CDP Public Cloud / Data Catalog

Cloudera Data Catalog Top Use Cases

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

https://docs.cloudera.com/

Legal Notice

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

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

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

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

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

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

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

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

Contents

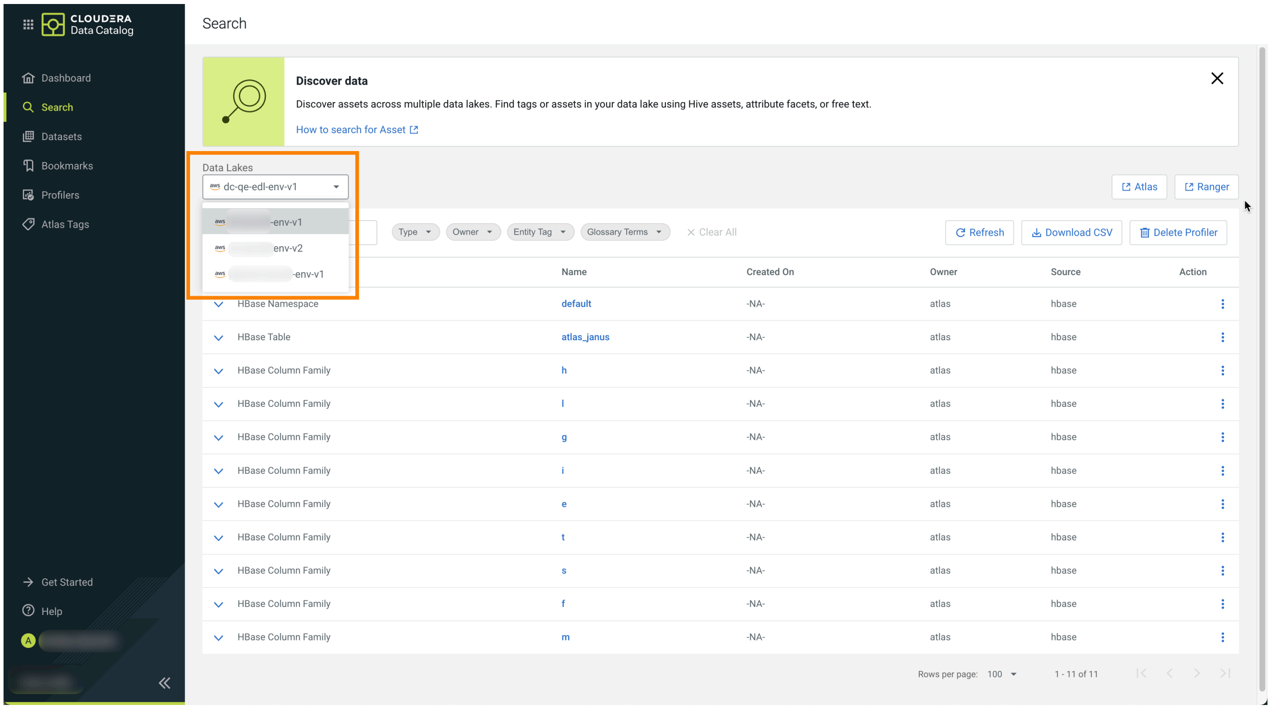
Search overview

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 in **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.

Accessing data lakes

Users have access to the lakes based on the permissions that are granted. You can choose the available lake by selecting the appropriate radio button.



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

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

	Name	Created On	Owner	Source	Action
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	⋮
HBase Column Family	f	-NA-	atlas	hbase	⋮
HBase Column Family	m	-NA-	atlas	hbase	⋮

Rows per page: 100 1 - 11 of 11

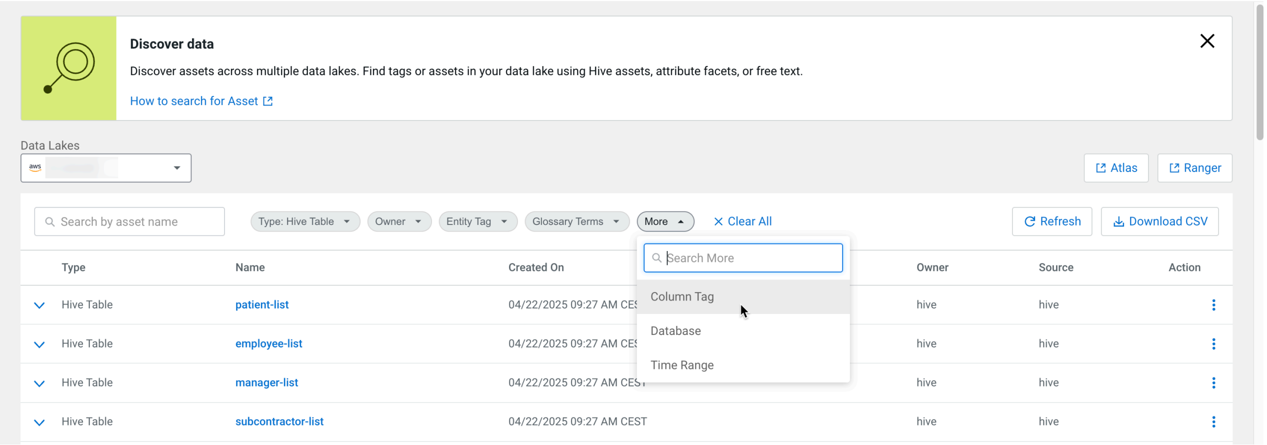
Note:

- You can search the assets of one data lake at a time.
- For the selected data lake, click the Atlas and Ranger links to go to the respective base cluster services in a new browser tab.

Using search filters

Use the search filter to fine-tune your results. By selecting a type, additional filter options become available and irrelevant filters are hidden. For example, after selecting Hive Table, the Column Tag, Database and Time Range filters are show under More.

Search



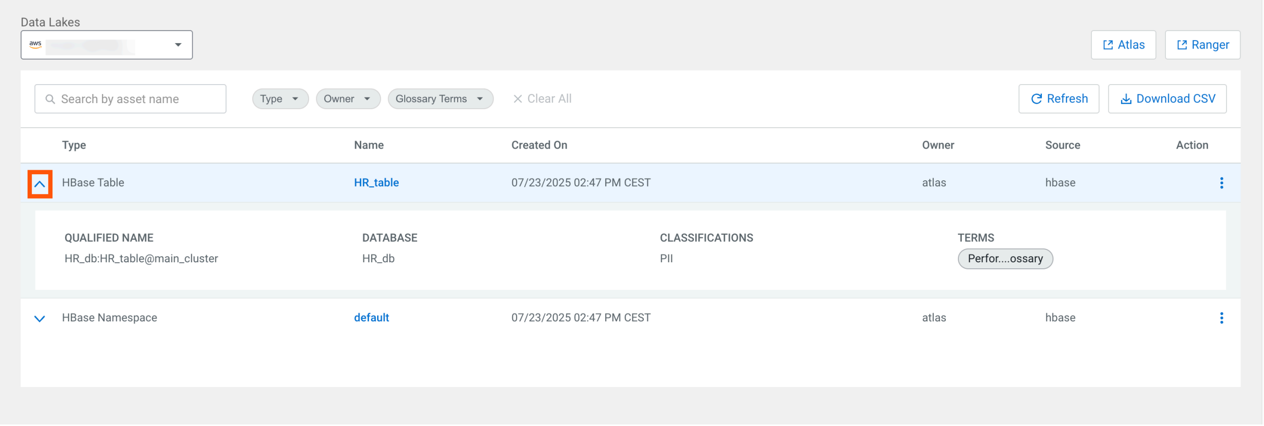
Type	Name	Created On	Owner	Source	Action
Hive Table	patient-list	04/22/2025 09:27 AM CES	hive	hive	⋮
Hive Table	employee-list	04/22/2025 09:27 AM CES	hive	hive	⋮
Hive Table	manager-list	04/22/2025 09:27 AM CEST	hive	hive	⋮
Hive Table	subcontractor-list	04/22/2025 09:27 AM CEST	hive	hive	⋮

Viewing Asset Details

Clicking the ∨ icon for a search result shows the most important data about an asset:

- **Qualified name**: - Qualified names are a unique identifier in Cloudera Data Catalog, identifying the asset with its context.

  A Hive table has the following qualified name patterns: *DATABASE_NAME.TABLE_NAME@CLUSTER_NAME*
- **Database**
- **Classifications** (Atlas tags)
- **Terms**



Type	Name	Created On	Owner	Source	Action
HBase Table	HR_table	07/23/2025 02:47 PM CEST	atlas	hbase	⋮
HBase Namespace	default	07/23/2025 02:47 PM CEST	atlas	hbase	⋮

QUALIFIED NAME	DATABASE	CLASSIFICATIONS	TERMS
HR_db:HR_table@main_cluster	HR_db	PII	Perfor....ossary

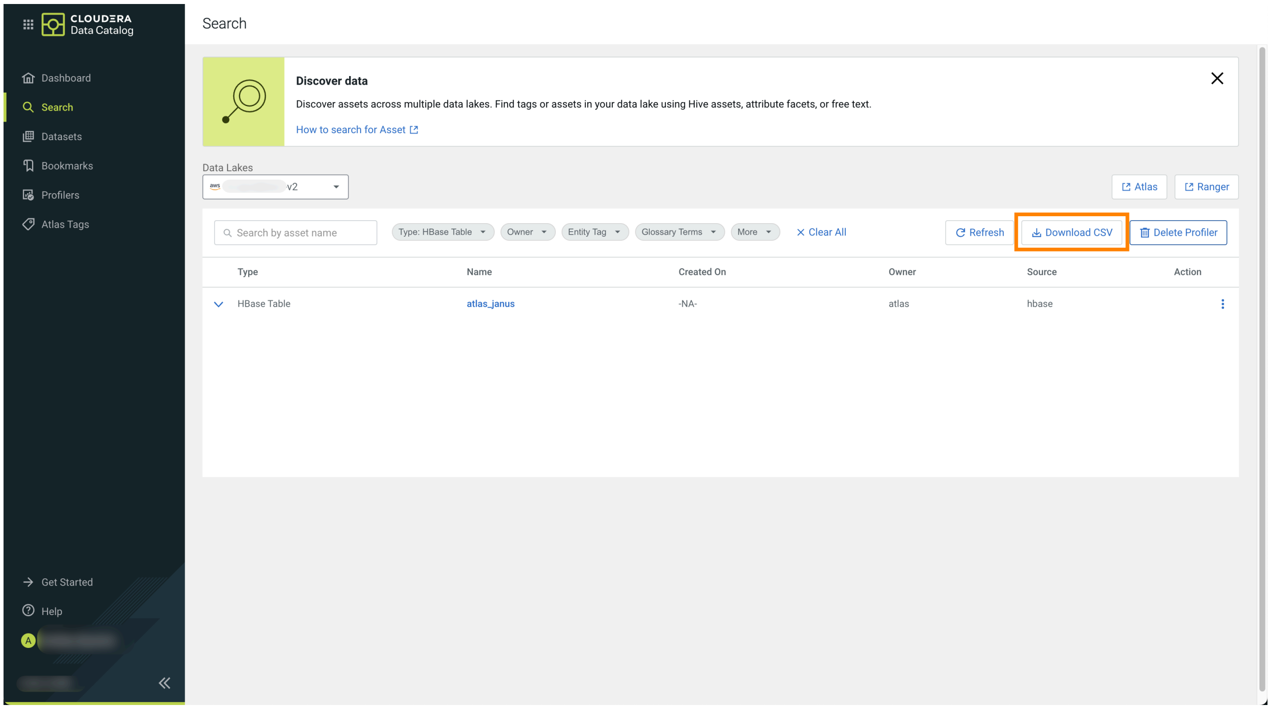
Clicking the Name of the entity will open its **Asset Details**.

Downloading search results as CSV files

You can also download the search result for the current query with the selected data lake. The feature allows you to download up to 10000 rows for the current search query.

The CSV file format does not conform to any specific order or continuation in the downloaded results. For example, a user can download 10000 assets and later downloads the results for the same query again, then the downloaded CSV files may not contain the search results in the same order as it was downloaded previously.

Click Download CSV to start your download:



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

Data Lakes: v2

Type	Name	Created On	Owner	Source	Action
HBase Table	atlas_janus	-NA-	atlas	hbase	⋮

Related Information

Datasets overview

Filters

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

Viewing Data Asset details

Launching profilers in Compute Cluster enabled environments

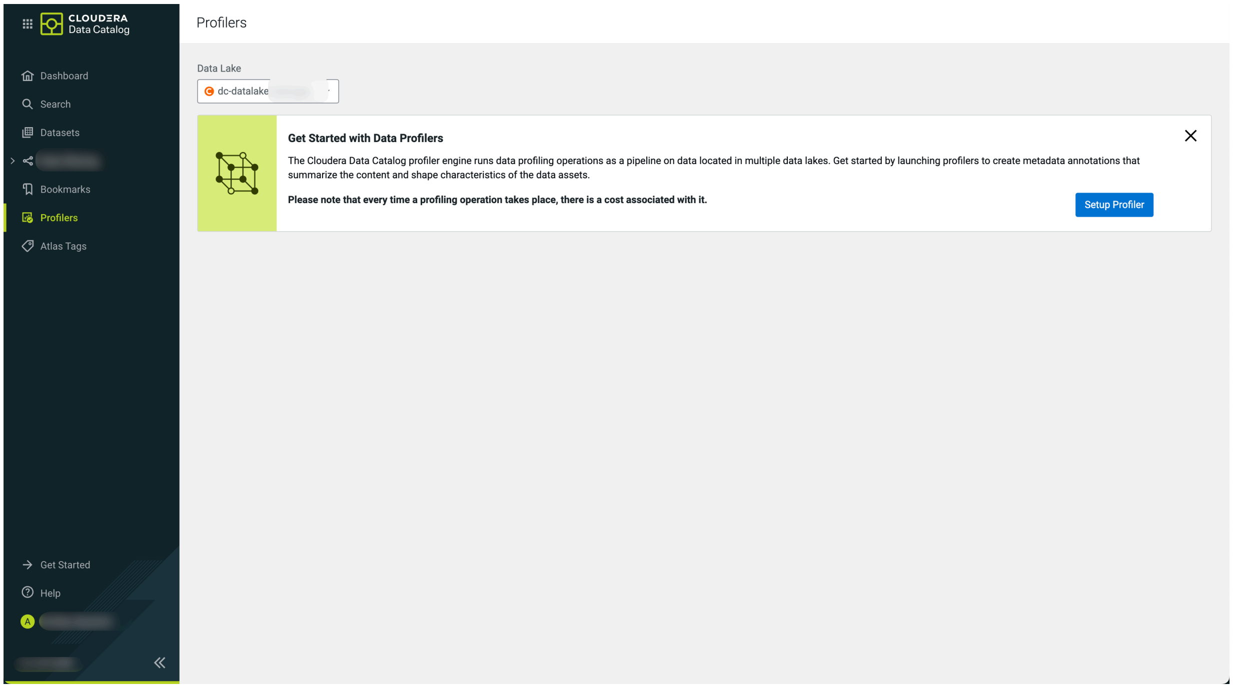
In Compute Cluster enabled environments, after you set up the Kubernetes profiler node group, the Profiler Launcher Service (PLS) keeps checking the availability of the node group automatically. Once the node group is ready, the PLS provisions the selected profilers by starting CRON jobs in the Kubernetes node groups.

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.



Profilers

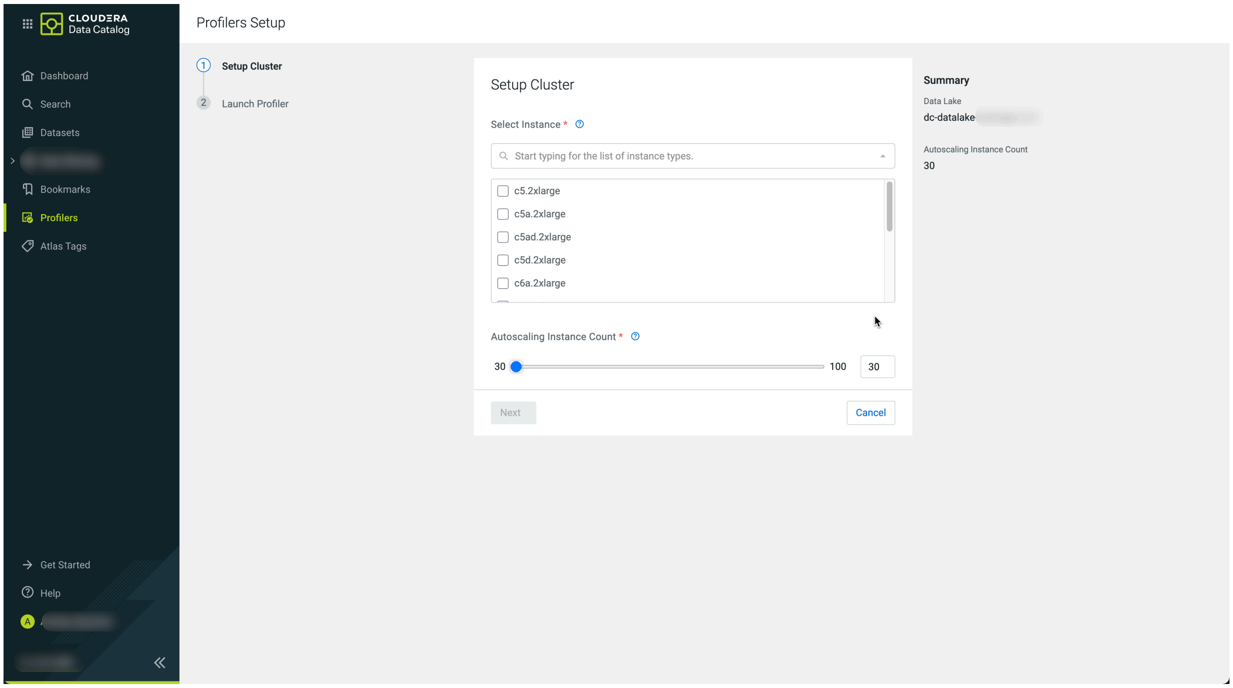
Data Lake: dc-datalake [redacted]

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 a profiling operation takes place, there is a cost associated with it.

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



Profilers Setup

Setup Cluster

Select Instance *

Start typing for the list of instance types.

□ c5.2xlarge
□ c5a.2xlarge
□ c5ad.2xlarge
□ c5d.2xlarge
□ c6a.2xlarge

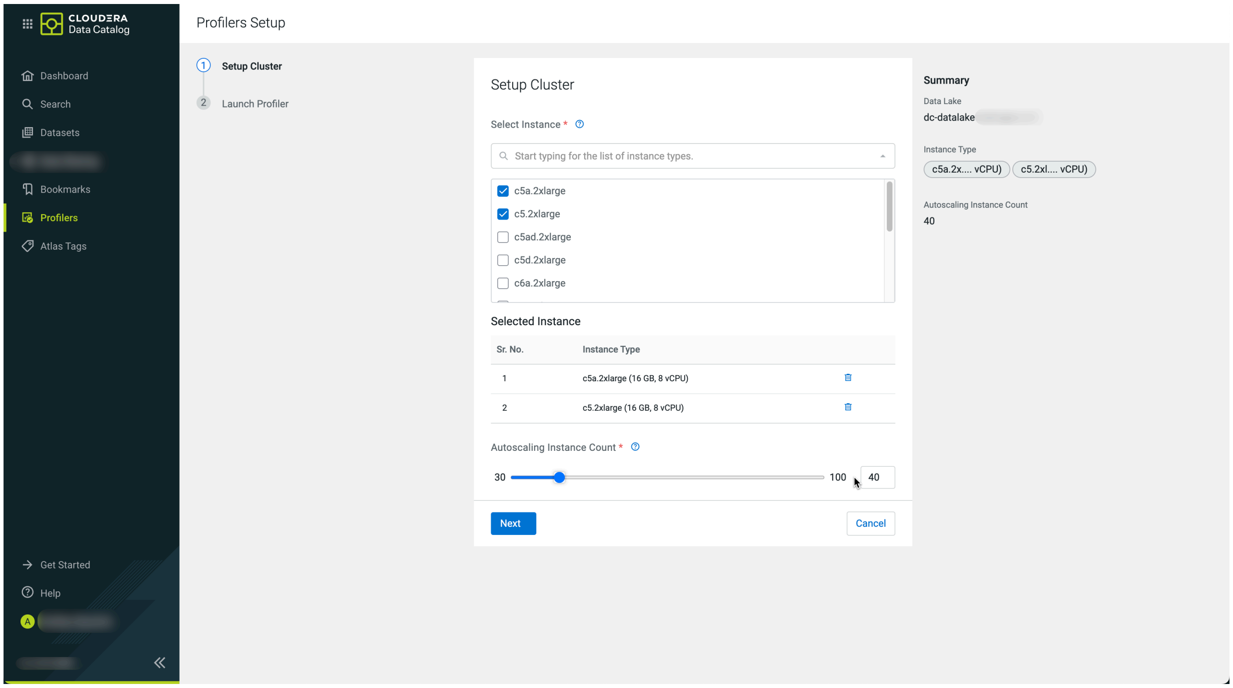
Autoscaling Instance Count *: 30 — 100, 30

Summary
Data Lake: dc-datalake [redacted]
Autoscaling Instance Count: 30

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.



Setup Cluster

Select Instance *

Start typing for the list of instance types.

☑ c5a.2xlarge
☑ c5.2xlarge
☐ c5ad.2xlarge
☐ c5d.2xlarge
☐ c6a.2xlarge

Selected Instance

Sr. No.	Instance Type
1	c5a.2xlarge (16 GB, 8 vCPU)
2	c5.2xlarge (16 GB, 8 vCPU)

Autoscaling Instance Count *

30 100 40

Summary

Data Lake
dc-datalake [redacted]

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

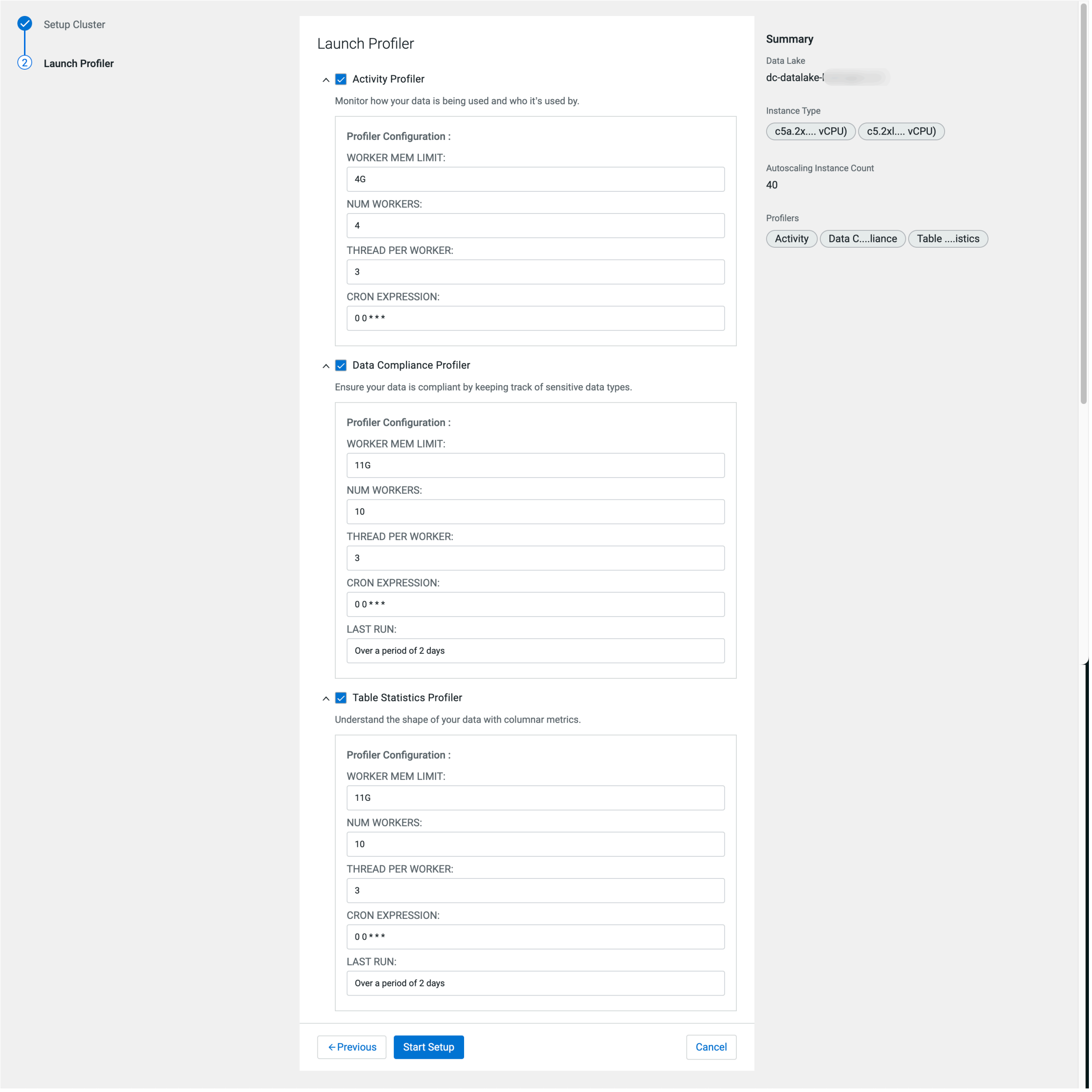
Autoscaling Instance Count
40

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



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-[redacted]

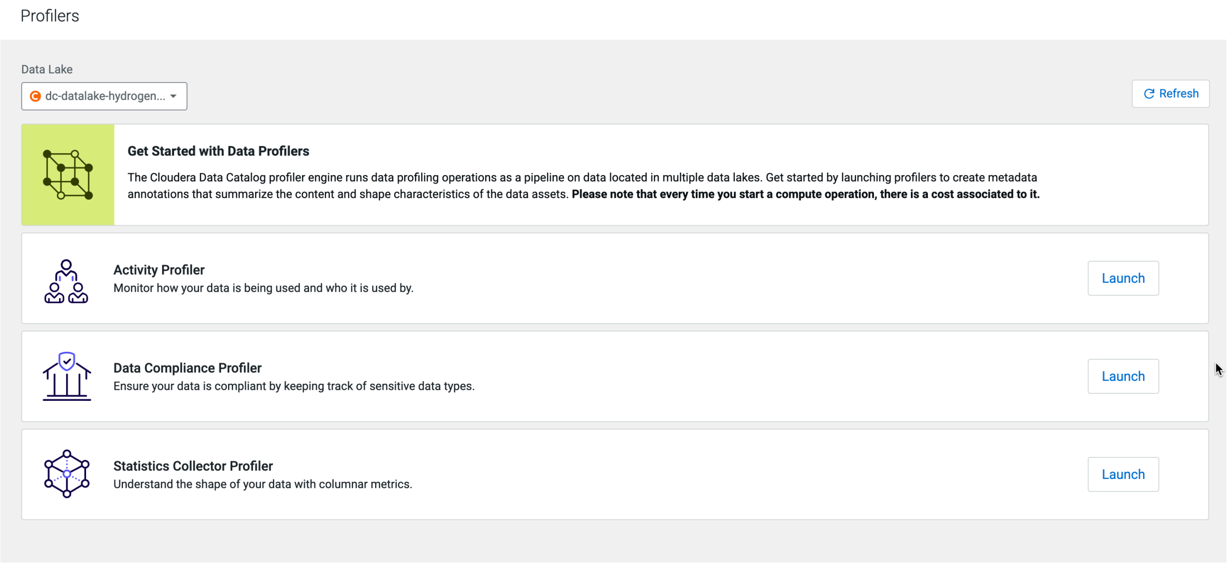
Instance Type: c5a.2x.... vCPU) c5.2xl.... vCPU)

Autoscaling Instance Count: 40

Profilers: Activity Data C....liance Table ....istics

7. Once the cluster is ready to accept Kubernetes profiler jobs, you can start the individual profilers by clicking Launch. If the profiler jobs were scheduled earlier, they will be automatically assigned to the finished Kubernetes node group.

   > **Note:** The readiness of the Kubernetes node group can be checked in Cloudera Management Console Environments <****YOUR_ENVIRONMENT****> Compute Clusters . The worker node group is created by the Liftie service. The expected setup time is around 15 to 30 minutes.


Profilers

Data Lake

dc-datalake-hydrogen...

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.**

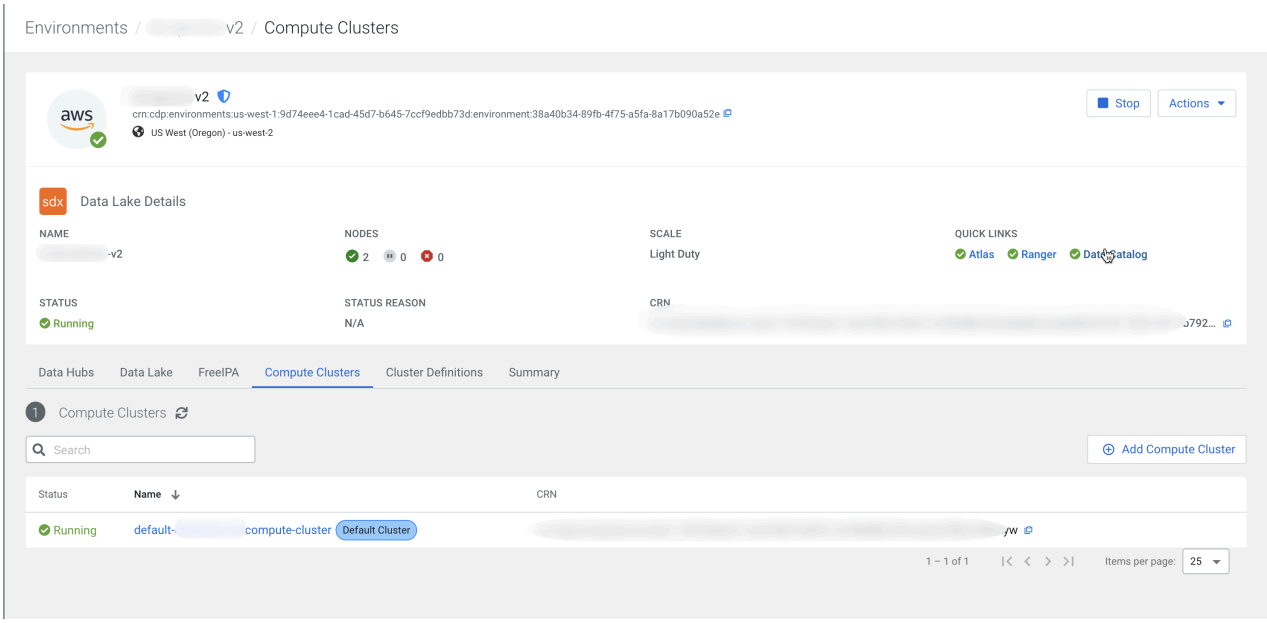
Activity Profiler
Monitor how your data is being used and who it is used by.

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

Statistics Collector Profiler
Understand the shape of your data with columnar metrics.

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 / [redacted]v2 / Compute Clusters

[redacted]v2

crn:cdp:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:environment:38a40b34-89fb-4f75-a5fa-8a17b090a52e

US West (Oregon) - us-west-2

sdx Data Lake Details

NAME: [redacted]-v2

NODES: 2 0 0

SCALE: Light Duty

QUICK LINKS: Atlas Ranger Data Catalog

STATUS: Running

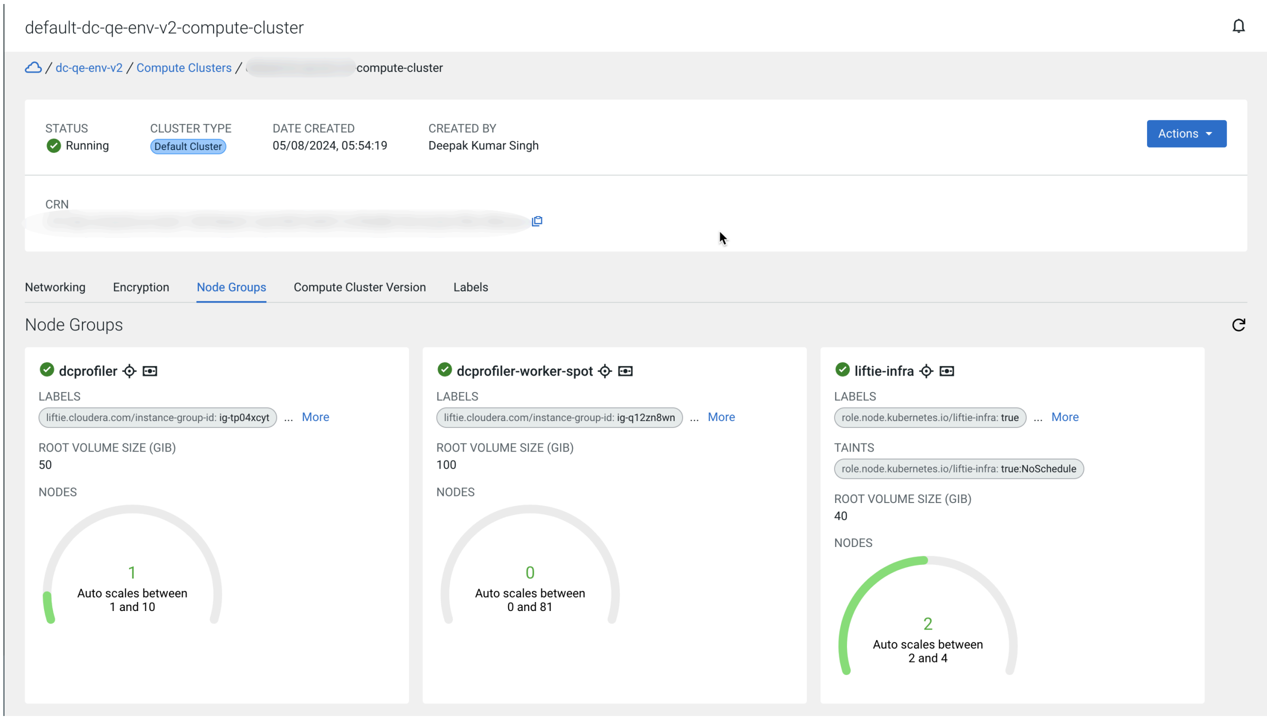
STATUS REASON: N/A

CRN: [redacted]5792...

1 Compute Clusters

Status	Name	CRN
Running	default-[redacted]compute-cluster Default Cluster	[redacted]yw

1 – 1 of 1 Items per page: 25



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

/ dc-qe-env-v2 / Compute Clusters / [redacted]compute-cluster

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

CRN: [redacted]

Node Groups

dcprofiler
LABELS: liftie.cloudera.com/instance-group-id: ig-tp04xcyt ... More
ROOT VOLUME SIZE (GIB): 50
NODES: 1 Auto scales between 1 and 10

dcprofiler-worker-spot
LABELS: liftie.cloudera.com/instance-group-id: ig-q12zn8wn ... More
ROOT VOLUME SIZE (GIB): 100
NODES: 0 Auto scales between 0 and 81

liftie-infra
LABELS: role.node.kubernetes.io/liftie-infra: true ... More
TAINTS: role.node.kubernetes.io/liftie-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:

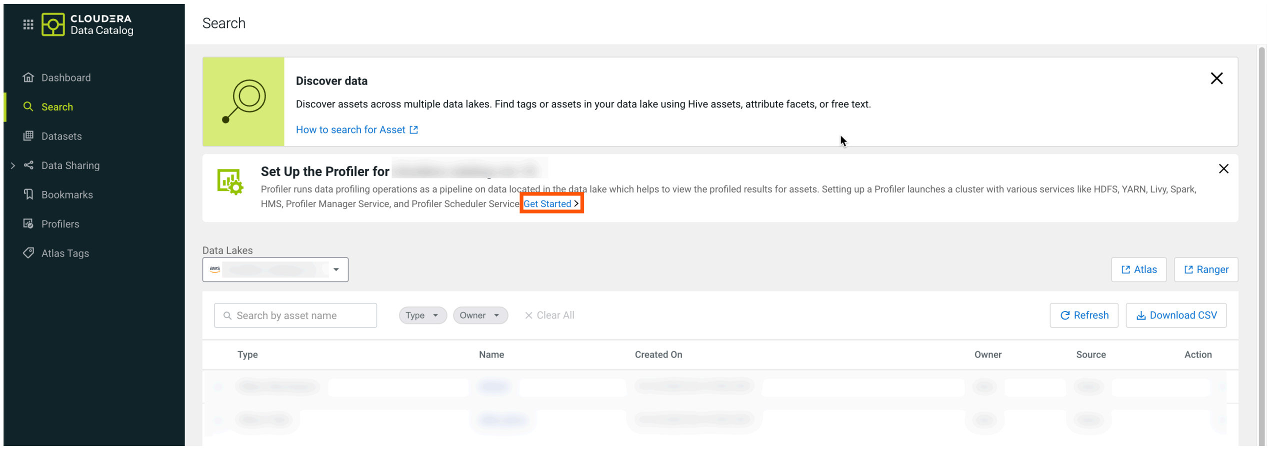
- Profiler Manager
- Profiler Scheduler

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.



Search

Discover data

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

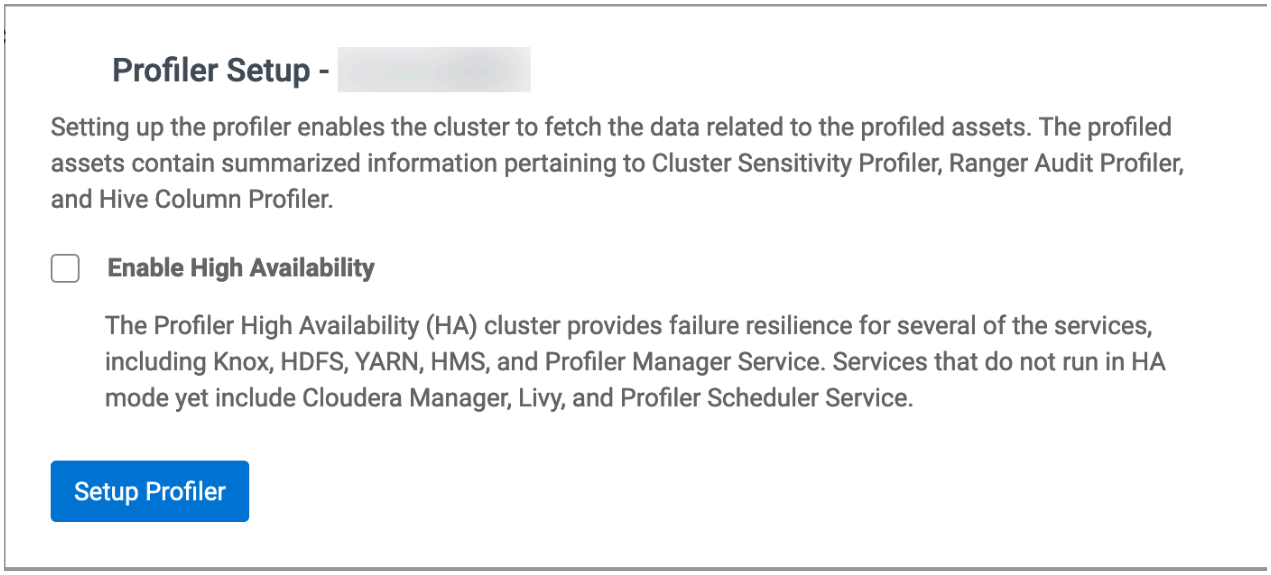
How to search for Asset

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 >

Data Lakes [redacted]

Type	Name	Created On	Owner	Source	Action



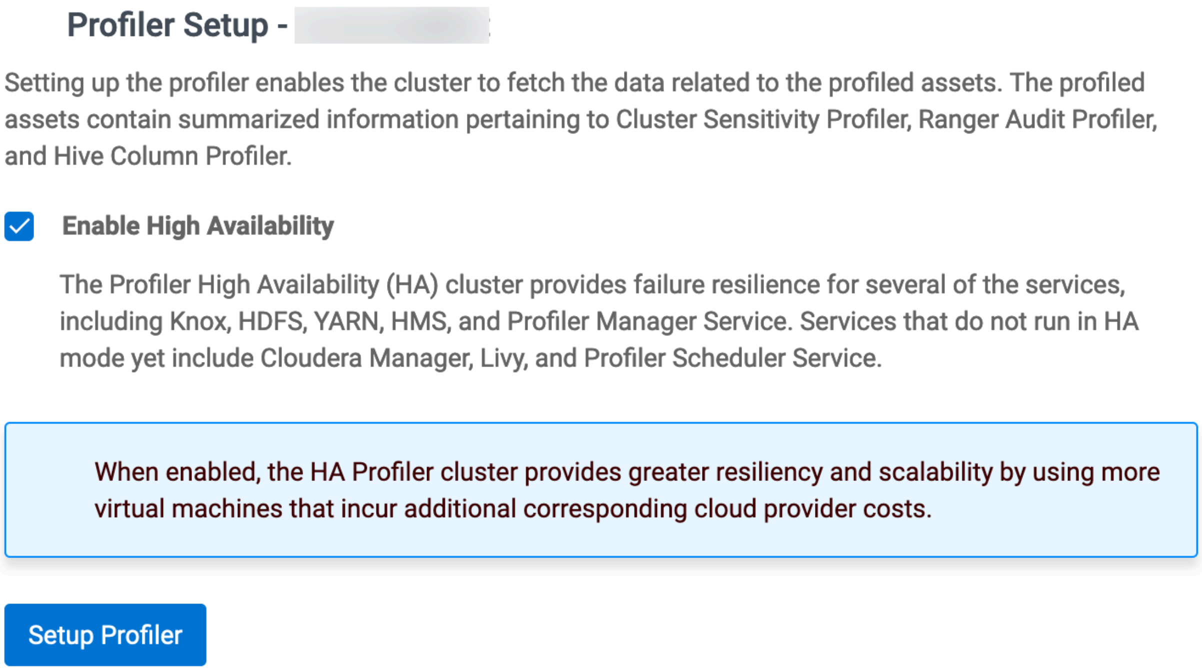
Profiler Setup - [redacted]

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]

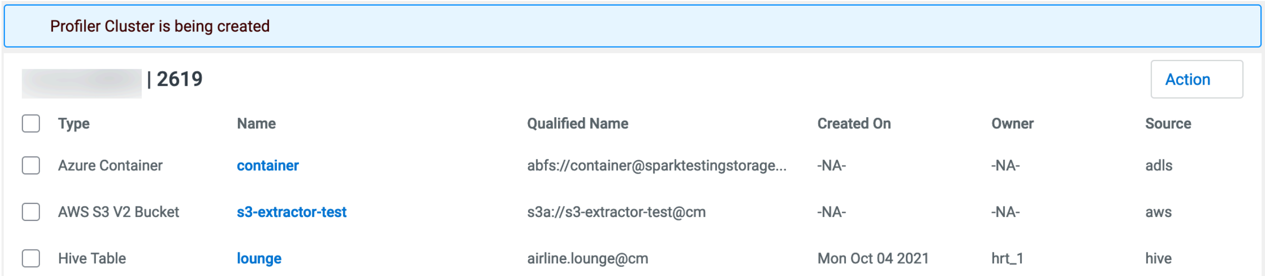
Setting up the profiler enables the cluster to fetch the data related to the profiled assets. The profiled assets contain summarized information pertaining to Cluster Sensitivity Profiler, Ranger Audit Profiler, and Hive Column Profiler.

☑ **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.

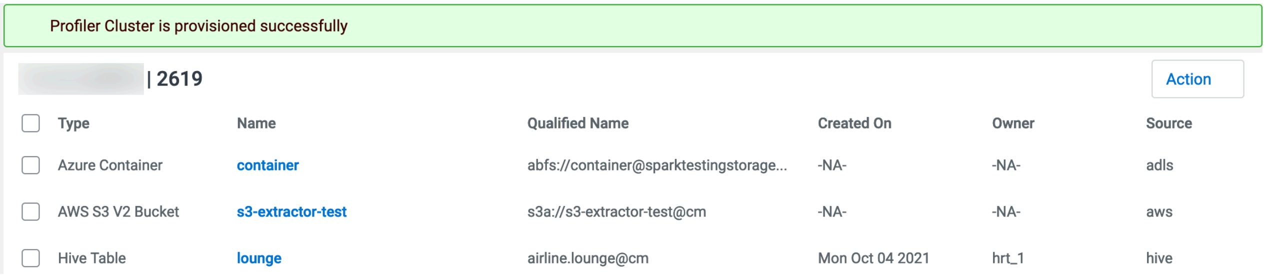


Profiler Cluster is being created

[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.



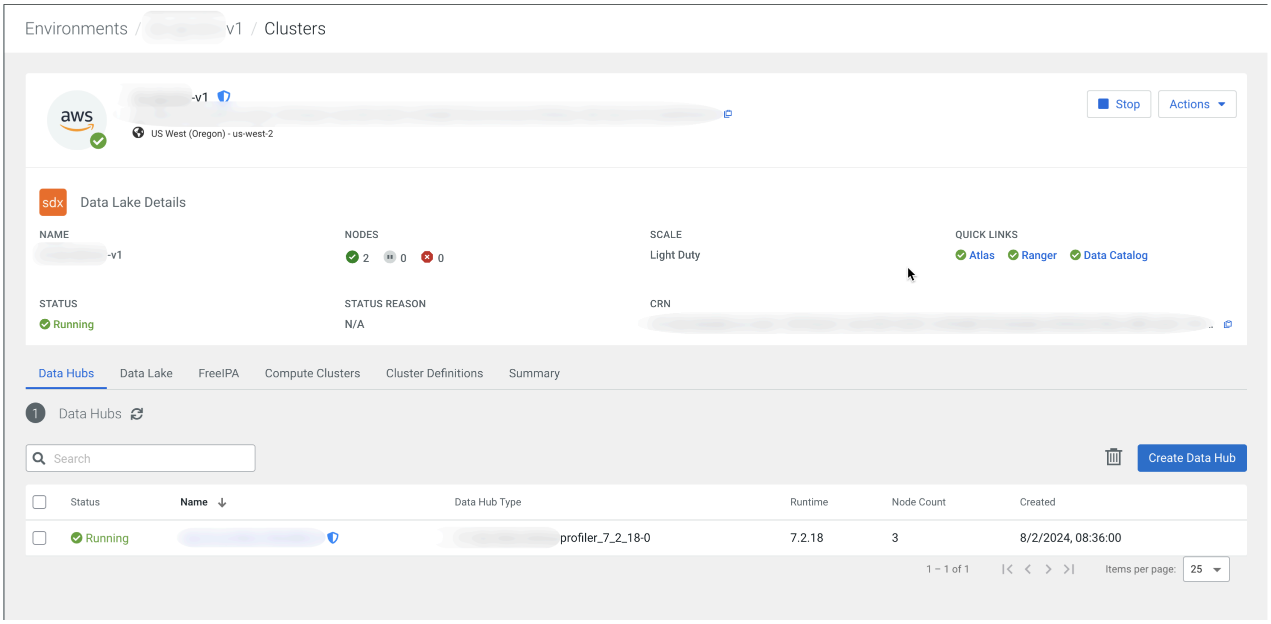
Profiler Cluster is provisioned successfully

[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

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

Configuring the Activity Profiler

Configure the scheduling and the available resources for your profiler.

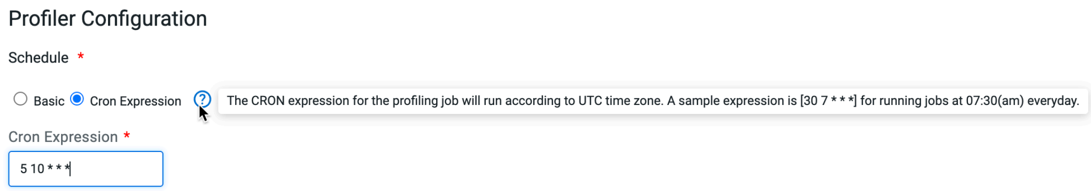
Procedure

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:** Both the Basic and Cron Expression scheduler (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



Profiler Configuration

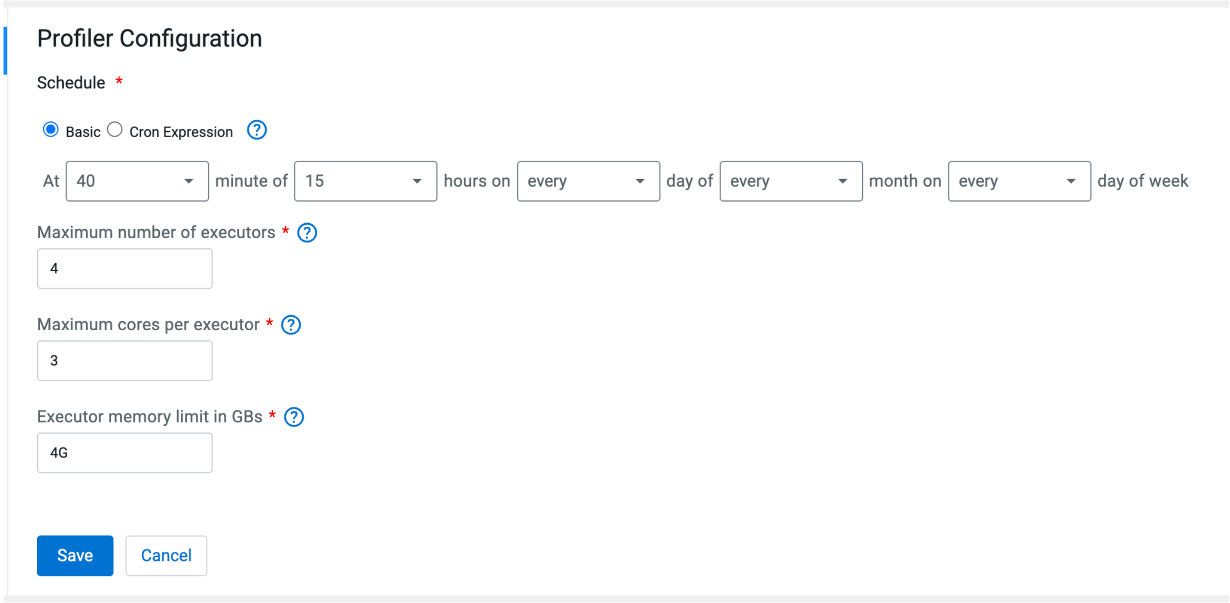
Schedule *

○ Basic ● Cron Expression ⓘ The CRON expression for the profiling job will run according to UTC time zone. A sample expression is [30 7 * * *] for running jobs at 07:30(am) everyday.

Cron Expression *

5 10 * * *

Figure 2: Profiler schedule with natural language



Profiler Configuration

Schedule *

● Basic ○ Cron Expression ⓘ

At 40 minute of 15 hours on every day of every month on every day of week

Maximum number of executors * ⓘ

4

Maximum cores per executor * ⓘ

3

Executor memory limit in GBs * ⓘ

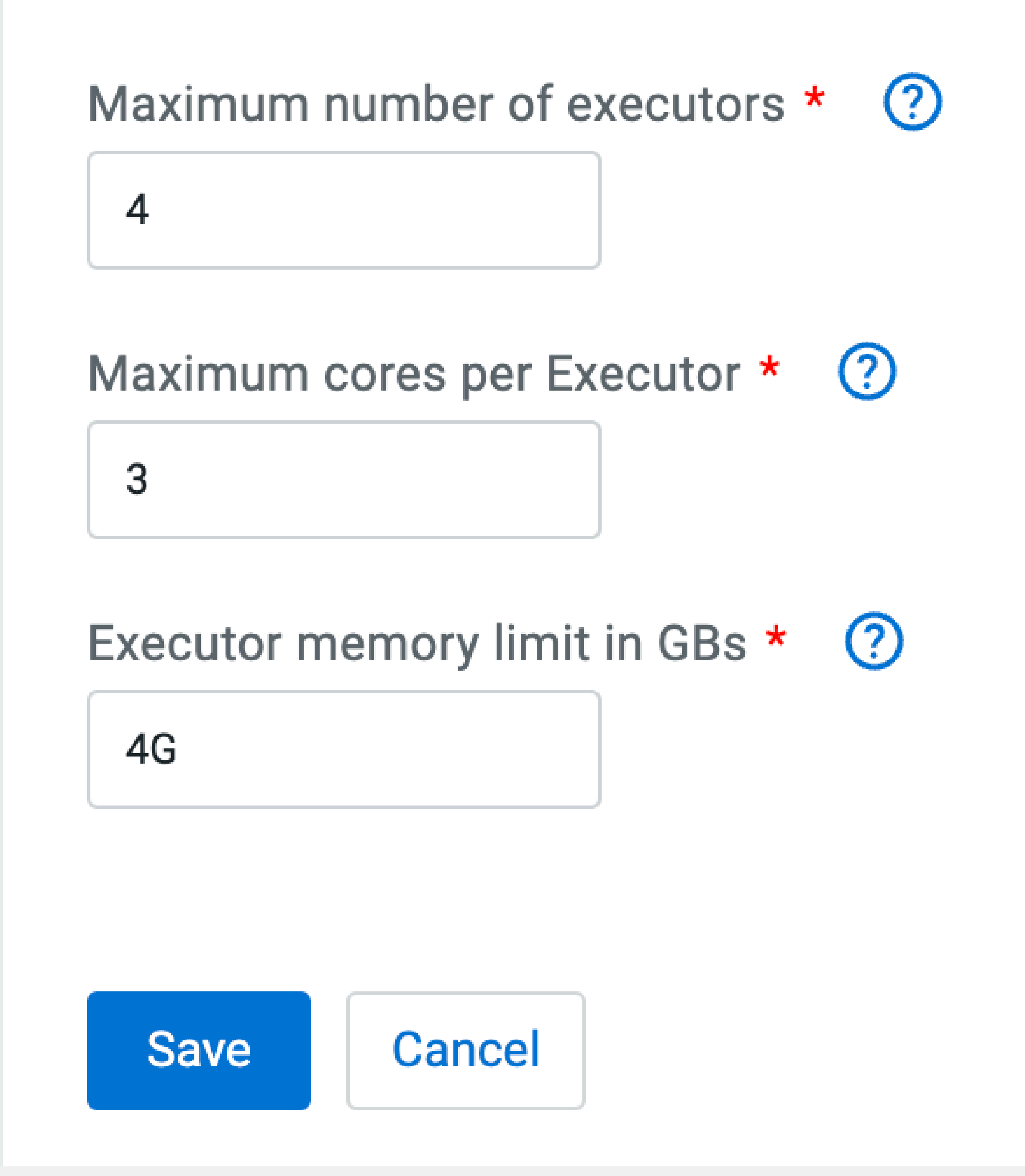
4G

> **Note:**
>
> Compute Cluster based profilers might hang if the underlying AWS cloud provider environment cannot provide the necessary memory for the executor instances. In this case, reconfigure your executors with 4-5 GB memory in Profiler Details Configuration .

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

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

Configuring the Ranger Audit Profiler

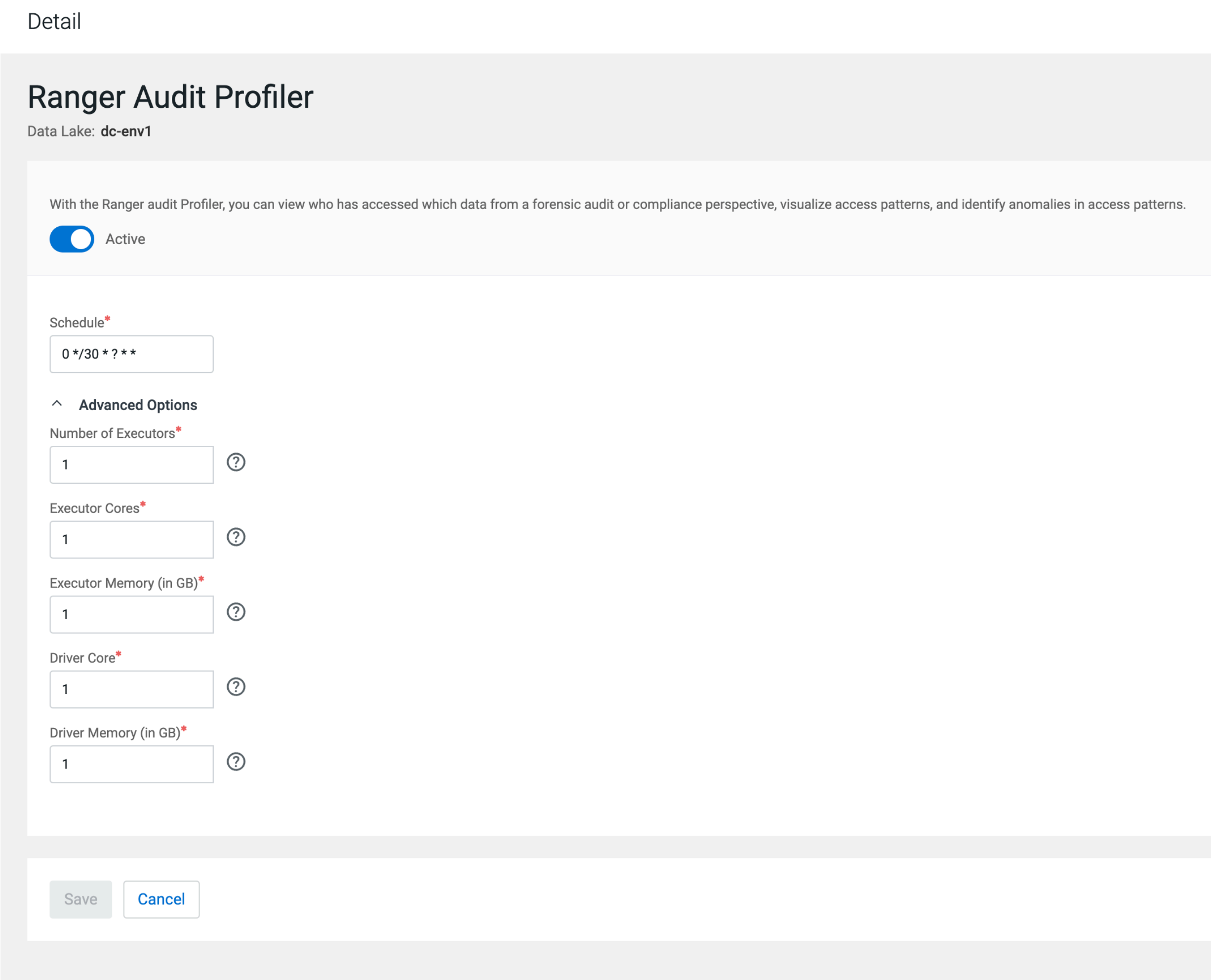
In addition to the generic configuration, there are additional parameters for the Ranger Audit Profiler that can be optionally edited.

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

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

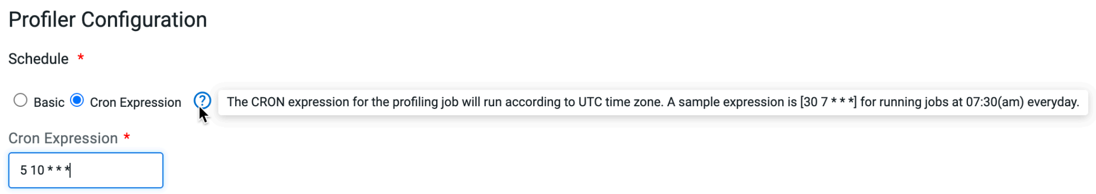
You can configure the scheduling and the available resources for your profiler.

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: Both the Basic and Cron Expression scheduler (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

   Profiler Configuration

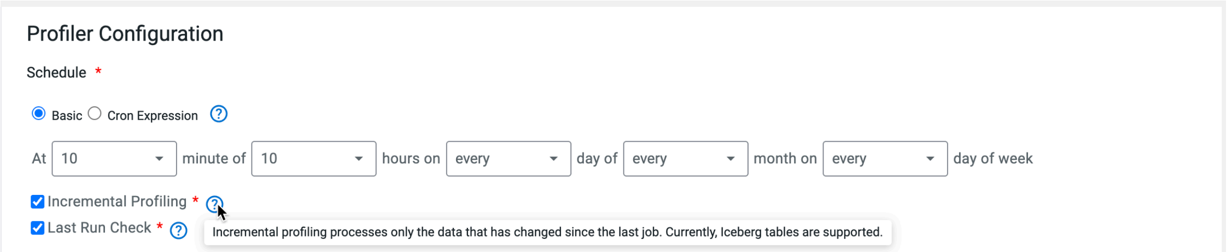
   Schedule *

   ○ Basic ● Cron Expression

   The CRON expression for the profiling job will run according to UTC time zone. A sample expression is [30 7 * * *] for running jobs at 07:30(am) everyday.

   Cron Expression *: 5 10 * * *

   ### Figure 4: Profiler schedule with natural language

   Profiler Configuration

   Schedule *

   ● Basic ○ Cron Expression

   At 10 minute of 10 hours on every day of every month on every day of week

   ☑ Incremental Profiling *

   ☑ Last Run Check *

   Incremental profiling processes only the data that has changed since the last job. Currently, Iceberg tables are supported.

4. Select Incremental Profiling when needed.

   Using Incremental Profiling can decrease the compute resources and the time needed for the profiling job by processing only the information (only Iceberg tables) updated or added since previous job.

   Using Incremental Profiling, you can refine the results from the Last Run Check. Incremental Profiling checks the data (rows) in assets, while Last Run Check filters complete assets.

5. 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.

   The Last Run Check precedes Incremental Profiling.

6. Continue with resource settings:

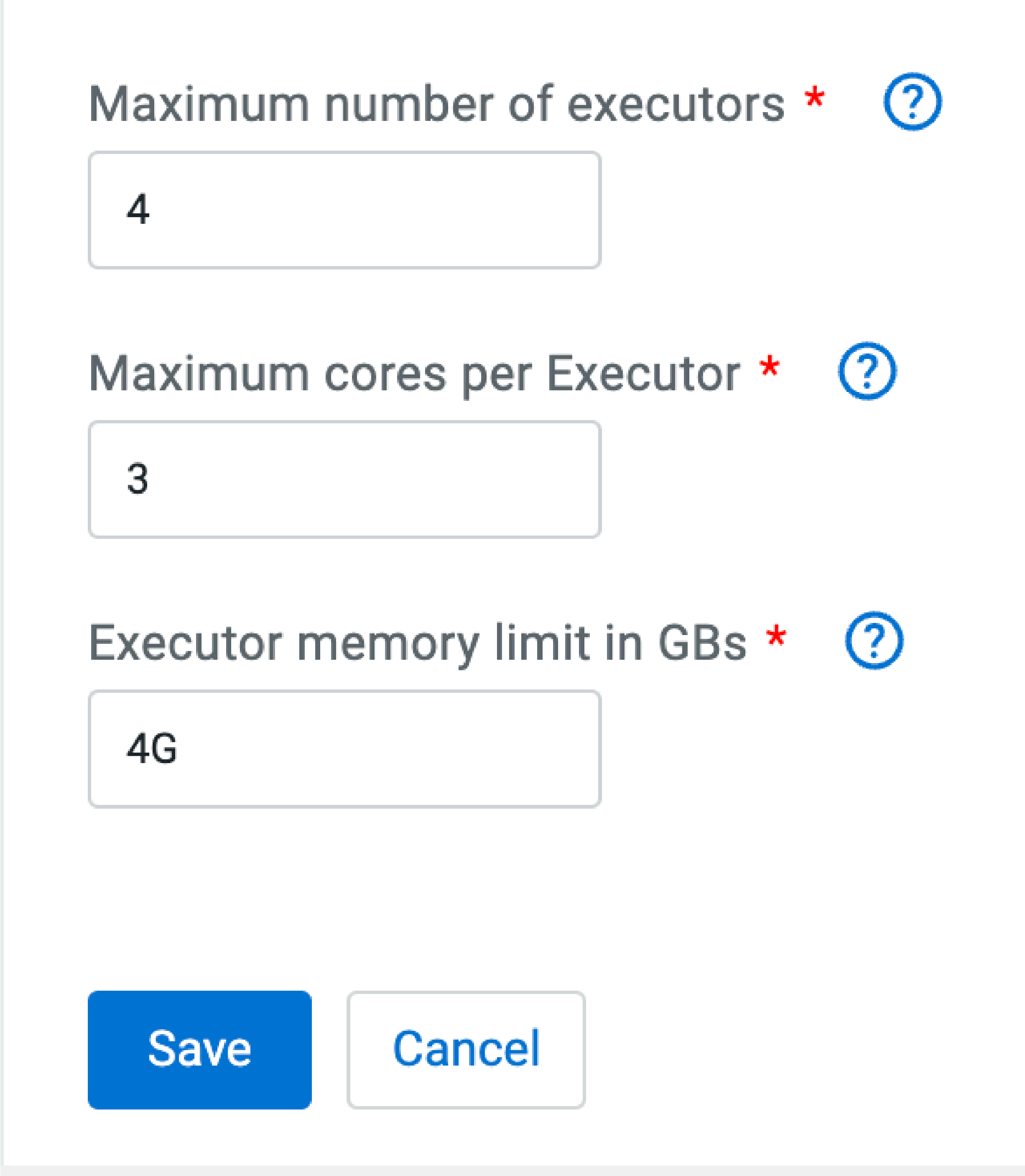
a) Set the Maximum number of executors

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

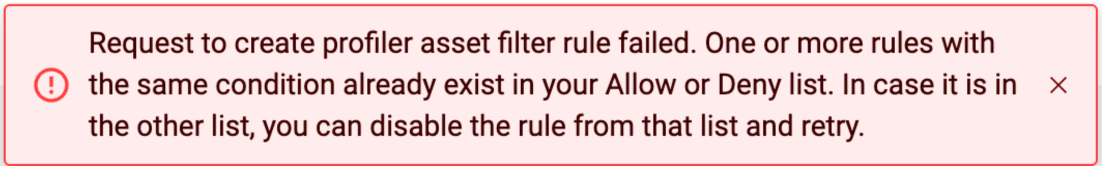
Note:

Compute Cluster based profilers might hang if the underlying AWS cloud provider environment cannot provide the necessary memory for the executor instances. In this case, reconfigure your executors with 4-5 GB memory in Profiler Details Configuration .

7. Click Save to apply the configuration changes to the selected profiler.
8. 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.
- 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 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**.

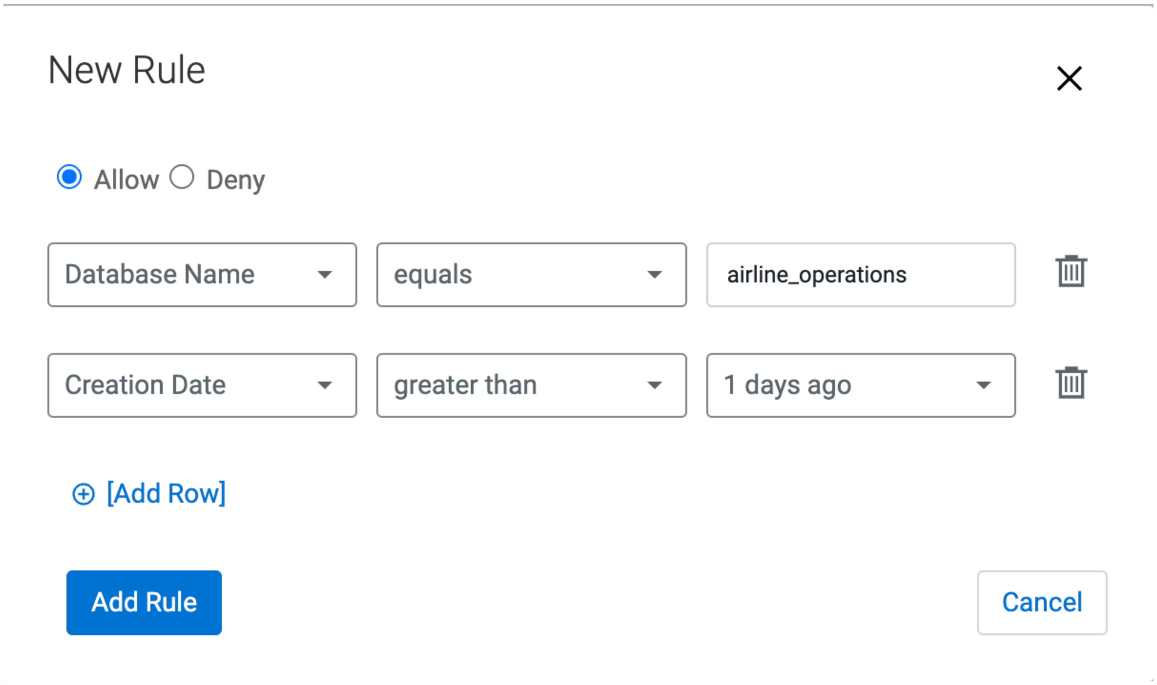
1. Click Add New Rule to define new rules.
2. Use the radio buttons to define your new rule for the Allow or Deny List.
3. Select the key from the drop-down list and the relevant operator. You can select from the following:

Key	Operator
Database name	• 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.



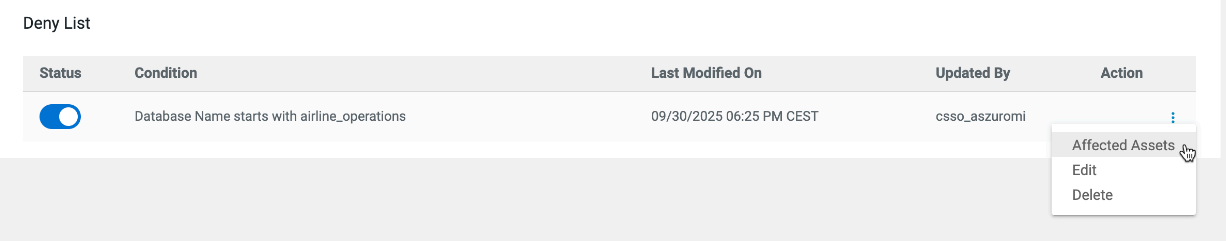
New Rule

● Allow ○ Deny

Database Name | equals | airline_operations

Creation Date | greater than | 1 days ago

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

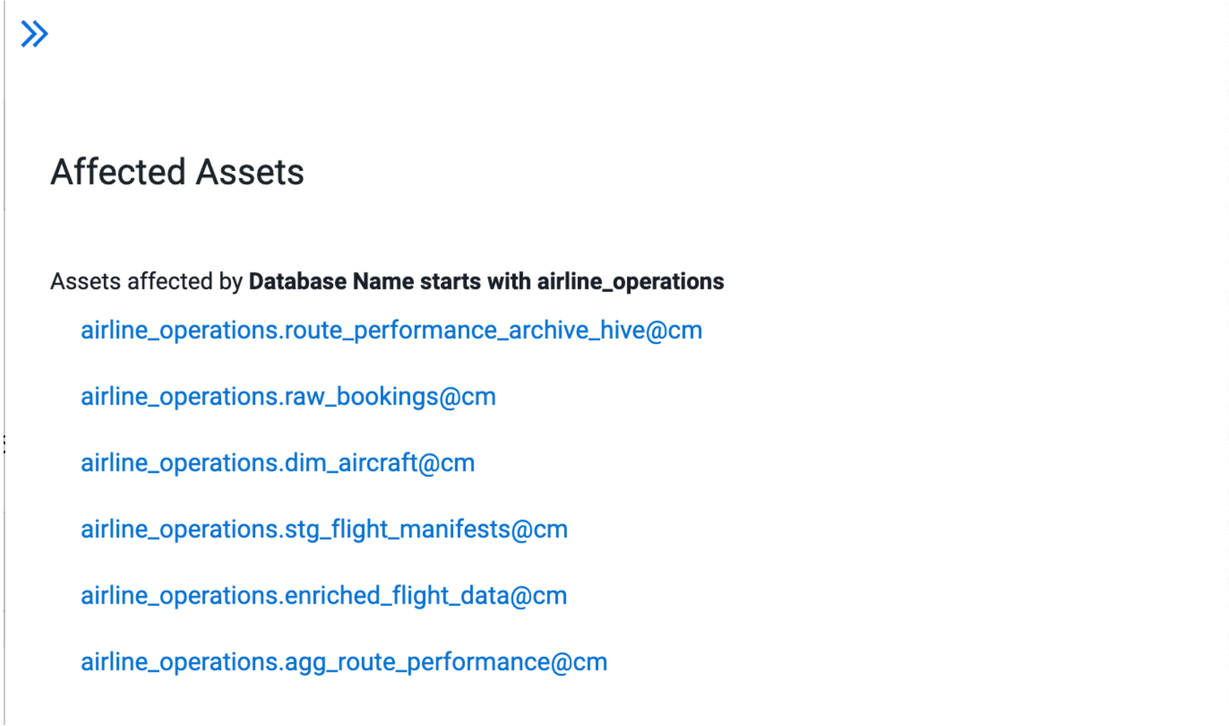


Deny List

Status	Condition	Last Modified On	Updated By	Action
●	Database Name starts with airline_operations	09/30/2025 06:25 PM CEST	csso_aszuromi	⋮

Figure 5: Affected Assets in Asset Filtering Rules configuration

[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.

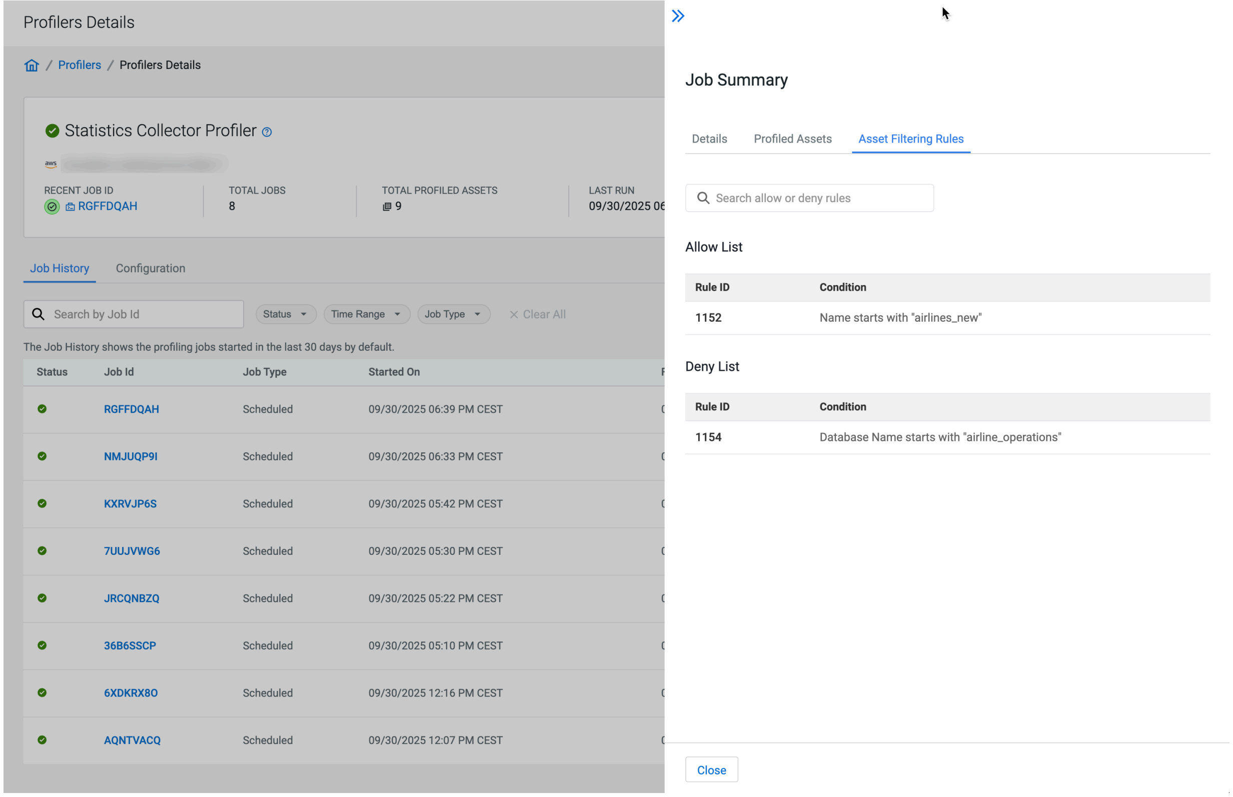


Affected Assets

Assets affected by **Database Name starts with airline_operations**

- airline_operations.route_performance_archive_hive@cm
- airline_operations.raw_bookings@cm
- airline_operations.dim_aircraft@cm
- airline_operations.stg_flight_manifests@cm
- airline_operations.enriched_flight_data@cm
- airline_operations.agg_route_performance@cm

Job Summary shows the asset filtering rules applied for the particular profiling job:



Profilers Details

Profilers / Profilers Details

Statistics Collector Profiler

RECENT JOB ID	TOTAL JOBS	TOTAL PROFILED ASSETS	LAST RUN
RGFFDQAH	8	9	09/30/2025 06

The Job History shows the profiling jobs started in the last 30 days by default.

Status	Job Id	Job Type	Started On
✓	RGFFDQAH	Scheduled	09/30/2025 06:39 PM CEST
✓	NMJUQP9I	Scheduled	09/30/2025 06:33 PM CEST
✓	KXRVJP6S	Scheduled	09/30/2025 05:42 PM CEST
✓	7UUJVWG6	Scheduled	09/30/2025 05:30 PM CEST
✓	JRCQNBZQ	Scheduled	09/30/2025 05:22 PM CEST
✓	36B6SSCP	Scheduled	09/30/2025 05:10 PM CEST
✓	6XDKRX8O	Scheduled	09/30/2025 12:16 PM CEST
✓	AQNTVACQ	Scheduled	09/30/2025 12:07 PM CEST

Job Summary

Allow List

Rule ID	Condition
1152	Name starts with "airlines_new"

Deny List

Rule ID	Condition
1154	Database Name starts with "airline_operations"

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.

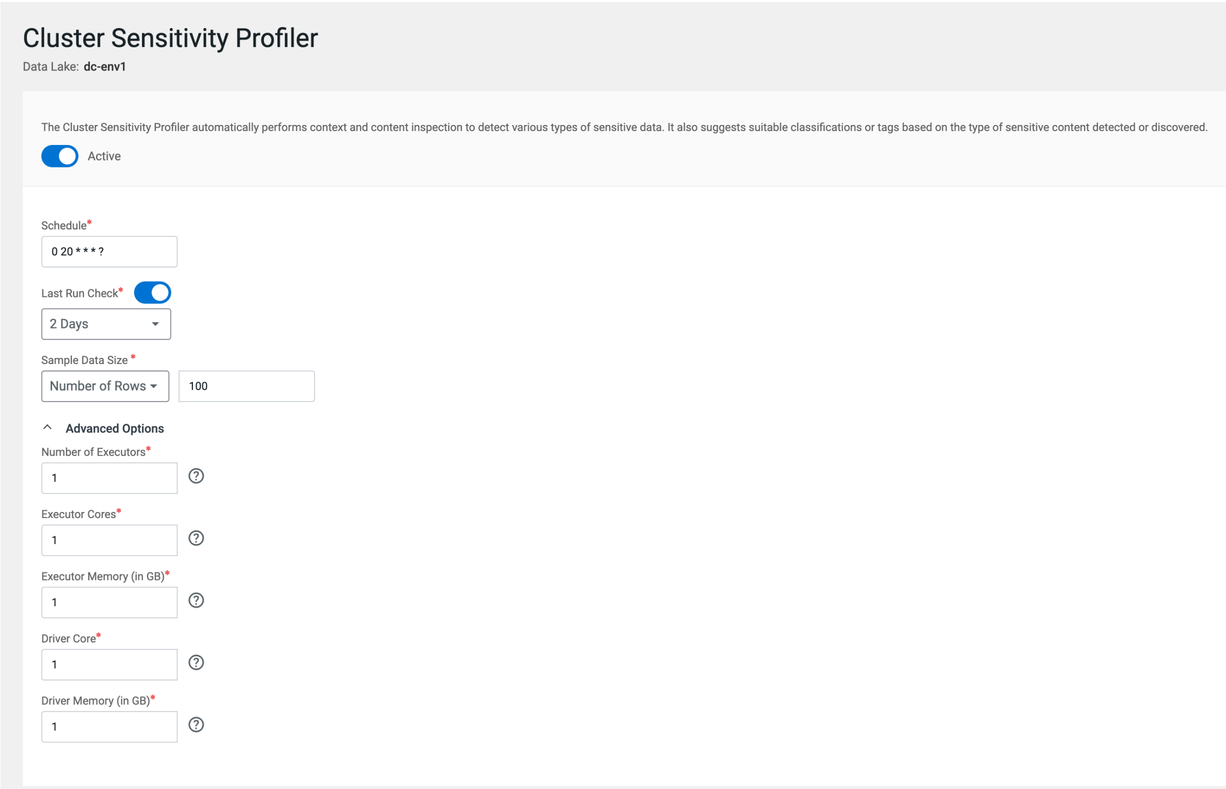
Before you begin

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.

   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**.

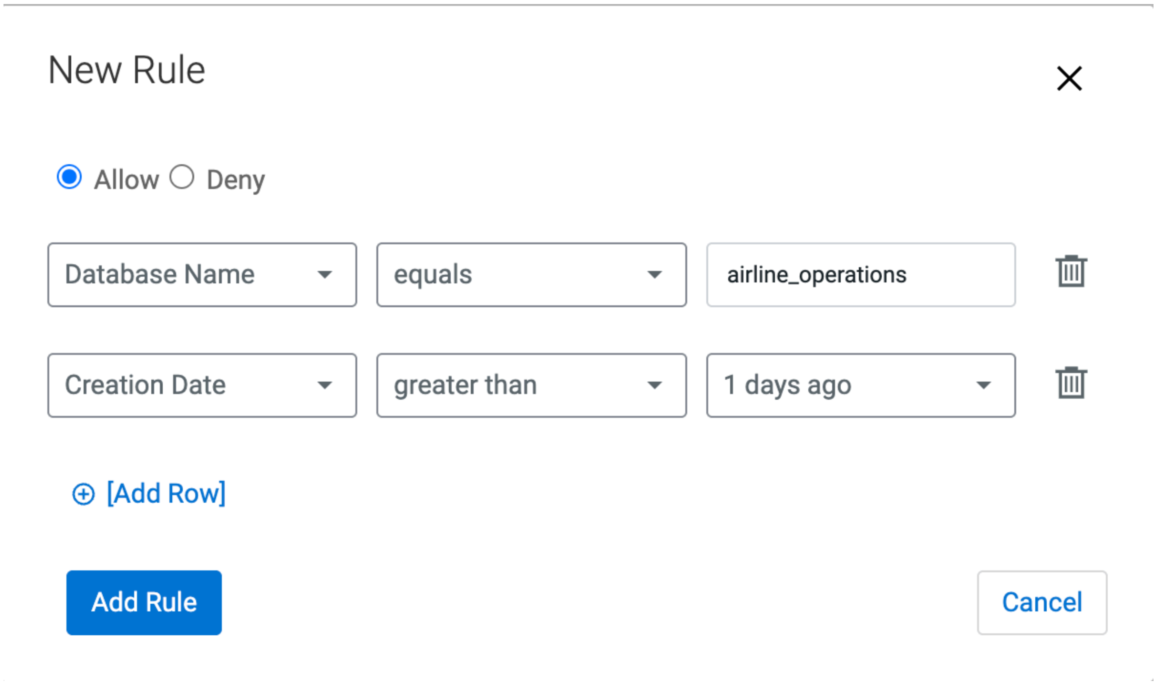
    1. Select the **Deny-list** or **Allow List** tab.
    2. Click Add New to define new rules.
    3. Select the key from the drop-down list and the relevant operator. You can select from the following:

Key	Operator
Database name	• equals • starts with • ends with

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

Configuring the Statistics Collector profiler

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

Procedure

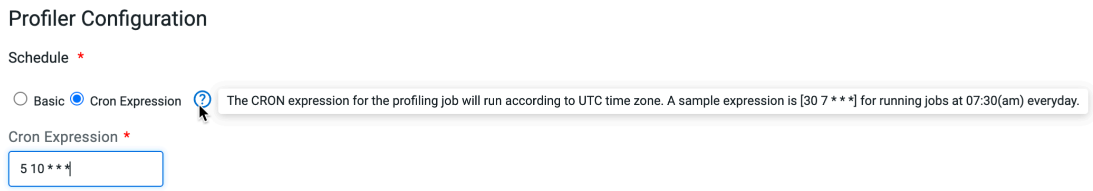
1. Go to **Profilers** and select your data lake.

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

   Note: Both the Basic and Cron Expression scheduler (Unix in Compute Cluster enabled environments cron jobs) use the UTC timezone instead of the local timezone of the user.

   Figure 6: Profiler schedule with cron expression

   Profiler Configuration

   Schedule *

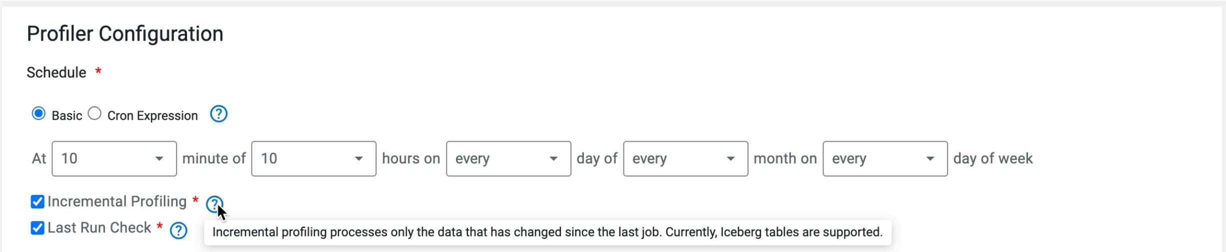
   ○ Basic ● Cron Expression ?

   The CRON expression for the profiling job will run according to UTC time zone. A sample expression is [30 7 * * *] for running jobs at 07:30(am) everyday.

   Cron Expression *

   5 10 * * *

   Figure 7: Profiler schedule with natural language

   Profiler Configuration

   Schedule *

   ● Basic ○ Cron Expression ?

   At 10 minute of 10 hours on every day of every month on every day of week

   ☑ Incremental Profiling * ?

   ☑ Last Run Check * ?

   Incremental profiling processes only the data that has changed since the last job. Currently, Iceberg tables are supported.

3. Select Incremental Profiling when needed.

   Using Incremental Profiling can decrease the compute resources and the time needed for the profiling job by processing only the information (only Iceberg tables) updated or added since previous job.

   Using Incremental Profiling, you can refine the results from the Last Run Check. Incremental Profiling checks the data (rows) in assets, while Last Run Check filters complete assets.

   Note: By Statistics Collector Profilers, the profiler compares the aggregated metrics between old and newly added data. Depending on the differences, this can slightly skew results. It is highly recommended to process the complete dataset time to time for the most accurate results.

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.

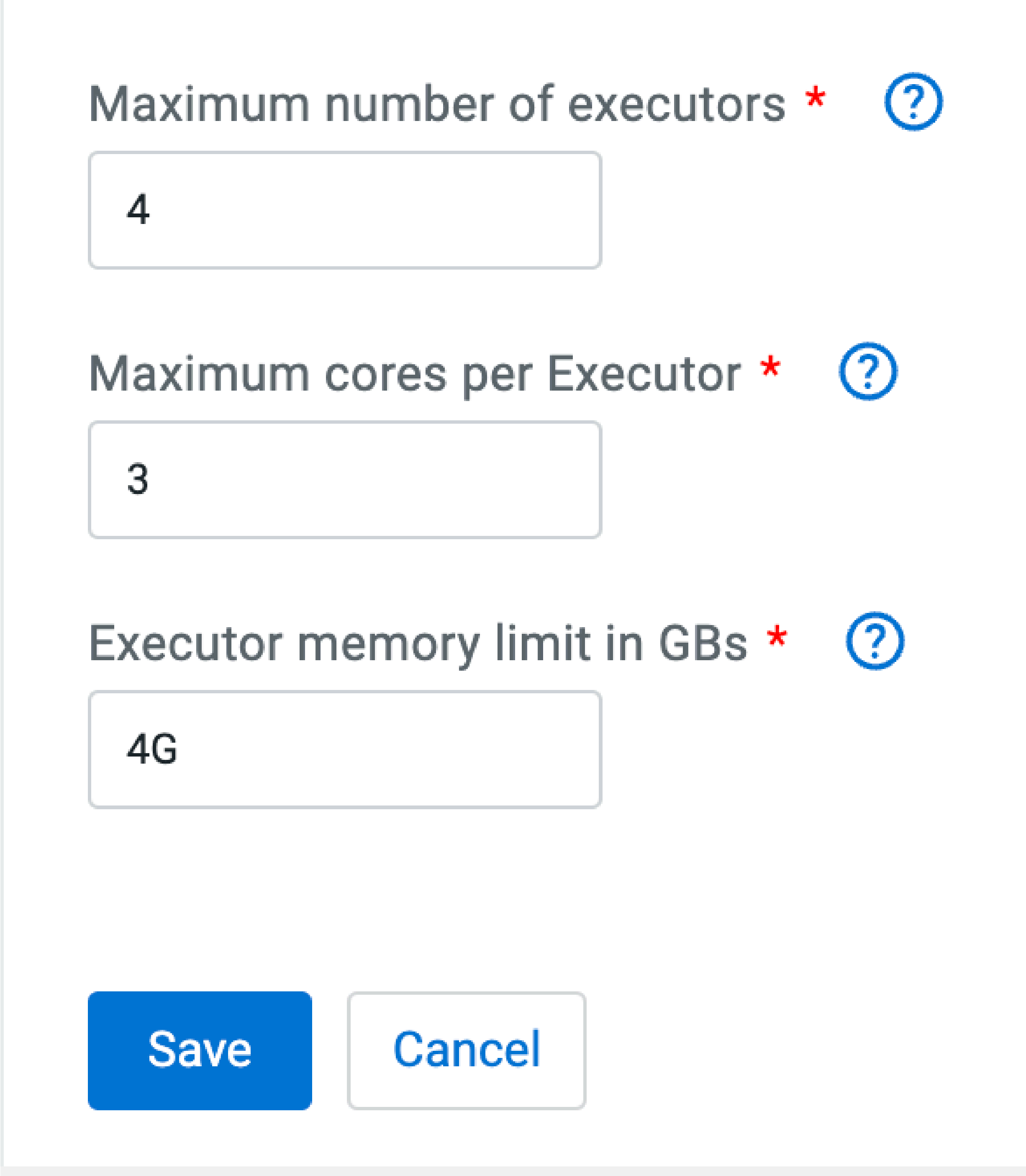
   If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.

   The Last Run Check precedes Incremental Profiling.

5. 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

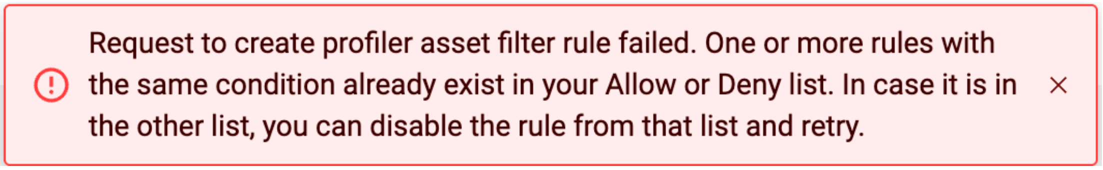
Executor memory limit in GBs * 4G

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

7. 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.
- 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 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**.

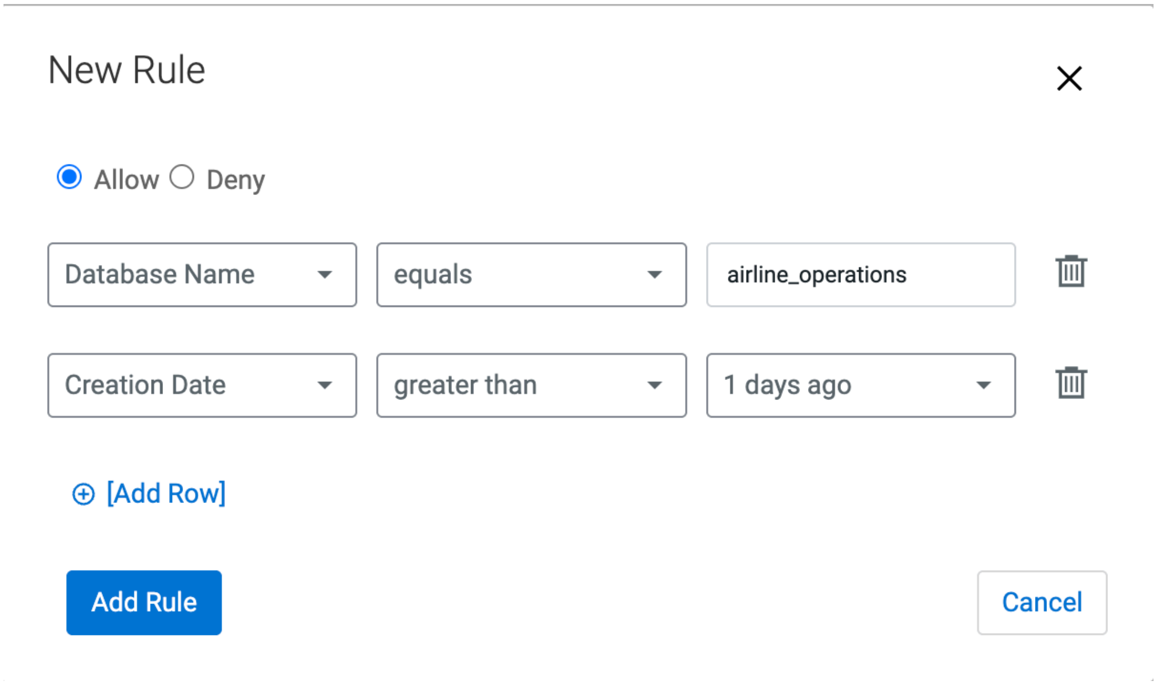
1. Click Add New Rule to define new rules.
2. Use the radio buttons to define your new rule for the Allow or Deny List.
3. Select the key from the drop-down list and the relevant operator. You can select from the following:

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

Key	Operator
Creation date[3]	• 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.

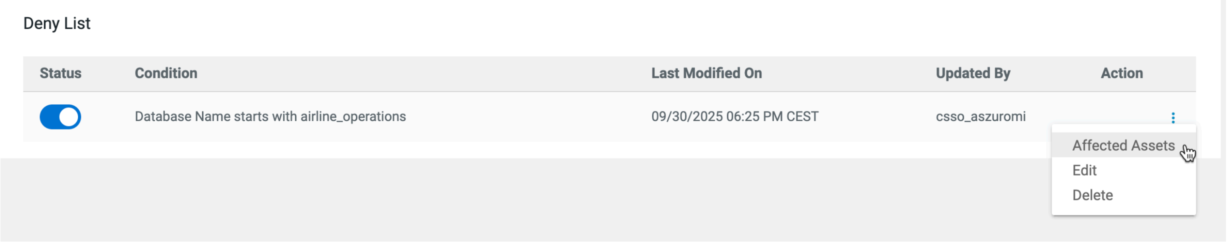


New Rule

● Allow ○ Deny

Database Name	equals	airline_operations
Creation Date	greater than	1 days ago

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

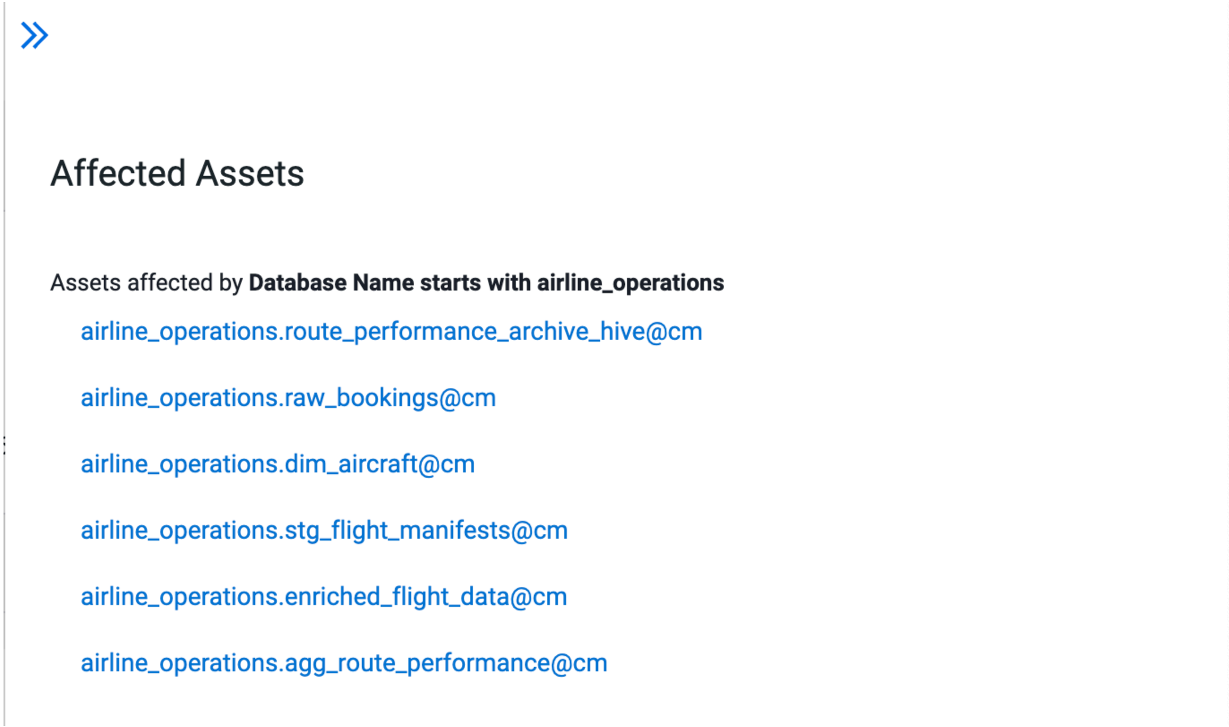


Deny List

Status	Condition	Last Modified On	Updated By	Action
●	Database Name starts with airline_operations	09/30/2025 06:25 PM CEST	csso_aszuromi	⋮

Figure 8: Affected Assets in Asset Filtering Rules configuration

[3] 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.

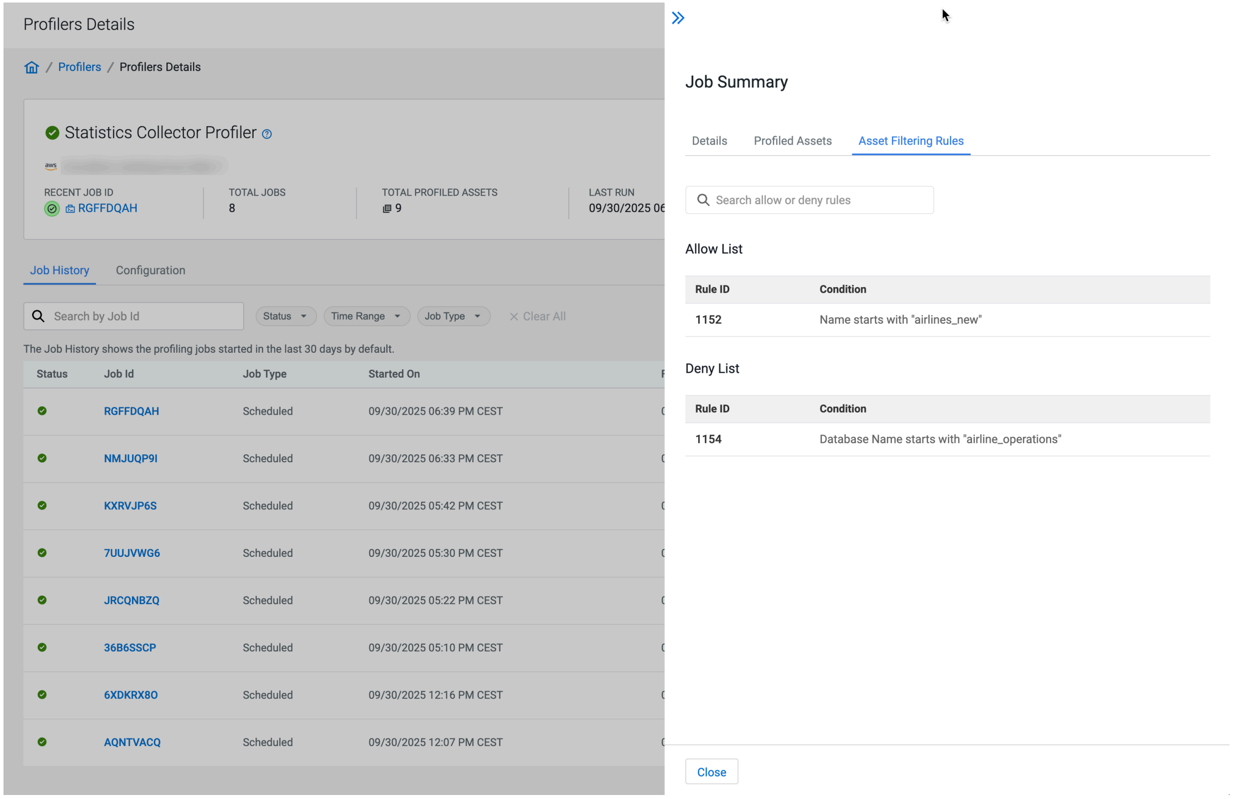


Affected Assets

Assets affected by **Database Name starts with airline_operations**

- airline_operations.route_performance_archive_hive@cm
- airline_operations.raw_bookings@cm
- airline_operations.dim_aircraft@cm
- airline_operations.stg_flight_manifests@cm
- airline_operations.enriched_flight_data@cm
- airline_operations.agg_route_performance@cm

Job Summary shows the asset filtering rules applied for the particular profiling job:



[redacted]

RECENT JOB ID	TOTAL JOBS	TOTAL PROFILED ASSETS	LAST RUN
RGFFDQAH	8	9	09/30/2025 06

The Job History shows the profiling jobs started in the last 30 days by default.

Status	Job Id	Job Type	Started On
✓	RGFFDQAH	Scheduled	09/30/2025 06:39 PM CEST
✓	NMJUQP9I	Scheduled	09/30/2025 06:33 PM CEST
✓	KXRVJP6S	Scheduled	09/30/2025 05:42 PM CEST
✓	7UUJVWG6	Scheduled	09/30/2025 05:30 PM CEST
✓	JRCQNBZQ	Scheduled	09/30/2025 05:22 PM CEST
✓	36B6SSCP	Scheduled	09/30/2025 05:10 PM CEST
✓	6XDKRX8O	Scheduled	09/30/2025 12:16 PM CEST
✓	AQNTVACQ	Scheduled	09/30/2025 12:07 PM CEST

Allow List

Rule ID	Condition
1152	Name starts with "airlines_new"

Deny List

Rule ID	Condition
1154	Database Name starts with "airline_operations"

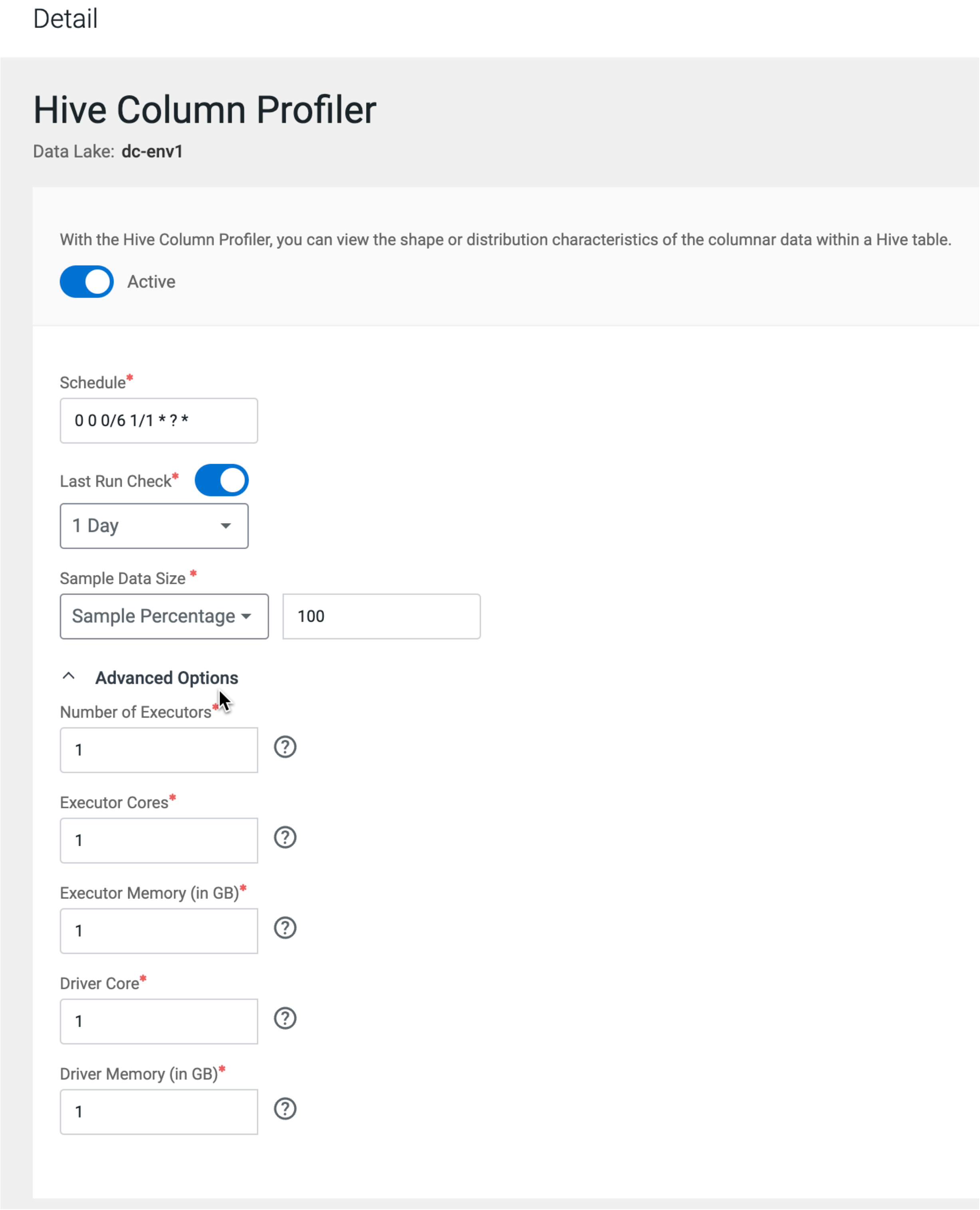
Configuring the Hive Column Profiler

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**.

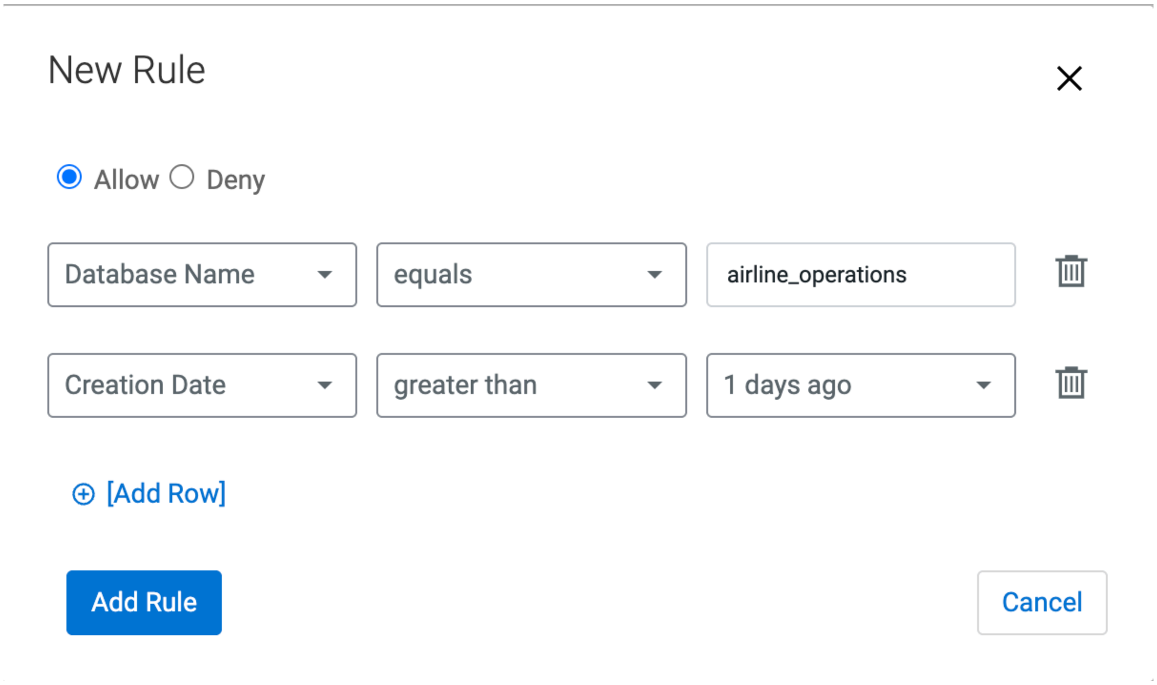
    1. Select the **Deny-list** or **Allow List** tab.
    2. Click Add New to define new rules.
    3. Select the key from the drop-down list and the relevant operator. You can select from the following:

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[4]	• 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

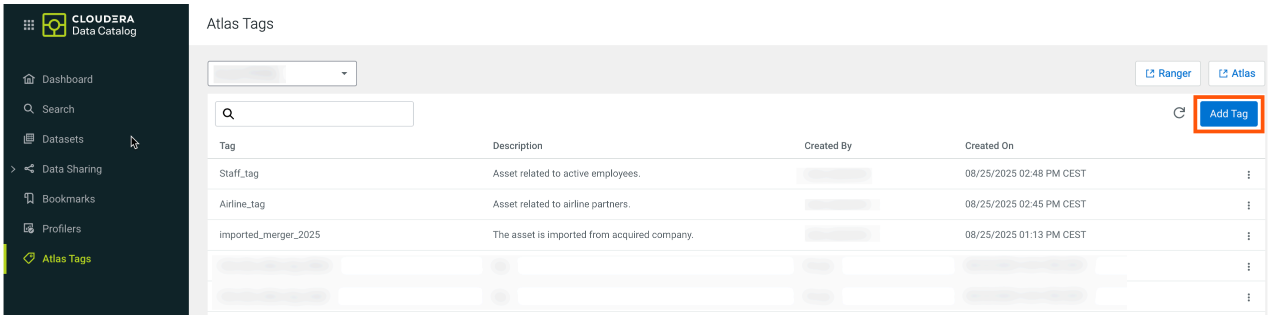
Atlas tag management

From the Atlas Tags menu, you can create, modify, and delete any of the Apache Atlas classifications to help data discovery and applying governance policies such as security and access control in Apache Ranger.

Creating Atlas tags

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.

[4] 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.



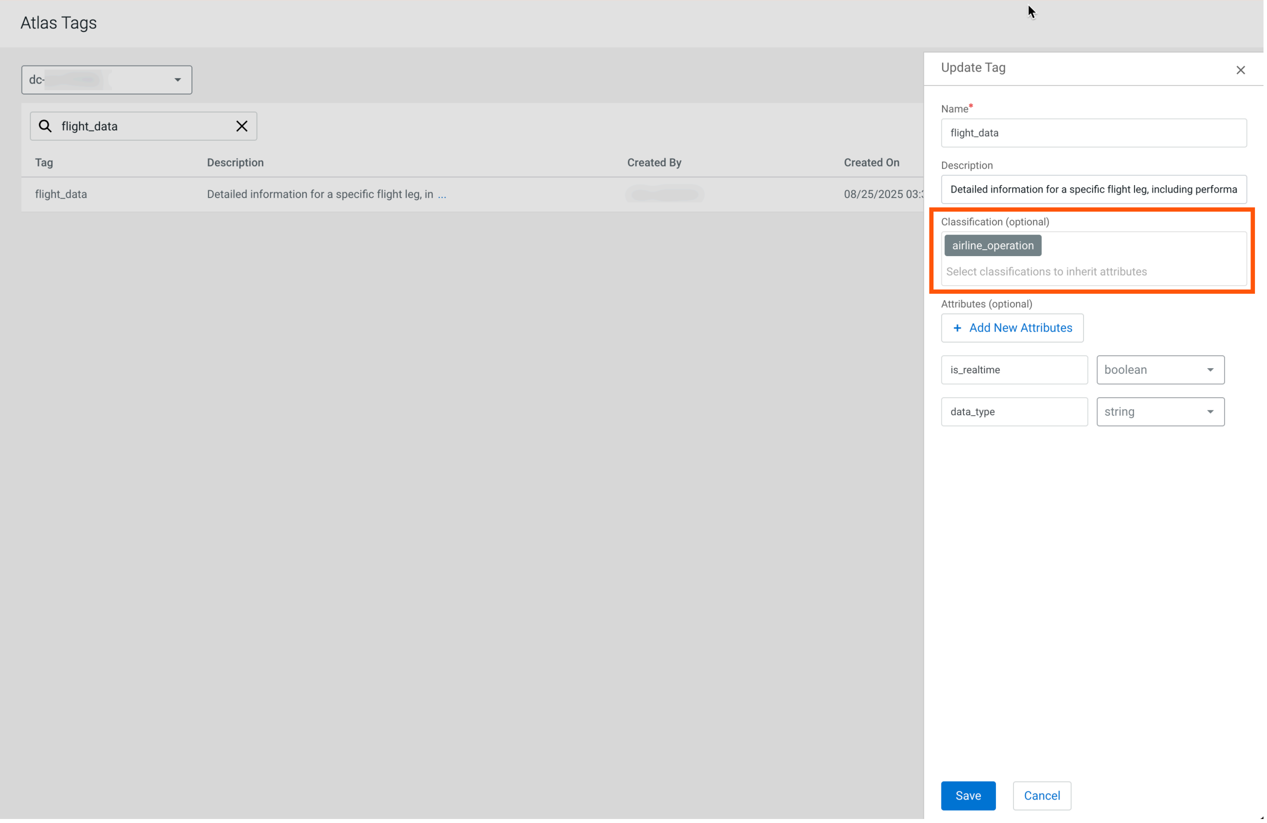
Atlas Tags

Tag	Description	Created By	Created On
Staff_tag	Asset related to active employees.	[redacted]	08/25/2025 02:48 PM CEST
Airline_tag	Asset related to airline partners.	[redacted]	08/25/2025 02:45 PM CEST
imported_merger_2025	The asset is imported from acquired company.	[redacted]	08/25/2025 01:13 PM CEST
[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]

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)

Note:

- Your classification still needs to be added to an asset in the **Search** or **Asset Details** menu.
- The inherited attributes are shown in Atlas. In Cloudera Data Catalog, you can only see the super-classification.



flight_data

Tag	Description	Created By	Created On
flight_data	Detailed information for a specific flight leg, in ...	[redacted]	08/25/2025 03:

Update Tag

Name*: flight_data

Description: Detailed information for a specific flight leg, including performa

Classification (optional): airline_operation

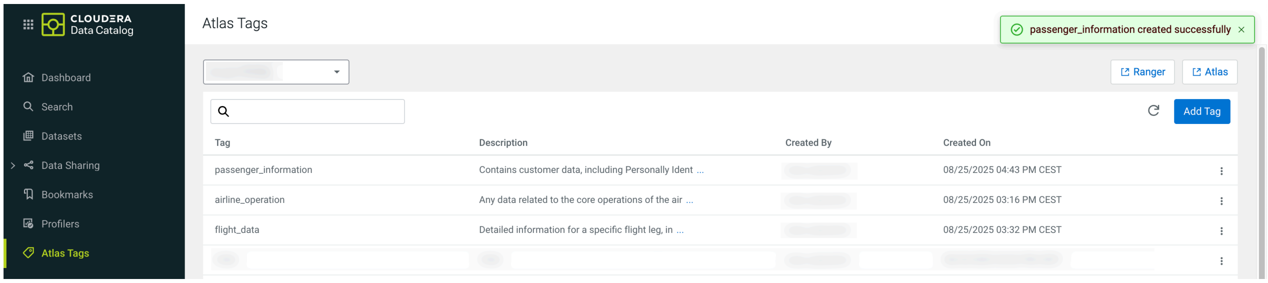
Select classifications to inherit attributes

Attributes (optional)

is_realtime: boolean

data_type: string

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



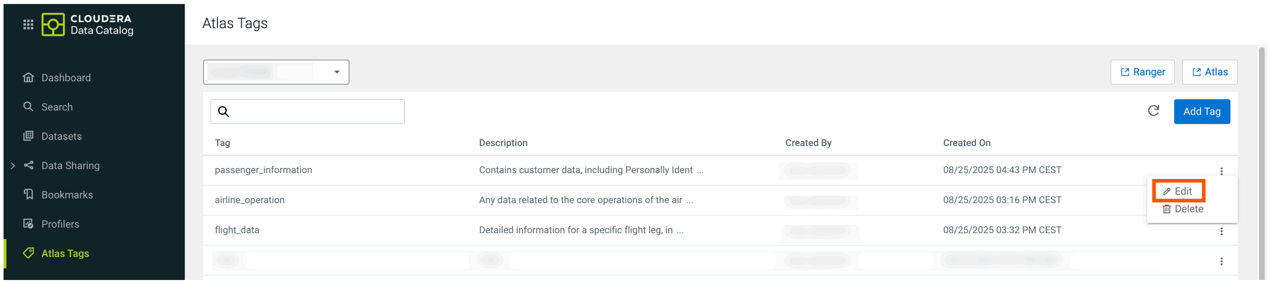
Atlas Tags

passenger_information created successfully

Tag	Description	Created By	Created On
passenger_information	Contains customer data, including Personally Ident ...	[redacted]	08/25/2025 04:43 PM CEST
airline_operation	Any data related to the core operations of the air ...	[redacted]	08/25/2025 03:16 PM CEST
flight_data	Detailed information for a specific flight leg, in ...	[redacted]	08/25/2025 03:32 PM CEST
[redacted]	[redacted]	[redacted]	[redacted]

Editing Atlas tags

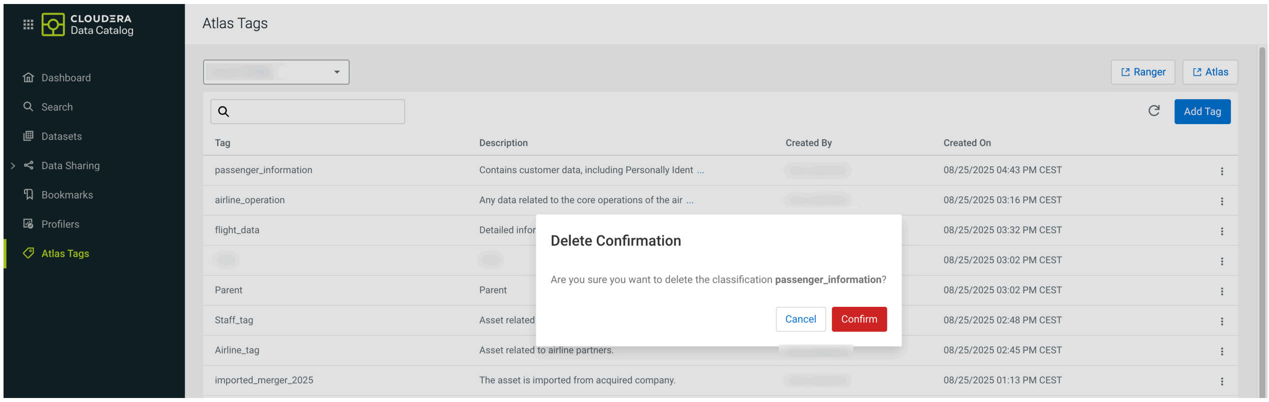
You can also edit or delete the Atlas tag as shown in the image. When you are editing the tag, you can only change the description or add new attributes.



Tag	Description	Created By	Created On
passenger_information	Contains customer data, including Personally Ident ...	[redacted]	08/25/2025 04:43 PM CEST
airline_operation	Any data related to the core operations of the air ...	[redacted]	08/25/2025 03:16 PM CEST
flight_data	Detailed information for a specific flight leg, in ...	[redacted]	08/25/2025 03:32 PM CEST
[redacted]	[redacted]	[redacted]	[redacted]

Deleting Atlas tags

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



Tag	Description	Created By	Created On
passenger_information	Contains customer data, including Personally Ident ...	[redacted]	08/25/2025 04:43 PM CEST
airline_operation	Any data related to the core operations of the air ...	[redacted]	08/25/2025 03:16 PM CEST
flight_data	Detailed infor		08/25/2025 03:32 PM CEST
[redacted]	[redacted]		08/25/2025 03:02 PM CEST
Parent	Parent		08/25/2025 03:02 PM CEST
Staff_tag	Asset related		08/25/2025 02:48 PM CEST
Airline_tag	Asset related to airline partners.	[redacted]	08/25/2025 02:45 PM CEST
imported_merger_2025	The asset is imported from acquired company.	[redacted]	08/25/2025 01:13 PM CEST

Delete Confirmation

Are you sure you want to delete the classification **passenger_information**?

Related Information

Propagated asset tagging

Creating tag rules in compute cluster environmentsin VM based environments

Creating tag rules in compute cluster environments

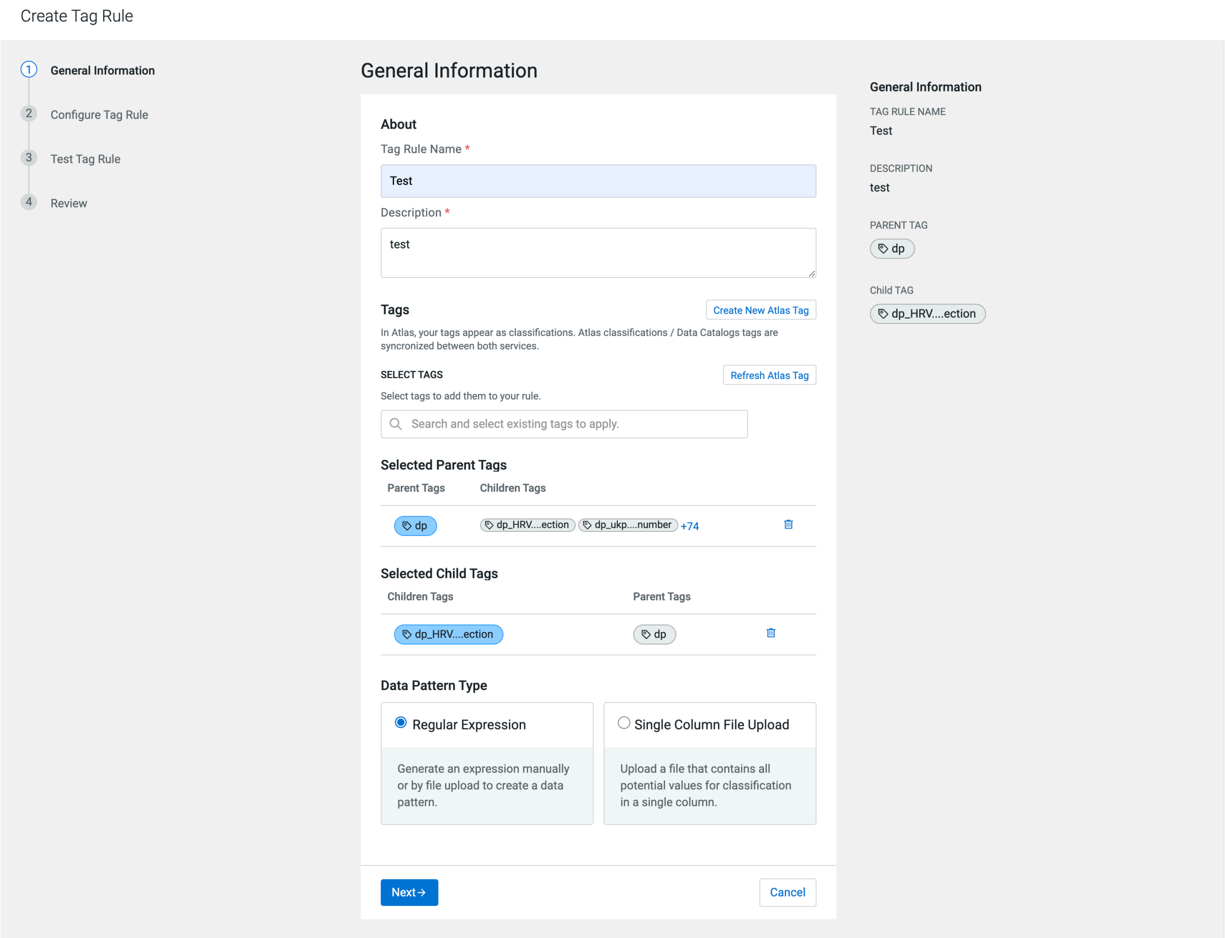
With tag rules, you can apply Apache Atlas classifications to your assets based on regex expressions or similarity to a set of values in a table.

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 *

Test

Description *

test

Tags

Create New Atlas Tag

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

SELECT TAGS

Refresh Atlas Tag

Select tags to add them to your rule.

Search and select existing tags to apply.

Selected Parent Tags

Parent Tags	Children Tags
dp	dp_HRV....ection dp_ukp....number +74

Selected Child Tags

Children Tags	Parent Tags
dp_HRV....ection	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 39.
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 40.

Creating regular expression based tag rule:

7. Define your regular expression for the table name.

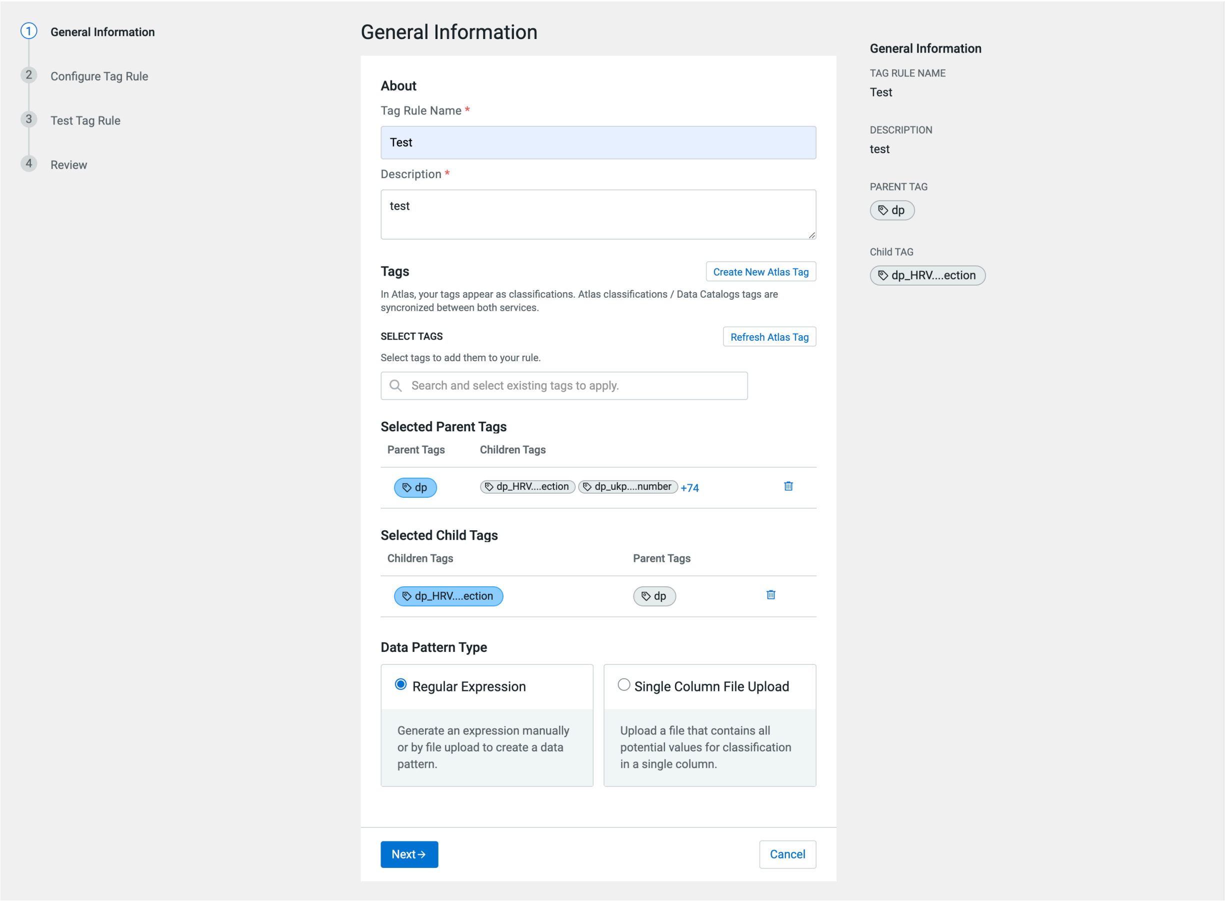
   Note: Cloudera recommends using PCRE2 compatible regular expressions. Non-compliant regular expressions may show reduced performance.

   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



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

General Information

About

Tag Rule Name *

Test

Description *

test

Tags

Create New Atlas Tag

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

SELECT TAGS

Refresh Atlas Tag

Select tags to add them to your rule.

Search and select existing tags to apply.

Selected Parent Tags

Parent Tags	Children Tags
dp	dp_HRV....ection dp_ukp....number +74

Selected Child Tags

Children Tags	Parent Tags
dp_HRV....ection	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

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.

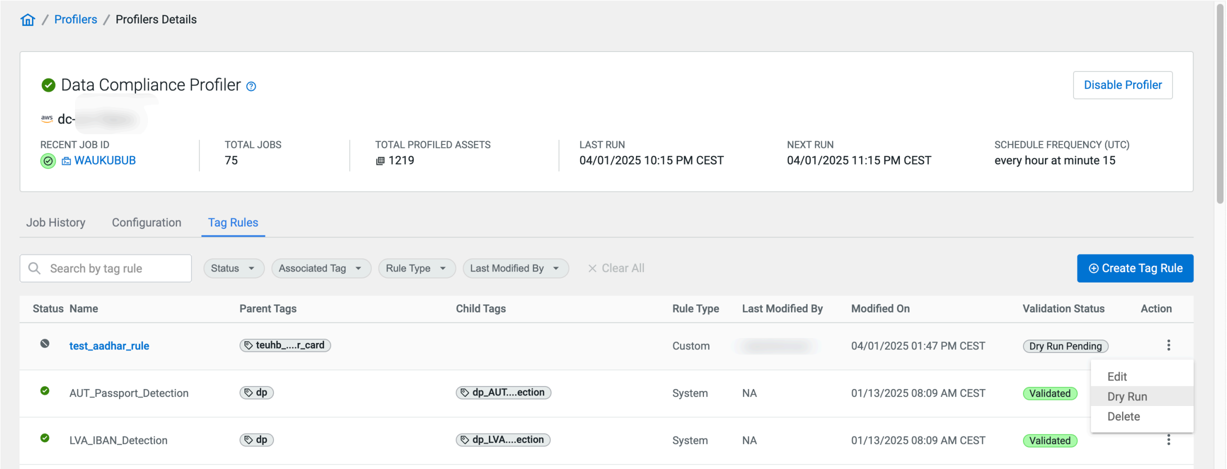
11. Review all your input before clicking Create Tag Rule.

a) Click Confirm to finalize your tag rule.

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

12. Click ⋮ > Dry Run.

Profilers Details



Data Compliance Profiler

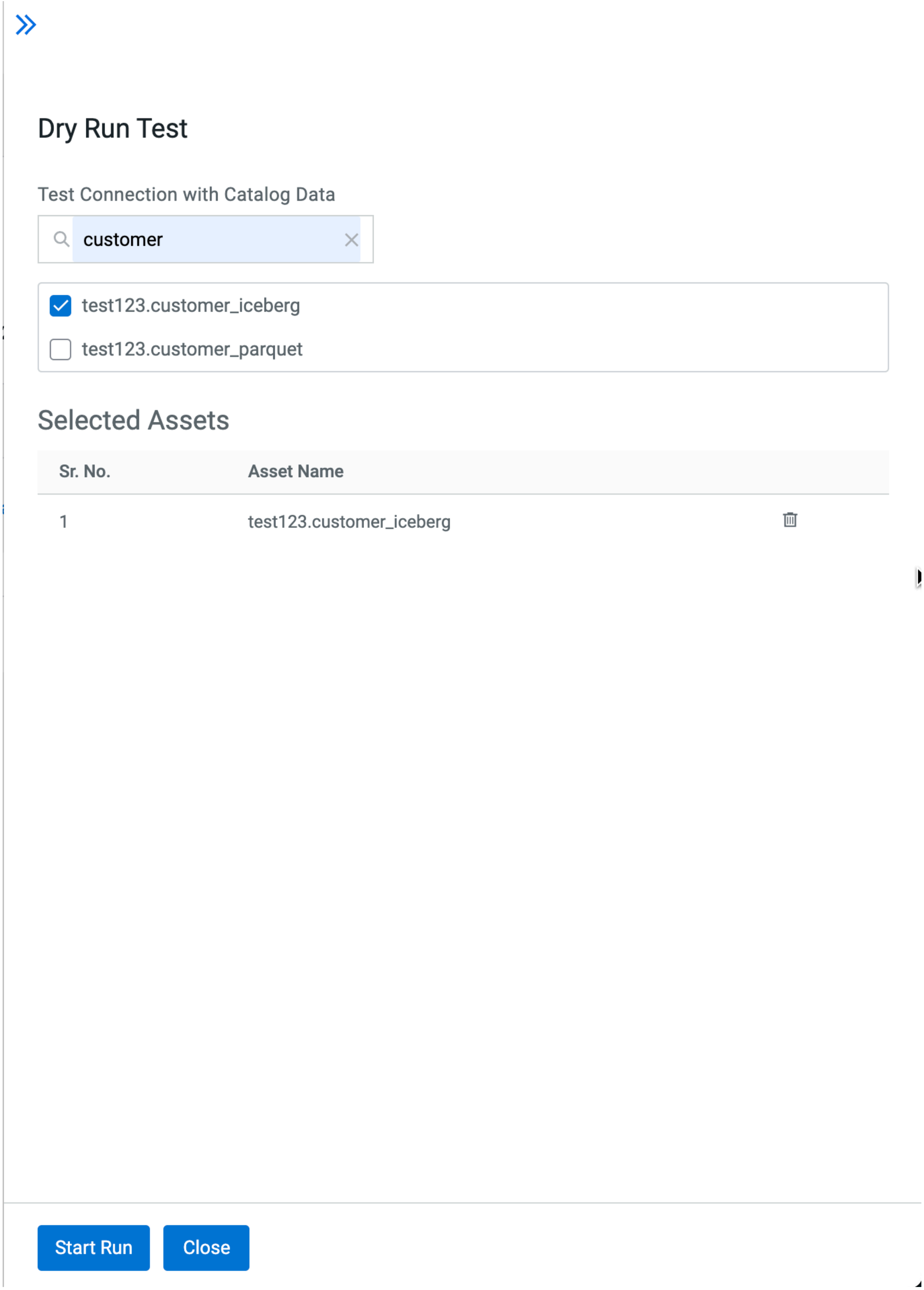
dc-[redacted]

RECENT JOB ID	TOTAL JOBS	TOTAL PROFILED ASSETS	LAST RUN	NEXT RUN	SCHEDULE FREQUENCY (UTC)
WAUKUBUB	75	1219	04/01/2025 10:15 PM CEST	04/01/2025 11:15 PM CEST	every hour at minute 15

Status	Name	Parent Tags	Child Tags	Rule Type	Last Modified By	Modified On	Validation Status	Action
⊘	test_aadhar_rule	teuhb_....r_card		Custom	[redacted]	04/01/2025 01:47 PM CEST	Dry Run Pending	⋮
✓	AUT_Passport_Detection	dp	dp_AUT....ection	System	NA	01/13/2025 08:09 AM CEST	Validated	
✓	LVA_IBAN_Detection	dp	dp_LVA....ection	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

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.

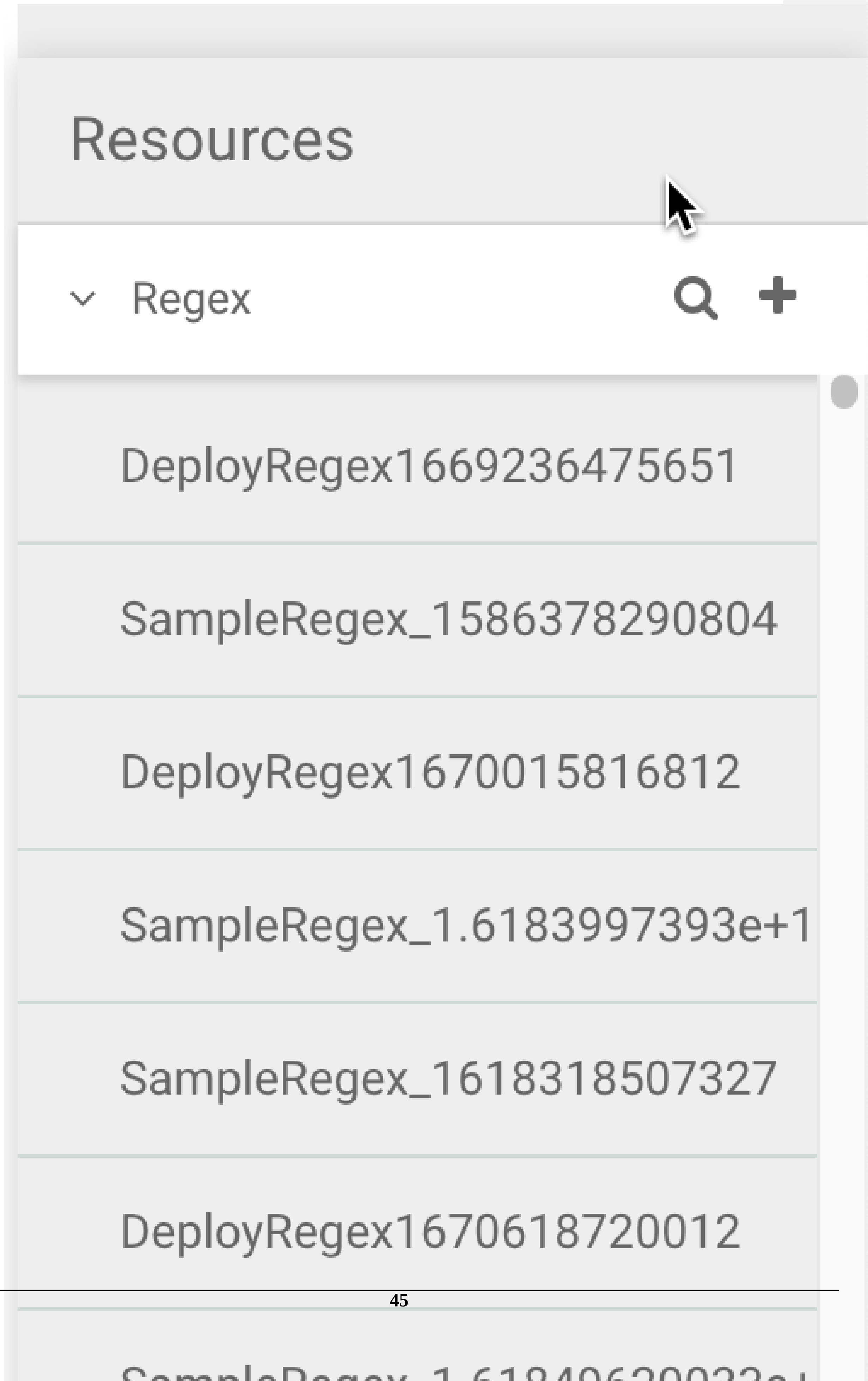
About this task

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

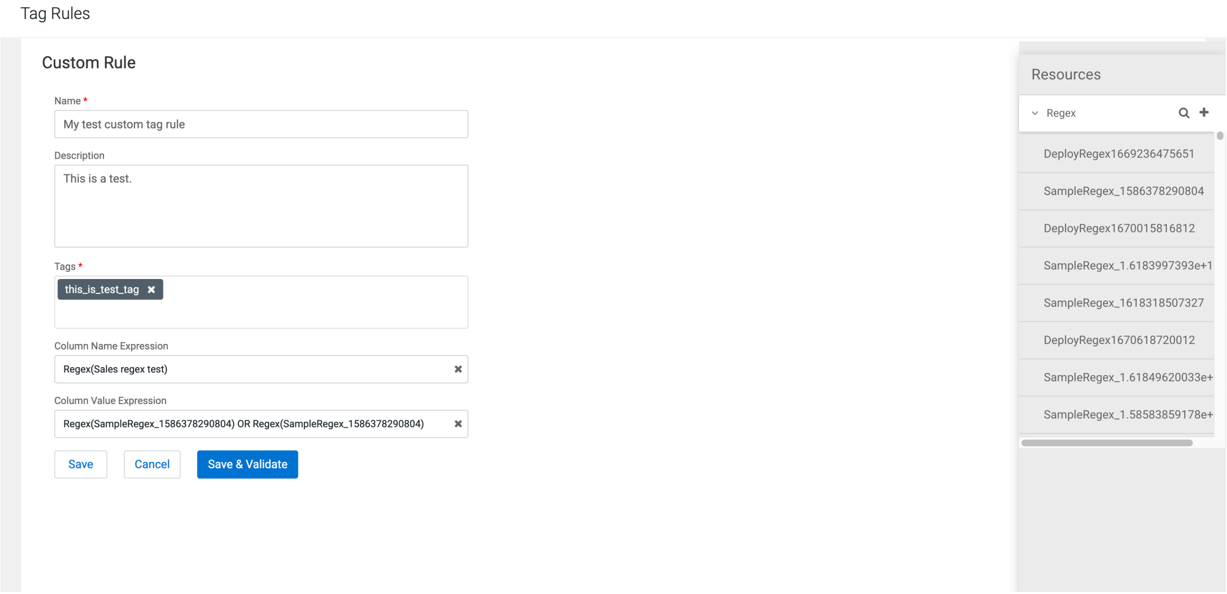
DeployRegex1670015816812

SampleRegex_1.6183997393e+1

SampleRegex_1618318507327

DeployRegex1670618720012

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) ✖

Resources

Regex

DeployRegex1669236475651

SampleRegex_1586378290804

DeployRegex1670015816812

SampleRegex_1.6183997393e+1

SampleRegex_1618318507327

DeployRegex1670618720012

SampleRegex_1.61849620033e+

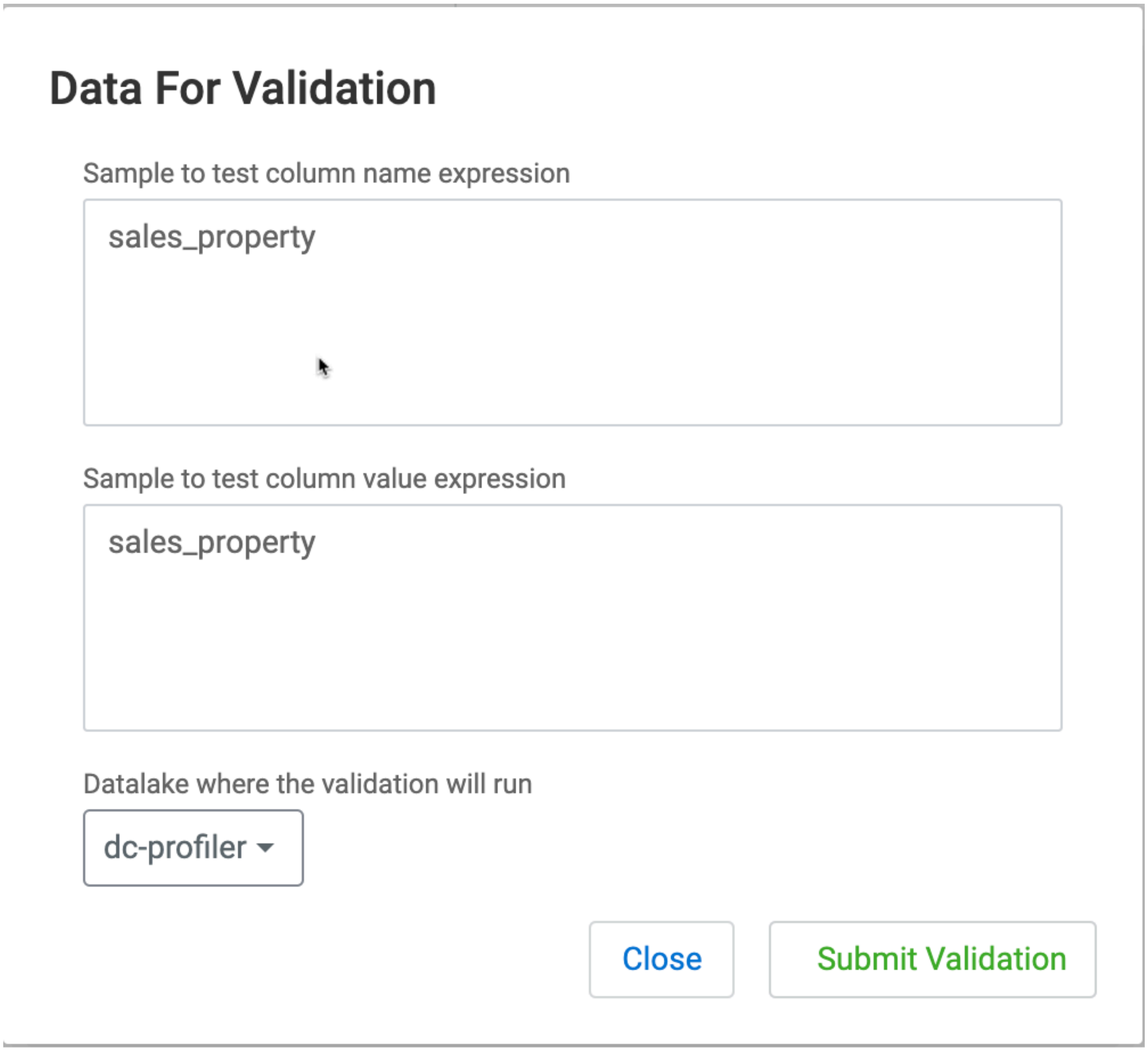
SampleRegex_1.58583859178e+

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.