

Cloudera Data Catalog Top Use Cases

Date published: 2019-11-14

Date modified: 2025-04-07

The Cloudera logo is displayed in a bold, orange, sans-serif font. The word "CLOUDERA" is written in all caps, with a stylized 'E' that has a horizontal bar extending to the right.

Legal Notice

© Cloudera Inc. 2025. All rights reserved.

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

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

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

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

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

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

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

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

Contents

Search for assets.....	4
Filters.....	4
Accessing Data Lakes.....	6
Searching for assets using Atlas glossaries.....	6
Using terms in Cloudera Data Catalog.....	7
Mapping glossary terms.....	8
Searching for assets using glossary terms.....	10
Additional search options for asset types.....	11
Searching for assets using additional search options.....	13
Accessing tables based on Ranger policies.....	14
Creating classifications for selected assets.....	15
Managing Profilers.....	17
Launching profilers in Compute Cluster enabled environments.....	17
Launching profilers using the command-line.....	22
Launching profilers in VM based environments.....	24
Launching profilers using the command-line.....	26
Enable or disable profilers in Compute cluster enabled environments.....	28
Enable or disable profilers in VM-based environments.....	29
Profiling table data in non-default buckets.....	29
Tracking profiler jobs in Compute cluster enabled environments.....	30
Tracking profiler jobs in VM-based environments.....	33
Viewing profiler configurations in Compute cluster enabled environments.....	34
Viewing profiler configurations in VM-based environments.....	37
Activity Profiler configuration.....	37
Ranger Audit Profiler configuration.....	39
Data Compliance profiler configuration.....	41
Cluster Sensitivity Profiler profiler configuration.....	43
Statistics Collector profiler configuration.....	45
Hive Column Profiler configuration.....	48
Backing up and restoring the profiler database.....	51
About the back up script.....	51
Running the back up script.....	52
Profiler tag rules in Compute Cluster enabled environments.....	53
Profiler tag rules in VM-based environments.....	55
Deleting profilers in Compute cluster enabled environments.....	57
Deleting profilers in VM-based environments.....	58
Atlas tag management.....	60

Search for assets

On the Cloudera Data Catalog **Search** page, select a data lake and enter a search string in the search box to view all the assets with details that contain the search string.

When you enter the search terms **Search**, you are looking up names, types, descriptions, and other metadata collected by Cloudera Data Catalog. The search index includes metadata (not data) about your environment and cluster data assets and operations. You can make the search more powerful by associating your own information (business metadata) to the stored assets.

**Note:**

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

Related Information

[Understanding datasets](#)

Filters

Use filters to refine the overview of all your available assets.

You must have access to at least one data lake to search and filter your results. By default, a data lake is already selected for you if you have access to it.

You can further refine your search results using filters as follows:

Owner

From all the owner names that appear, you can select the owner to further refine the results and display those search results with the selected owner.

Type

Select an entity type to view all the assets stored in that type of database.

- Azure BLOB
- Azure Container
- Azure Directory
- AWS S3 Bucket
- AWS S3 Object
- AWS S3 Pseudo Dir
- AWS S3 V2 Bucket
- AWS S3 V2 Directory
- AWS S3 V2 Object
- Hbase Column Family
- Hbase Namespace
- Hbase Table
- HDFS path
- Hive Column
- Hive DB
- Hive Table
- Iceberg Column

¹ Iceberg assets are discoverable in VM-based environments but they can be profiled only in Compute Cluster enabled environments.

- Iceberg Table¹
- Impala Column Lineage
- Impala Process
- Impala Process Execution
- Kafka topic
- ML Model Build
- ML Model Deployment
- ML Project
- RDBMS Column
- RDBMS DB
- RDBMS Foreign key
- RDBMS Index
- RDBMS Table
- Spark Application
- Spark Column
- Spark Column Lineage
- Spark DB
- Spark ML Directory
- Spark ML Model
- Spark ML Pipeline
- Spark Process
- Spark Process Execution
- Spark Table



Note: After selecting an entity type, further filters related to that type will be available under the More filter. For example, selecting the Hive Table type will enable the Column Tag filter.

Entity Tag

Use entity tags to refine your search results. You can add business metadata as entity tags in Atlas as classifications, or in the **Atlas Tags** menu. Use these tags to refine your search results and view the details of the required data asset.

Time Range

You can filter your assets by the **Created On** date (if provided by Atlas) after selecting an asset Type. Use the calendar widget to select a range and click Apply.

Glossary Terms

You can filter assets based on business glossary terms. You can search for any asset without any entity type restrictions.



Note: This filter appears only if Atlas has terms set up.

Click Cancel for any filter to clear the selection or Clear All to reset all your filters.

In the resulting list of your matching assets, you can click a row and see the following:

- **Qualified name**
- **Database**
- **Classification**
- **Terms**

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

Accessing Data Lakes

In the **Search** page, the accessible data lakes are displayed in a drop-down.

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.

For example, in the following diagram, the logged in user has access to all the listed data lakes.



Note: You can search the assets of one data lake at a time.

The screenshot shows the Cloudera Data Catalog interface. On the left is a navigation sidebar with options like Dashboard, Search, Datasets, Bookmarks, Profilers, and Atlas Tags. The main area is titled 'Search' and features a 'Discover data' section with a search prompt. Below this is a 'Data Lakes' dropdown menu, which is highlighted with an orange box. The dropdown shows several options, including 'dc-qe-edl-env-v1', 'env-v1', 'env-v2', and 'env-v1'. Below the dropdown is a table of HBase assets. The table has columns for Name, Created On, Owner, Source, and Action. The first row is 'HBase Namespace' with name 'default', created on '-NA-', owner 'atlas', and source 'hbase'. Subsequent rows are 'HBase Table' (atlas_janus), and several 'HBase Column Family' entries with names like 'h', 'l', 'g', 'i', 'e', 't', 's', 'f', and 'm'. All column families have 'Created On' as '-NA-' and 'Source' as 'hbase'. At the bottom right of the table, it says 'Rows per page: 100' and '1 - 11 of 11'.

Related Information

[Introduction to data lakes](#)

[Understanding data lake details](#)

Searching for assets using Atlas glossaries

Use Apache Atlas glossaries to define a common set of search terms that data users across your organization use to describe their data.

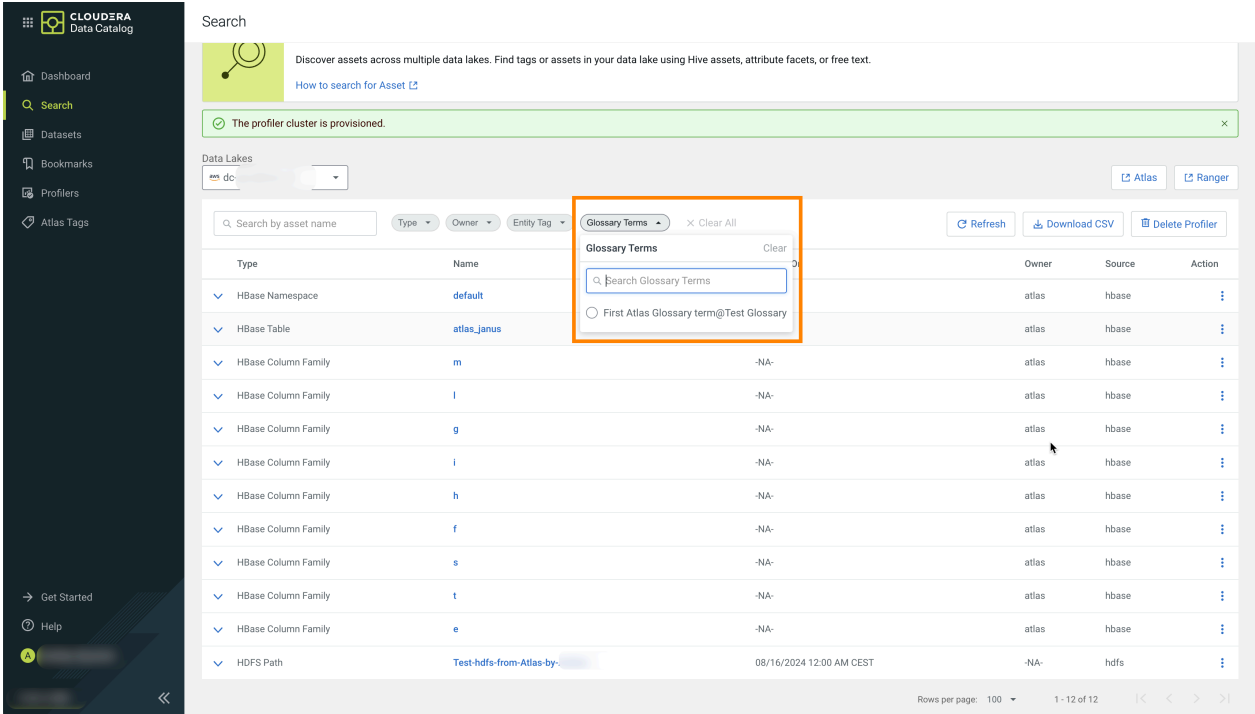
Data can describe a wide variety of content: lists of names or text or columns full of numbers. You can use algorithms to describe data as having a specific pattern, of being within a range or having wide variation, but what's missing from these descriptions is what does the data mean in a given business context and what is it used for? Is this column of integers the count of pallets that entered a warehouse on a given day or number of visitors for each room in a conference center?

The glossary is a way to organize the context information that your business uses to make sense of your data beyond what can be figured out just by looking at the content. The glossary holds the terms you've agreed upon across your organization so business users can use familiar terms to find what they are looking for.

Glossaries enable you to define a hierarchical set of business terms that represents your business domain.

Glossary terms can be thought of as of a flat (but searchable) list of business terms organized by glossaries. Unlike classifications, terms are not propagated through lineage relationships: the context of the term is what's important, so propagation may or may not make sense.

You can search for the datasets using the Glossary Terms filter available on the **Search** page.



The screenshot displays the Cloudera Data Catalog Search page. A sidebar on the left contains navigation links for Dashboard, Search, Datasets, Bookmarks, Profilers, and Atlas Tags. The main content area features a search bar and several filter options: Type, Owner, Entity Tag, and Glossary Terms. The Glossary Terms filter is currently selected and expanded, showing a search input field and a radio button option labeled "First Atlas Glossary term@Test Glossary". Below the filters is a table of assets. The table has columns for Type, Name, Owner, Source, and Action. The assets listed are primarily HBase Column Families with names like 'default', 'atlas_janus', 'm', 'l', 'g', 'i', 'h', 'f', 's', 't', and 'e'. The last row is an HDFS Path. At the bottom right, there are pagination controls showing "Rows per page: 100" and "1 - 12 of 12".

Using terms in Cloudera Data Catalog

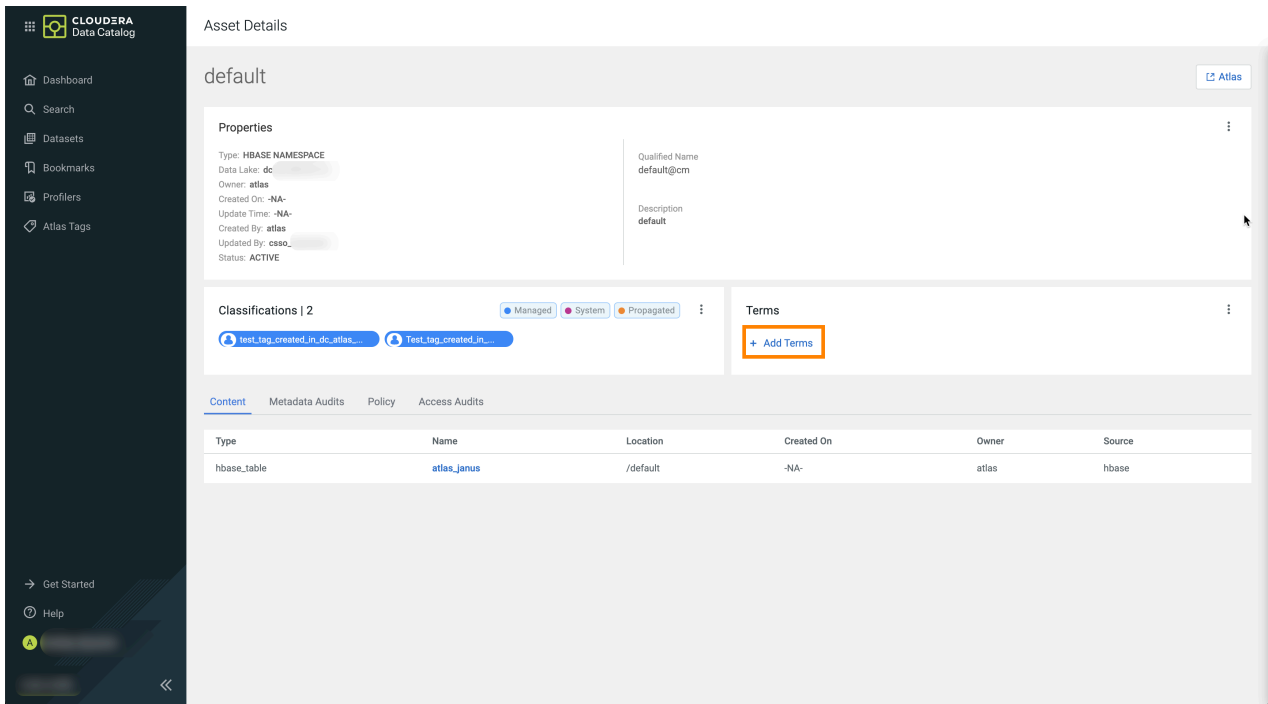
You can use the Asset Details page to add or modify Apache Atlas glossary terms for your selected assets.

Use Atlas to define rich glossary vocabularies using the natural terminology (technical terms and/or business terms) of your industry. You can also create semantic relationships between your terms. Then, in Cloudera Data Catalog, use the **Terms** widget in the **Asset Details** page to map assets to glossary terms.

You can use terms in Cloudera Data Catalog to search for entities, filter them by glossary term(s), and also search for entities associated with them in Atlas.



Note: When you work with terms in Cloudera Data Catalog and map them to your assets, you can search for the same datasets in Atlas by using the corresponding terms.



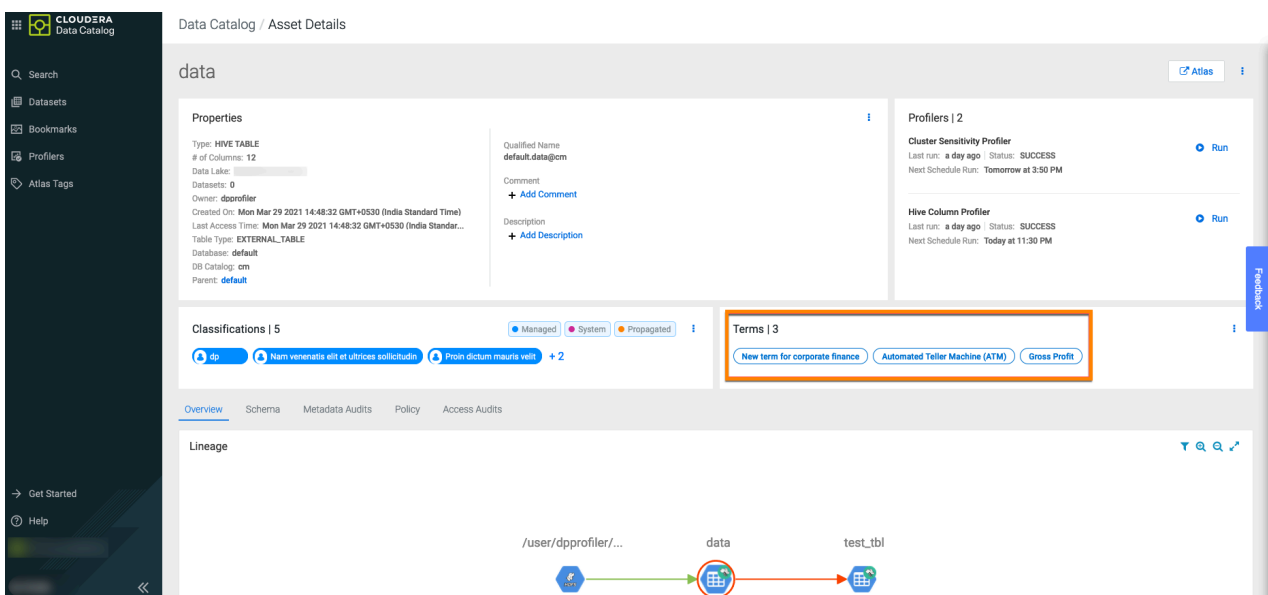
Mapping glossary terms

Cloudera Data Catalog contains the glossary terms that are created in Apache Atlas.

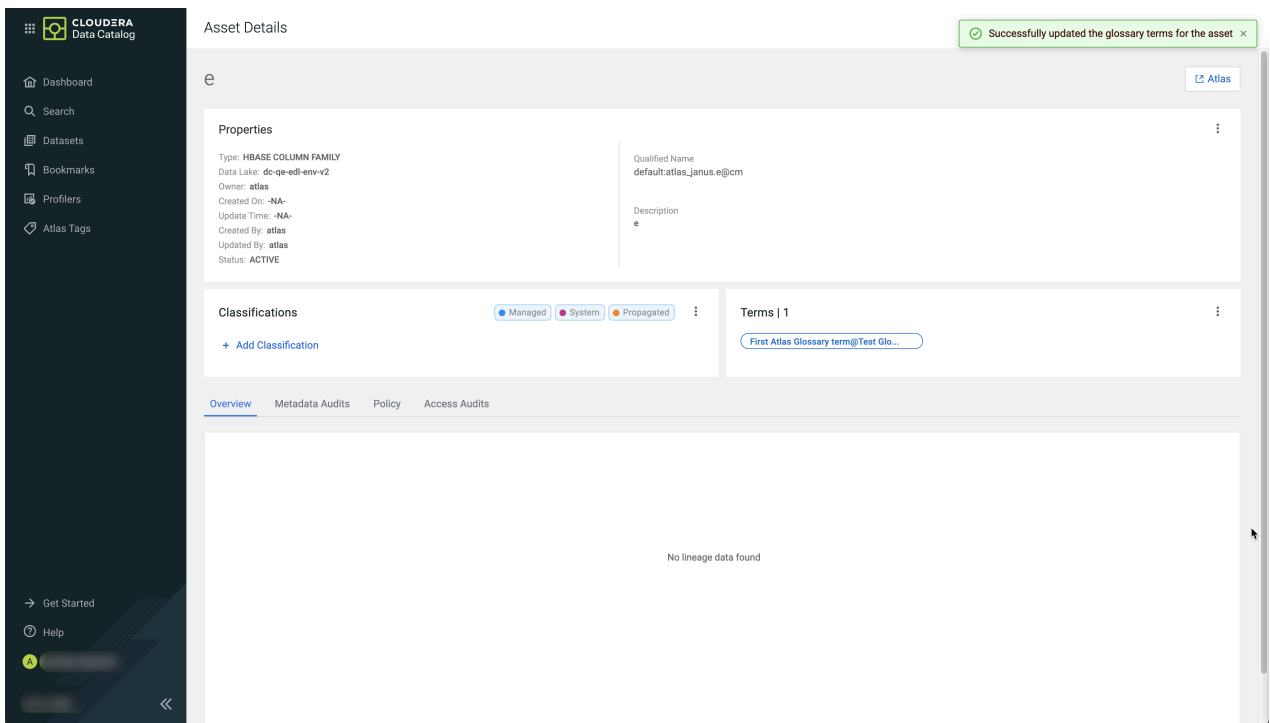
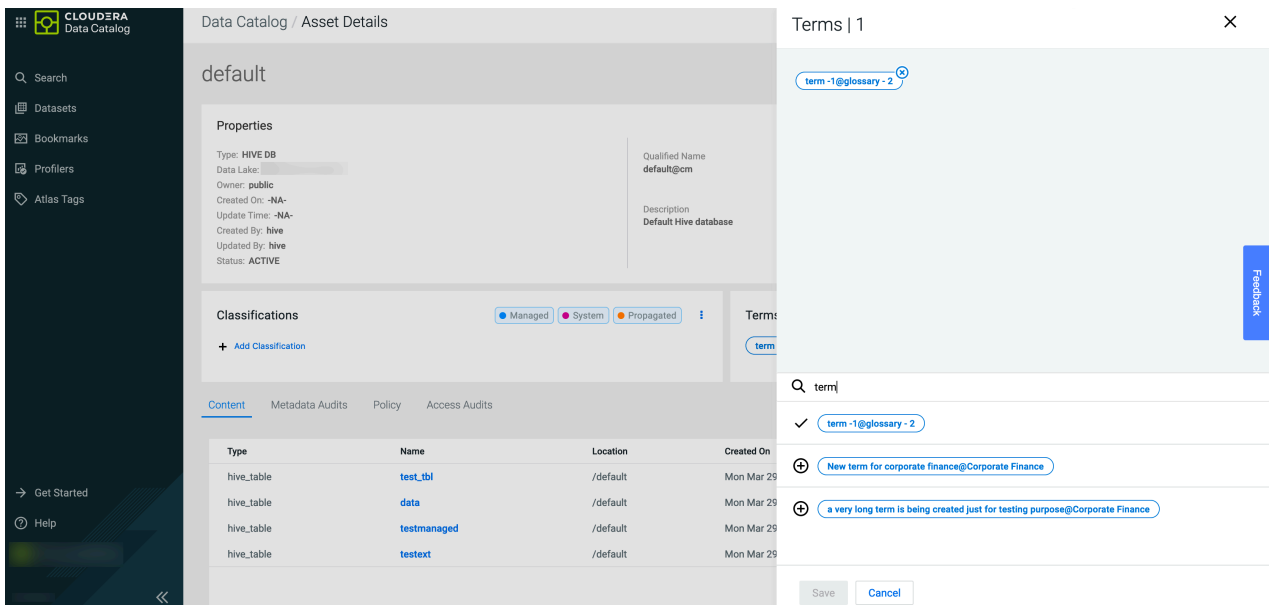
You can search for those terms in Cloudera Data Catalog and map specific terms with assets. You can also search for terms to delete them from the selected asset. The selected asset displays the total number of terms associated or mapped accordingly.

When you map a specific term for your dataset, the term is displayed in the following format:

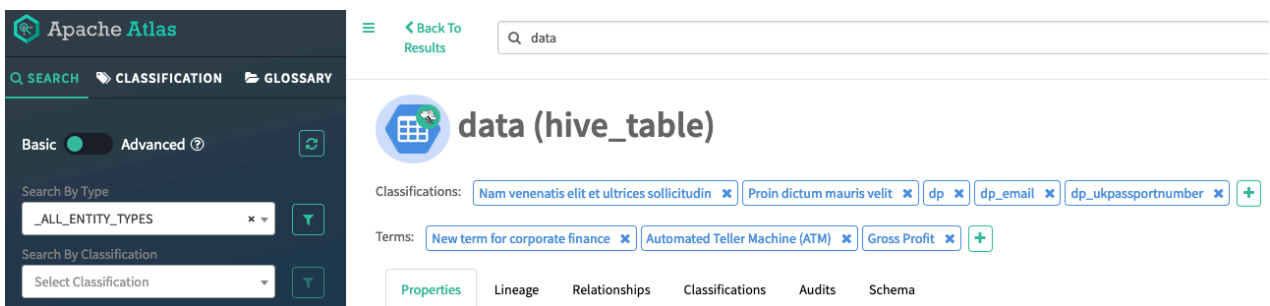
```
<termname>@glossaryname>
```



You can use the icon in the **Terms** widget on the **Asset Details** page to add new terms for your assets. Click Save to save the changes.



You can search for the same asset in the corresponding Atlas environment as shown in the example image.



When you select a Hive table asset and navigate to the **Asset Details** page, under the **Schema** tab, you can view the list of terms associated with the asset.

Chart Type	Column Name	Type	Unique Values *	Null Values	Max	Min	Mean	Comment	Classifications	Terms
▼	age	int	21	18	49	1	23.66		Nam venenatis elit et. + 1	
▼	cabin	string	9	0					Nam venenatis elit et. + 1	Accounting Rate of ... + 1
▼	embarked	string	3	0					dp_ukpassportnumbe + 2	Compound Annual G... + 1
▼	fare	float	35	0	262.38		23.78		dp_ukpassportnumbe + 1	New term for corpor... + 5
▼	name	string	54	0					dp_ukpassportnumbe + 6	New term for corpor... + 5
▼	parch	int	3	0	2		0.42		dp_ukpassportnumbe + 1	New term for corpor... + 6
▼	passengerid	int	50	0	53	1	27		dp_ukpassportnumbe + 2	New term for corpor... + 2
▼	pclass	int	3	0	3	1	2.42		dp_ukpassportnumbe	New term for corpor... + 5
▼	sex	string	2	0					dp_ukpassportnumbe	a very long term is b... + 6
▼	sibsp	int	4	0	8		0.43			
▼	survived	int	2	0	1		0.72			
▼	ticket	string	48	0						

You can add or update the terms for the associated datasets by clicking the Edit button.

The screenshot shows the 'Data Catalog / Asset Details' page. The 'Edit' button in the top right corner of the table is highlighted with a red arrow. A dropdown menu is open over the 'Terms' column, showing a list of terms and their counts. The 'Terms' column in the table below shows terms like 'a very long term is b...' with counts of +7.

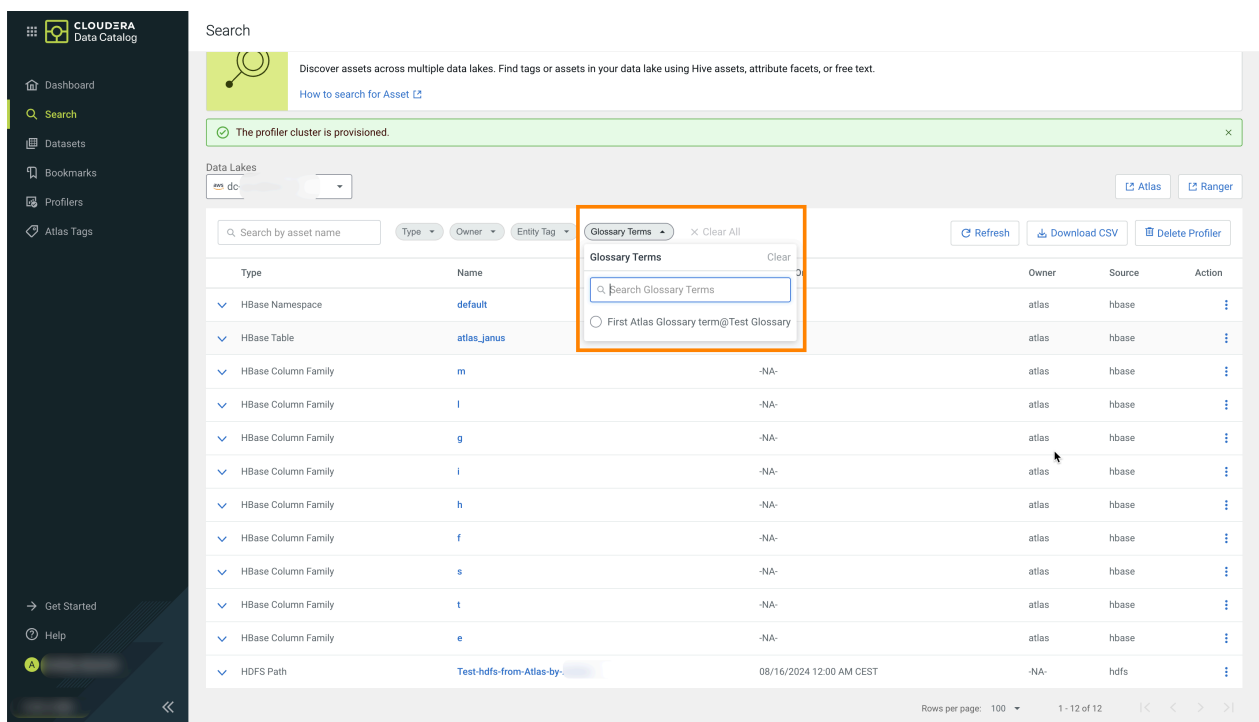
Searching for assets using glossary terms

You can search for the datasets using the Glossary Terms filter available on the Search page.



Note: The option for searching based on Glossary terms appears only if there are terms available in Apache Atlas.

1. Go to **Search**.
2. Select your data lake.
3. Click the Glossary Terms drop-down and selected the term to be searched.



Additional search options for asset types

Using Cloudera Data Catalog, you can add or edit asset description values to search for data assets across both Cloudera Data Catalog and Apache Atlas services by using the asset content. These values can be searched.

Adding comments and descriptions to assets

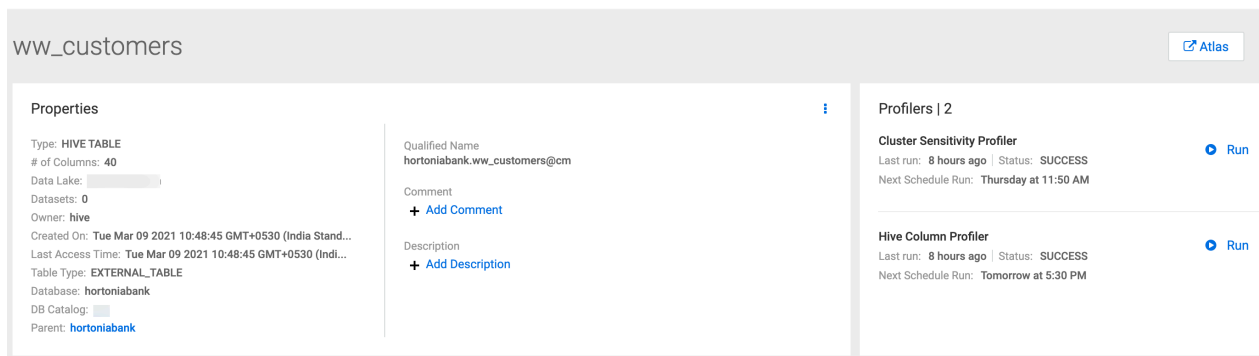
In the **Asset Details** page for each asset type that you select, you can add or edit **comment** or **description** fields. Including these values for the selected asset helps you to identify your chosen asset.

Using the same set of values (comment or description), you can also search for the asset types in Atlas.



Note: The comment and description options are supported only for Hive table and Hive Column assets. For other asset types, only the description option is supported.

Data Catalog / Asset Details



Click **+ Add Comment** or **+ Add Description** fields to include the respective values.

Data Catalog / Asset Details

ww_customers [Atlas](#)

Properties

Type: HIVE TABLE
of Columns: 40
Data Lake:
Datsets: 0
Owner: hive
Created On: Tue Mar 09 2021 10:48:45 GMT+0530 (India Stand...
Last Access Time: Tue Mar 09 2021 10:48:45 GMT+0530 (Indi...
Table Type: EXTERNAL_TABLE
Database: hortoniabank
DB Catalog:
Parent: hortoniabank

Qualified Name
hortoniabank.ww_customers@cm

Comment
passport_number

Description
visa_number

Profilers | 2

Cluster Sensitivity Profiler
Last run: 9 hours ago | Status: SUCCESS
Next Schedule Run: Thursday at 11:50 AM

Hive Column Profiler
Last run: 8 hours ago | Status: SUCCESS
Next Schedule Run: Tomorrow at 5:30 PM

[Cancel](#) [Save](#) [Run](#) [Run](#)

Click Save to save your changes.

Data Catalog / Asset Details

Asset details were updated successfully.

ww_customers [Atlas](#)

Properties

Type: HIVE TABLE
of Columns: 40
Data Lake:
Datsets: 0
Owner: hive
Created On: Tue Mar 09 2021 10:48:45 GMT+0530 (India Stand...
Last Access Time: Tue Mar 09 2021 10:48:45 GMT+0530 (Indi...
Table Type: EXTERNAL_TABLE
Database: hortoniabank
DB Catalog:
Parent: hortoniabank

Qualified Name
hortoniabank.ww_customers@cm

Comment
passport_number

Description
visa_number


Profilers | 2

Cluster Sensitivity Profiler
Last run: 9 hours ago | Status: SUCCESS
Next Schedule Run: Thursday at 11:50 AM


Hive Column Profiler
Last run: 8 hours ago | Status: SUCCESS
Next Schedule Run: Tomorrow at 5:30 PM

[Run](#) [Run](#)



Note: You can also edit the already saved valued by clicking the  icon.

Clicking on the Atlas button will navigate to the corresponding Atlas asset page as shown:

 **ww_customers (hive_table)**

Classifications: [+](#)

Terms: [+](#)

[Properties](#) [Lineage](#) [Relationships](#) [Classifications](#) [Audits](#) [Schema](#)

Technical properties

columns (40)
title
givenname
middleinitial

comment passport_number

createTime 03/09/2021 10:48:45 AM (IST)

db hortoniabank

dcProfiledData
{
samplePercent: "100.0",
rowCount: 50000,
}

description visa_number

[User-defined properties](#) [Add](#)

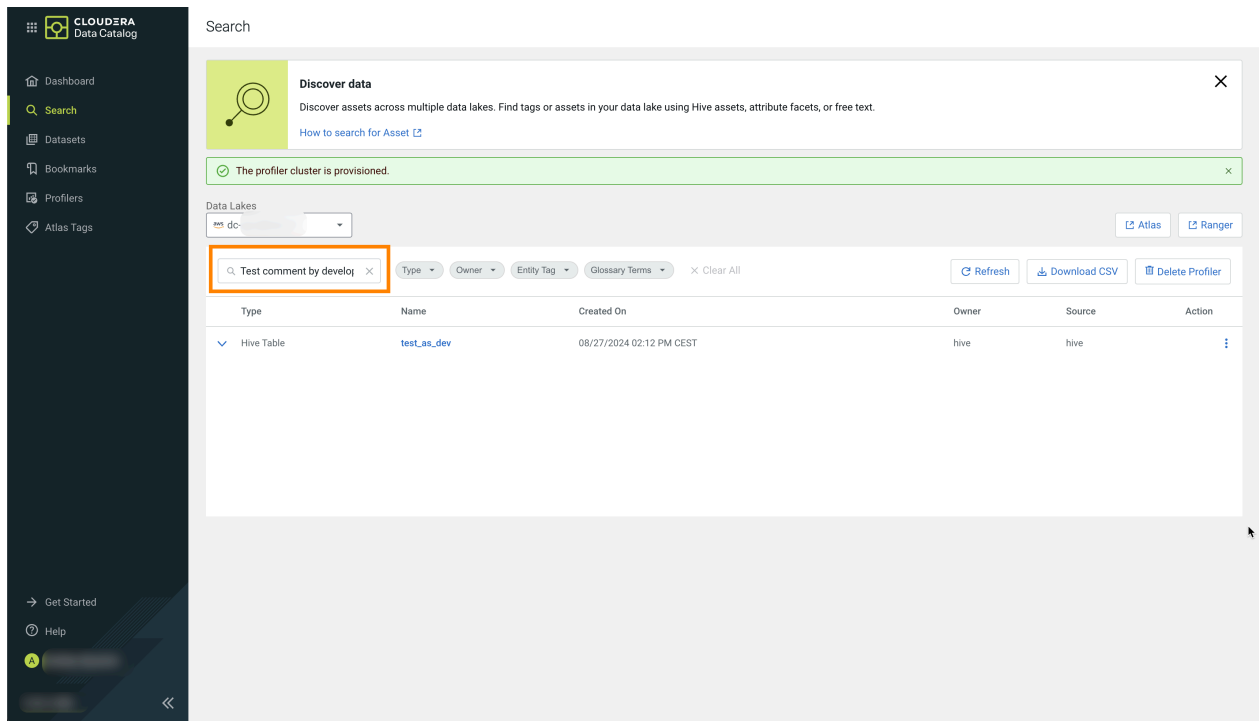
[Labels](#) [Add](#)

[Business Metadata](#) [Add](#)

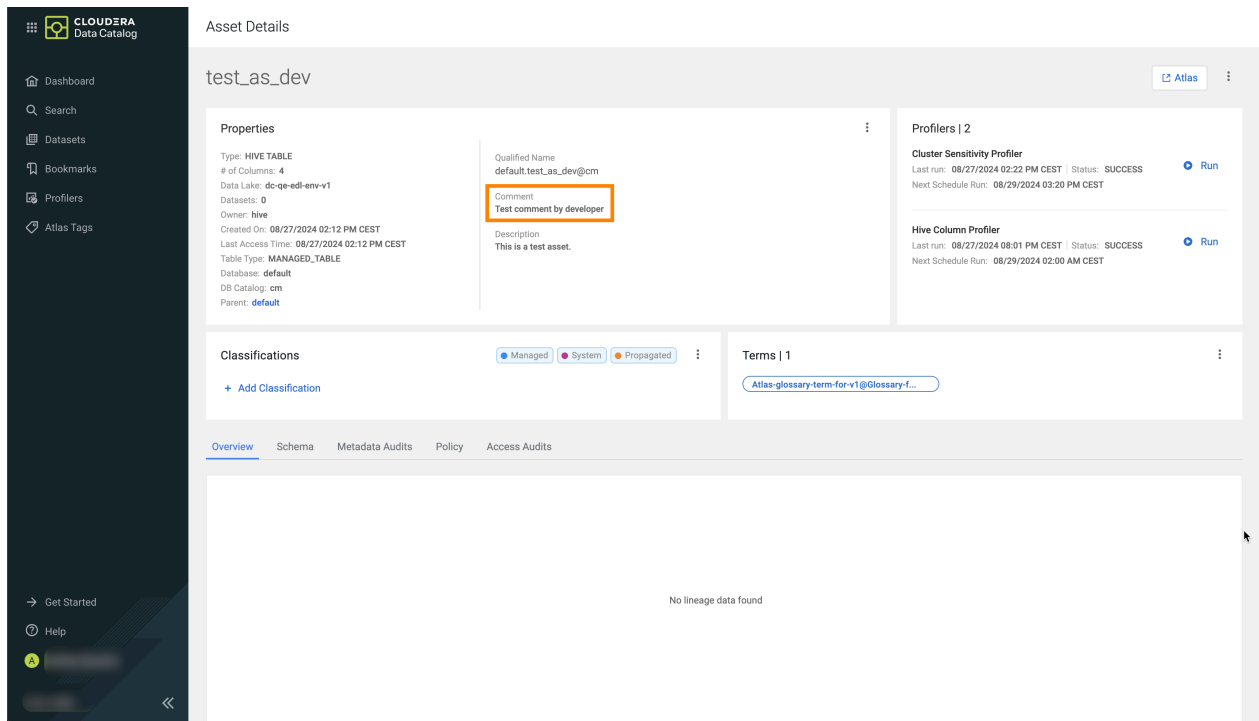
[Switch to Beta UI](#)

Searching comments and descriptions

The values of the **Comment** or **Description** fields can be searched in the **Search** menu. The result page displays the assets where you added your comments and descriptions without the use of filters.



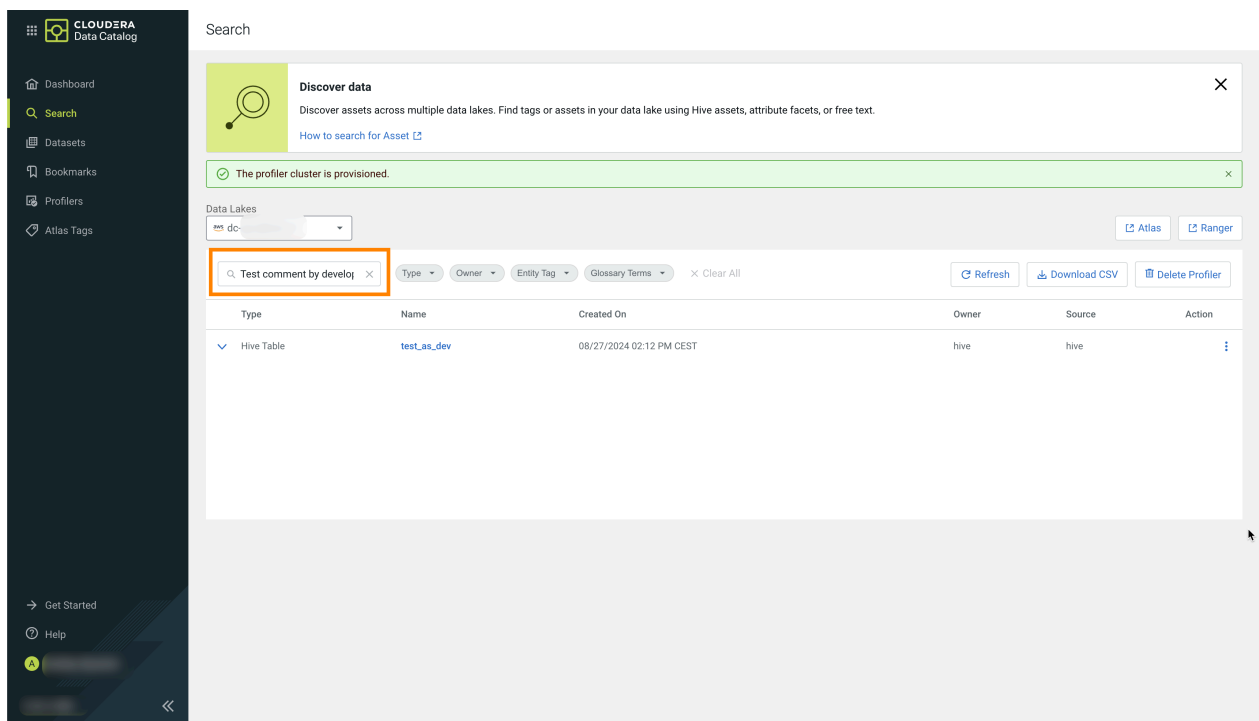
Clicking on the asset type displays the comment and description values.



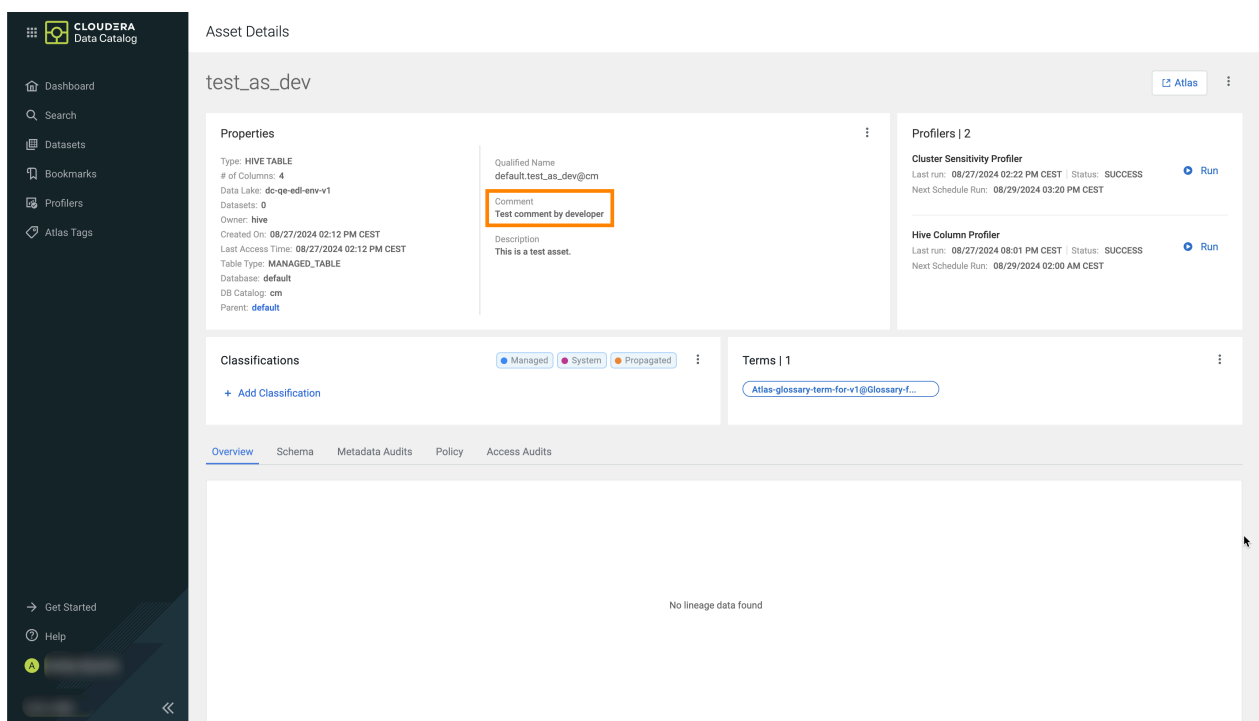
Searching for assets using additional search options

In Cloudera Data Catalog, you can select a data asset type and under the Asset Details page, to insert a comment and to provide a description for the selected asset.

The values of the **Comment** or **Description** fields can be searched in the **Search** menu. The result page displays the assets where you added your comments and descriptions without the use of filters.



Clicking on the asset type displays the comment and description values.

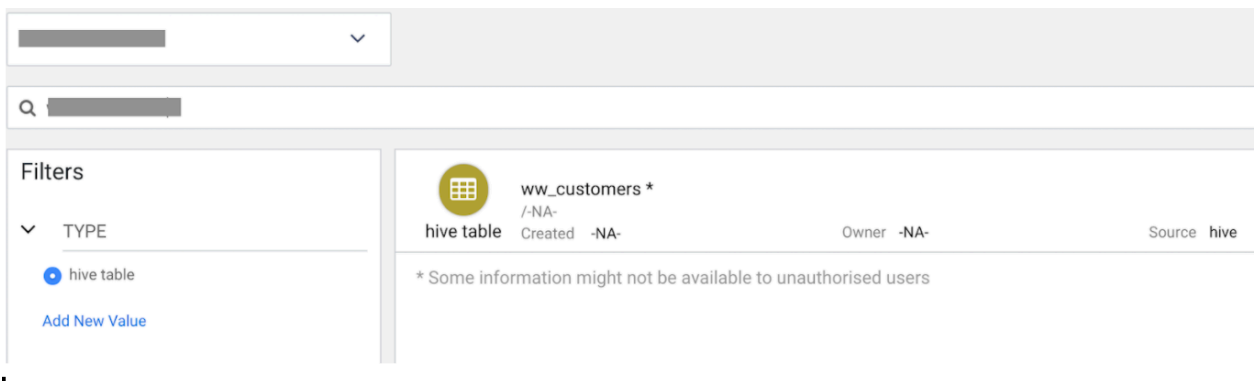


Accessing tables based on Ranger policies

When a table (in blue color link) is clicked, the Asset Details view page is displayed.

If a user is not authorized to click or view table details, it implies that the user permissions have not been set up in the Apache Ranger.

As seen in the following diagram, if users are not able to view the table details, a message appears next to the same table "Some information might not be available to unauthorised users".



In the next example diagram, tables that have the permissions to view are displayed with a blue color link. The ones that do not have read permissions are visible in grey.


Filter	Table Name	Created	Owner	Source
CREATED BEFORE Clear <input type="radio"/> Last 1 day <input type="radio"/> Last 7 days <input type="radio"/> Last 15 days Add New Value	scheduled_queries	Tue Apr 07 2020	hive	hive
	schemata	Tue Apr 07 2020	hive	hive
	table_stats_view	Tue Apr 07 2020	hive	hive
	scheduled_executions	Tue Apr 07 2020	hive	hive
	andromeda	-	-	hive
	milky	-	-	hive
	bear	-	-	hive
	n170	-	-	hive
	umajor5	-	-	hive

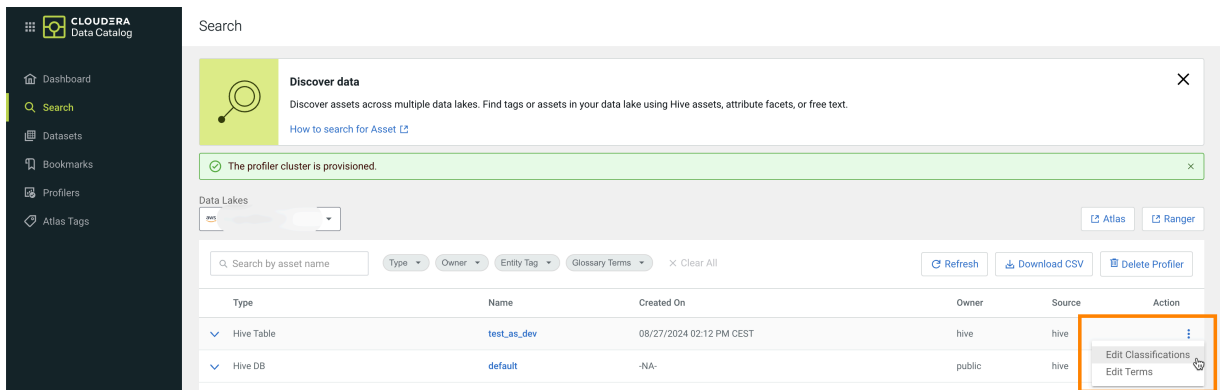
Creating classifications for selected assets

You can create classifications in multiple pages. These classifications can be associated with an asset. Then, you can use these classifications to filter your assets both in Cloudera Data Catalog and Apache Atlas.

Creating a classification from the Search page


1. Navigate to the **Search** page.

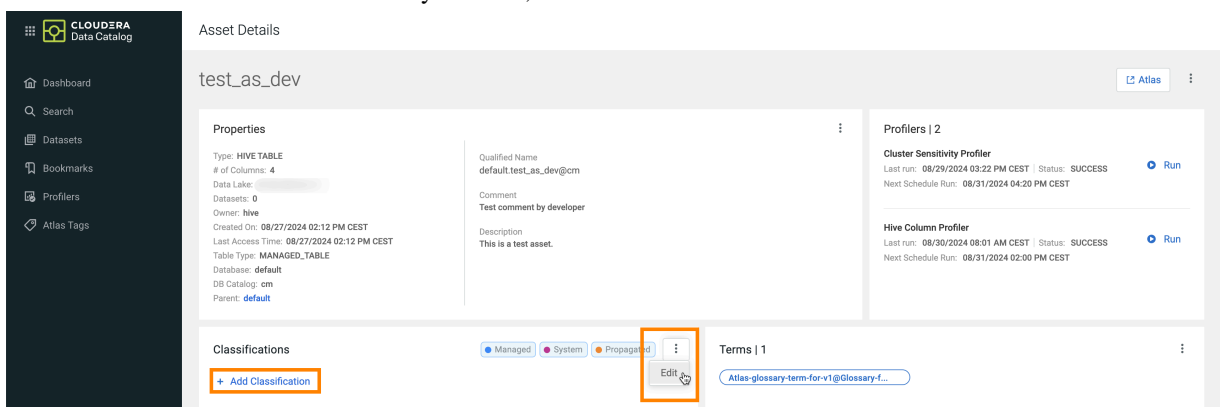
2. Click the  icon by an asset, then select Edit Classifications.



3. Search for a previously created classification or create a new one.
4. Click Save to finalize your changes.

Creating a classification from Asset Details

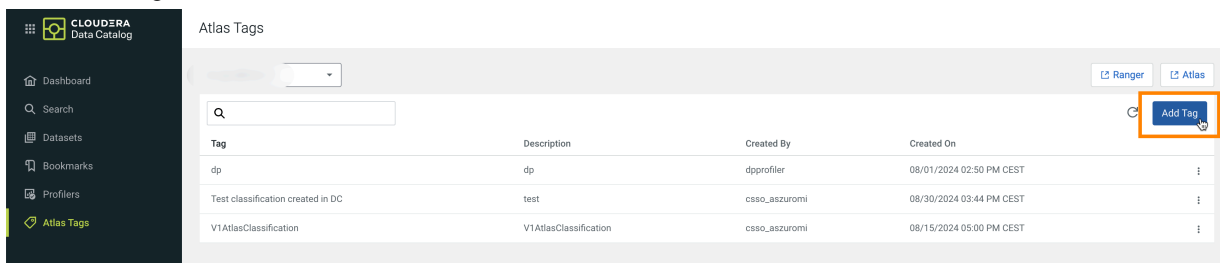
1. Navigate to the **Asset Details** page of an asset.
2. Click Add Classification or  icon by an asset, then select Edit.



3. Search for a previously created classification or create a new one.
4. Click Save to finalize your changes.

Creating a classification in Atlas Tags

1. Navigate to **Atlas Tags**.
2. Click Add Tag.



3. Fill in the details and Save your changes.



Note: Your classification still needs to be added to an asset in the **Search** or **Asset Details** menu.



Note: Classifications are synchronized between Apache Atlas and Cloudera Data Catalog.

Managing Profilers

The Cloudera Data Catalog profiler engine runs data profiling operations on data located in multiple data lakes. These profilers create metadata annotations that summarize the content and shape characteristics of the data assets.

Table 1: List of built-in profilers

Profiler Name in VM-based environments	Profiler Name in Compute Cluster enabled environments	Description
Cluster Sensitivity Profiler	Data Compliance	A sensitive data profiler- PII, PCI, HIPAA, etc.
Ranger Audit Profiler	Activity Profiler	A Ranger audit log summarizer.
Hive Column Profiler	Statistics Collector	Provides summary statistics like Maximum, Minimum, Mean, Unique, and Null values at the Hive column level.

Limitations

- In VM-based environments, profilers do not support Iceberg tables. However, Iceberg tables are discoverable. In Compute Cluster enabled environments, Iceberg tables can be profiled.
- In Compute Cluster enabled environments, profilers only support tables which are stored on AWS S3 storage.
- Supported file formats:
 - VM-based environments:
 - CSV
 - Compute Cluster enabled environments:
 - Statistics Collector profilers and Data Compliance profilers
 - CSV
 - Parquet
 - Iceberg tables
 - ORC



Note: Text format tables are not supported in Compute Cluster enabled environments. Profilers skip tables containing text and continue with the next selected asset. The status SKIPPED is shown in Profiler Details Job History Job Summary Profiled Assets .

Related Information

[Understanding the Cloudera Data Catalog Profiler](#)

[Understanding the Cluster Sensitivity Profiler](#)

[Understanding the Ranger Audit Profiler](#)

Launching profilers in Compute Cluster enabled environments

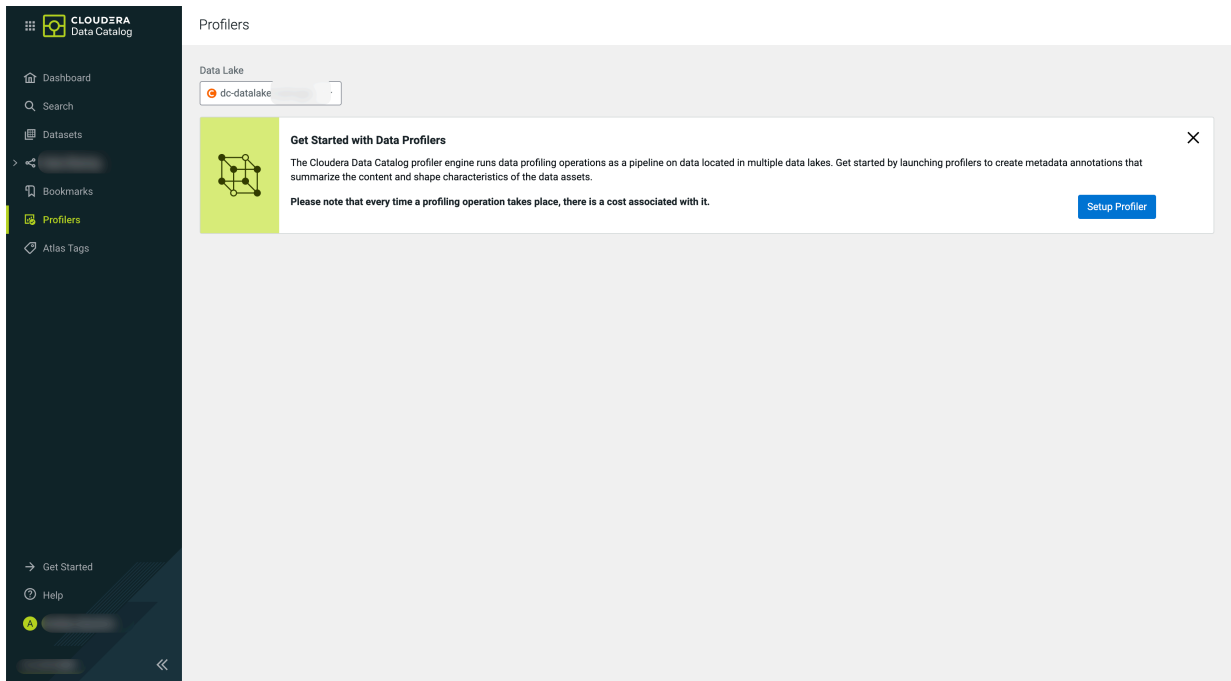
In Compute Cluster enabled environments, after you set up the profiler, the Profiler Launcher Services automatically starts the profiler Kubernetes containers.



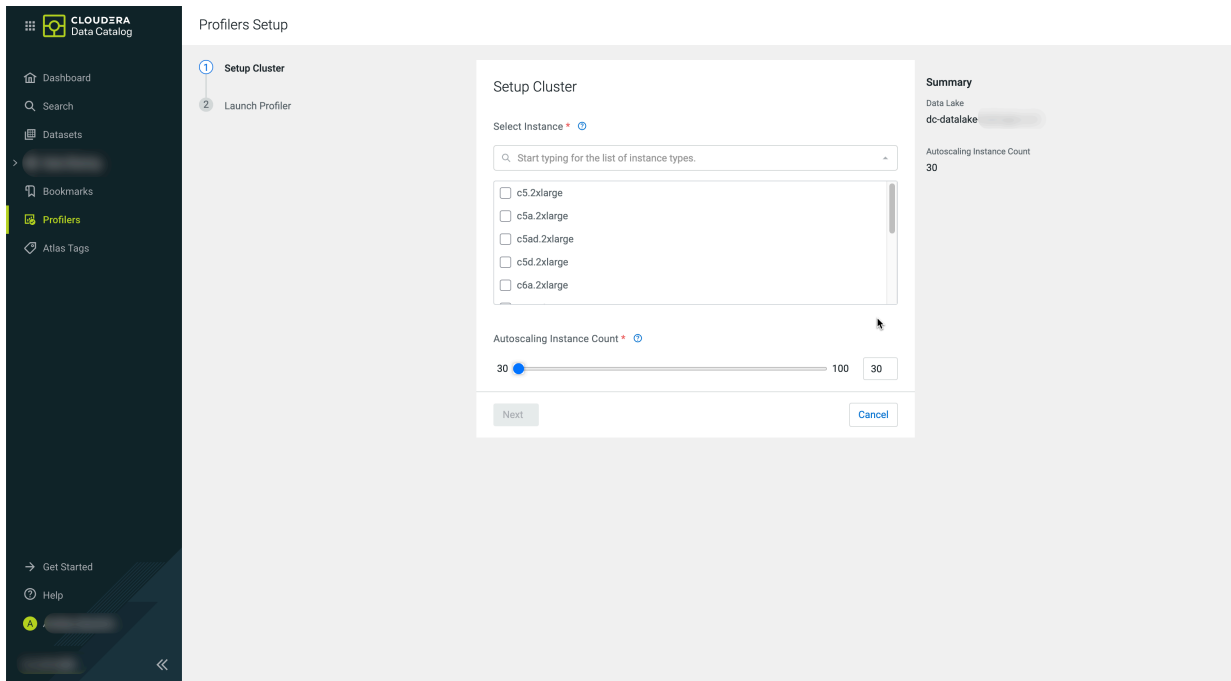
Note: You must be a Power User to launch a profiler cluster.

How to launch the profiler for Compute Cluster enabled environments

1. On the **Profilers** page, select the data lake from which you want to launch the profiler cluster.
2. Click Setup Profiler, to start the profiler cluster setup.



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

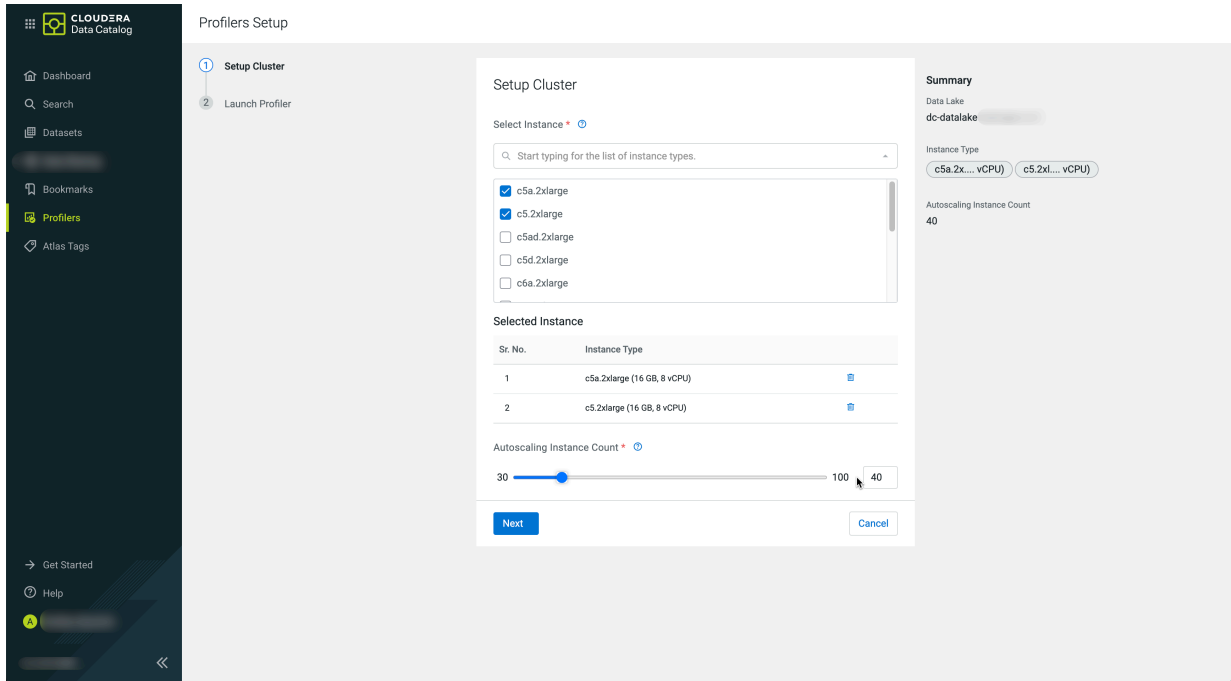


The available instance types depend on the cloud provider of the underlying environment. Choose from them based on your performance and cost requirements.



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

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



5. Click Next.

6. Select the necessary profilers to be launched.



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

Profilers Setup

1 Setup Cluster

2 Launch Profiler

Launch Profiler

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

Profiler Configuration :

WORKER MEM LIMIT: 4G

NUM WORKERS: 4

THREAD PER WORKER: 3

CRON EXPRESSION: 0 0 ***

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

Profiler Configuration :

WORKER MEM LIMIT: 11G

NUM WORKERS: 10

THREAD PER WORKER: 3

CRON EXPRESSION: 0 0 ***

LAST RUN: Over a period of 2 days

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

Profiler Configuration :

WORKER MEM LIMIT: 11G

NUM WORKERS: 10

THREAD PER WORKER: 3

CRON EXPRESSION: 0 0 ***

LAST RUN: Over a period of 2 days

[← Previous](#) [Start Setup](#) [Cancel](#)

Summary

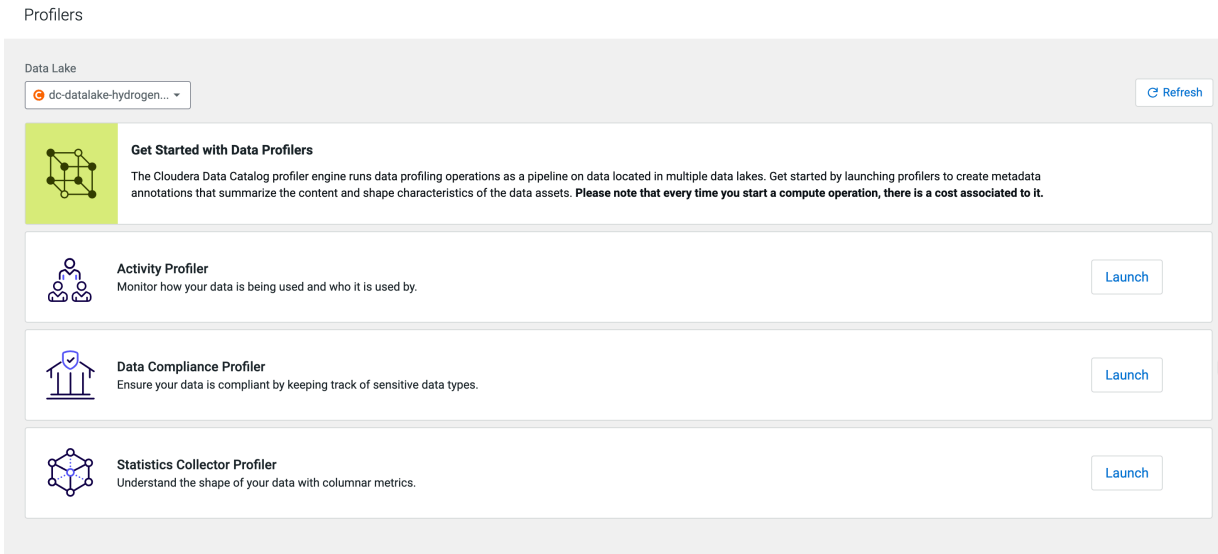
Data Lake: dc-datalake-1

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

Autoscaling Instance Count: 40

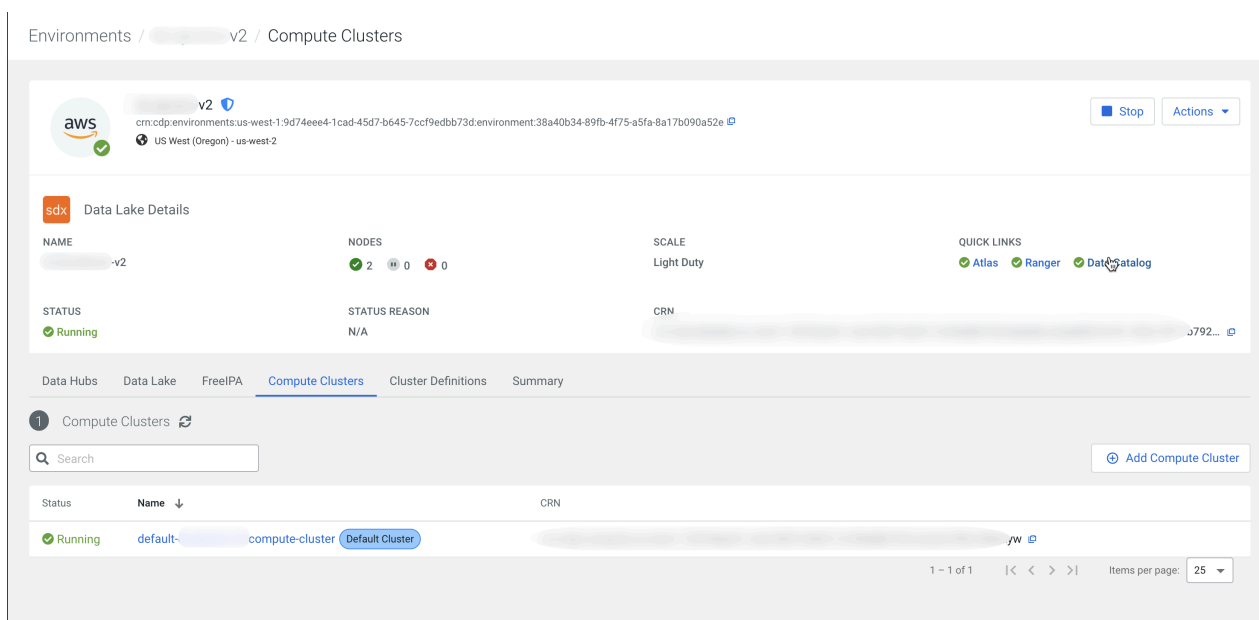
Profilers: Activity, Data C...liance, Table ...istics

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



The screenshot displays the Cloudera Data Catalog interface for a compute cluster. At the top, the cluster name is 'default-dc-qe-env-v2-compute-cluster'. Below this, a summary row shows the cluster is 'Running', of type 'Default Cluster', created on '05/08/2024, 05:54:19' by 'Deepak Kumar Singh'. An 'Actions' button is present. The 'Node Groups' section is expanded, showing three groups:

- dcprofiler**: Root volume size 50 GIB, 1 node, auto-scales between 1 and 10.
- dcprofiler-worker-spot**: Root volume size 100 GIB, 0 nodes, auto-scales between 0 and 81.
- liftie-infra**: Root volume size 40 GIB, 2 nodes, 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
```

You get additional information about this command by using:

```
cdp datacatalog launch-profilers --help
```

NAME

```
launch-profilers -
```

DESCRIPTION

```
Launches DataCatalog profilers in a given datalake.
```

SYNOPSIS

```
launch-profilers
--datalake <value>
[--cli-input-json <value>]
[--generate-cli-skeleton]
```

OPTIONS

```
--datalake (string)
```

The 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 values 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 input can be used as an argument for --cli-input-json.

OUTPUT

```
.
datahubCluster -> (object)
  Information about a cluster.
clusterName -> (string)
  The name of the cluster.
crn -> (string)
  The CRN of the cluster.
creationDate -> (datetime)
  The date when the cluster was created.
clusterStatus -> (string)
  The status of the cluster.
nodeCount -> (integer)
  The cluster node count.
workloadType -> (string)
  The workload type for the cluster.
cloudPlatform -> (string) The cloud platform.
imageDetails -> (object)
  The details of the image used for cluster instances.
  name -> (string)
    The name of the image used for cluster instances.
  id -> (string)
    The ID of the image used for cluster instances.
    This is internally generated by the cloud provider to Uniquely identify the image.
catalogUrl -> (string)
  The image catalog URL.
catalogName -> (string)
  The image catalog name.
environmentCrn -> (string)
  The CRN of the environment.
credentialCrn -> (string)
  The CRN of the credential.
datalakeCrn -> (string)
  The CRN of the attached datalake.
```

```
clusterTemplateCrn -> (string)
The CRN of the cluster template used for the cluster creation.
```

Launching the profiler

You can use the following CLI command to launch the data profiler:

```
cdp datacatalog launch-profilers --datalake [***DATALAKE CRN***]
```

Example:

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

Launching profilers in VM based environments

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



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

Profiler cluster in VM based environments

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



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



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

- AWS - m5.4xlarge
- Azure - Standard_D16_v3
- GCP - e2-standard-16

There are two types of Profiler Services:

- Profiler Manager
- Profiler Scheduler

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



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

How to launch the profiler cluster for VM based environments

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

Profiler Setup - [Redacted]

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

Enable High Availability

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

[Setup Profiler](#)

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

Profiler Setup - [Redacted]

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

Enable High Availability

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

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

[Setup Profiler](#)

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

Profiler Cluster is being created						
[Redacted] 2619						Action
<input type="checkbox"/> Type	Name	Qualified Name	Created On	Owner	Source	
<input type="checkbox"/> Azure Container	container	abfs://container@sparktestingstorage...	-NA-	-NA-	adls	
<input type="checkbox"/> AWS S3 V2 Bucket	s3-extractor-test	s3a://s3-extractor-test@cm	-NA-	-NA-	aws	
<input type="checkbox"/> Hive Table	lounge	airline.lounge@cm	Mon Oct 04 2021	hrt_1	hive	

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

Profiler Cluster is provisioned successfully

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

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

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

Environments / v1 / Clusters

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

Stop Actions

sdx Data Lake Details

NAME	NODES	SCALE	QUICK LINKS
v1	2 0 0	Light Duty	Atlas Ranger Data Catalog

STATUS: Running STATUS REASON: N/A CRN: [redacted]

Data Hubs Data Lake FreelPA Compute Clusters Cluster Definitions Summary

Data Hubs

Search Create Data Hub

Status	Name	Data Hub Type	Runtime	Node Count	Created
Running	v1	profiler_7.2.18-0	7.2.18	3	8/2/2024, 08:36:00

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

Launching profilers using the command-line

Cloudera Data Catalog supports launching profilers using the Command-Line Interface (CLI) option.

The CLI is one executable and does not have any external dependencies. You can execute some operations in the Cloudera Data Catalog service using the Cloudera CLI commands.

Users must have valid permissions to launch profilers on a data lake.

For more information about the access details, see [Prerequisites to access Cloudera Data Catalog](#).

Prerequisites

You must have the following entitlement granted to use this feature:

DATA_CATALOG_ENABLE_API_SERVICE

For more information about the Cloudera command-line interface and setting up the same, see [Cloudera CLI](#).

The Cloudera Data Catalog CLI

In your Cloudera CLI environment, enter the following command to get started in the CLI mode.

```
cdp datacatalog --help
```

This command provides information about the available commands in Cloudera Data Catalog for Cloudera on cloud 7.2.18, and earlier versions.

The output is displayed as:

```
NAME
datacatalog
DESCRIPTION
Cloudera Data Catalog Service is a web service, using this service user can
  execute operations like launching profilers in Data Catalog.
AVAILABLE SUBCOMMANDS
launch-profilers
```

You get additional information about this command by using:

```
cdp datacatalog launch-profilers --help
```

```
NAME
launch-profilers -
DESCRIPTION
Launches DataCatalog profilers in a given datalake.
```

```
SYNOPSIS
  launch-profilers
  --datalake <value>
  [--cli-input-json <value>]
  [--generate-cli-skeleton]
OPTIONS
  --datalake (string)
The 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 values will override th
e 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 input can be
used as an argument for --cli-input-json.
```

```
OUTPUT
.
datahubCluster -> (object)
Information about a cluster.
clusterName -> (string)
The name of the cluster.
crn -> (string)
The CRN of the cluster.
creationDate -> (datetime)
The date when the cluster was created.
clusterStatus -> (string)
The status of the cluster.
nodeCount -> (integer)
The cluster node count.
workloadType -> (string)
The workload type for the cluster. cloudPlatform -> (string) The cloud plat
form.
imageDetails -> (object)
The details of the image used for cluster instances.
name -> (string)
The name of the image used for cluster instances.
id -> (string)
The ID of the image used for cluster instances.
This is internally generated by the cloud provider to Uniquely identify the
image.
```

```

catalogUrl -> (string)
The image catalog URL.
catalogName -> (string)
The image catalog name.
environmentCrn -> (string)
The CRN of the environment.
credentialCrn -> (string)
The CRN of the credential.
datalakeCrn -> (string)
The CRN of the attached datalake.
clusterTemplateCrn -> (string)
The CRN of the cluster template used for the cluster creation.

```

Launching the profiler

You can use the following CLI command to launch the data profiler:

```
cdp datacatalog launch-profilers --datalake [***DATALAKE CRN***]
```

Example:

```


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

```

Enable or disable 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 [Refresh](#)

Activity Profiler	FREQUENCY (UTC)	NEXT RUN	TOTAL EXECUTIONS	
 Activity Profiler	-NA-	03/09/2025 01:00 AM CET	-NA-	⋮
JOB ID	COMPLETED AT	JOB DURATION	ASSETS PROFILED	
-NA-	-NA-		-NA-	<ul style="list-style-type: none"> Details Pause Profiler Delete Profiler

3. Click Confirm.

- You can click Resume Profiler to continue using it.

Profilers

Data Lake: dc

Activity Profiler	FREQUENCY (UTC)	NEXT RUN	TOTAL EXECUTIONS
	-NA-	-NA-	-NA-

JOB ID	COMPLETED AT	JOB DURATION	ASSETS PROFILED
-NA-	-NA-		-NA-

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

Enable or disable profilers in VM-based environments

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

Procedure

- Go to Profilers Configs .
- Select the profiler to proceed further.

Profilers / Configs

Jobs **Configs** Tag Rules

Profiler Configuration

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

- Switch the toggle to the desired state.

Profilers / Configs / Detail

Ranger Audit Profiler

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

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

Active

Schedule*

0 *30 *7**

Advance Options

Save Cancel

Profiling table data in non-default buckets

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

Procedure

- In Cloudera Data Catalog, make a note of your environment's name in the **Search** menu.
- Go Cloudera Management Console Environments

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

The screenshot shows the Cloudera Manager Configuration Search interface. The search term "Profiler Scheduler Spark conf" is entered in the search bar. The results show a configuration snippet for "Profiler Scheduler Spark conf" under the "profiler_scheduler" group. The snippet includes the following properties:

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

The interface also shows a sidebar with navigation options like Clusters, Hosts, Diagnostics, Charts, and Administration. The top navigation bar includes Status, All Health Issues, Configuration (26), and All Recent Commands. The date and time are shown as Sep 19, 1:37 PM UTC.

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

Profiler	FREQUENCY (UTC)	NEXT RUN	TOTAL EXECUTIONS								
Activity Profiler	at 00:00	03/15/2025 01:00 AM CET	9								
<table border="1"> <thead> <tr> <th>JOB ID</th> <th>COMPLETED AT</th> <th>JOB DURATION</th> </tr> </thead> <tbody> <tr> <td>KGVVOETJ-igt6</td> <td>03/14/2025 01:01 AM CET</td> <td>2 seconds</td> </tr> </tbody> </table>				JOB ID	COMPLETED AT	JOB DURATION	KGVVOETJ-igt6	03/14/2025 01:01 AM CET	2 seconds		
JOB ID	COMPLETED AT	JOB DURATION									
KGVVOETJ-igt6	03/14/2025 01:01 AM CET	2 seconds									
Data Compliance Profiler	at 01:01	03/15/2025 02:01 AM CET	5								
<table border="1"> <thead> <tr> <th>JOB ID</th> <th>COMPLETED AT</th> <th>JOB DURATION</th> <th>ASSETS PROFILED</th> </tr> </thead> <tbody> <tr> <td>DXNMSUSN</td> <td>03/14/2025 02:01 AM CET</td> <td>3 seconds</td> <td>-NA-</td> </tr> </tbody> </table>				JOB ID	COMPLETED AT	JOB DURATION	ASSETS PROFILED	DXNMSUSN	03/14/2025 02:01 AM CET	3 seconds	-NA-
JOB ID	COMPLETED AT	JOB DURATION	ASSETS PROFILED								
DXNMSUSN	03/14/2025 02:01 AM CET	3 seconds	-NA-								
Statistics Collector Profiler	at 00:00	03/15/2025 01:00 AM CET	24								
<table border="1"> <thead> <tr> <th>JOB ID</th> <th>COMPLETED AT</th> <th>JOB DURATION</th> <th>ASSETS PROFILED</th> </tr> </thead> <tbody> <tr> <td>FGUOEWTS</td> <td>03/14/2025 01:01 AM CET</td> <td>1 seconds</td> <td>-NA-</td> </tr> </tbody> </table>				JOB ID	COMPLETED AT	JOB DURATION	ASSETS PROFILED	FGUOEWTS	03/14/2025 01:01 AM CET	1 seconds	-NA-
JOB ID	COMPLETED AT	JOB DURATION	ASSETS PROFILED								
FGUOEWTS	03/14/2025 01:01 AM CET	1 seconds	-NA-								


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
- Modes²
- **Dry Run**³

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.

- **On Demand**
- **Scheduled**

Profilers Details

Data Compliance Disable Profiler

dc


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 | Clear All

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-SE8EFFH3	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-DNAFHZBE	Dry Run	02/26/2025 09:56 AM CET	02/26/2025 09:56 AM CET	-NA-
●	dryrun-ZZ3FJ9JK	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-N4BRIVUT	Dry Run	02/26/2025 09:18 AM CET	02/26/2025 09:18 AM CET	-NA-

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.

² Only available for the Data Compliance and Statistics Collector profilers.

³ Only available for the Data Compliance profiler.

The screenshot shows the 'Data Compliance' profiler details. The 'Job History' table lists several jobs with their statuses and start times. The 'Job Summary' panel on the right provides key metrics for the selected job.

Status	Job Id	Job Type	Started On
○	YPTT8MTM	Scheduled	02/26/2025 03:04 PM CET
●	VETBZKVB	On Demand	02/26/2025 12:25 PM CET
●	X9NZACUT	On Demand	02/26/2025 12:21 PM CET
●	dryrun-5E8EFH3	Dry Run	02/26/2025 10:10 AM CET
●	dryrun-ZUDTAMYU	Dry Run	02/26/2025 10:07 AM CET
●	dryrun-DNAFHZ8E	Dry Run	02/26/2025 09:56 AM CET
●	dryrun-ZZ3PJ9JK	Dry Run	02/26/2025 09:53 AM CET
●	dryrun-X2JU902V	Dry Run	02/26/2025 09:47 AM CET
●	dryrun-N4BRVUT	Dry Run	02/26/2025 09:18 AM CET

Job ID	STARTED ON	FINISHED ON	ASSETS PROFILED
YPTT8MTM	02/26/2025 03:04 PM CET	-NA-	32

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

The screenshot shows the 'Data Compliance Profiler' details. The 'Job History' table lists several jobs. The 'Job Summary' panel on the right shows a list of 'Profiled Assets' with a search filter and status indicators. A tooltip is visible over a skipped asset.

Status	Asset Name
●	airline.flight
●	airline.lounge
●	airline.lounge_classic
●	airline.lounge_premium
⊛	claim_provider_summary
⊛	cost_savings.claim_savings
⊛	default.new_data_tablesb2
⊛	default.new_data_table6ap
●	default.datagen_table_sensitive_326__1
⊛	default.new_data_tablezy8
●	finance.tax_2015
●	hortoniabank.eu_countries
●	hortoniabank.us_customers
●	hortoniabank.ww_customers



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

Tracking profiler jobs in VM-based environments

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

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

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

Figure 2: Profiling jobs in a VM-based environment

Profilers / Jobs

Jobs Configs Tag Rules

Filters [Clear All](#)

Job Status

Finished 65

Running 0

Failed 0

Profilers

Cluster Sensitivity Profiler 0

Ranger Audit Profiler 65

Hive Column Profiler 0

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

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

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

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

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

- Scheduler Service - The part of Profiler Admin which queues the profiler requests.
- Livy - This service is managed by YARN and is used to submit the Apache Spark jobs after which the actual asset profiling takes place.
- Metrics Service - Reads the profiled data files and publishes them.

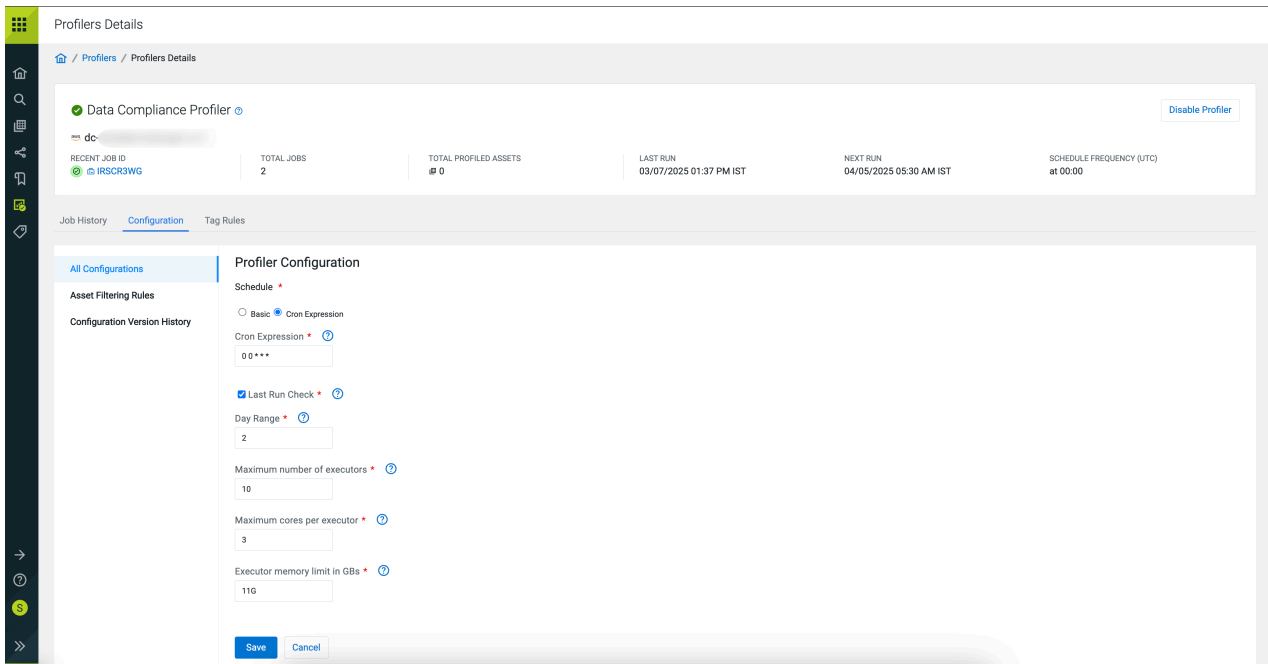


Note: More than one occurrence of Scheduler Service or Livy indicates that there could be more assets to be profiled. For example, if an HBase schedule has about 80 assets to be profiled, the first 50 assets would be profiled in the first Livy batch and the other assets get profiled in the next batch.

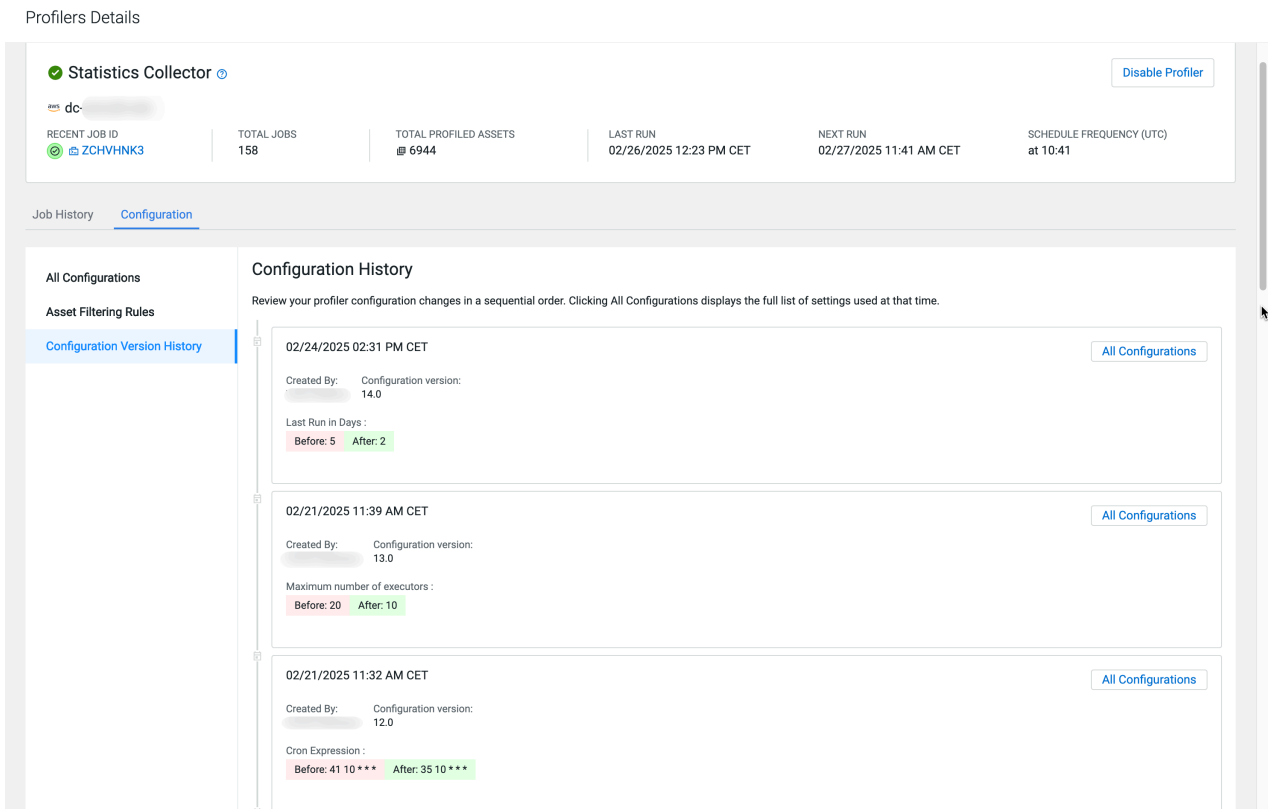
Viewing profiler configurations in Compute cluster enabled environments

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.



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



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

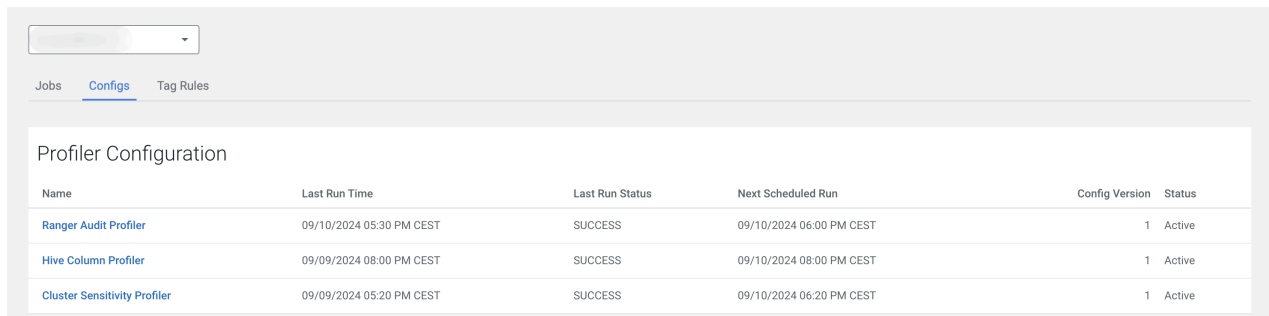
Executor memory limit in GBs :
11G

Close

Viewing profiler configurations in VM-based environments

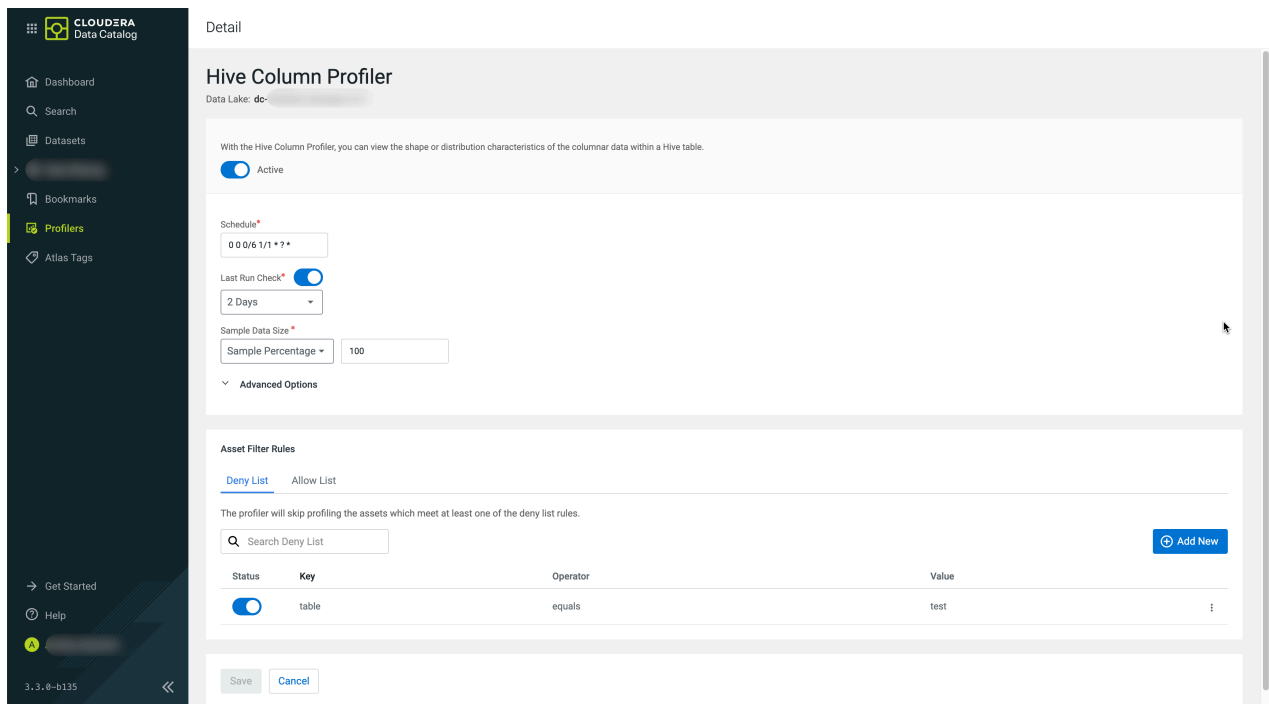
You can monitor the last status of individual profilers by viewing them in Profiler > Configs. Also, you can change their resources, sensitivity and scheduling.

Profilers / Configs



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

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



Detail

Hive Column Profiler

Data Lake: de- [redacted]

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

Active

Schedule*
0 0/6 1/1 * * *

Last Run Check*
2 Days

Sample Data Size*
Sample Percentage 100

Advanced Options

Asset Filter Rules

[Deny List](#) [Allow List](#)

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

Search Deny List [Add New](#)

Status	Key	Operator	Value
<input checked="" type="checkbox"/>	table	equals	test

[Save](#) [Cancel](#)

Monitoring the profiler configurations has the following uses:

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

Activity Profiler configuration

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.

Profiler Configuration

Schedule *

Basic Cron Expression

Cron Expression * [?](#)

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

Profiler Configuration

Schedule *

Basic Cron Expression

At minute of hours on st day of month on day of week

Time Zone:

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 * 

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

Ranger Audit Profiler configuration


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.

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*

^ **Advanced Options**

Number of Executors*
 ?

Executor Cores*
 ?

Executor Memory (in GB)*
 ?

Driver Core*
 ?

Driver Memory (in GB)*
 ?

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.

Data Compliance profiler configuration

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.

Profiler Configuration

Schedule *

Basic Cron Expression

Cron Expression * [?](#)

Profiler Configuration

Schedule *

Basic Cron Expression

At minute of hours on st day of month on day of week

Time Zone:

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



Note:

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

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

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

5. Continue with resource settings:
 - a) Set the Maximum number of executors

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

- b) Set the Maximum cores per executor

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

- c) Set the Executor memory limit in GBs

Maximum number of executors * 

Maximum cores per Executor * 

Executor memory limit in GBs * 



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

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



Note:

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

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

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


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

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

Key	Operator
Database name	<ul style="list-style-type: none"> • 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
Creation date	<ul style="list-style-type: none"> • greater than • less than

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



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

Cluster Sensitivity Profiler profiler configuration

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

Procedure

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

3. Select Cluster Sensitivity Profiler.

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

Detail

Cluster Sensitivity Profiler

Data Lake: dc-env1

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

Active

Schedule*

Last Run Check*

Sample Data Size*

^ Advanced Options

Number of Executors*
 ⓘ

Executor Cores*
 ⓘ

Executor Memory (in GB)*
 ⓘ

Driver Core*
 ⓘ

Driver Memory (in GB)*
 ⓘ

4.



Use the toggle button to enable or disable the profiler.

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

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

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



Note:

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

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

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

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

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

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

8. Continue with the resource settings.

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

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



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

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

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



Note:

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

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

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

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

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

Key	Operator
Database name	<ul style="list-style-type: none"> • 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
Creation date	<ul style="list-style-type: none"> • greater than • less than

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.

Statistics Collector profiler configuration

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

Profiler Configuration

Schedule *

Basic Cron Expression

Cron Expression * [?](#)

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

Profiler Configuration

Schedule *

Basic Cron Expression

At minute of hours on st day of month on day of week

Time Zone:

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



Note:

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

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

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

4. Continue with resource settings:

a) Set the Maximum number of executors

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

b) Set the Maximum cores per executor

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

c) Set the Executor memory limit in GBs

Maximum number of executors * 

Maximum cores per Executor * 

Executor memory limit in GBs * 

Save



Cancel

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

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

**Note:**

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

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

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


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

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

Key	Operator
Database name	<ul style="list-style-type: none"> • 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
Creation date	<ul style="list-style-type: none"> • greater than • less than

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



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

Hive Column Profiler configuration

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*

Last Run Check*

Sample Data Size *

^ **Advanced Options**

Number of Executors* ?

Executor Cores* ?

Executor Memory (in GB)* ?

Driver Core* ?

Driver Memory (in GB)* ?

- 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:

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

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

8. Continue with the resource settings.

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

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



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

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

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



Note:

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

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

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

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

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

Key	Operator
Database name	<ul style="list-style-type: none"> • equals • starts with • ends with

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

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.

Backing up and restoring the profiler database

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



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

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

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

- profiler_agent
- profiler_metrics



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

About the back up script

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

Scenarios for using the script

- When you upgrade the data lake cluster and want to preserve profiler data in the Cloudera Data Hub cluster.
- When you want to delete the Cloudera Data Hub cluster but preserve the profiler data.
- When you want to relaunch the profiler and access the older processed data.



Note: For users using Cloudera Data Catalog on Cloudera Runtime 7.2.14 version, note the following:

- No user action or manual intervention needed after the upgrading Cloudera Data Hub cluster to the 7.2.14 version.
- Also, as an example use case scenario, in case a new profiler cluster is launched that contains Custom Sensitivity Profiler tags and which is deleted and relaunched later, the changes are retained and no further action is required.
- No user action is required to backup and restore the profiler data. The changes are automatically restored.

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

Go to the following locations to pick up your scripts:

Back up

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

Restore

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

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

Go to the following locations to pick up your scripts:

Back up

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

Restore

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

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

Navigate to the following location to pick up your scrips:

Back up

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

Restore

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

Running the back up script

Running the profiler Backup and Restore script has multiple phases.

About this task

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

Backing up the profiler database

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



Attention: Users of Cloudera Runtime below 7.2.8 version should contact [Cloudera Support](#).



Note:

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

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

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

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

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

Restoring the profiler database

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



Attention: Users of Cloudera Runtime below 7.2.8 version should contact [Cloudera Support](#).



Note:

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

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

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

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

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



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

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

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.



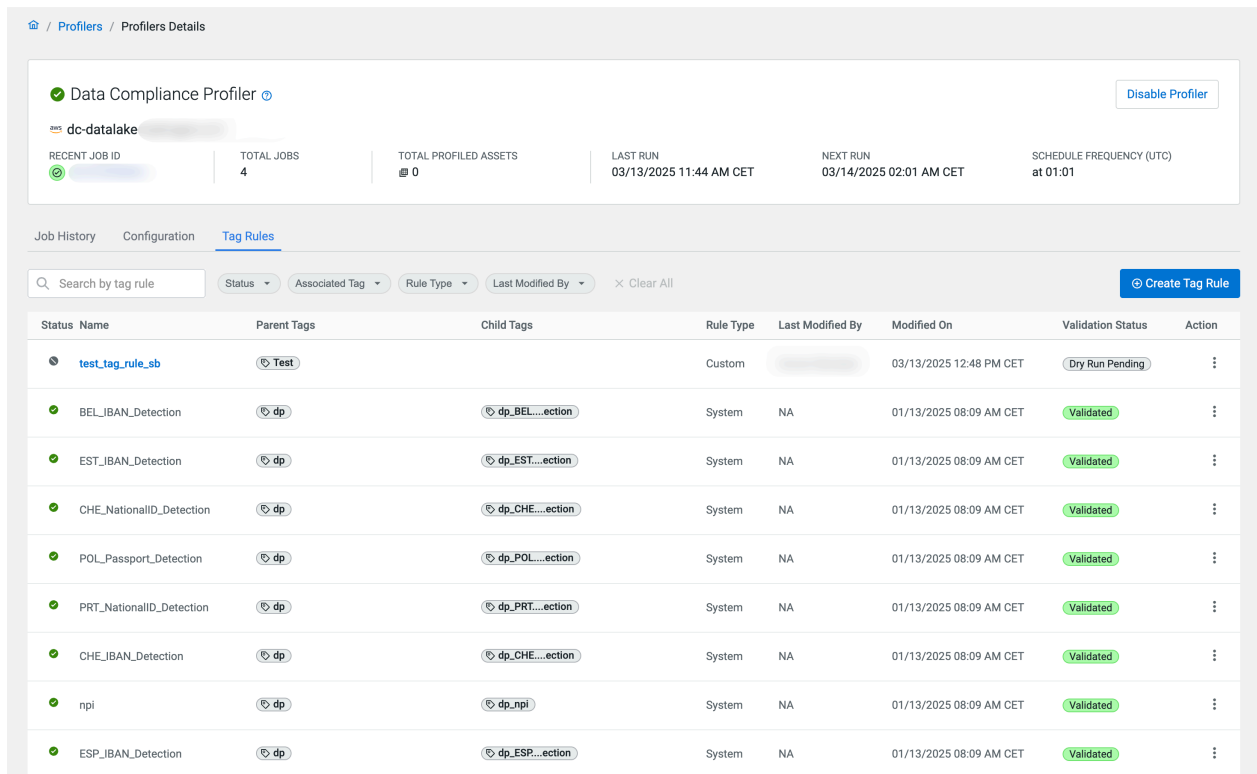
Note:

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



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.

Profiler tag rules in VM-based environments


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

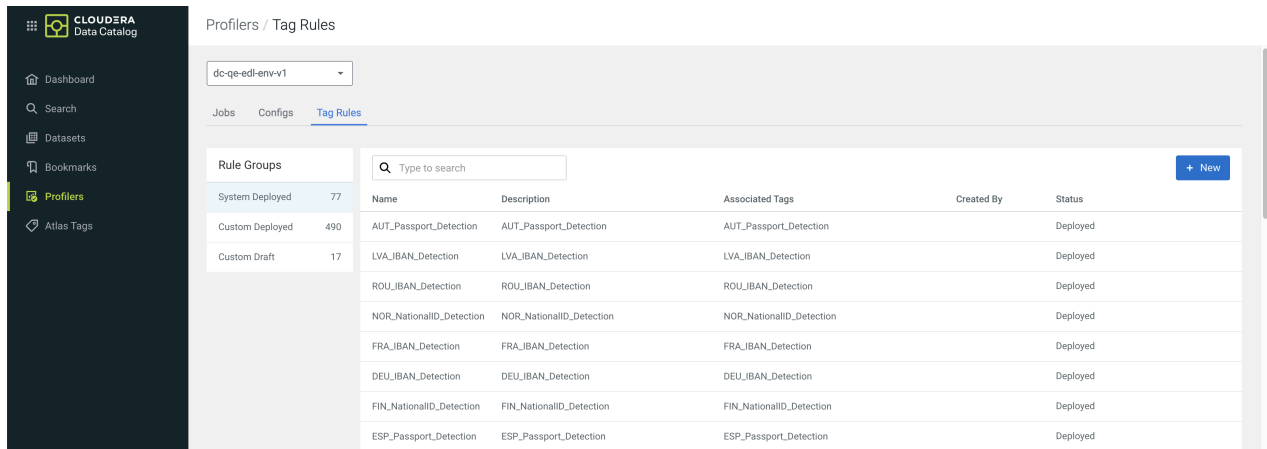


Note: The improved tag rules are available for Compute Cluster enabled environments. In VM-based environments, tag rules are valid for all data lakes, while tag rules in Compute Cluster enabled environments are data lake specific.

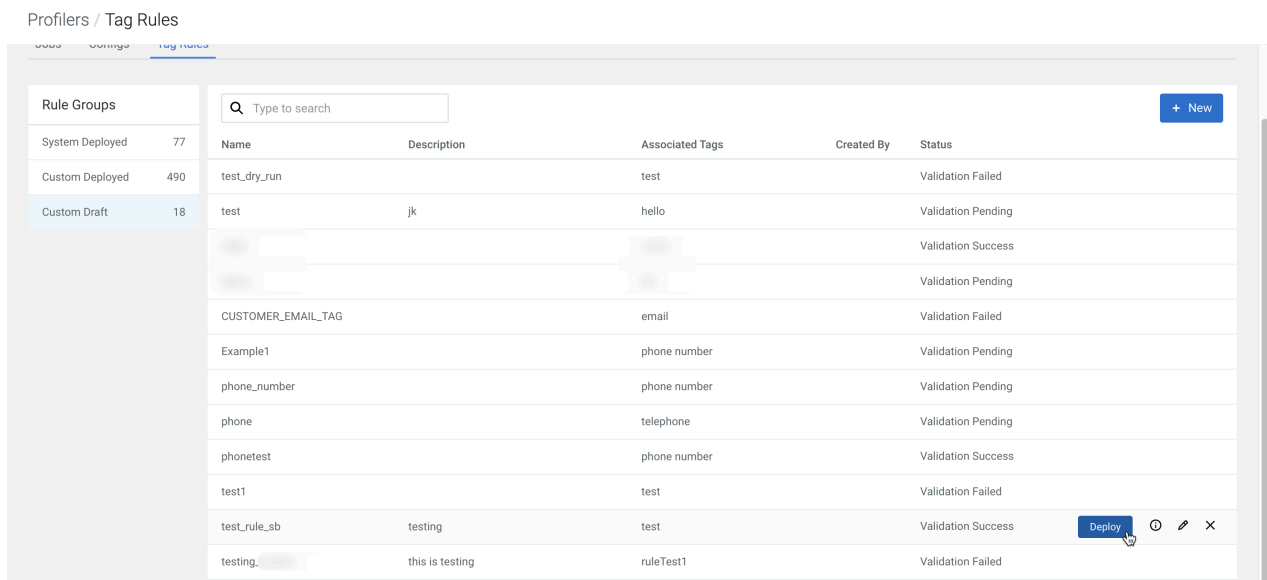
Tag rule types

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

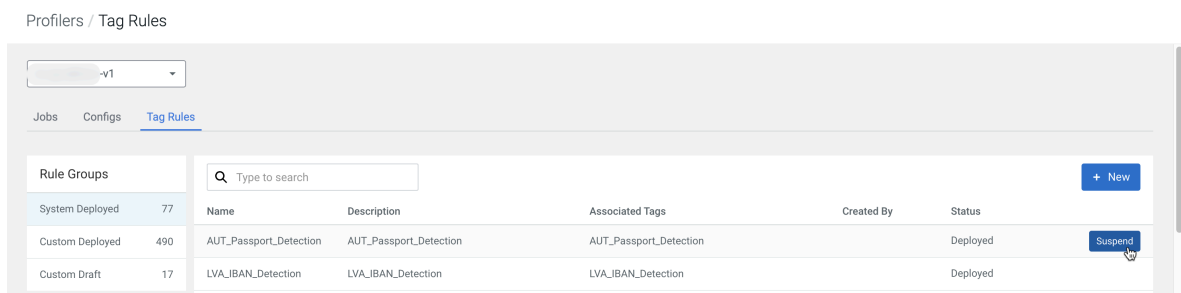
- **System Deployed:** These are built-in rules that cannot be edited. You can only enable or disable them for your data.
- **Custom Deployed:** Tag rules that you create, edit and deploy on clusters after validation will appear under this category. Click the  icon in the **Action** column to enable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



After creating your rule, you have to validate them. 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

Input type	VM based environments	Compute Cluster enabled environments
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

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

Tag rule testing

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

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

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:

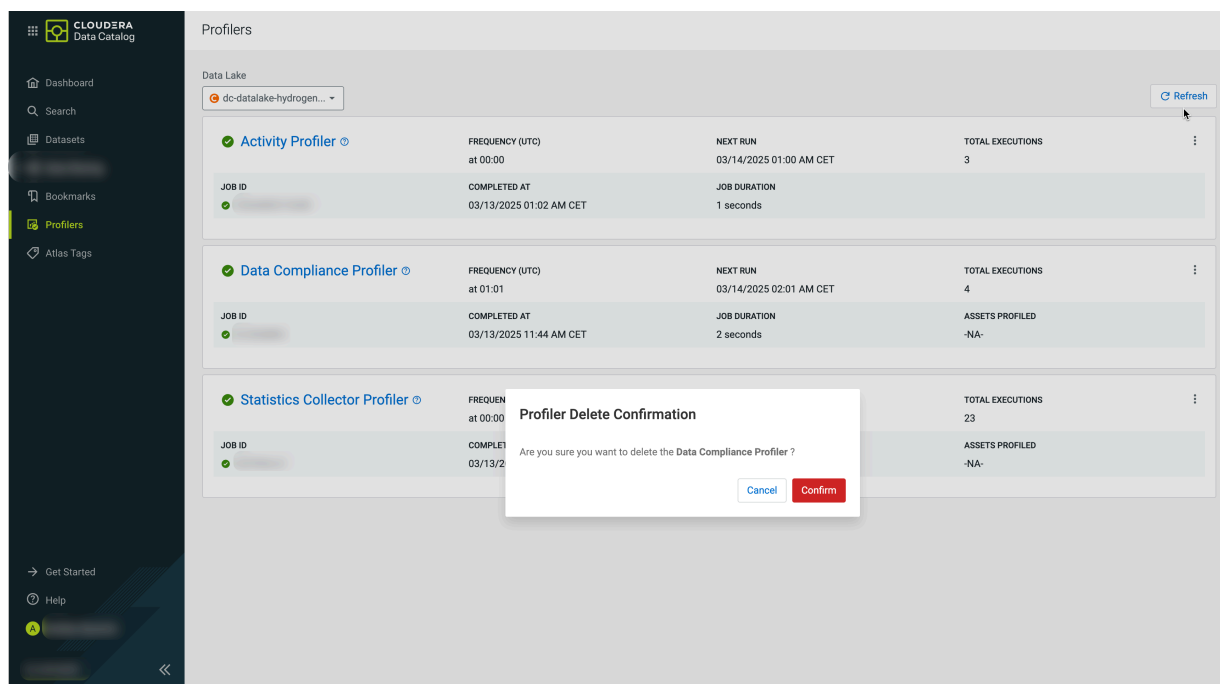
- In a Compute Cluster enabled environment, when you delete the scheduled jobs, the associated Kubernetes cron job object is deleted from the Kubernetes cluster.
- The associated data of the profilers from the Cloudera Management Console database is also deleted for the specified data lake.

Procedure

- On the **Profilers** page, select the data lake from the drop-down.
- Click Delete Profiler in the action menu () for the profiler you want to delete.

- Confirm the deletion in the message dialog box.

Figure 3: Deleting a profiler in a Compute Cluster enabled environment



- 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.

Deleting profilers in VM-based environments

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

About this task

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

- If your profiler cluster or profiler jobs have rules that are not changed or updated, you can directly delete them or the profiler cluster.
- If the rules were modified or updated, you have an option to download the modified rules along with deletion. The modified rules consist of the suspended system rules and the deployed custom rules. Using the downloaded rules, you can manually add or modify them to your newly added profiler jobs or the profiler cluster.

Procedure


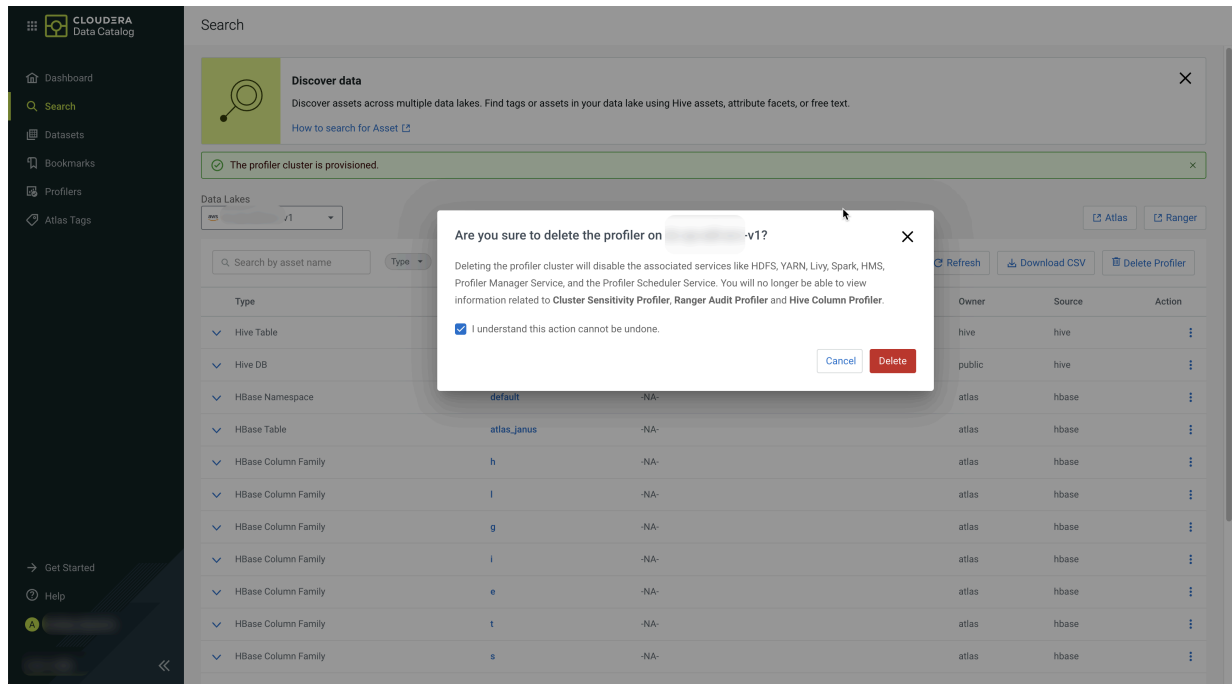
1. On the **Profilers** page, select the data lake from the drop-down.
2. Click Delete Profiler in the action menu () for the profiler you want to delete.
3. If you agree, select the warning message I understand this action cannot be undone.

Figure 4: Deleting a profiler in a VM-based environment

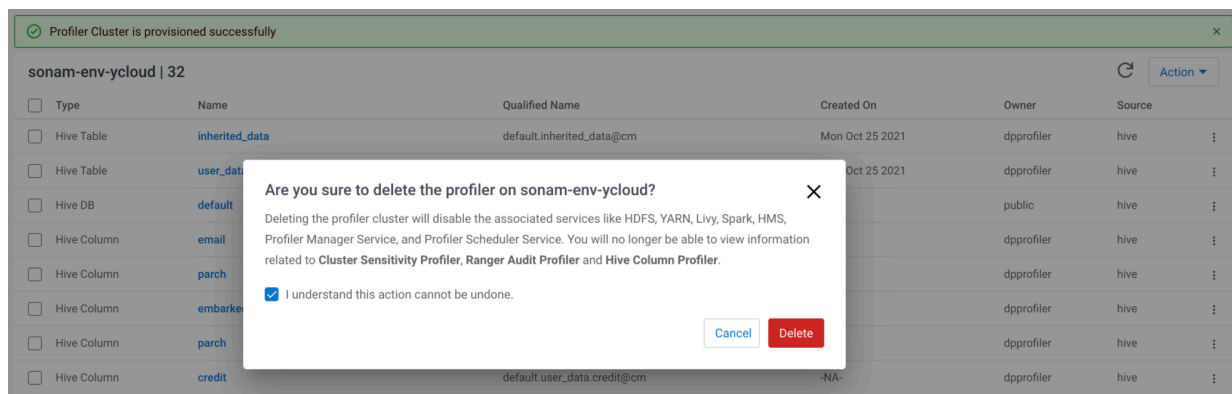


4. Click Delete.

The application displays the following message.



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



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



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

The profiler cluster is deleted successfully.

Atlas tag management

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

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

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

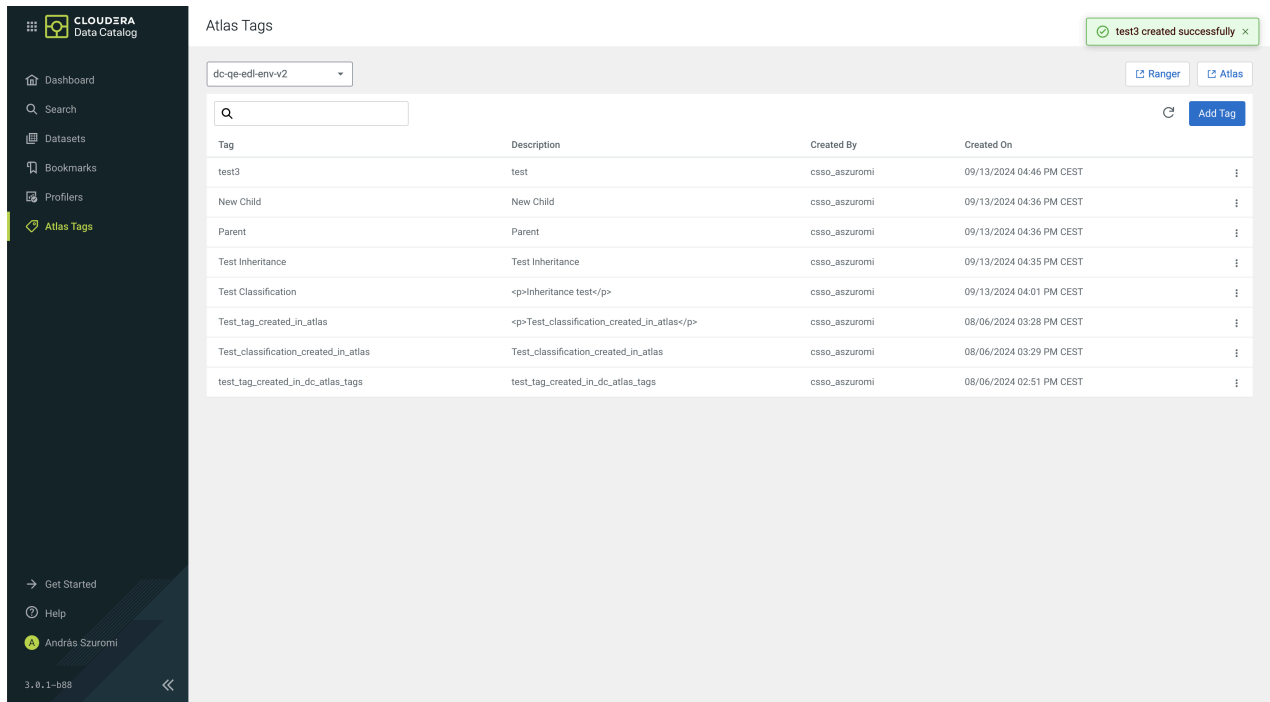
You can create a new Cloudera Data Catalog tag in the **Atlas Tags**, which are synced to Atlas. Click **Add Tag** to open the **Create a new tag** page.

Tag	Description	Created By	Created On
Test_classification_created_in_atlas	Test_classification_created_in_atlas	csso_aszuromi	08/06/2024 03:29 PM CEST
Test_tag_created_in_atlas	<p>Test_classification_created_in_atlas</p>	csso_aszuromi	08/06/2024 03:28 PM CEST
test_tag_created_in_dc_atlas_tags	test_tag_created_in_dc_atlas_tags	csso_aszuromi	08/06/2024 02:51 PM CEST

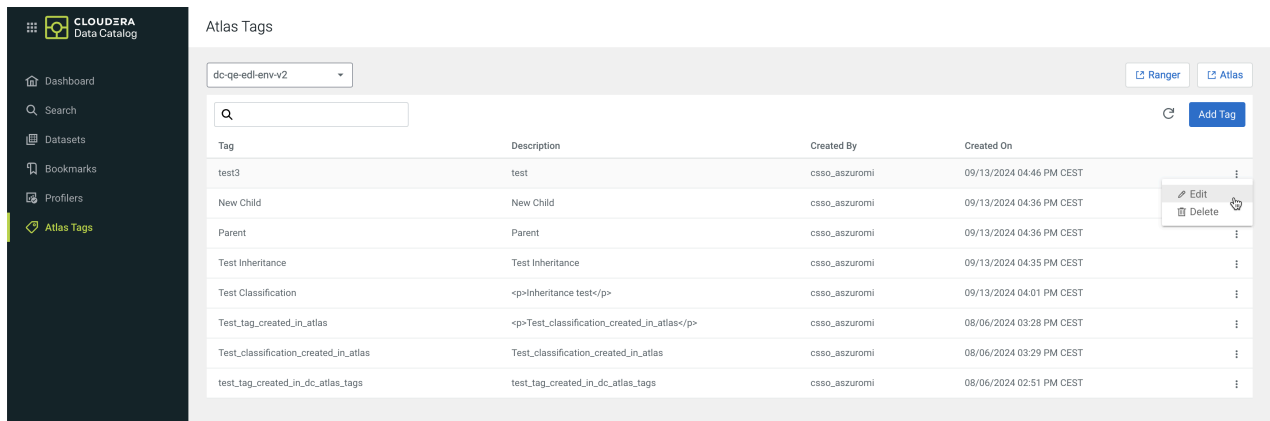
In **Create New Tag**, you can define the tag name, description and the "super-classification" from which the attributes are inherited for the sub-classification (or tag in Cloudera Data Catalog)

Tag	Description	Created By	Created On
test_tag_created_in_dc_atlas_tags	test_tag_created_in_dc_atlas_tags	csso_aszuromi	08/06/2024 02:51 PM CEST
Test_tag_created_in_atlas	<p>Test_classification_created_in_atlas</p>	csso_aszuromi	08/06/2024 03:28 PM CEST
Test_classification_created_in_atlas	Test_classification_created_in_atlas	csso_aszuromi	08/06/2024 03:29 PM CEST
Test Inheritance	Test Inheritance	csso_aszuromi	09/06/2024 03:29 PM CEST
Test Classification	<p>inheritance test</p>	csso_aszuromi	09/06/2024 03:29 PM CEST

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



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



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

