

Cloudera Data Catalog Top Use Cases

Date published: 2019-11-14

Date modified: 2024-11-12

CLUSTERA

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.....	5
Searching for assets using Atlas glossaries.....	6
Using terms in Cloudera Data Catalog.....	7
Mapping glossary terms.....	7
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.....	14
Additional entity type selection for searching assets.....	16
 Managing Profilers.....	 17
Profiler data testing.....	18
Launching profilers.....	18
Launching profilers using the command-line.....	22
Deleting profilers.....	24
On-Demand Profilers.....	27
Profiling table data in non-default buckets.....	28
Tracking profiler jobs.....	29
Viewing profiler configurations.....	31
Ranger Audit Profiler configuration.....	31
Cluster Sensitivity Profiler configuration.....	33
Hive Column Profiler configuration.....	35
Understanding the Cron Expression generator.....	37
Backing up and restoring the profiler database.....	38
About the back up script.....	38
Running the back up script.....	39
Enable or disable profilers.....	40
Profiler Tag Rules.....	41
 Tag management.....	 42

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
- Iceberg Table
- 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.

Type	Owner	Entity Tag	Glossary Terms	Refresh	Download CSV	Delete Profiler
▼	HBase Namespace	default	-NA-	atlas	hbase	⋮
▼	HBase Table	atlas_janus	-NA-	atlas	hbase	⋮
▼	HBase Column Family	h	-NA-	atlas	hbase	⋮
▼	HBase Column Family	l	-NA-	atlas	hbase	⋮
▼	HBase Column Family	g	-NA-	atlas	hbase	⋮
▼	HBase Column Family	i	-NA-	atlas	hbase	⋮
▼	HBase Column Family	e	-NA-	atlas	hbase	⋮
▼	HBase Column Family	t	-NA-	atlas	hbase	⋮
▼	HBase Column Family	s	-NA-	atlas	hbase	⋮
▼	HBase Column Family	f	-NA-	atlas	hbase	⋮
▼	HBase Column Family	m	-NA-	atlas	hbase	⋮

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.

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.

Asset Details

default [Atlas](#)

Properties

Type: HBASE NAMESPACE
 Data Lake: dc
 Owner: atlas
 Created On: -NA-
 Update Time: -NA-
 Created By: atlas
 Updated By: casso,
 Status: ACTIVE

Qualified Name: default@cm
 Description: default

Classifications | 2 Managed System Propagated

[test_tag_created_in_dc_atlas...](#) [Test_tag_created_in...](#)

Terms [+ Add Terms](#)

Content Metadata Audits Policy Access Audits

Type	Name	Location	Created On	Owner	Source
hbase_table	atlas_janus	/default	-NA-	atlas	hbase

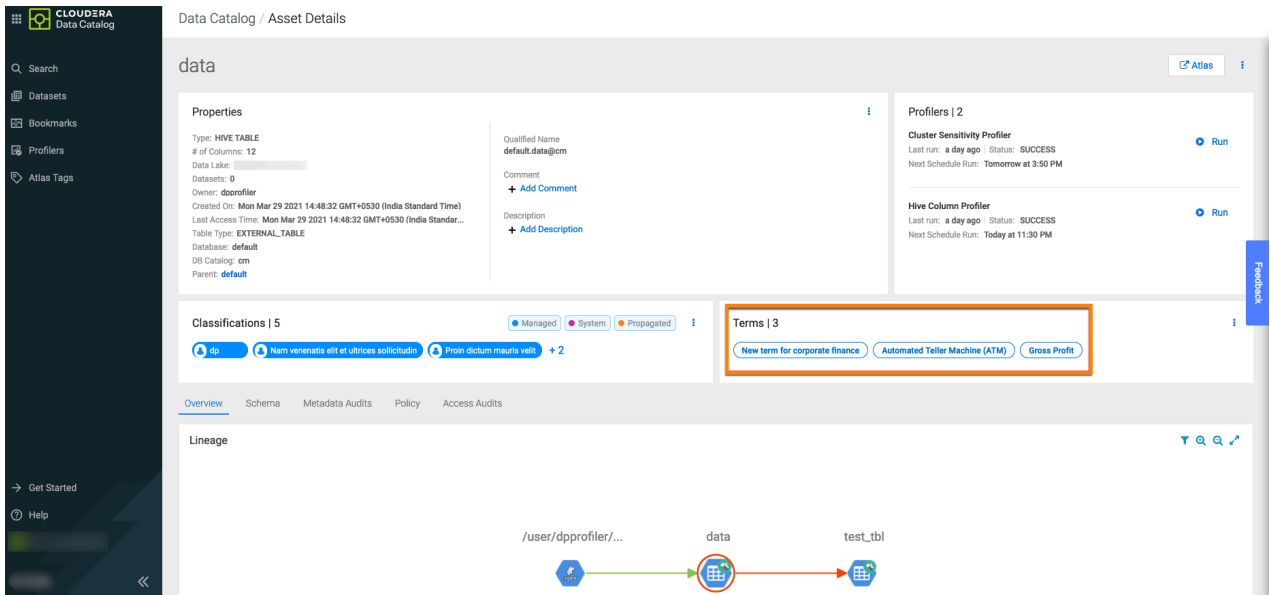
Mapping glossary terms

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

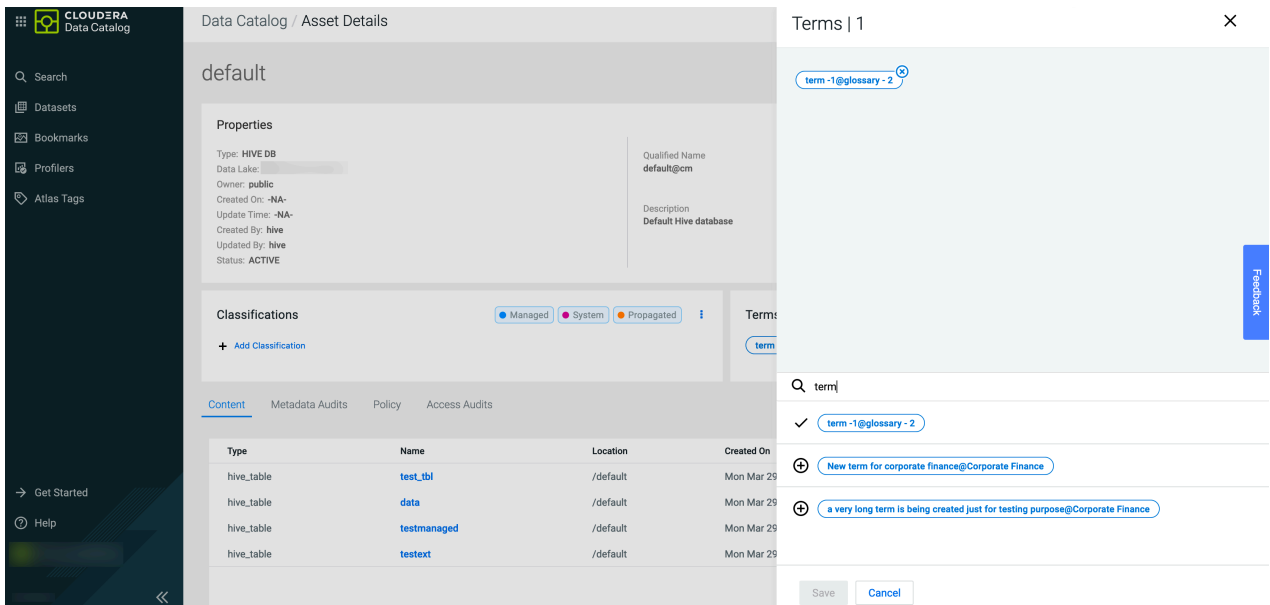
You can search for those terms in Cloudera Data Catalog and map specific terms with assets. You can also search for terms to delete them from the selected asset. The selected asset displays the total number of terms associated or mapped accordingly.

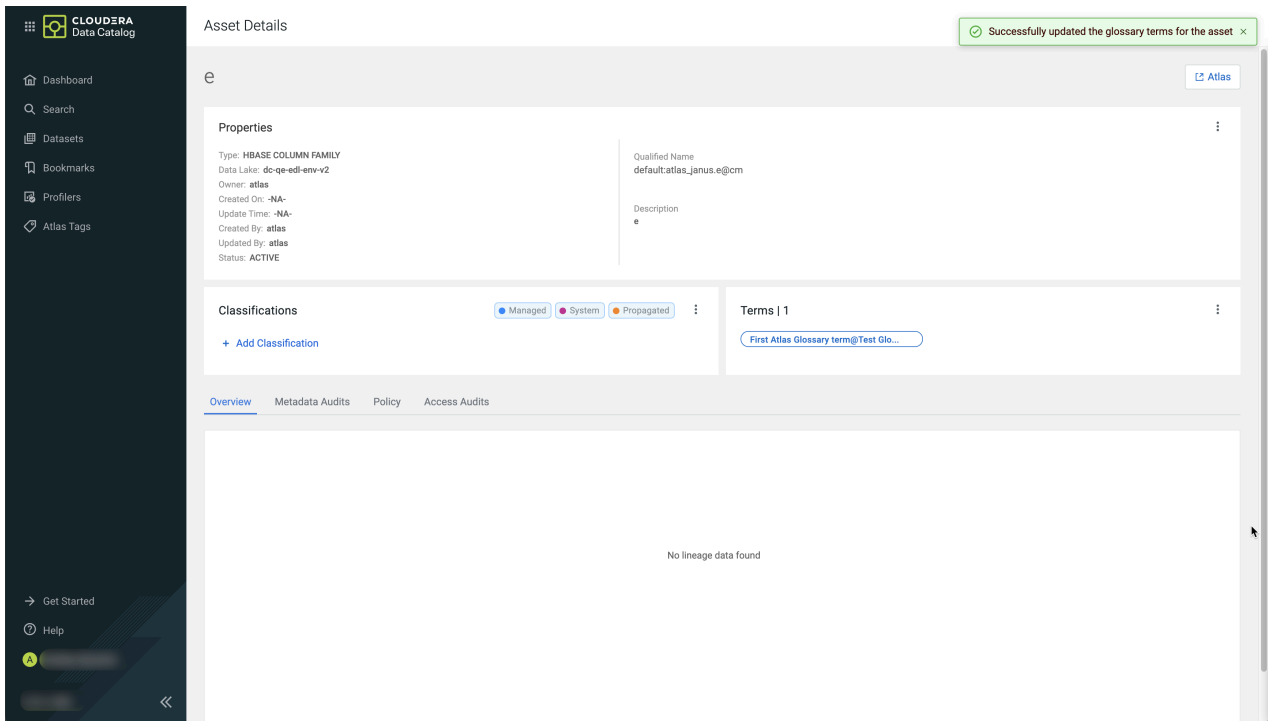
When you map a specific term for your dataset, the term is displayed in the following format:

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

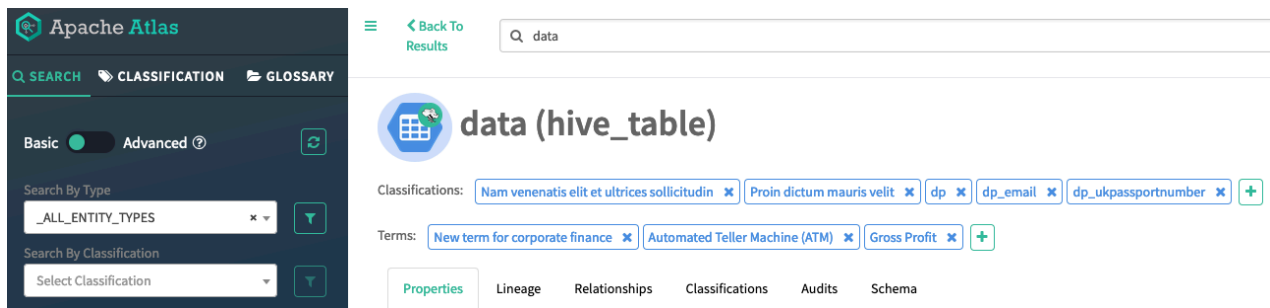


You can use the icon in the **Terms** widget on the **Asset Details** page to add new terms for your assets. Click Save to save the changes.





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.

Overview **Schema** Metadata Audits Policy Access Audits

Search Column Edit

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						

Rows per page: 20 1 - 12 of 12

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

CloudERA Data Catalog

Data Catalog / Asset Details

Overview **Schema** Metadata Audits Policy Access Audits

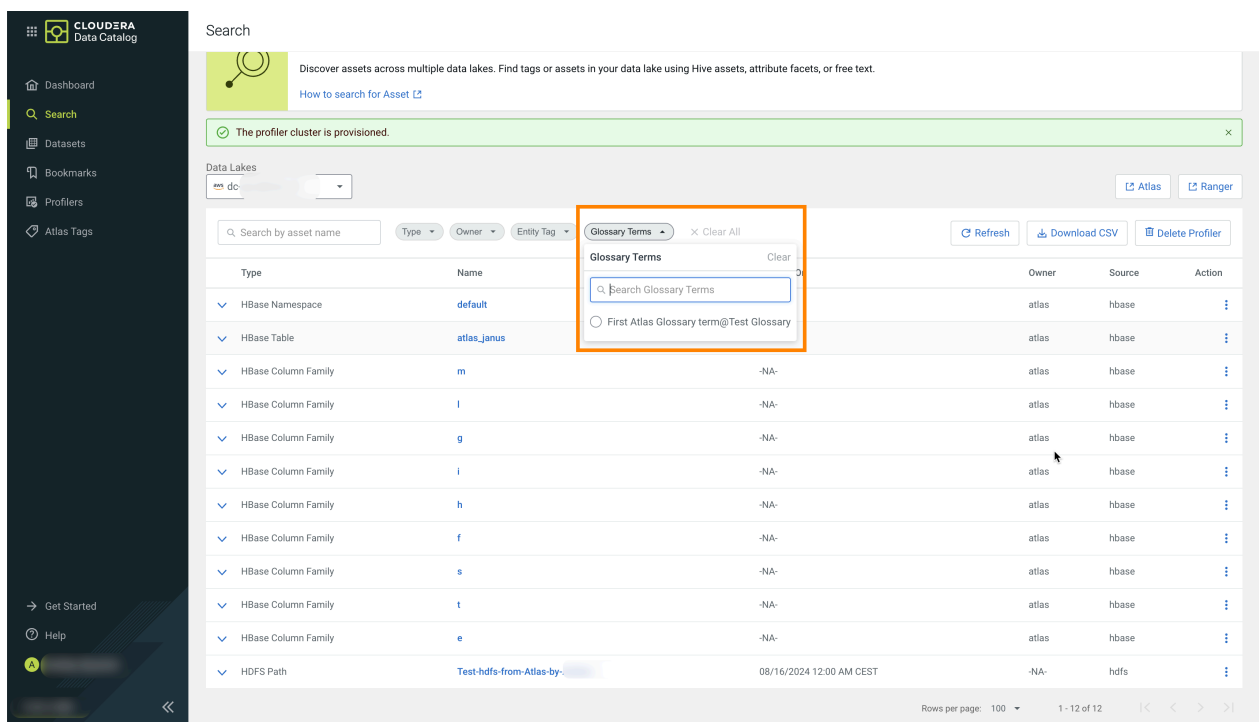
Search Column Edit

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	a very long term is b... +7
▼	cabin	string	9	0					Nam venenatis elit et. +1	a very long term is b... +7
▼	embarked	string	3	0					dp_ukpassportnumbe +2	a very long term is b... +7
▼	fare	float	35	0	262.38		23.78		dp_ukpassportnumbe +1	a very long term is b... +7
▼	name	string	54	0					dp_ukpassportnumbe +6	a very long term is b... +7
▼	parch	int	3	0	2		0.42		dp_ukpassportnumbe +1	a very long term is b... +7
▼	passengerid	int	50	0	53	1	27		dp_ukpassportnumbe +2	a very long term is b... +7
▼	pclass	int	3	0	3	1	2.42		dp_ukpassportnumbe	a very long term is b... +7
▼	sex	string	2	0					dp_ukpassportnumbe	a very long term is b... +7
▼	sibsp	int	4	0	8		0.43			a very long term is b... +7
▼	survived	int	2	0	1		0.72			a very long term is b... +7
▼	ticket	string	48	0						a very long term is b... +7

Rows per page: 20 1 - 12 of 12

Searching for assets using glossary terms

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



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.

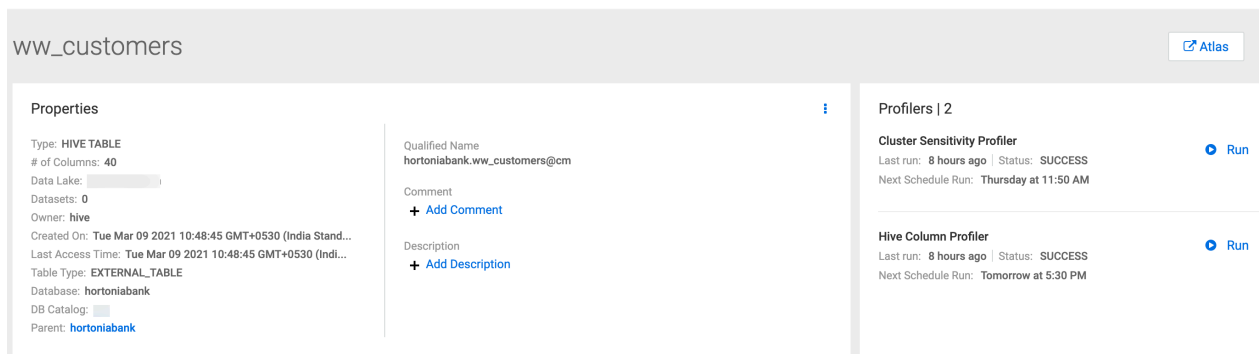
In the **Asset Details** page for each asset type that you select, you can add or edit **comment** or **description** fields. Including these values for the selected asset helps you to identify your chosen asset.

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

Description

Profilers | 2

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

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

Cancel Save

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

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



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 Toggle

columns (40) Dropdown
`title`
`givenname`
`middleinitial`

comment passport_number

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

db Dropdown
hortoniabank

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

description visa_number

User-defined properties Add

Labels Add

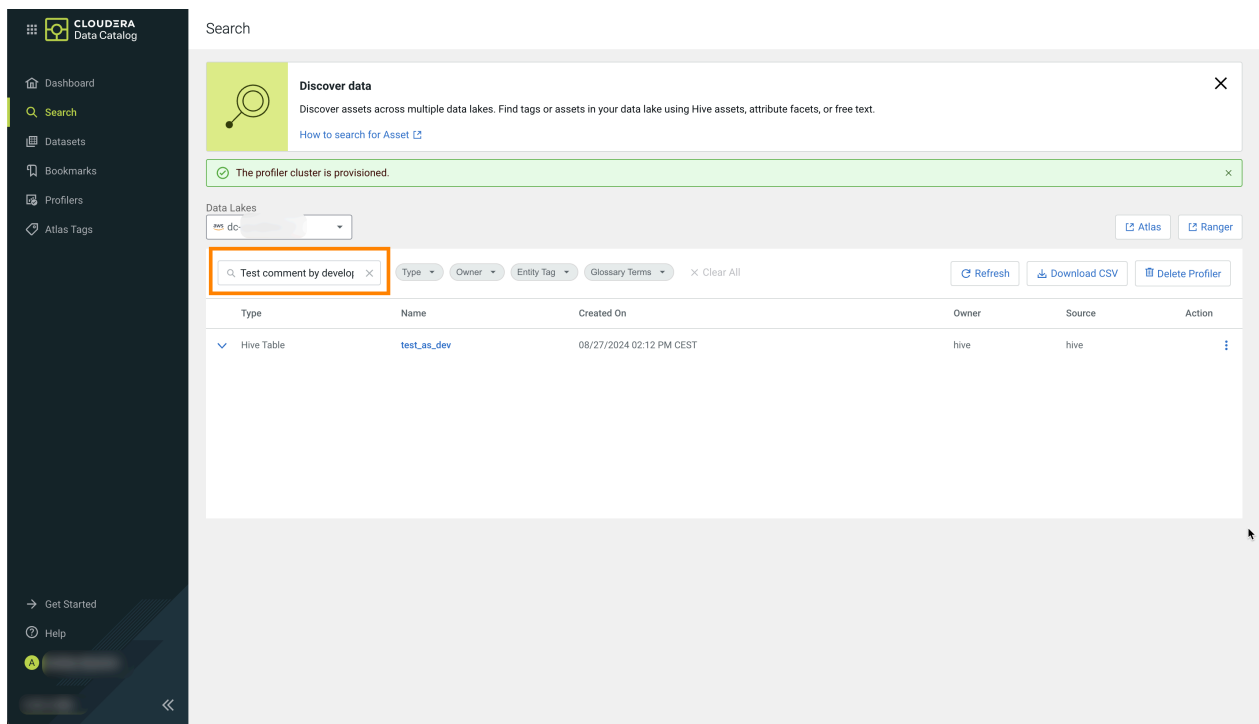
Business Metadata Add

[Switch to Beta UI](#)

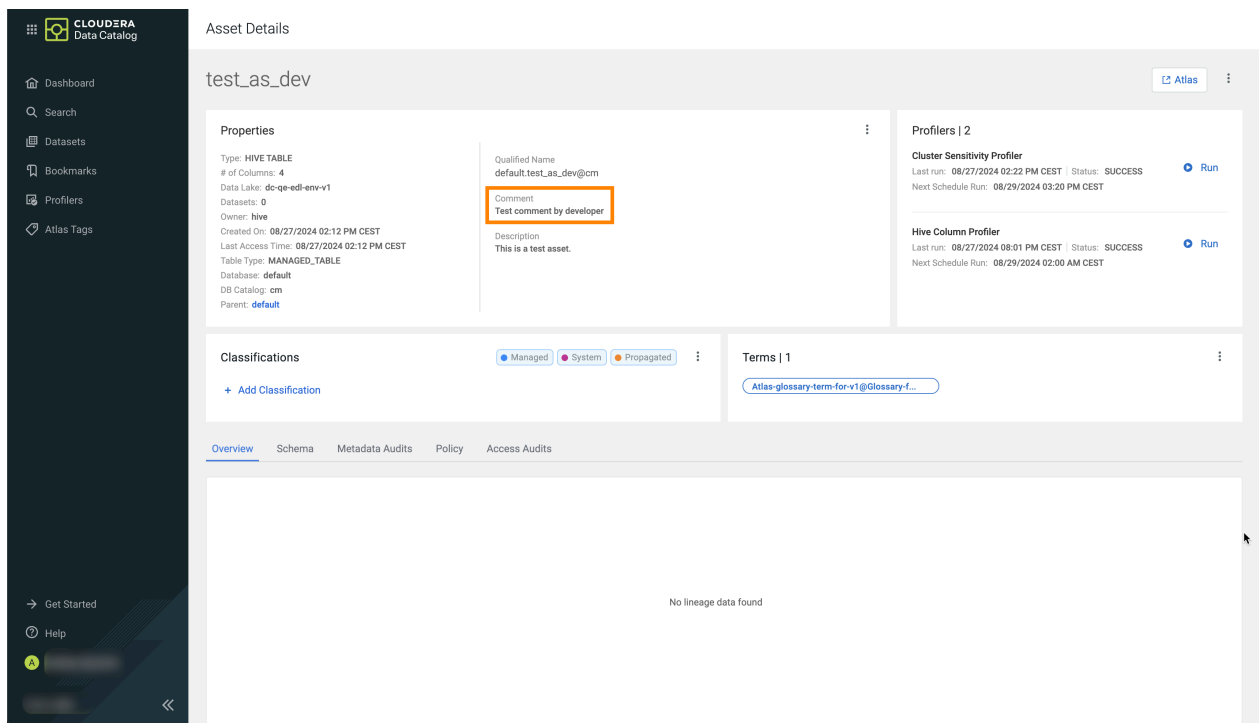
Searching for assets using additional search options

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.

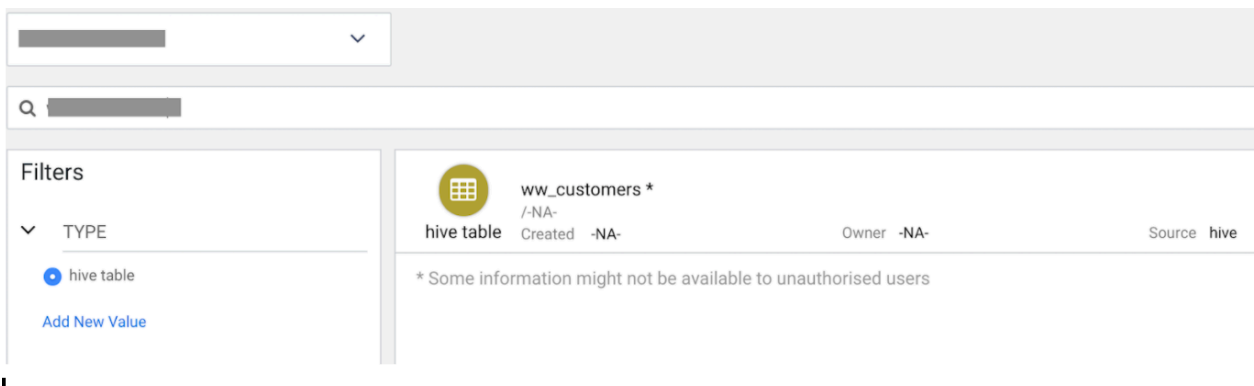


Accessing tables based on Ranger policies

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.

The screenshot shows a list of tables in a data catalog. A 'CREATED BEFORE' filter is active on the left. The table list includes:


Type	Path	Created	Owner	Source
Hive Table	/sys	Tue Apr 07 2020	Owner	hive
Hive Table	/information_schema	Tue Apr 07 2020	Owner	hive
Hive Table	/information_schema	Tue Apr 07 2020	Owner	hive
Hive Table	/sys	Tue Apr 07 2020	Owner	hive
Hive Table	/information_schema	Tue Apr 07 2020	Owner	hive
Hive Table	/	-	Owner	hive
Hive Table	/-	-	Owner	hive
Hive Table	/-	-	Owner	hive
Hive Table	/-	-	Owner	hive
Hive Table	/-	-	Owner	hive

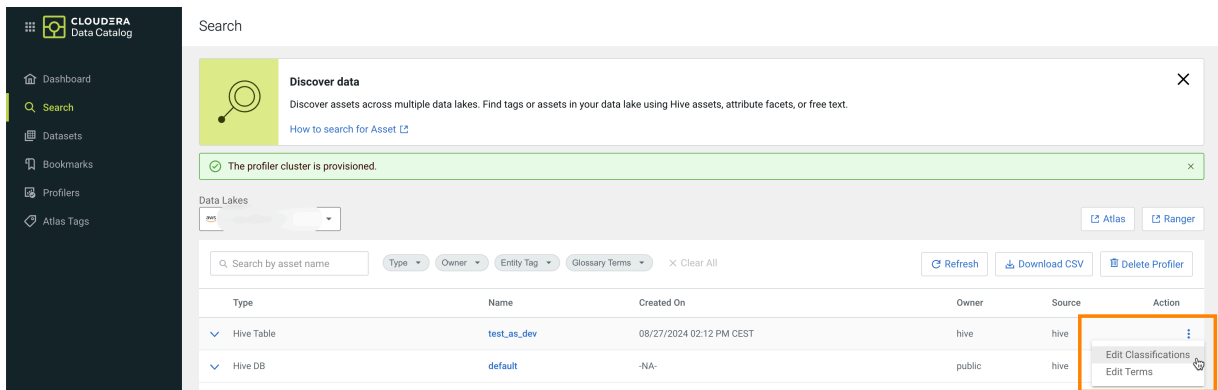
Creating classifications for selected assets

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.

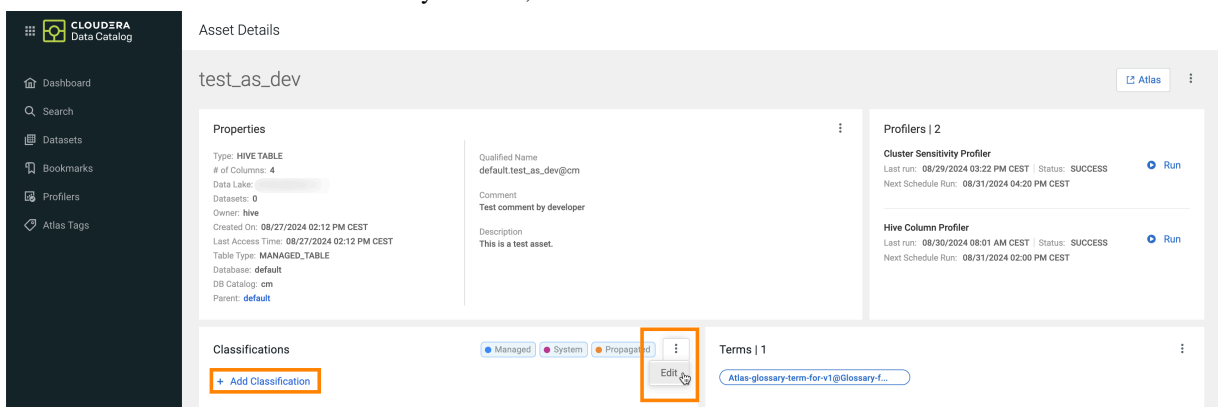
2. Click the  icon by an asset, then select Edit Classifications.



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

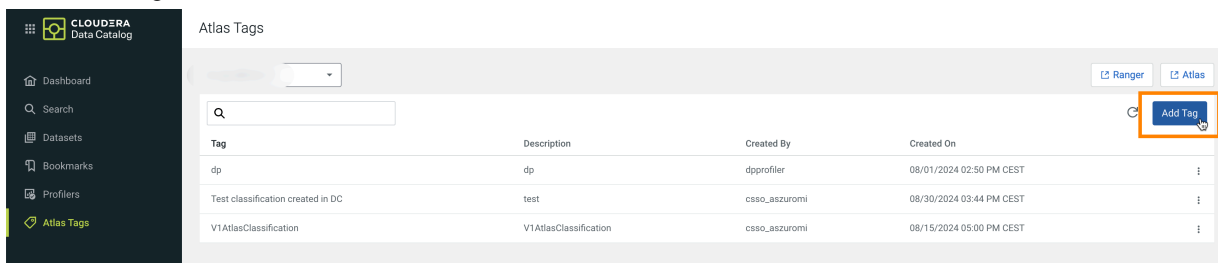
1. Navigate to the **Asset Details** page of an asset.
2. Click Add Classification or  icon by an asset, then select Edit.



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.

Additional entity type selection for searching assets

Using the Cloudera Data Catalog service, you can search for assets by using entity types.

Cloudera Data Catalog users can search and discover different asset types.

Supported entity types include the following:

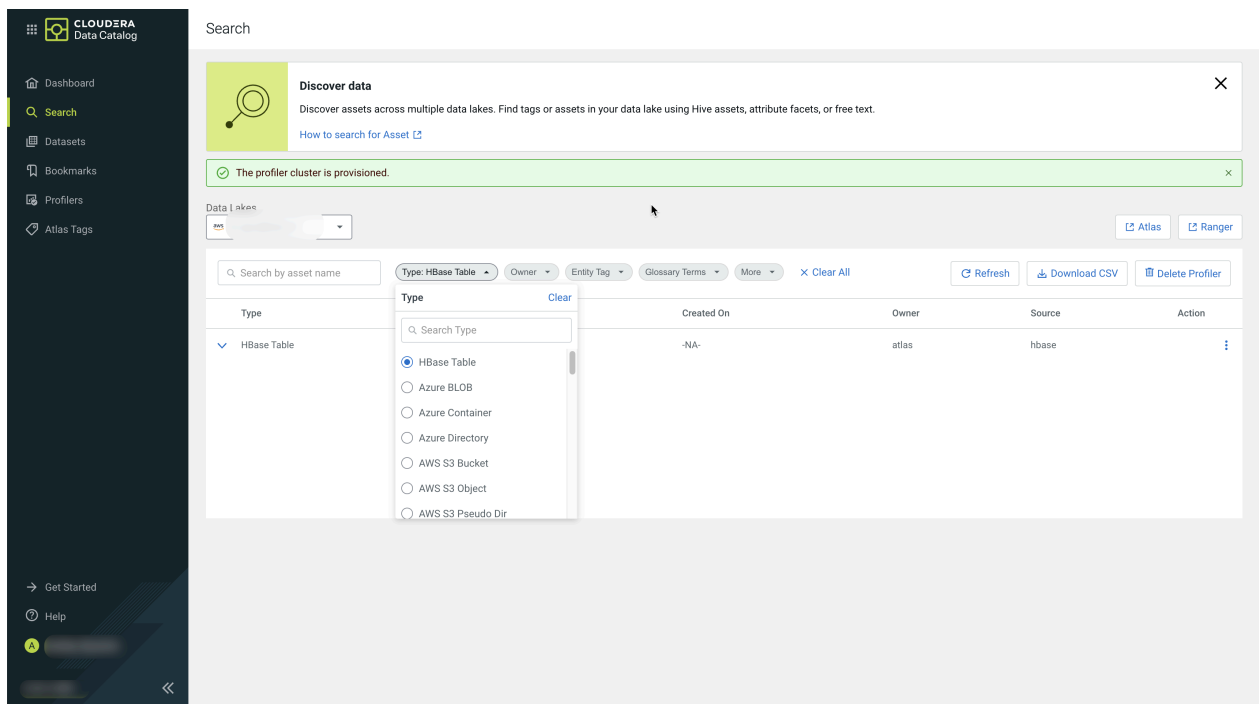
- Azure BLOB
- Azure Container
- Azure Directory
- AWS S3 Object
- AWS S3 V2 Object
- AWS S3 Bucket
- AWS S3 V2 Bucket
- AWS S3 Pseudo Dir
- AWS S3 V2 Directory
- HBase Table
- HBase Column Family
- HBase Namespace
- HDFS Path
- Hive DB
- Hive Table
- Hive Column
- ML Project
- ML Model Build
- ML Model Deployment
- NiFi Flow
- NiFi Data
- Iceberg Column¹
- Iceberg Table¹
- Impala Process
- Impala Column Lineage
- Impala Process Execution
- Kafka Topic
- RDBMS DB
- RDBMS Column
- RDBMS Foreign Key
- RDBMS Index
- RDBMS Instance
- RDBMS Table
- Spark Process
- Spark Application
- Spark Column
- Spark Column Lineage
- Spark DB

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

- Spark ML Directory
- Spark ML Model
- Spark ML Pipeline
- Spark Process Execution
- Spark Table

Selecting a type triggers a search query for that type. The **Owner** of the asset is derived from the response received from the type based queries.

The following examples depict the entity type selection search results:



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	Description
Cluster Sensitivity Profiler	A sensitive data profiler- PII, PCI, HIPAA, etc.
Ranger Audit Profiler	A Ranger audit log summarizer.
Hive Column Profiler	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:
 - Hive Column Profilers and Cluster Sensitivity Profilers
 - CSV
 - Parquet
 - Iceberg tables

Related Information

[Understanding the Cloudera Data Catalog Profiler](#)

[Understanding the Cluster Sensitivity Profiler](#)

[Understanding the Ranger Audit Profiler](#)

Profiler data testing

You must note the important information about profiler services.



Note: The Cloudera Data Catalog profilers are not tested at par with the Hive scale.

The following dataset has been validated and works as expected for VM-based environments:

- DataHub Master: m5.4xlarge
- Hive tables: 3000 Hive assets
- Total Number of assets (including Hive columns, tables, databases): 1,000,000
- Total Data Size = 1 GB
- Partitions on Hive tables: Around 5000 partitions spread across five tables



Note: For Compute Cluster enabled environments, more detailed testing information will be provided in the next release of Cloudera Data Catalog.

The following dataset has been validated and works as expected for Compute Cluster enabled environments:

- Total Data Size = 300 GB
- Sampling profiler size = 50% (150 GB)

Launching profilers

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 and datasets. 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.

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



Note: This is applicable from the following build of Cloudera Data Catalog: 2.0.17: 2.0.17-b26.

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.



Note: The HA functionality is being supported only from Cloudera Runtime 7.2.10 release onwards. If you are using a Cloudera Runtime version below 7.2.10, you are not able to use the HA feature when launching the profiler services.

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						
[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	

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

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

How to launch the profiler for Compute Cluster enabled environments

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

Search

Discover data

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

How to search for Asset

Set Up the Profiler for v2

Profiler runs profiling operations on assets' data located in the data lake. Setting up profilers results in new cron jobs in Kubernetes which require an additional 12 cores and 24 GB RAM is required to run them efficiently. Get Started

Profiler Setup v2

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.

Setup Profiler

Type	Owner	Source	Action
HBase Namespace	atlas	hbase	
HBase Table	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	
HBase Column Family	atlas	hbase	

Click Setup Profiler, Cloudera Data Catalog processes the request and the profiler creation will start.

Next, you can verify that the profiler jobs are running under the Cloudera Management Console Environments Compute Clusters Node Groups pane.

Environments / v2 / Compute Clusters

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

Stop
Actions

sdx Data Lake Details

NAME
-v2

STATUS
Running

NODES
2 0 0

STATUS REASON
N/A

SCALE
Light Duty

CRN
j792...

QUICK LINKS
Atlas Ranger DataCatalog

Data Hubs
Data Lake
FreeIPA
Compute Clusters
Cluster Definitions
Summary

1 Compute Clusters

Status	Name	CRN
Running	default- compute-cluster Default Cluster	j792...

Add Compute Cluster

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

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

dc-qe-env-v2 / Compute Clusters / compute-cluster
🔔

STATUS
Running

CLUSTER TYPE
Default Cluster

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

CREATED BY
Deepak Kumar Singh

Actions

CRN
j792...

Networking
Encryption
Node Groups
Compute Cluster Version
Labels

dcprofiler

LABELS
lifitie.cloudera.com/instance-group-id: ig-tp04xcyt ... More

ROOT VOLUME SIZE (GIB)
50

NODES
1
Auto scales between 1 and 10

dcprofiler-worker-spot

LABELS
lifitie.cloudera.com/instance-group-id: ig-q12zn8wn ... More

ROOT VOLUME SIZE (GIB)
100

NODES
0
Auto scales between 0 and 81

lifitie-infra

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

TAINTS
role.node.kubernetes.io/lifitie-infra: true:NoSchedule

ROOT VOLUME SIZE (GIB)
40

NODES
2
Auto scales between 2 and 4

Related Information

[Understanding the Cloudera Data Catalog Profiler](#)

[Understanding the Cluster Sensitivity Profiler](#)

[Understanding the Ranger Audit Profiler](#)

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 Public 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 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
}

```

Deleting profilers

In VM-based environments, deleting the profiler cluster (or in Compute Cluster enabled environments deleting the profiler jobs) removes all the Custom Sensitivity 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) the profiler jobs by Compute Cluster enabled environments, there is a provision to retain the status of the Custom 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.

**Note:**

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

Procedure

1. On the **Search** page, select the data lake from the drop-down.
2. Click Delete Profiler.

- If you agree, select the warning message I understand this action cannot be undone.

Figure 1: Deleting a profiler in a VM-based environment

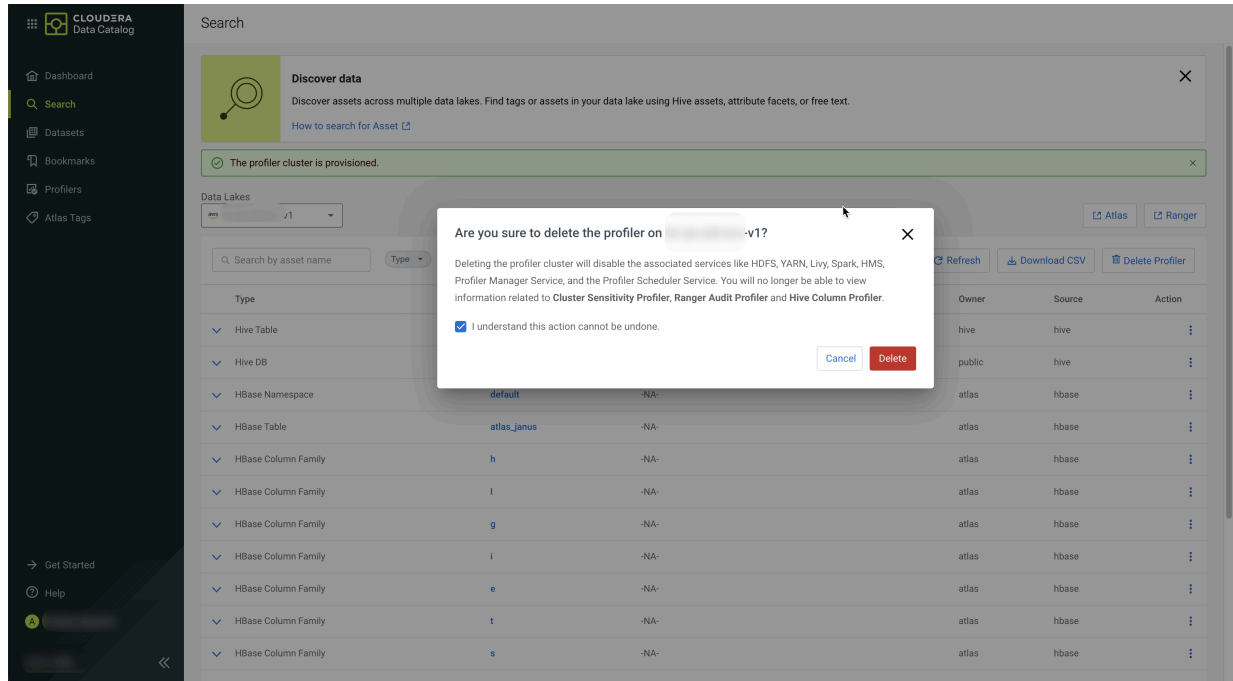
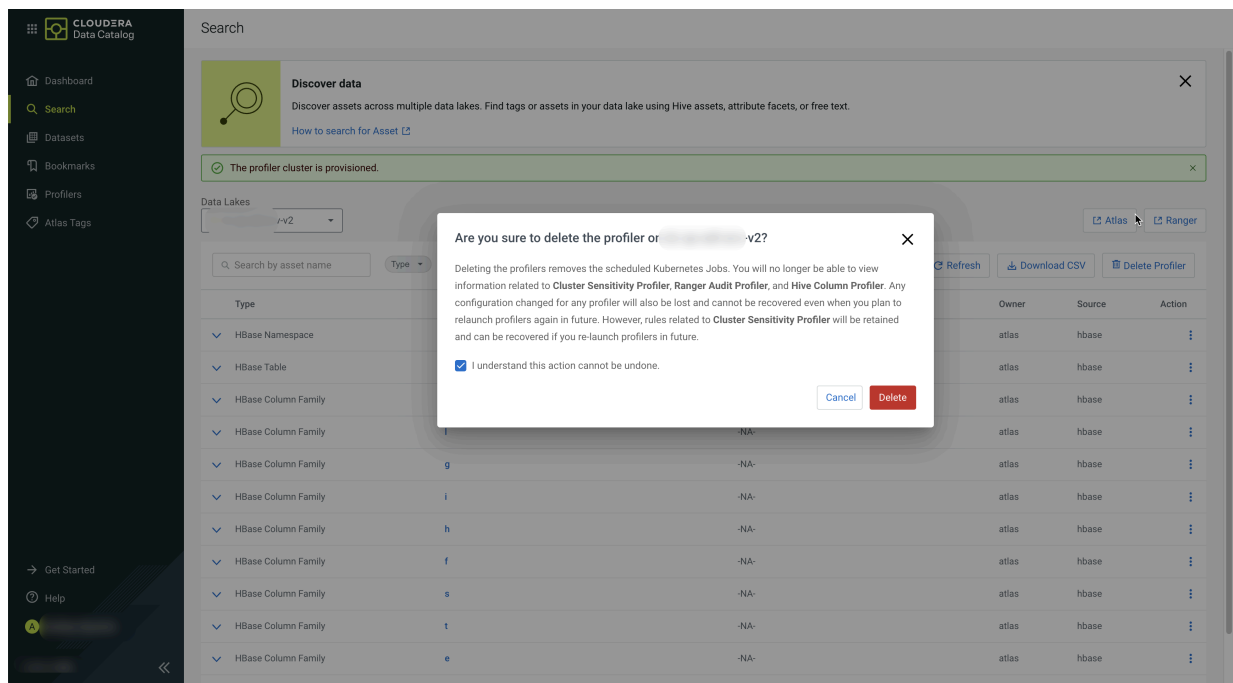
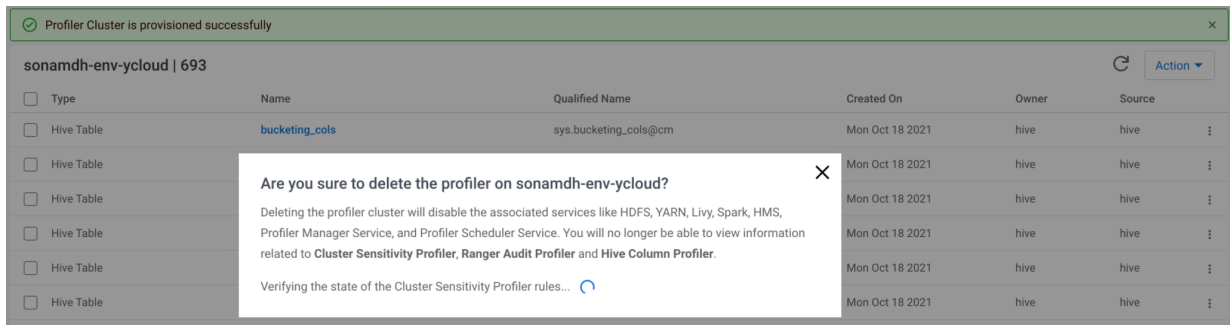


Figure 2: Deleting a profiler in a Compute Cluster enabled environment

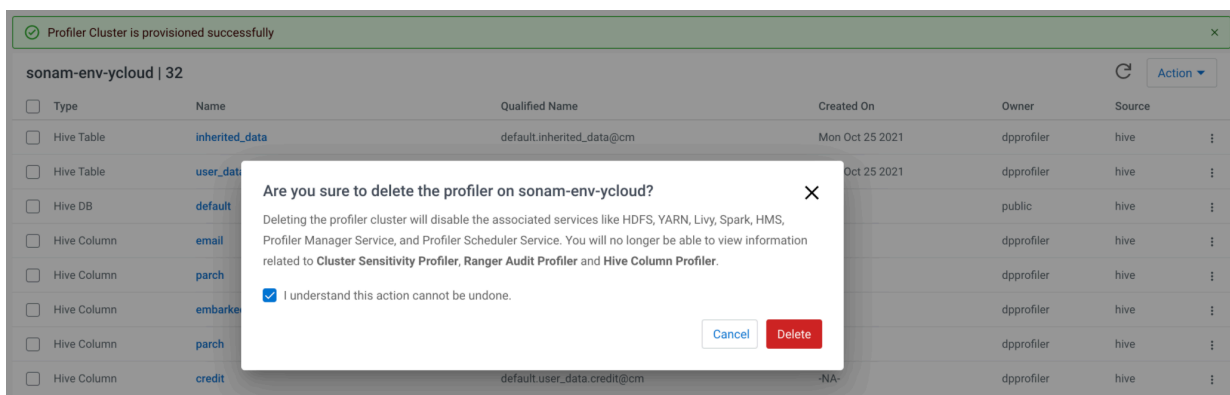


4. Click Delete.

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.



Additionally, note that you can delete the profiler cluster in these situations, when:

- Profiler cluster is up and running
- Profiler cluster is created but stopped
- Profiler cluster creation failed but is registered with the data lake
- Profiler cluster is down and inaccessible



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.

On-Demand Profilers

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

The following image shows the **Asset Details** page of an asset. You can run an On-Demand Profiler for Hive Column Profiler and Cluster Sensitivity Profiler by clicking on the appropriate Run button. The next scheduled run provides details about the next scheduled profiling for the respective profilers.



Note: You can use the On-Demand Profiler feature to profile both external and managed tables.

Profilers | 2

Hive Column Profiler ▶ Run

Last run: 10 mins ago | Status: SUCCESS

Next Schedule Run: Today at 11:30 PM

Cluster Sensitivity Profiler ▶ Run

Last run: 12 mins ago | Status: SUCCESS

Next Schedule Run: NA, Profiler is Disabled.



Note: In Compute Cluster-enabled environments, Iceberg tables can be also profiled with the On-Demand Profiler.

Profiling table data in non-default buckets

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

Procedure

1. In Cloudera Data Catalog, make note of your environment name in the **Search** menu.
2. Go Cloudera Management Console Environments
3. Search for your environment, then switch to the **Data hubs** tab.
4. Open your 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 interface. The left sidebar contains navigation options: Clusters, Hosts, Diagnostics, Charts, Administration, Parcels, Running Commands, and Support. The main content area is titled 'Home' and shows the 'Configuration Search' results for 'Profiler Scheduler Spark conf'. The search results are filtered by SCOPE (Profiler Scheduler Agent: 1), CATEGORY (Main: 1), and STATUS (Error: 0, Warning: 0, Edited: 0, Non-Default: 0, Include Overrides: 0). The configuration details for 'Profiler Scheduler Spark conf' are displayed, showing a list of properties under the 'profiler_scheduler > Profiler Scheduler Agent Default Group'.

Property Name	Value
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.uris)
spark.sql.hive.hiveserver2.jdbc.url.principal	\$(hive.server2.authentication.kerberos.principal)
spark.sql.hive.hiveserver2.jdbc.url	\$(beeline.hs2.jdbc.url.hive_on_tez)

The interface also includes a 'Save Changes (CTRL+S)' button at the bottom right.

Tracking profiler jobs

Use the Profilers > Jobs page for tracking 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.

Figure 3: Profiling jobs in a Compute Cluster enabled environment

Profilers / Jobs

Jobs Configs Tag Rules

Filters [Clear All](#)

Job Status

- Finished 6
- Running 2
- Failed 4

Profilers

- Ranger Audit Profiler 6
- Hive Column Profiler 2
- Cluster Sensitivity Profiler 4

Profiler	Status	Job ID	Start On	Last Updated On
Cluster Sensitivity	Failed	2CAY7TA9	09/10/2024 01:33 PM CEST	09/10/2024 01:33 PM CEST
Ranger Audit	Finished	IP6W2BUV-Jnp4	09/10/2024 01:32 PM CEST	09/10/2024 01:32 PM CEST
Ranger Audit	Finished	IP6W2BUV-JVE2	09/10/2024 01:32 PM CEST	09/10/2024 01:32 PM CEST
Cluster Sensitivity	Failed	XBLVQ52T	09/10/2024 01:32 PM CEST	09/10/2024 01:32 PM CEST
Ranger Audit	Finished	IP6W2BUV-5j7q	09/10/2024 01:32 PM CEST	09/10/2024 01:32 PM CEST
Table Stats	Running	MD6V9C6U	09/10/2024 01:30 PM CEST	09/10/2024 01:30 PM CEST
Cluster Sensitivity	Failed	AWXWB2UX	09/10/2024 02:04 AM CEST	09/10/2024 02:04 AM CEST
Cluster Sensitivity	Failed	RUUSD4GS	09/10/2024 02:03 AM CEST	09/10/2024 02:03 AM CEST
Table Stats	Running	P5TAWKGO	09/10/2024 02:01 AM CEST	09/10/2024 02:01 AM CEST
Ranger Audit	Finished	DNGSG9VR-A9Zd	09/10/2024 02:01 AM CEST	09/10/2024 02:01 AM CEST
Ranger Audit	Finished	DNGSG9VR-eeGg	09/10/2024 02:01 AM CEST	09/10/2024 02:01 AM CEST
Ranger Audit	Finished	DNGSG9VR-NFFz	09/10/2024 02:01 AM CEST	09/10/2024 02:01 AM CEST

Rows per page: 50 1 - 12 of 12

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

Figure 4: 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**
- **Stage** (for VM-based environments)
- **Job Status**
- **Job ID**
- **Start Time**
- **Last Update On time**

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

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

- Scheduler Service - The part of Profiler Admin which queues the profiler requests.
- Livy - This service is managed by YARN and where 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.

In case of Ranger Audit profiling, there could be a “NA” status for the total number of assets profiled. It indicates that the auditing that happens is dependent on the Ranger policies. In other words, the Ranger policies are actually profiled and not the assets.

Related Information

[Understanding the Cloudera Data Catalog Profiler](#)

[Understanding the Cluster Sensitivity Profiler](#)

[Understanding the Ranger Audit Profiler](#)

Viewing profiler configurations

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

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.



Note: You can also filter your profilers by job status, type for the last day, week and month.

Related Information

[Understanding the Cloudera Data Catalog Profiler](#)

[Understanding the Cluster Sensitivity Profiler](#)

[Understanding the Ranger Audit Profiler](#)

Ranger Audit Profiler configuration

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

Procedure

1. Go to Profilers Configs .
2. Select your data lake.
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. This is implemented as a quartz cron expression.
6. Continue with the resource settings.

For VM-based environment

- 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](#).

For Compute Cluster enabled environment

- a. In **Pod Configurations**, set the Kubernetes job resources.

Pod configurations specify the resources that would be allocated to a pod when the profiler job starts to run. As all profilers are submitted as Kubernetes jobs, you must decide if you want to add or reduce resources to handle workload of various sizes.

- **Pod CPU Limit:** Indicates the maximum number of cores that can be allocated to a Pod. The accepted values range from one through eight.
- **Pod CPU Requirement:** This is the minimum number of CPUs that will be allocated to a Pod when its provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The accepted values range from one through eight.
- **Pod Memory Limit:** Maximum amount of memory can be allocated to a Pod. The accepted values range from 1 through 256.
- **Pod Memory Requirement:** This is the minimum amount of RAM that will be allocated to a Pod when it is provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a

container is not allowed to use more than its resource limit. The accepted values range from 1 through 256.

b. In **Executor Configurations**, update the following:

- **Number of workers:** Indicates the number of processes that are used by the distributed computing framework. The accepted values range from one through eight.
- **Number of threads per worker:** Indicates the number of threads used by each worker to complete the job. The accepted values range from one through eight.
- **Worker Memory limit in GB:** To avoid over utilization of memory, this parameter forces an upper threshold memory usage for a given worker. For example, if you have a 8 GB Pod and 4 threads, the value of this parameter must be 2 GB. The accepted values range from one through four.

Executor configurations are the runtime configurations. These configuration must be changed if you are changing the pod configurations and when there is a requirement for additional compute power.

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

Cluster Sensitivity Profiler configuration

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


Procedure

1. Go to Profilers Configs .
2. Select your data lake.
3. Select Cluster Sensitivity Profiler.

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

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](#) on page 37.
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.



Note: In Compute Cluster enabled environments, skip to step 8 on page 34.

8. Continue with the resource settings.

For VM-based environment

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](#).

For Compute Cluster enabled environment

a. In **Pod Configurations**, update the following:

- **Pod CPU Limit:** Indicates the maximum number of cores that can be allocated to a Pod. The accepted values range from one through eight.
- **Pod CPU Requirement:** This is the minimum number of CPUs that will be allocated to a Pod when it is provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The accepted values range from one through eight.
- **Pod Memory Limit:** Maximum amount of memory can be allocated to a Pod. The accepted values range from 1 through 256.
- **Pod Memory Requirement:** This is the minimum amount of RAM that will be allocated to a Pod when it is provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The accepted values range from 1 through 256.

b. In **Executor Configurations**, update the following:

- **Number of workers:** Indicates the number of processes that are used by the distributed computing framework. The accepted values range from one through eight.
- **Number of threads per worker:** Indicates the number of threads used by each worker to complete the job. The accepted values range from one through eight.
- **Worker Memory limit in GB:** To avoid over utilization of memory, this parameter forces an upper threshold memory usage for a given worker. For example, if you have a 8 GB Pod and 4 threads, the value of this parameter must be 2 GB. The accepted values range from one through four.

Executor configurations are the runtime configurations. These configuration must be changed if you are changing the pod configurations and when there is a requirement for additional compute power.

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

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 include rules.

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

- Database name
- Asset name
- Asset owner
- Path to the asset
- Created date

4. Select the operator from the drop-down list. Depending on the keys selected, you can select an operator such as equals, contains. For example, you can select the name of assets that contain a particular string.

5. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.

6. Click Done. Once rule is added, you can toggle the state of the new rule to enable it or disable it as needed.

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

Related Information

[Understanding the Cloudera Data Catalog Profiler](#)

[Understanding the Cluster Sensitivity Profiler](#)

Hive Column Profiler configuration

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

Procedure

1. Go to Profilers Configs .

2. Select your data lake.

3. Select Hive Column 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. This is implemented as a quartz cron expression.

For more information, see [Understanding the Cron Expression generator](#) on page 37.

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.



Note: In Compute Cluster enabled environments, skip to step 8 on page 36.

8. Continue with the resource settings.

For VM-based environment

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](#).

For Compute Cluster enabled environment

a. In **Pod Configurations**, update the following:

- **Pod CPU Limit:** Indicates the maximum number of cores that can be allocated to a Pod. The accepted values range from one through eight.
- **Pod CPU Requirement:** This is the minimum number of CPUs that will be allocated to a Pod when it is provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The accepted values range from one through eight.
- **Pod Memory Limit:** Maximum amount of memory can be allocated to a Pod. The accepted values range from 1 through 256.
- **Pod Memory Requirement:** This is the minimum amount of RAM that will be allocated to a Pod when it is provisioned. If the node where a Pod is running has enough resources available, it is possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The accepted values range from 1 through 256.

b. In **Executor Configurations**, update the following:

- **Number of workers:** Indicates the number of processes that are used by the distributed computing framework. The accepted values range from one through eight.
- **Number of threads per worker:** Indicates the number of threads used by each worker to complete the job. The accepted values range from one through eight.
- **Worker Memory limit in GB:** To avoid over utilization of memory, this parameter forces an upper threshold memory usage for a given worker. For example, if you have a 8 GB Pod and 4 threads, the value of this parameter must be 2 GB. The accepted values range from one through four.

Executor configurations are the runtime configurations. These configuration must be changed if you are changing the pod configurations and when there is a requirement for additional compute power.

9. Add **Asset Filter Rules** as needed to customize the selection and deselection of assets which the profiler profiles.
 - 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 include rules.
 3. Select the key from the drop-down list. You can select from the following:
 - Database name
 - Asset name
 - Asset owner
 - Path to the asset
 - Created date
 4. Select the operator from the drop-down list. Depending on the keys selected, you can select an operator such as equals, contains. For example, you can select the name of assets that contain a particular string.
 5. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
 6. Click Done. Once rule is added, you can toggle the state of the new rule to enable it or disable it as needed.
10. Click Save to apply the configuration changes to the selected profiler.

Understanding the Cron Expression generator

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

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

Each * in the cron represents a unique value.

For VM-based environments

The ? character is used for undefined day of the month and the day of the week.

Schedule: * * * * ? *

For example, consider a cron with the following values:

1 2 3 2 5 ? 2021

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



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

For Compute Cluster enabled environments

Cron Expression: 0 18 * * *

In this format the * characters represent the following units:Minute hour day(month) month day(week)

For example, consider a cron with the following values:

CRON Expression: 30 10 15 5 *

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



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

Consider another example:

30 10 * 5 7

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



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

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

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 scripts:

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)

Enable or disable profilers

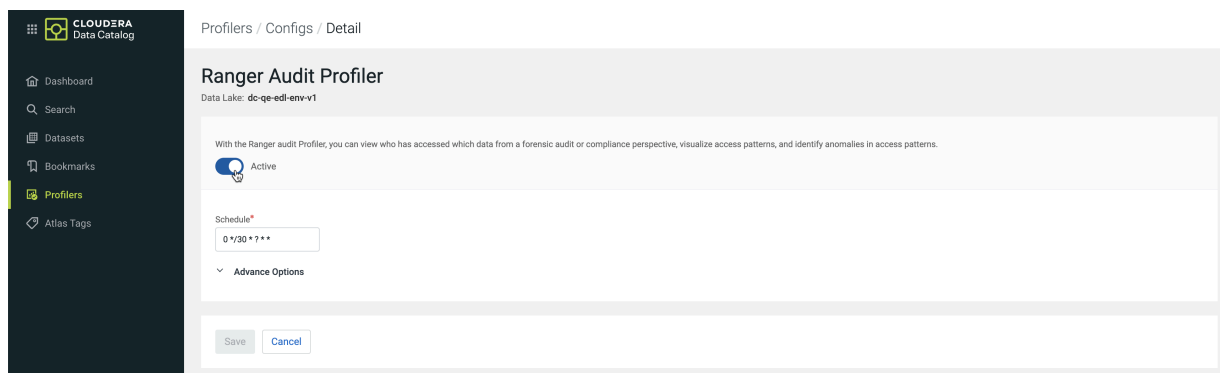
By default, profilers are enabled and run every 30 minutes. If you want to disable (or re-enable) a profiler, you can do this by selecting the appropriate profiler from the Configs tab.

Procedure

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

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

3. Switch the toggle to the desired state.

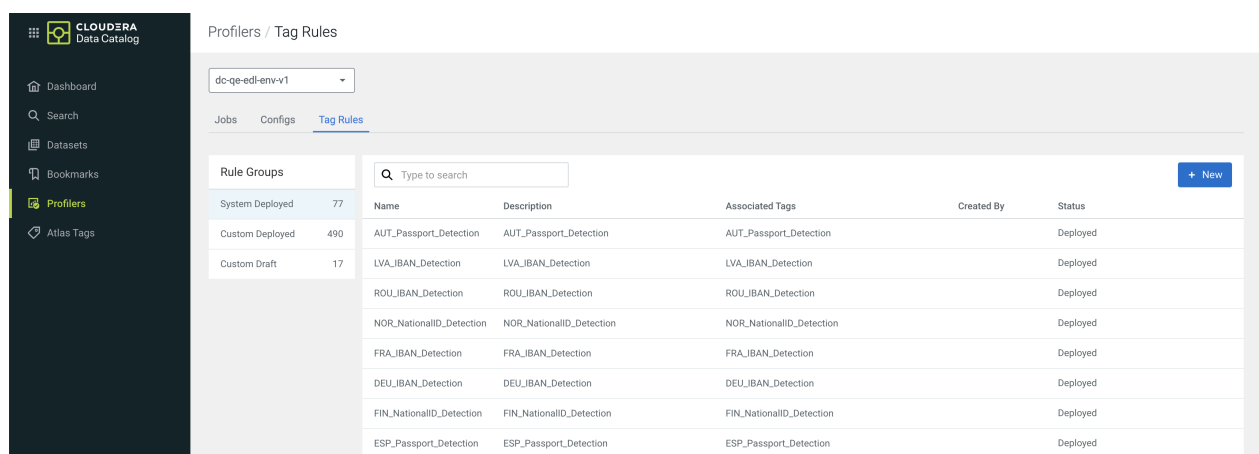


Profiler Tag Rules

You can use preconfigured tag rules or create new rules based on regular expressions to limit the number of assets to be profiled.

Rules are categorized into following groups:

- **System Deployed:** These are built-in rules that cannot be edited.
- **Custom Deployed:** Tag rules that you create and deploy on clusters after validation will appear under this category. Hover your mouse over the tag rules to deploy or suspend them as needed. 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. Such rules appear under this category.



After creating your rule, you have to validate them with test data and, then Deploy them from **Custom Draft**.

Profilers / Tag Rules

Rule Groups	Name	Description	Associated Tags	Created By	Status
System Deployed 77	test_dry_run		test		Validation Failed
Custom Deployed 490	test	jk	hello		Validation Pending
Custom Draft 18					Validation Success
					Validation Pending
	CUSTOMER_EMAIL_TAG		email		Validation Failed
	Example1		phone number		Validation Pending
	phone_number		phone number		Validation Pending
	phone		telephone		Validation Pending
	phonetest		phone number		Validation Success
	test1		test		Validation Failed
	test_rule_sb	testing	test		Validation Success
	testing_	this is testing	ruleTest1		Validation Failed



Note: Tag Rules can be temporarily suspended.

Profilers / Tag Rules

Rule Groups	Name	Description	Associated Tags	Created By	Status
System Deployed 77	AUT_Passport_Detection	AUT_Passport_Detection	AUT_Passport_Detection		Deployed
Custom Deployed 490	LVA_IBAN_Detection	LVA_IBAN_Detection	LVA_IBAN_Detection		Deployed
Custom Draft 17					

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:

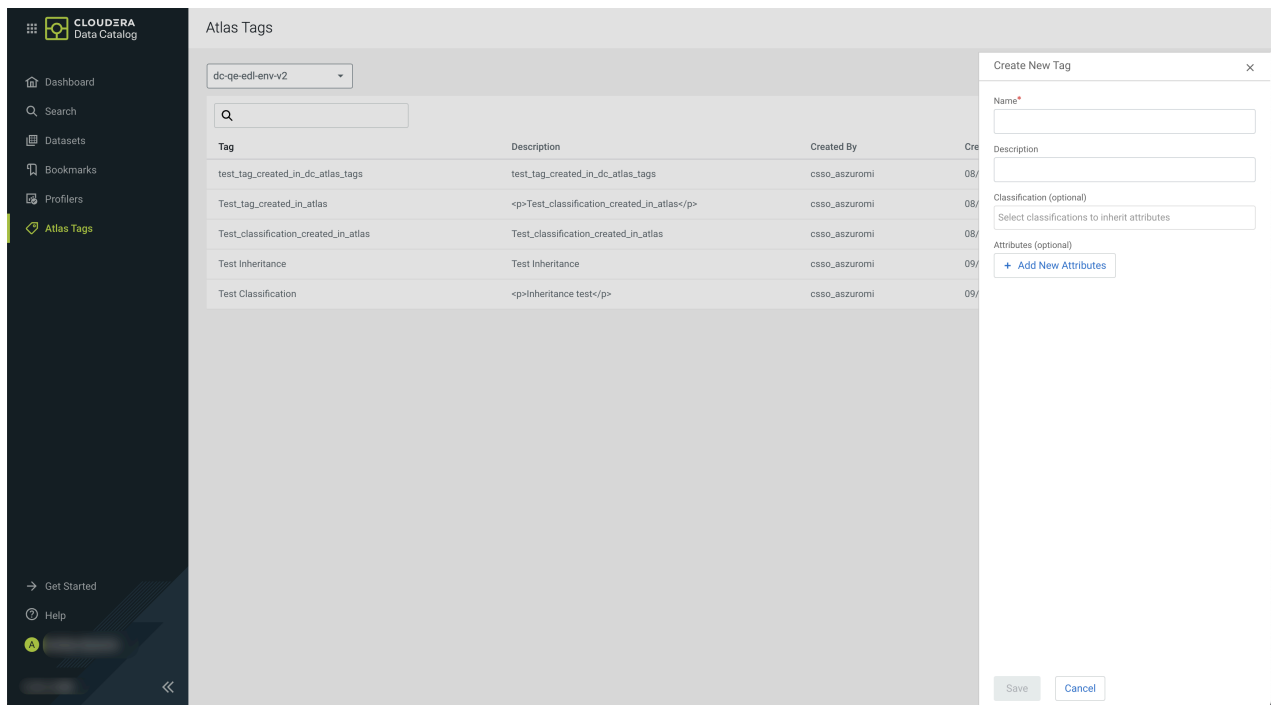
- Selecting a data lake
- Searching for a tag
- Adding a tag
- Editing a tag
- Deleting a tag

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.

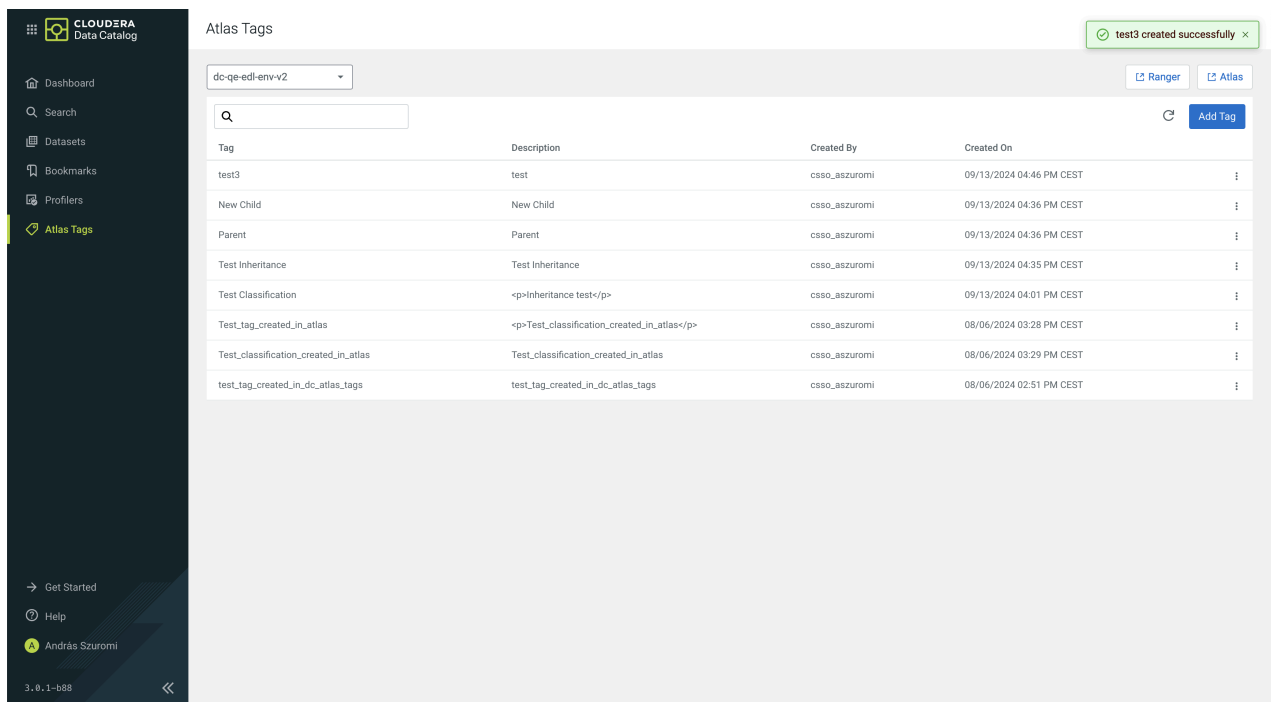
Atlas Tags

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

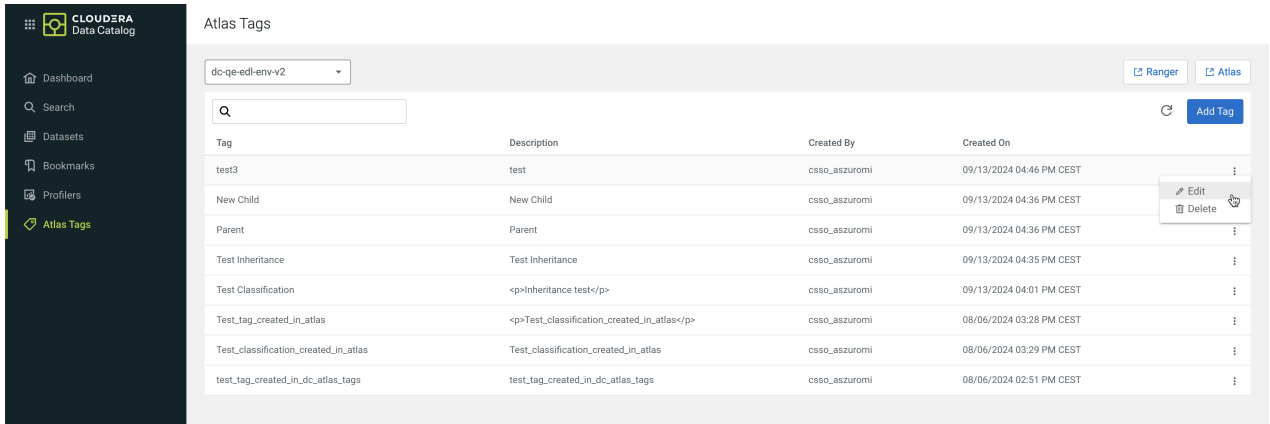
In **Create New Tag**, you can define the tag name, description and the "super-classification" from which the attributes are inherited for the sub-classification (or tag in Cloudera Data Catalog)



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.

