

Profilers in Compute Cluster Enabled Environments

Date published: 2019-11-14

Date modified: 2025-10-17



Legal Notice

© Cloudera Inc. 2026. All rights reserved.

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

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

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

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

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

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

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

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

Contents

Launching profilers in Compute Cluster enabled environments.....	4
Launching profilers using the command-line.....	8
Enabling or disabling profilers in Compute cluster enabled environments.....	11
Tracking profiler jobs in Compute cluster enabled environments.....	12
Viewing profiler configurations in Compute cluster enabled environments.....	15
Configuring the Activity Profiler.....	18
Configuring the Data Compliance profiler.....	20
Profiler tag rules in Compute Cluster enabled environments.....	25
Creating tag rules in compute cluster environments.....	26
Approving Data Compliance Profiler tags.....	31
Configuring the Statistics Collector profiler.....	35
Understanding the Cron Expression generator.....	40
On-Demand Profilers in compute cluster environments.....	41
Deleting profilers in Compute cluster enabled environments.....	43

Launching profilers in Compute Cluster enabled environments

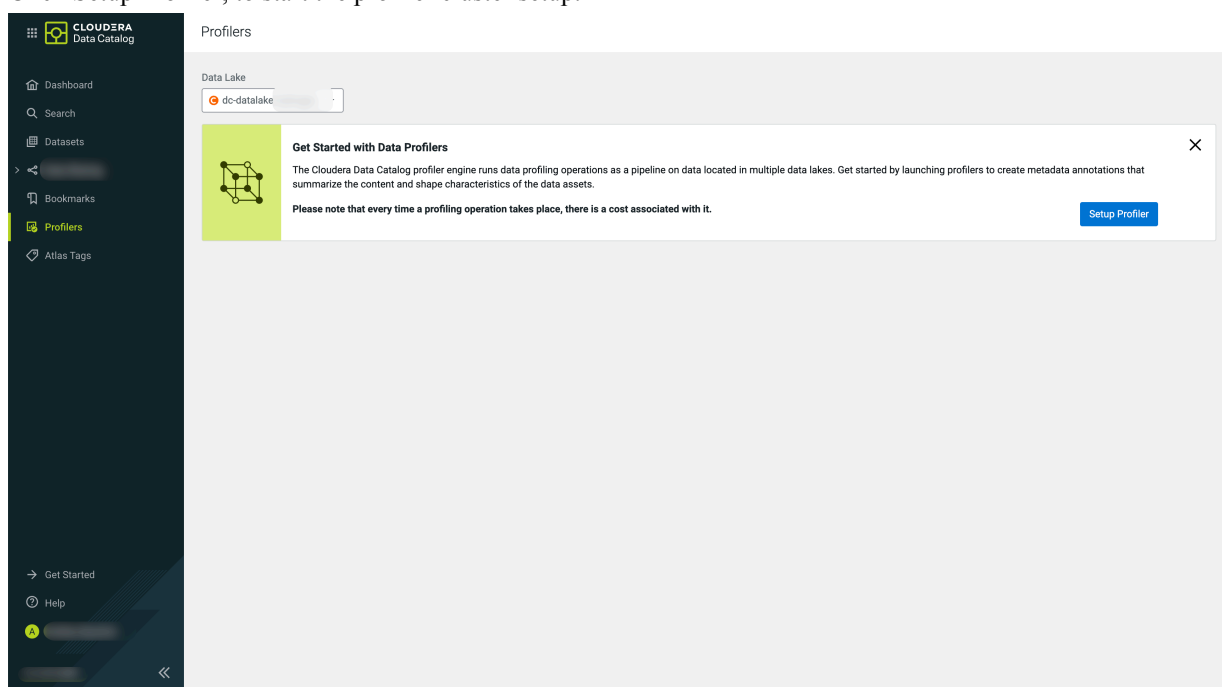
In Compute Cluster enabled environments, after you set up the Kubernetes profiler node group, the Profiler Launcher Service (PLS) keeps checking the availability of the node group automatically. Once the node group is ready, the PLS provisions the selected profilers by starting CRON jobs in the Kubernetes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



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

Profilers Setup

1 Setup Cluster

2 Launch Profiler

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

Next Cancel

Summary

Data Lake

dc-datalake

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.

Profilers Setup

1 Setup Cluster

2 Launch Profiler

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

Next Cancel

Summary

Data Lake

dc-datalake

Instance Type

c5a.2x... vCPU c5.2x... 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

Setup Cluster

Launch Profiler

Launch Profiler

Activity Profiler

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

Profiler Configuration :

WORKER MEM LIMIT:

4G

NUM WORKERS:

4

THREAD PER WORKER:

3

CRON EXPRESSION:

0 0 ***

Data Compliance Profiler

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

Profiler Configuration :

WORKER MEM LIMIT:

11G

NUM WORKERS:

10

THREAD PER WORKER:

3

CRON EXPRESSION:

0 0 ***

LAST RUN:

Over a period of 2 days

Table Statistics Profiler

Understand the shape of your data with columnar metrics.

Profiler Configuration :

WORKER MEM LIMIT:

11G

NUM WORKERS:

10

THREAD PER WORKER:

3

CRON EXPRESSION:

0 0 ***

LAST RUN:

Over a period of 2 days

Summary

Data Lake

dc-datalake-l

Instance Type

c5a.2x... vCPU

c5.2xl... vCPU

Autoscaling Instance Count

40

Profilers

Activity

Data C...liance

Table ...istics

Previous

Start Setup

Cancel

7. Once the cluster is ready 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 Lifte service. The expected setup time is around 15 to 30 minutes.

Profilers

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

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

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

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

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

Verifying the profiler cluster for Compute Cluster enabled environments

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

Environments / v2 / Compute Clusters

aws v2 [Stop](#) [Actions](#)

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

Data Lake Details

NAME	NODES	SCALE	QUICK LINKS
v2	2 0 0	Light Duty	Atlas Ranger Data Catalog

STATUS: Running STATUS REASON: N/A CRN: [j792...](#)

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

1 Compute Clusters [Add Compute Cluster](#)

Status	Name	CRN
Running	default- compute-cluster Default Cluster	jw

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

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

compute-cluster

STATUS: Running

CLUSTER TYPE: Default Cluster

DATE CREATED: 05/08/2024, 05:54:19

CREATED BY: Deepak Kumar Singh

CRN: [redacted]

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

Node Groups

- dcprofiler**
 - LABELS: lifitf.cloudera.com/instance-group-id: ig-tp04xcty ... More
 - ROOT VOLUME SIZE (GIB): 50
 - NODES: 1 (Auto scales between 1 and 10)
- dcprofiler-worker-spot**
 - LABELS: lifitf.cloudera.com/instance-group-id: ig-q12zn8wn ... More
 - ROOT VOLUME SIZE (GIB): 100
 - NODES: 0 (Auto scales between 0 and 81)
- lifitf-infra**
 - LABELS: role.node.kubernetes.io/lifitf-infra: true ... More
 - TAINTS: role.node.kubernetes.io/lifitf-infra: true:NoSchedule
 - ROOT VOLUME SIZE (GIB): 40
 - NODES: 2 (Auto scales between 2 and 4)

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 datalake.

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-scaling
        node group, in the range of 30 to 100 (both inclusive). (Applicable
        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-skeleton
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.

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 [***DATA LAKE 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 Compute cluster enabled environments

Profilers can be temporarily paused to save resources.

Procedure

- 1. Go to **Profilers**.
- 2. Click  > Pause Profiler.

Profilers

Data Lake

dc-

Refresh

<div><div>Activity Profiler</div></div>	FREQUENCY (UTC) <div>-NA-</div>	NEXT RUN <div>03/09/2025 01:00 AM CET</div>	TOTAL EXECUTIONS <div>-NA-</div>	<div><div>Details</div><div>Pause Profiler</div><div>Delete Profiler</div></div>
JOB ID <div>-NA-</div>	COMPLETED AT <div>-NA-</div>	JOB DURATION	ASSETS PROFILED <div>-NA-</div>	

- 3. Click Confirm.
- 4. You can click Resume Profiler to continue using it.

Profilers

Data Lake

dc

Refresh

<div><div>Activity Profiler</div></div>	FREQUENCY (UTC) <div>-NA-</div>	NEXT RUN <div>-NA-</div>	TOTAL EXECUTIONS <div>-NA-</div>	<div><div>Details</div><div>Resume Profiler</div><div>Delete Profiler</div></div>
JOB ID <div>-NA-</div>	COMPLETED AT <div>-NA-</div>	JOB DURATION	ASSETS PROFILED <div>-NA-</div>	

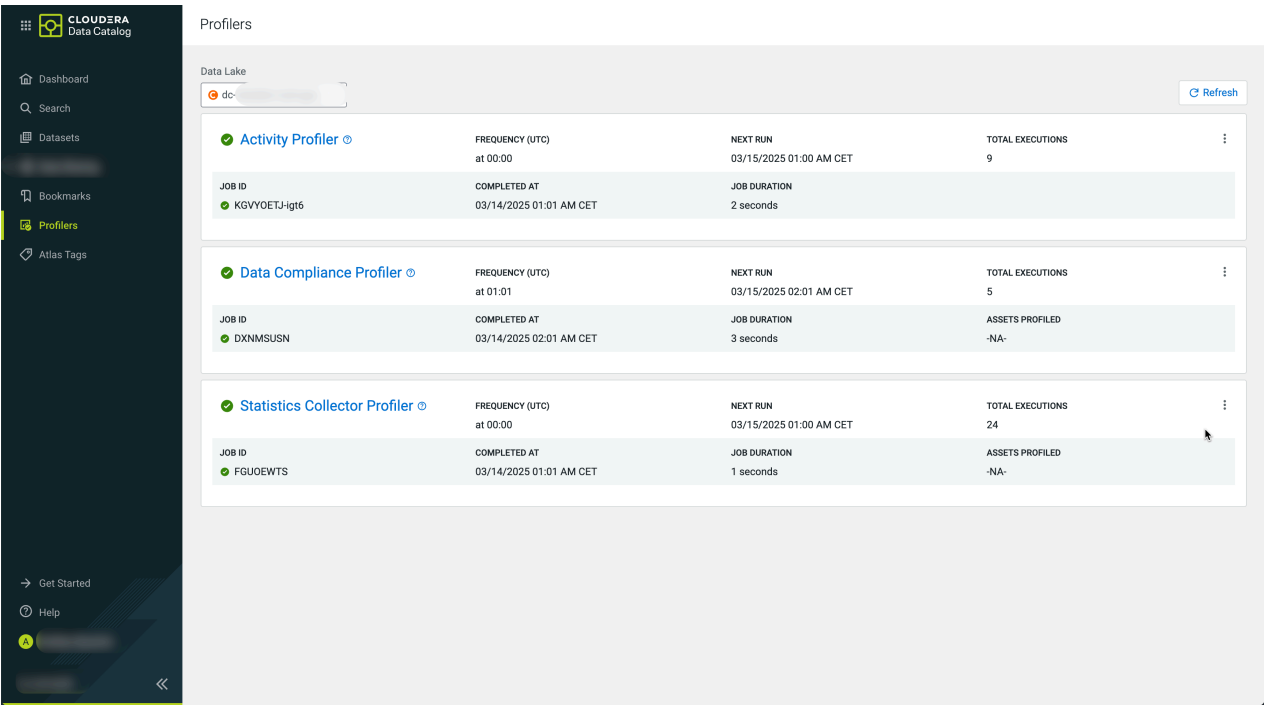
Activity Profiler was disabled on 03/08/2025 07:58 PM CET by András Szuromi

Tracking profiler jobs in Compute cluster enabled environments







In Profilers, you can see the status and statistics of your profilers.

Under Profilers , you can have an overview of your profiler since their launch and some basic information of the last jobs. Use this page to quickly check if your profiler jobs are failing.

Figure 1: Profiling jobs in a Compute Cluster enabled environment




For each profiler, you can view the details about:

- **Profiler type**
- Profiler **Status** for the last job (as an icon  ,  , )
- **Frequency (UTC)**
- **Next Run** (in your local timezone)
- **Total Executions** (since the launch of the profiler)
- Job status is marked with icons ( ,  , )
 - Running (Successfully launched)
 - Paused
 - Creation in Progress
- **JOB ID** of the last job
- **COMPLETED AT**
- **JOB DURATION** (of the last job)
- **ASSETS PROFILED** (by the last job)



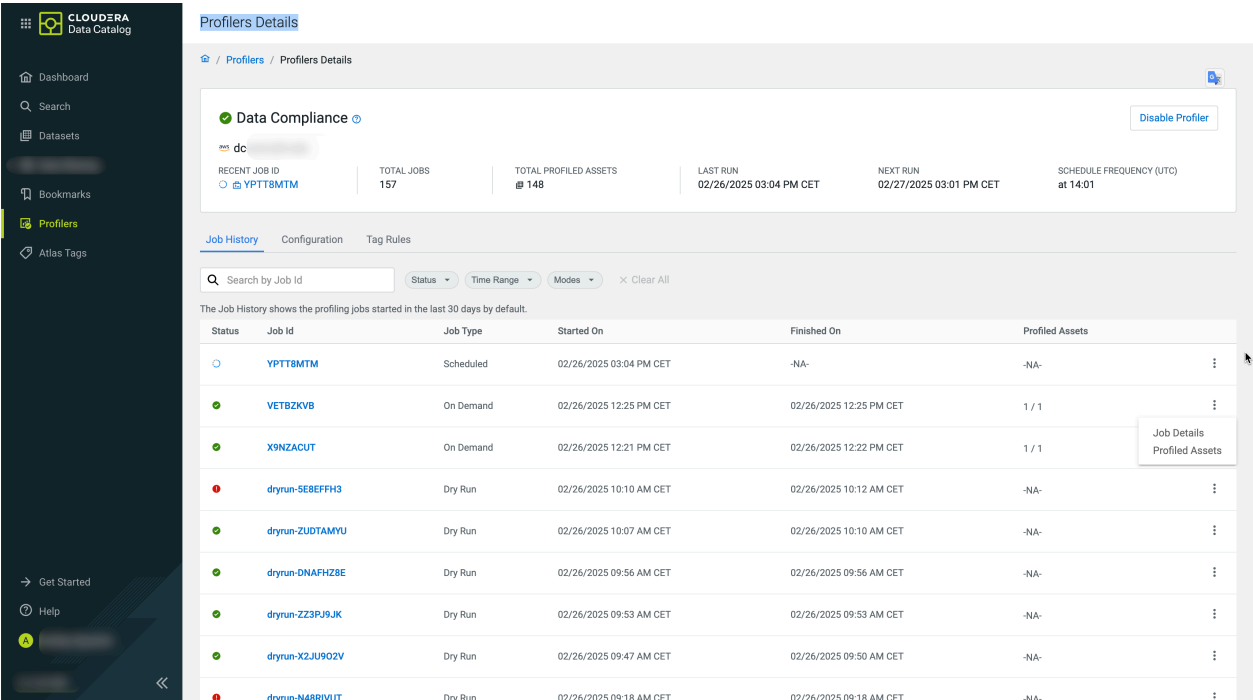
Note: Not available for the Activity Profiler.

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.

Click  >Details to gain more information about your profiler jobs in **Profilers** **Profilers Details**.


Use the following filters to screen your profiler jobs:

- Status:
 - Failed
 - Running
 - Finished
- Time Range
- Modes¹
 - **Dry Run**²
 - **On Demand**
 - **Scheduled**



Profilers Details

Dashboard / Search / Datasets / Bookmarks / **Profilers** / Atlas Tags










Data Compliance  [Disable Profiler](#)

RECENT JOB ID: [YPTT8MTM](#) | TOTAL JOBS: 157 | TOTAL PROFILED ASSETS: 148 | LAST RUN: 02/26/2025 03:04 PM CET | NEXT RUN: 02/27/2025 03:01 PM CET | SCHEDULE FREQUENCY (UTC): at 14:01


Job History | Configuration | Tag Rules

Search by Job Id: Status: Time Range: Modes:

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

Status	Job Id	Job Type	Started On	Finished On	Profiled Assets
	YPTT8MTM	Scheduled	02/26/2025 03:04 PM CET	-NA-	-NA-
	VETBZKVB	On Demand	02/26/2025 12:25 PM CET	02/26/2025 12:25 PM CET	1 / 1
	X9NZACUT	On Demand	02/26/2025 12:21 PM CET	02/26/2025 12:22 PM CET	1 / 1
	dryrun-SE8EFH3	Dry Run	02/26/2025 10:10 AM CET	02/26/2025 10:12 AM CET	-NA-
	dryrun-ZUDTAMYU	Dry Run	02/26/2025 10:07 AM CET	02/26/2025 10:10 AM CET	-NA-
	dryrun-DNAFH2BE	Dry Run	02/26/2025 09:56 AM CET	02/26/2025 09:56 AM CET	-NA-
	dryrun-ZZ3PJ9JK	Dry Run	02/26/2025 09:53 AM CET	02/26/2025 09:53 AM CET	-NA-
	dryrun-X2JU902V	Dry Run	02/26/2025 09:47 AM CET	02/26/2025 09:50 AM CET	-NA-
	dryrun-N48RVUT	Dry Run	02/26/2025 09:18 AM CET	02/26/2025 09:18 AM CET	-NA-

Job Details Profiled Assets

By clicking  by the individual jobs in **Profilers Details**, you can drill further down to **Job Summary** and **Profiled Assets**.

The **Job Summary** shows you the specific configuration applied for that particular job run.

¹ Only available for the Data Compliance and Statistics Collector profilers.

² Only available for the Data Compliance profiler.

The Dry Run mode refers to the first on-demand profiler jobs, triggered to validate custom tag rules. These test the tag rule on a subset of entities from the data lake.

CLOUDERA

Data Catalog

Dashboard

Search

Datasets

Bookmarks

Profilers

Atlas Tags

Get Started

Help

Profilers Details

Profilers / Profilers Details

✓ Data Compliance Profiler

RECENT JOB ID

PNVZG29V

TOTAL JOBS

12

TOTAL PROFILED ASSETS

45

LAST RUN

10/15/2025 05:04 PM CEST

Job History

Configuration

Tag Rules

Search by Job Id

Status

Time Range

Job Type

Clear All

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

Status	Job Id	Job Type	Started On
✓	PNVZG29V	Scheduled	10/15/2025 05:40 PM CEST
✓	Z2FKKSX8	Scheduled	10/15/2025 04:49 PM CEST
✓	GKDTJA9V	On Demand	10/15/2025 04:39 PM CEST
✓	9C6M3JKY	On Demand	10/15/2025 04:38 PM CEST
✓	YDZVTJPR	On Demand	10/15/2025 04:36 PM CEST
✓	UP998XNU	Scheduled	10/15/2025 03:49 PM CEST
✓	EBVSCF4F	Scheduled	10/15/2025 02:49 PM CEST
✓	D56XMTT8	Scheduled	10/15/2025 01:49 PM CEST
✓	FMVJH9XT	Scheduled	10/15/2025 12:49 PM CEST

Job Summary

Details

Profiled Assets

Asset Filtering Rules

JOB ID	STARTED ON	FINISHED ON	ASSETS PROFILED
✓ Z2FKKSX8	10/15/2025 04:49 PM CEST	10/15/2025 04:50 PM CEST	7

WORKER MEMORY LIMIT

6G

THREADS PER WORKER

3

LAST RUN CHECK

Disabled

CRON EXPRESSION

49 ****

NUMBER OF WORKERS

10

false

Close

The **Profiled Assets** not only gives you a list of entities that were selected by your profiler to be profiled, but it let's you filter them.

CLOUDERA

Data Catalog

Dashboard

Search

Datasets

Bookmarks

Profilers

Atlas Tags

Get Started

Help

Profilers Details

Profilers / Profilers Details

✓ Data Compliance Profiler

RECENT JOB ID

PNVZG29V

TOTAL JOBS

12

TOTAL PROFILED ASSETS

45

LAST RUN

10/15/2025 05:04 PM CEST

Job History

Configuration

Tag Rules

Search by Job Id

Status

Time Range

Job Type

Clear All

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

Status	Job Id	Job Type	Started On
✓	PNVZG29V	Scheduled	10/15/2025 05:40 PM CEST
✓	Z2FKKSX8	Scheduled	10/15/2025 04:49 PM CEST
✓	GKDTJA9V	On Demand	10/15/2025 04:39 PM CEST
✓	9C6M3JKY	On Demand	10/15/2025 04:38 PM CEST
✓	YDZVTJPR	On Demand	10/15/2025 04:36 PM CEST
✓	UP998XNU	Scheduled	10/15/2025 03:49 PM CEST
✓	EBVSCF4F	Scheduled	10/15/2025 02:49 PM CEST
✓	D56XMTT8	Scheduled	10/15/2025 01:49 PM CEST
✓	FMVJH9XT	Scheduled	10/15/2025 12:49 PM CEST

Job Summary

Details

Profiled Assets

Asset Filtering Rules


Search by Asset Name

Status

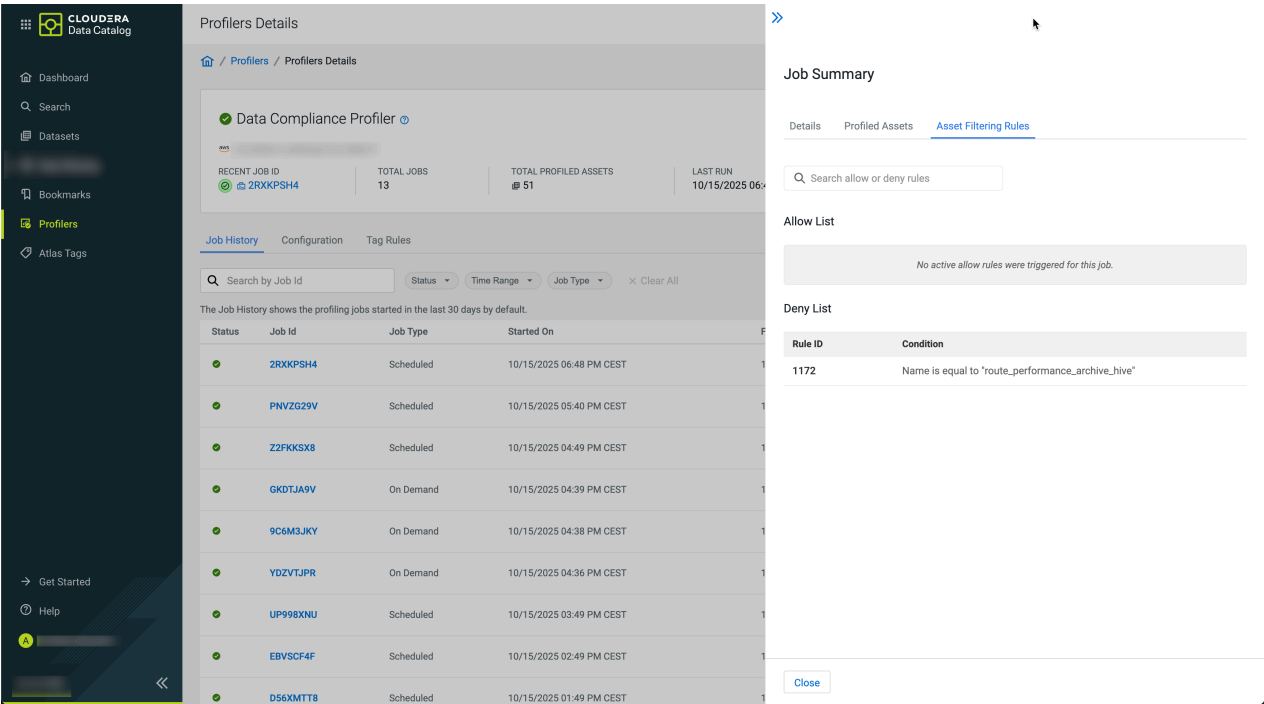
Clear All

Status	Asset Name	Suggested Tags
✓	airline_operations.route_performance.archive_hive	N/A
✓	airline_operations.raw_bookings	N/A
✓	airline_operations.stg_flight_manifests	N/A
✓	airline_operations.enriched_flight_data	N/A
✓	airline_operations.agg_route_performance	N/A
✓	airline_operations.dim_aircraft	N/A
✓	airline_operations.airlines_new	11

Close

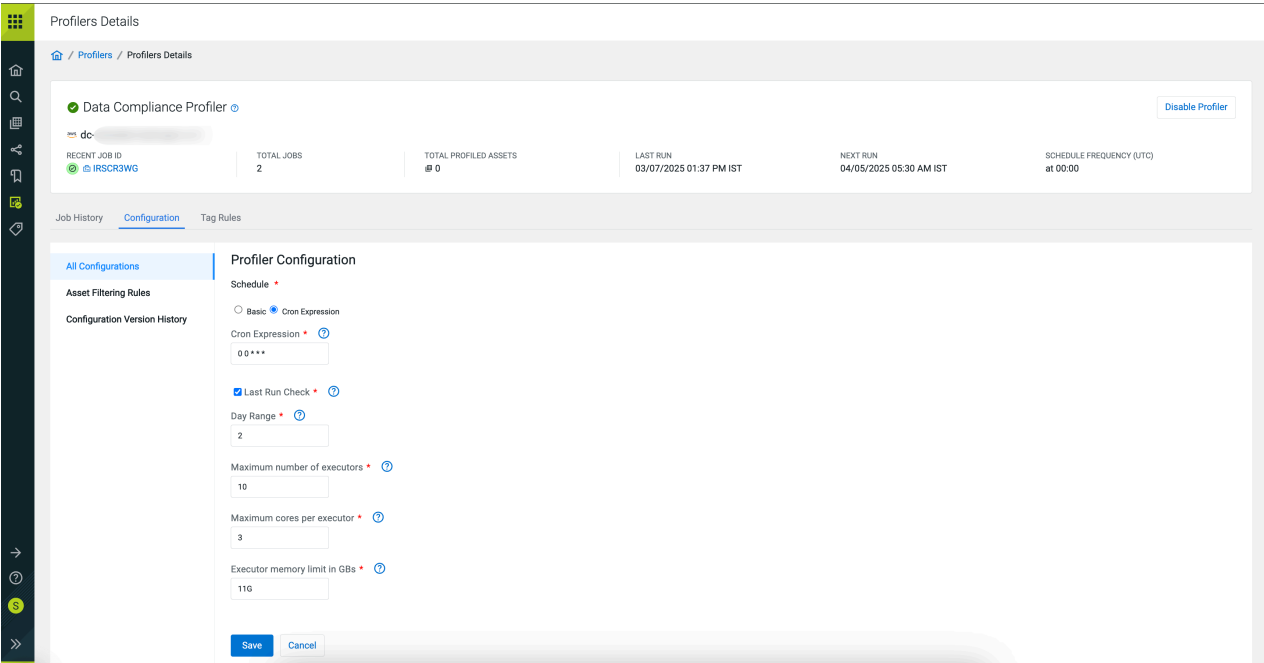
 **Note:** Hovering over the skipped assets shows the reason for not including the particular asset.

The **Asset Filtering Rules** tab summarizes the active filtering rules for your profiling job.



Viewing profiler configurations in Compute cluster enabled environments

You can check the configuration and its changes for your profiles in Profilers > Profiler Details > Configuration. In Profilers Profiler Details Configuration All Configurations you can set the scheduling and resources of your profilers.



Configuration Version History lets you check your changes to your settings.

Profilers Details

Statistics Collector

dc-

RECENT JOB ID
ZCHVHNK3

TOTAL JOBS
158

TOTAL PROFILED ASSETS
6944

LAST RUN
02/26/2025 12:23 PM CET

NEXT RUN
02/27/2025 11:41 AM CET

SCHEDULE FREQUENCY (UTC)
at 10:41

Disable Profiler

Job History

Configuration

All Configurations

Asset Filtering Rules

Configuration Version History

Configuration History

Review your profiler configuration changes in a sequential order. Clicking All Configurations displays the full list of settings used at that time.

02/24/2025 02:31 PM CET

Created By: Configuration version: 14.0

Last Run in Days :
Before: 5 After: 2

All Configurations

02/21/2025 11:39 AM CET

Created By: Configuration version: 13.0

Maximum number of executors :
Before: 20 After: 10

All Configurations

02/21/2025 11:32 AM CET

Created By: Configuration version: 12.0

Cron Expression :
Before: 41 10 * * * After: 35 10 * * *

All Configurations

Clicking **All Configurations** shows all settings at the time, including the unchanged options.

16



Configuration version 14.0

Profiler Configuration

Cron Expression :
41 10 ***

Cron Expression :
41 10 ***

Last Run in Days : 5	Last Run in Days : 2
-------------------------	-------------------------

Last Run Enabled :
true

Last Run Enabled :
true

Executor Configurations

Maximum number of executors :
20

Maximum number of executors :
20

Maximum cores per Executor :
3

Maximum cores per Executor :
3

Executor memory limit in GBs :
11G

Executor memory limit in GBs :
11G

Close

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 2: 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 3: 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

Save

Cancel



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 * 

Maximum cores per Executor * 

Executor memory limit in GBs * 

Save**Cancel****5. 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 4: 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 5: 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 * 

Maximum cores per Executor * 

Executor memory limit in GBs * 

**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	<ul style="list-style-type: none"> • equals • starts with • ends with
Name (of asset)	<ul style="list-style-type: none"> • equals
Owner (of asset)	<ul style="list-style-type: none"> • contains • starts with • ends with

Key	Operator
Creation date ³	<ul style="list-style-type: none">greater thanless 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

airline_operations

equals

greater than

Creation Date

1 days ago

✕

✕

⊕ [Add Row]

Add Rule

Cancel



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	<div>Affected Assets Edit Delete</div>

Figure 6: 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:

Profilers Details

Profilers / Profilers Details

Statistics Collector Profiler

RECENT JOB ID

RGFFDQAH

TOTAL JOBS

8

TOTAL PROFILED ASSETS

9

LAST RUN

09/30/2025 06:39 PM CEST

Job History

Configuration

Search by Job Id

Status

Time Range

Job Type

Clear All

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
✔	KXRVP6S	Scheduled	09/30/2025 05:42 PM CEST
✔	7UJJVWG6	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
✔	6XDKRX80	Scheduled	09/30/2025 12:16 PM CEST
✔	AQNTVACQ	Scheduled	09/30/2025 12:07 PM CEST

Job Summary

Details

Profiled Assets

Asset Filtering Rules

Search allow or deny rules

Allow List

Rule ID	Condition
1152	Name starts with "airlines_new"

Deny List

Rule ID	Condition
1154	Database Name starts with "airline_operations"

Close

Profiler tag rules in Compute Cluster enabled 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 Data Compliance. 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 Defined:** These are built-in rules that cannot be edited. You can only enable or disable them for your data.




Note:

Calculation for System Defined tag rules:

The match threshold is set to 70% for column values with the given regex. The column value matching is given a weightage of 85% in the final score and the remaining 15% is associated with the column name matching.

- **Custom:** 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.

Profilers Details

🏠 / Profilers / Profilers Details

Data Compliance Profiler
 Disable Profiler

dc-datalake

RECENT JOB ID

TOTAL JOBS

TOTAL PROFILED ASSETS

LAST RUN

NEXT RUN

SCHEDULE FREQUENCY (UTC)

4

0

03/13/2025 11:44 AM CET

03/14/2025 02:01 AM CET

at 01:01

Job History

Configuration

Tag Rules

Status

Associated Tag

Rule Type

Last Modified By

×

Clear All

Create Tag Rule

Status	Name	Parent Tags	Child Tags	Rule Type	Last Modified By	Modified On	Validation Status	Action
	test_tag_rule_sb	Test		Custom		03/13/2025 12:48 PM CET	Dry Run Pending	
	BEL_IBAN_Detection	dp	dp_BEL.....action	System	NA	01/13/2025 08:09 AM CET	Validated	
	EST_IBAN_Detection	dp	dp_EST.....action	System	NA	01/13/2025 08:09 AM CET	Validated	
	CHE_NationalID_Detection	dp	dp_CHE.....action	System	NA	01/13/2025 08:09 AM CET	Validated	
	POL_Passport_Detection	dp	dp_POL.....action	System	NA	01/13/2025 08:09 AM CET	Validated	
	PRT_NationalID_Detection	dp	dp_PRT.....action	System	NA	01/13/2025 08:09 AM CET	Validated	
	CHE_IBAN_Detection	dp	dp_CHE.....action	System	NA	01/13/2025 08:09 AM CET	Validated	
	npi	dp	dp_npi	System	NA	01/13/2025 08:09 AM CET	Validated	
	ESP_IBAN_Detection	dp	dp_ESP.....action	System	NA	01/13/2025 08:09 AM CET	Validated	

After creating your rule, you have to validate them with test data by completing a Dry Run and, only then you can click Enable.



Note: Tag Rules can be temporarily suspended.

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	<ul style="list-style-type: none"> Manually entered regex pattern Uploaded regex pattern
Column value	Manually entered regex pattern	<ul style="list-style-type: none"> 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		<ul style="list-style-type: none"> Manually entered regex pattern Uploaded regex pattern

Match thresholds and weightage

In Compute Cluster enable environments, you can adjust the **Column Value Weightage** for tag rules defined with regex patterns. The column value weightage percentage complements the column name weightage to 100%. This means that if you set the column value weightage to 80%, the column name adds to the final match score either 20 or zero, The reason for this is that column name matching can have only binary results (match or no match), while column value match is the number of matching values (rows) from all values in the column.

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 Compute Cluster enabled environments, review them with data uploaded in a file, then save them to reach the Dry Run Pending status. Tag rules in this status must be also tested with a Dry Run on a subset of your data (up to 10 tables) in the data lake before deploying them. A Dry Run is a special on-demand profiling job.

Tag handling by tag rules

Successfully tested and enabled tag rules apply Atlas classifications or synchronized Cloudera Data Catalog tags to tables, columns.

In Compute cluster enabled environments, the parent-child tag relationships are respected. When the column value matches a child tag, the table receives the parent tag.



Note:

Tags created in Cloudera Data Catalog automatically receive a status attribute. This is can be used to identify the association of the tag with the asset.

Creating tag rules in compute cluster environments

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

About this task

Procedure

1. To start applying tags, go to **Profilers** and select your data lake.

2. Go to Profilers Data Compliance Tag Rules .
3. Click + Create Tag Rule.
4. Name your tag rule and add a description to it in **General Information**.

Create Tag Rule

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

General Information

About

Tag Rule Name *

Description *

Tags

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

[Create New Atlas Tag](#)

SELECT TAGS

Select tags to add them to your rule.

[Refresh Atlas Tag](#)

Selected Parent Tags

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

Selected Child Tags

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

Data Pattern Type

☒ **Regular Expression**

Generate an expression manually or by file upload to create a data pattern.

☐ **Single Column File Upload**

Upload a file that contains all potential values for classification in a single column.

General Information

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 27.

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 28.

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

The screenshot shows the 'General Information' step of the 'Create Tag Rule' wizard. On the left, a progress bar indicates four steps: 1. General Information (active), 2. Configure Tag Rule, 3. Test Tag Rule, and 4. Review. The main form area is titled 'General Information' and contains several sections:

- About:** 'Tag Rule Name' is set to 'Test' and 'Description' is set to 'test'.
- Tags:** A section explaining that tags appear as classifications in Atlas and Data Catalogs, with links to 'Create New Atlas Tag' and 'Refresh Atlas Tag'.
- SELECT TAGS:** A search bar with the placeholder 'Search and select existing tags to apply.'
- Selected Parent Tags:** A table with two columns: 'Parent Tags' and 'Children Tags'. It lists 'dp', 'dp_HRV...action', and 'dp_ukp...number +74'.
- Selected Child Tags:** A table with two columns: 'Children Tags' and 'Parent Tags'. It lists 'dp_HRV...action' and 'dp'.
- Data Pattern Type:** Two radio buttons: 'Regular Expression' (selected) and 'Single Column File Upload'. The 'Regular Expression' option has a sub-note: 'Generate an expression manually or by file upload to create a data pattern.' The 'Single Column File Upload' option has a sub-note: 'Upload a file that contains all potential values for classification in a single column.'

At the bottom, there are 'Next ->' and 'Cancel' buttons.



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

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

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



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

Tag rule testing:

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



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

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

a) Click Confirm to finalize your tag rule.

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

- Click > Dry Run.

Profilers Details

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

Data Compliance Profiler

[Disable Profiler](#)

dc-

RECENT JOB ID

WALKUBUB

TOTAL JOBS
75

TOTAL PROFILED ASSETS

1219

LAST RUN
04/01/2025 10:15 PM CEST

NEXT RUN
04/01/2025 11:15 PM CEST

SCHEDULE FREQUENCY (UTC)
every hour at minute 15

[Job History](#)
[Configuration](#)
[Tag Rules](#)

Status
Associated Tag
Rule Type
Last Modified By
Clear All

[Create Tag Rule](#)

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

Edit
Dry Run
Delete

The **Dry Run Test** pane opens.

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

>>

Dry Run Test

Test Connection with Catalog Data

customer

×

☒ test123.customer_iceberg

☐ test123.customer_parquet


Selected Assets

Sr. No.	Asset Name	
1	test123.customer_iceberg	

Start Run

Close

Your tag rule becomes VALIDATED after a successful dry run.

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

Approving Data Compliance Profiler tags

You are able to check and approve tags created by the Data Compliance Profiler assigned to your assets before applying them.

About this task

Once your Data Compliance Profiler job completes, you can check your profiled assets one-by-one and decide to keep the suggested tags or remove them from asset before they are synchronized to Apache Atlas.


This gives you the ability, for example, to correct the tagging of assets mistakenly marked as PII. Incorrectly applied tags can affect your tag-based policies, unexpectedly changing the access of your users.

Before you begin



Note: This feature is only available for the Data Compliance Profiler in the Compute Cluster enabled environment.

Procedure

1. Once your profiler job is completed, go to Profilers Data Compliance Profiler Job History Profiled Assets .
 - a) Select the row with the relevant *JOB ID* and click  , then open **Profiled Assets**.

Dashboard
Search
Datasets
Data Sharing
All Shares
Manage Users
Bookmarks
Profilers
Atlas Tags

Profilers Details

Data Compliance Profiler

Disable Profiler

RECENT JOB ID

VJBKG9PK

TOTAL JOBS

166

TOTAL PROFILED ASSETS

2345

LAST RUN

07/18/2025 04:14 PM CEST

NEXT RUN

07/18/2025 05:12 PM CEST

SCHEDULE FREQUENCY (UTC)

every hour at minute 12

Job History

Configuration

Tag Rules

Status

Time Range

Job Type

Clear All

Refresh

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

Status	Job Id	Job Type	Started On	Finished On	Profiled Assets
	VJBKG9PK	Scheduled	07/18/2025 04:14 PM CEST	07/18/2025 04:21 PM CEST	56 / 56
	XQG3KEEP	Scheduled	07/18/2025 03:14 PM CEST	07/18/2025 03:21 PM CEST	56 / 56
	KPBWFFSK	Scheduled	07/18/2025 02:14 PM CEST	07/18/2025 02:21 PM CEST	56 / 56

Job Details

Profiled Assets

2. Check the content of either of the following columns:

Option

Asset Name

Opens the **Asset Details**.

Option

Suggested Tags Count

Opens the list of identified tags for the asset. If the asset has columns, you can see the tags assigned to each of them.

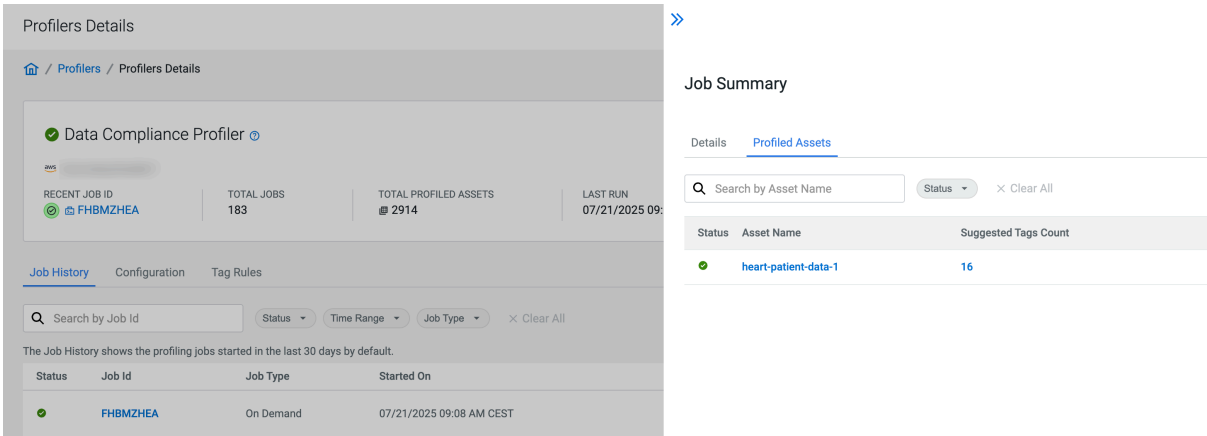
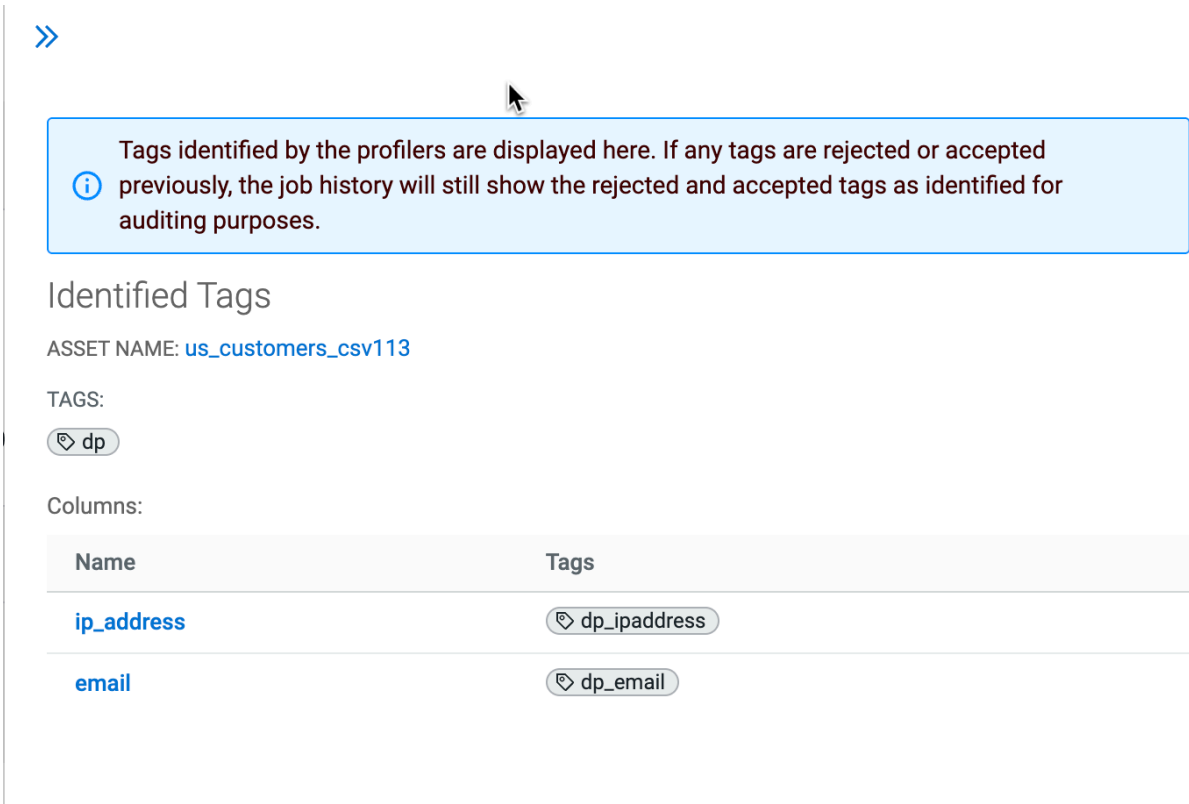




Figure 7: List of tags in Suggested Tags Count





- 3. Click the asset whose suggested tags you want to review.


4. Click Asset Details Classifications  Edit .

Tags (Apache Atlas classifications) pending approval are marked with the  icon.

Approved tags are marked with the  icon.

Asset Details

heart-patient-data-1  

Properties 


Type: **HIVE TABLE**
of Columns: 5
Data Lake: **recoverycluster1**
Datasets: 0
Owner: **hive**
Created On: 07/16/2025 03:37 PM CEST
Last Updated At: 07/21/2025 12:15 PM CEST
Table Type: **MANAGED_TABLE**
Database: **default**
DB Catalog: **cm**
Parent: **default**


Qualified Name
heart-patient-data-1@cm





Comment
[+ Add Comment](#)






Description
[+ Add Description](#)


Profilers | 2

Data Compliance Profiler
Last run: 07/21/2025 08:22 AM CEST | Status: **SUCCESS**  [Run](#)
Next Schedule Run: 07/22/2025 07:10 AM CEST


Statistics Collector Profiler
Last run: 07/21/2025 12:15 PM CEST | Status: **SUCCESS**  [Run](#)
Next Schedule Run: 07/21/2025 01:14 PM CEST

Classifications | 9    

 PII  oper...  heart  input...  Insura... [+ 4](#)

Terms 

[+ Add Terms](#)

You can remove suggested and already applied tags by clicking the  icon.

Classifications | 9

Managed

System

Propagated

PII

operation

heart

inpatient

insurance

Search and add classifications

Create

Save

Cancel

5. Click Save to apply your changes.

Configuring the Statistics Collector profiler

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

Procedure

1. Go to **Profilers** and select your data lake.
2. Select a schedule to run profiler using either UNIX Cron Expression or the Basic scheduler



Note: 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 8: 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 9: 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

Save



Cancel

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

7. Add **Asset Filtering Rules** as needed to customize the selection 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	<ul style="list-style-type: none"> • equals • starts with • ends with
Name (of asset)	<ul style="list-style-type: none"> • equals
Owner (of asset)	<ul style="list-style-type: none"> • contains • starts with • ends with

Key	Operator
Creation date ⁴	<ul style="list-style-type: none">greater thanless 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

airline_operations

equals

greater than

1 days ago

Creation Date

⊕ [Add Row]

Add Rule

Cancel



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	<div>Affected Assets Edit Delete</div>

Figure 10: Affected Assets in Asset Filtering Rules configuration

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.



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

Home / Profilers / Profilers Details

Statistics Collector Profiler

RECENT JOB ID

RGFFDQAH

TOTAL JOBS

8

TOTAL PROFILED ASSETS

9

LAST RUN

09/30/2025 06:39 PM CEST

Job History

Configuration

Search by Job Id

Status

Time Range

Job Type

Clear All

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

Status	Job Id	Job Type	Started On
Success	RGFFDQAH	Scheduled	09/30/2025 06:39 PM CEST
Success	NMJUQP9I	Scheduled	09/30/2025 06:33 PM CEST
Success	KXRVP6S	Scheduled	09/30/2025 05:42 PM CEST
Success	7UJJVWG6	Scheduled	09/30/2025 05:30 PM CEST
Success	JRCQNBZQ	Scheduled	09/30/2025 05:22 PM CEST
Success	36B6SSCP	Scheduled	09/30/2025 05:10 PM CEST
Success	6XDKRX80	Scheduled	09/30/2025 12:16 PM CEST
Success	AQNTVACQ	Scheduled	09/30/2025 12:07 PM CEST

Job Summary

Details

Profiled Assets

Asset Filtering Rules

Search allow or deny rules

Allow List

Rule ID	Condition
1152	Name starts with "airlines_new"

Deny List

Rule ID	Condition
1154	Database Name starts with "airline_operations"

Close

Understanding the Cron Expression generator

In Profilers > Profilers Details > Configuration > All Configurations, a cron expression can be used to define 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)
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 May.”



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 compute cluster 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.

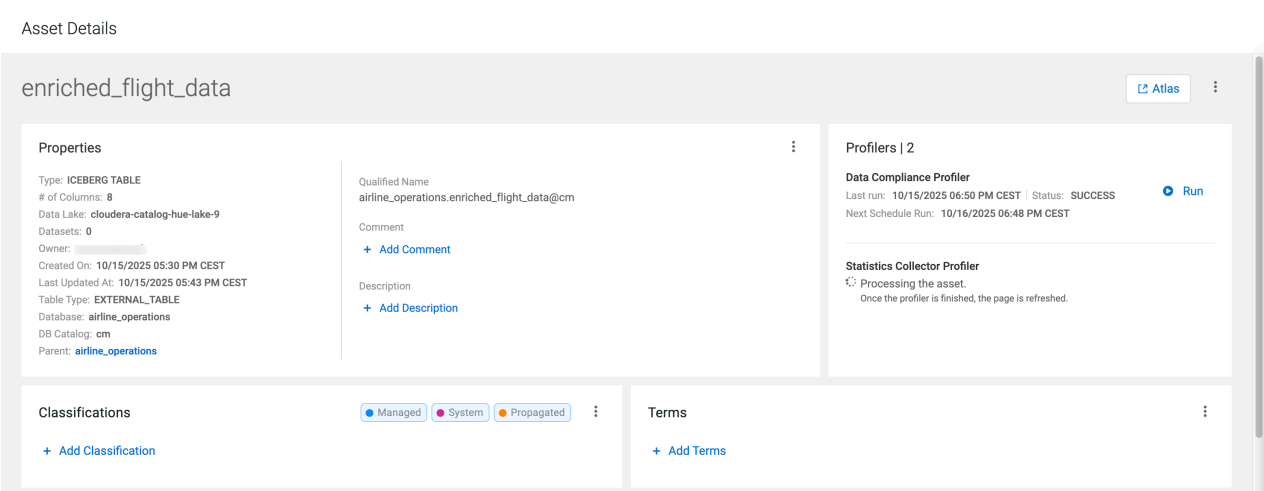
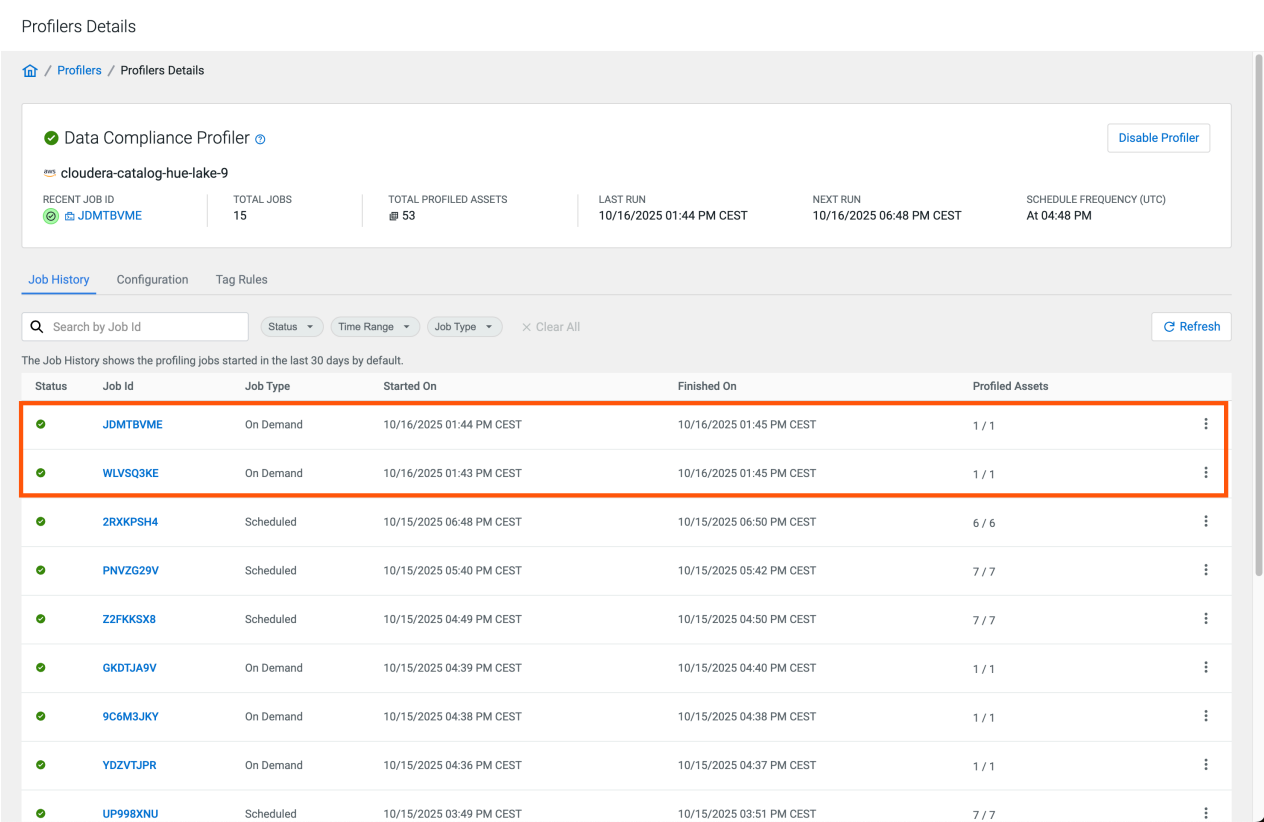



Figure 11: Tracking on-demand profilers



-  **Note:**
- In Compute Cluster-enabled environments, Iceberg tables can be also profiled with the On-Demand Profiler.
 - Profiler configurations apply to both scheduled and on-demand profiler jobs.

Deleting profilers in Compute cluster enabled environments

In Compute Cluster enabled environments deleting the profiler jobs 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



Note:

- In a Compute Cluster enabled environment, when you delete the scheduled jobs, the associated Kubernetes cron job object is deleted from the Kubernetes cluster.
- The associated data of the profilers from the Cloudera Management Console database is also deleted for the specified data lake.

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. Confirm the deletion in the message dialog box.

Figure 12: Deleting a profiler in a Compute Cluster enabled environment

The screenshot shows the Cloudera Data Catalog interface. On the left is a sidebar with navigation links: Dashboard, Search, Datasets, Bookmarks, Profilers (highlighted), and Atlas Tags. The main content area is titled 'Profilers' and shows a table of profilers for the data lake 'dc-datalake-hydrogen...'. A modal dialog titled 'Profiler Delete Confirmation' is open, asking 'Are you sure you want to delete the Data Compliance Profiler?'. The dialog has 'Cancel' and 'Confirm' buttons. The background table lists three profilers: Activity Profiler, Data Compliance Profiler, and Statistics Collector Profiler. The Data Compliance Profiler is the one being targeted for deletion.

Profiler	FREQUENCY (UTC)	NEXT RUN	TOTAL EXECUTIONS
Activity Profiler	at 00:00	03/14/2025 01:00 AM CET	3
Data Compliance Profiler	at 01:01	03/14/2025 02:01 AM CET	4
Statistics Collector Profiler	at 00:00		23

4. Click Confirm and repeat the step for each profiler.



Note: It might take a couple of minutes until all profilers are deleted as a running profiler cannot be stopped. Periodically click Refresh to update the status.



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



Note: By deleting the last profiler, you also delete the namespace and underlying infrastructure associated with the profilers.

The profiler cluster is deleted successfully.