profiler Scheduler Spark conf

profiler_scheduler_spark_conf

profiler_scheduler > Profiler Scheduler Agent Default Group

```
spark.sql.extensions=com.qubole.spark.hiveacid.HiveAcidAutoConvertExtension
spark.kryo.registrator=com.qubole.spark.hiveacid.util.HiveAcidKyroRegistrator
spark.sql.hive.hwc.execution.mode=spark
spark.datasource.hive.warehouse.read.via.llap=false
spark.datasource.hive.warehouse.metastoreUri=${hive.metastore.uris}
spark.sql.hive.hiveserver2.jdbc.url.principal=${hive.server2.authentication.kerberos.principal}
spark.sql.hive.hiveserver2.jdbc.url=${beeline.hs2.jdbc.url.hive_on_tez}
```

Deleting profilers in VM-based environments

In VM-based environments, deleting the profiler cluster removes all the Data Compliance profiler rules and other updates to the specific cluster. It could also cause loss of data specific to currently applied rules on the deleted profiler cluster.

About this task

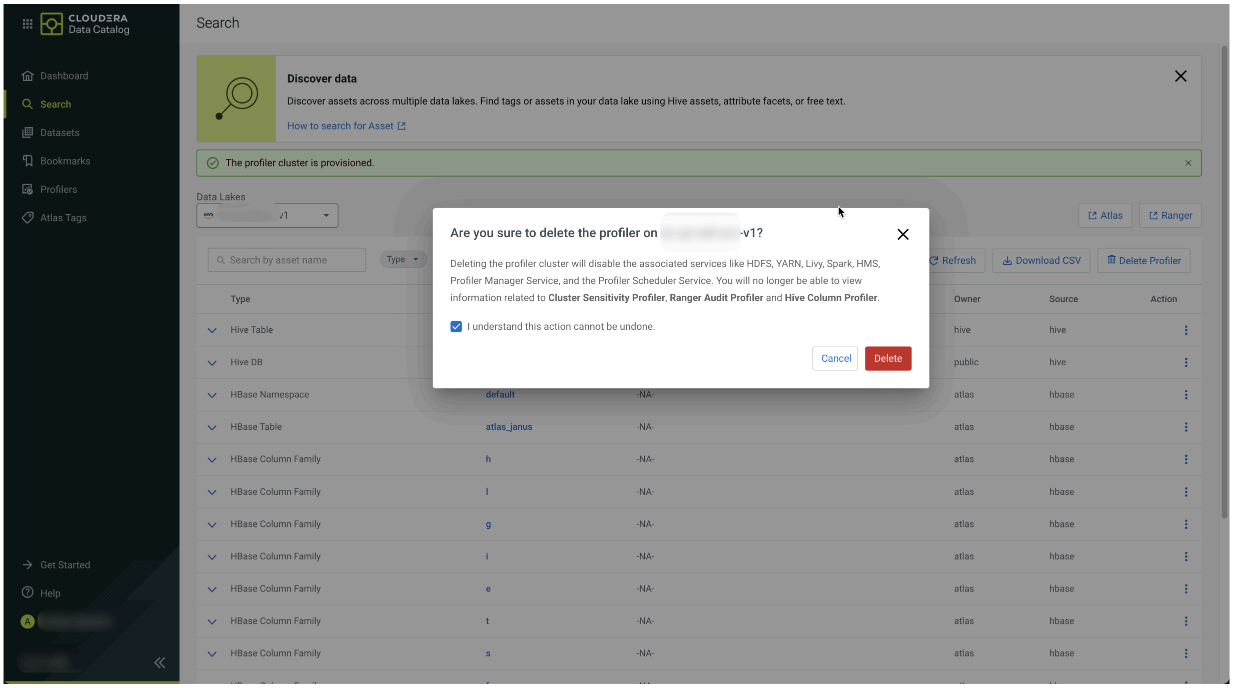
To overcome this situation, when you decide to delete the profiler cluster or (in VM-based environments), there is a provision to retain the status of the Cluster Sensitivity Profiler rules:

- If your profiler cluster or profiler jobs have rules that are not changed or updated, you can directly delete them or 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 jobs or the profiler cluster.

Procedure

1. On the **Profilers** page, select the data lake from the drop-down.
2. Click Delete Profiler in the action menu (⋮) for the profiler you want to delete.
3. If you agree, select the warning message: I understand this action cannot be undone.

Figure 3: Deleting a profiler in a VM-based environment



Discover data

Discover assets across multiple data lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text.

How to search for Asset

The profiler cluster is provisioned.

Data Lakes: aws [redacted]-v1

Are you sure to delete the profiler on [redacted]-v1?

Deleting the profiler cluster will disable the associated services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and the Profiler Scheduler Service. You will no longer be able to view information related to **Cluster Sensitivity Profiler**, **Ranger Audit Profiler** and **Hive Column Profiler**.

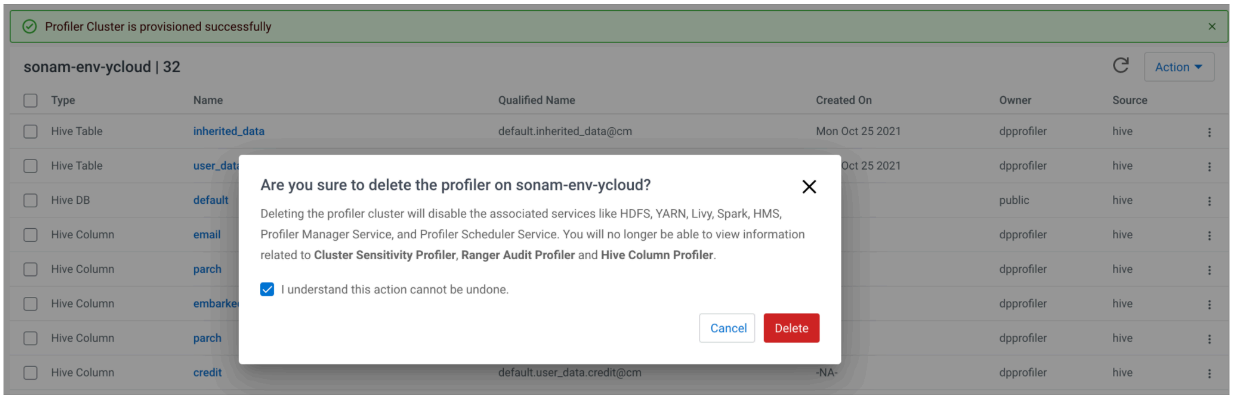
☑ I understand this action cannot be undone.

Type			Owner	Source	Action
Hive Table			hive	hive	⋮
Hive DB			public	hive	⋮
HBase Namespace	default	-NA-	atlas	hbase	⋮
HBase Table	atlas_janus	-NA-	atlas	hbase	⋮
HBase Column Family	h	-NA-	atlas	hbase	⋮
HBase Column Family	l	-NA-	atlas	hbase	⋮
HBase Column Family	g	-NA-	atlas	hbase	⋮
HBase Column Family	i	-NA-	atlas	hbase	⋮
HBase Column Family	e	-NA-	atlas	hbase	⋮
HBase Column Family	t	-NA-	atlas	hbase	⋮
HBase Column Family	s	-NA-	atlas	hbase	⋮

4. Click Delete.

   The application displays the following message.

   Note: When you launch Cloudera Data Catalog in Cloudera Runtime version 7.2.14, and later if the profiler cluster is deleted, the following message is displayed.



Profiler Cluster is provisioned successfully

sonam-env-ycloud | 32

	Type	Name	Qualified Name	Created On	Owner	Source
□	Hive Table	inherited_data	default.inherited_data@cm	Mon Oct 25 2021	dpprofiler	hive
□	Hive Table	user_dat		Oct 25 2021	dpprofiler	hive
□	Hive DB	default			public	hive
□	Hive Column	email			dpprofiler	hive
□	Hive Column	parch			dpprofiler	hive
□	Hive Column	embarke			dpprofiler	hive
□	Hive Column	parch			dpprofiler	hive
□	Hive Column	credit	default.user_data.credit@cm	-NA-	dpprofiler	hive

Are you sure to delete the profiler on sonam-env-ycloud?

Deleting the profiler cluster will disable the associated services like HDFS, YARN, Livy, Spark, HMS, Profiler Manager Service, and Profiler Scheduler Service. You will no longer be able to view information related to **Cluster Sensitivity Profiler**, **Ranger Audit Profiler** and **Hive Column Profiler**.

☑ I understand this action cannot be undone.

Note: You cannot delete a profiler while it is running.

Note: In VM-based environments, if the profiler cluster is not registered with the data lake, Cloudera Data Catalog cannot locate or trace the profiler cluster. You have to delete the profiler cluster from the Cloudera Data Hub page (Cloudera Management Console).

The profiler cluster is deleted successfully.