Cloudera Runtime

Data Catalog Top Use Cases

Date published: 2022-08-30
Date modified: 2023-03-10

https://docs.cloudera.com/

Legal Notice

Contents

Search for Assets

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

When you enter the search terms in Data Catalog Search, you are looking up names, types, descriptions, and other metadata collected by 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 assets that Data Catalog stores.

Filters

When you select a property value, a filter breadcrumb shows above the search results.

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.
- Database - Select the database to view all the assets stored in that database. This filter is applicable to Hive and HBase tables only.

  Note: For information purposes, Database filter is displayed as Namespace in case of HBase tables.

- Entity Tag - Use entity tags to refine your search results. You can add business metadata as entity tags in Atlas and use these tags to refine your search results and view the details of the required data asset.
- Created Within - You can choose to refine your search results of assets within the data lake to view the data assets created within the last 7 days, 15 days, or 30 days. You can also add custom values such as 5 days or 10 days to view specific information.
- Created Before - Depending on the time when the assets were created, you can choose to refine the search results and view data assets created before 1 day, 7 days, or 15 days. You can add custom values to view data assets created before the days of your preference such as 8 days or 12 days.

  Note: These two filters (Created Within and Created Before) are applicable only when Atlas provides the created time for the assets.

- Column Tag - You can search for Hive and HBase table assets by tags that have been applied on their children entities, that is, columns or column families using the column tags filter.
- Glossary - You can filter assets based on business glossary terms. You can search for any asset without any entity type restrictions.

Click Clear for any filter to clear the selection. You can use a combination of filters to view the required data assets.

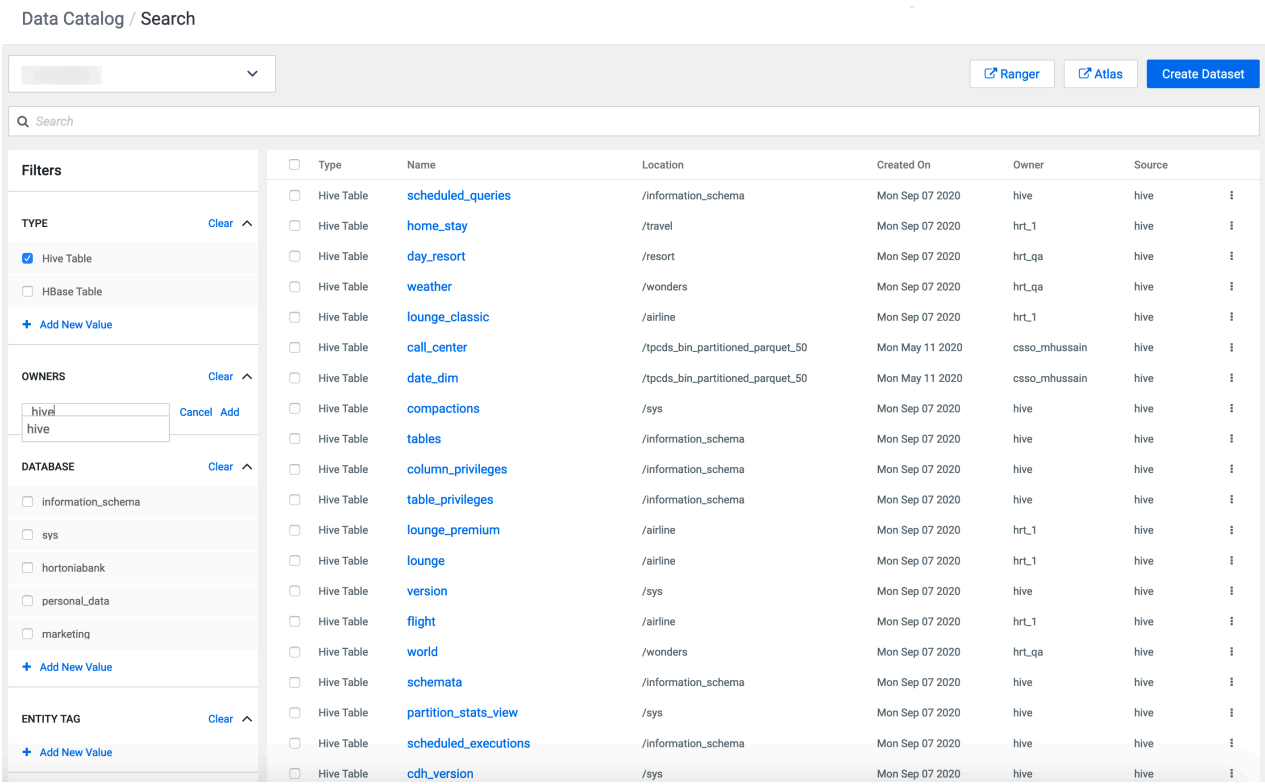
Prepopulating Asset Owners

In Data Catalog, under the search page, you can filter for assets based on the owners.

Rather than having to type in the owners manually, the available asset owners are listed in drop down. Select the record from the list and add it as a filter criteria

For example, in the following diagram, the selected asset TYPE is “Hive”.

For the selected TYPE the owner “hive” is available in the drop-down and based on this condition, the assets can be filtered in the search page.



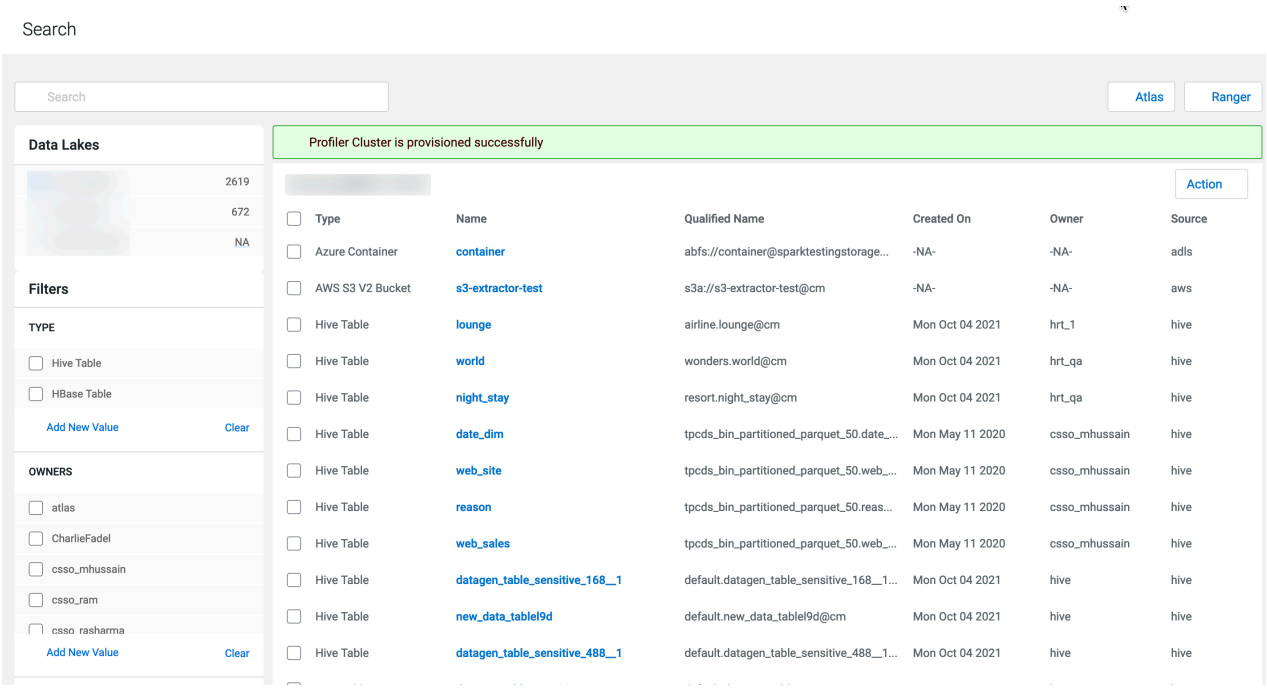
Accessing Data Lakes

In the Data Catalog search dashboard, the accessible data lakes are displayed under the search panel.

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 more information about the user permissions, see Prerequisite to access Data Catalog Service.

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



Searching for assets across multiple data lakes

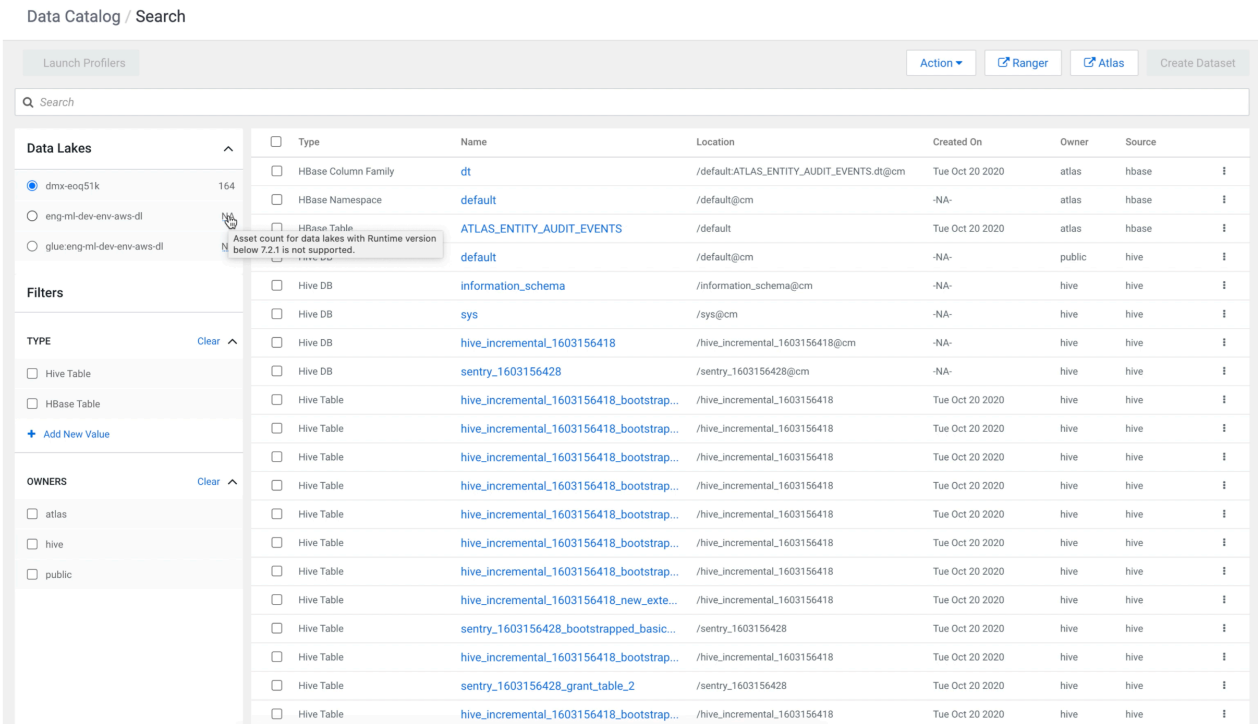
In Data Catalog, the data lake search capabilities has been enhanced with the way you can search for the assets. You can now view the complete list of data lakes that are available in a specific Data Catalog instance.

Previously, to search for assets using a data lake, a drop-down menu was available to select a data lake. You can now use the radio button to select a specific data lake.

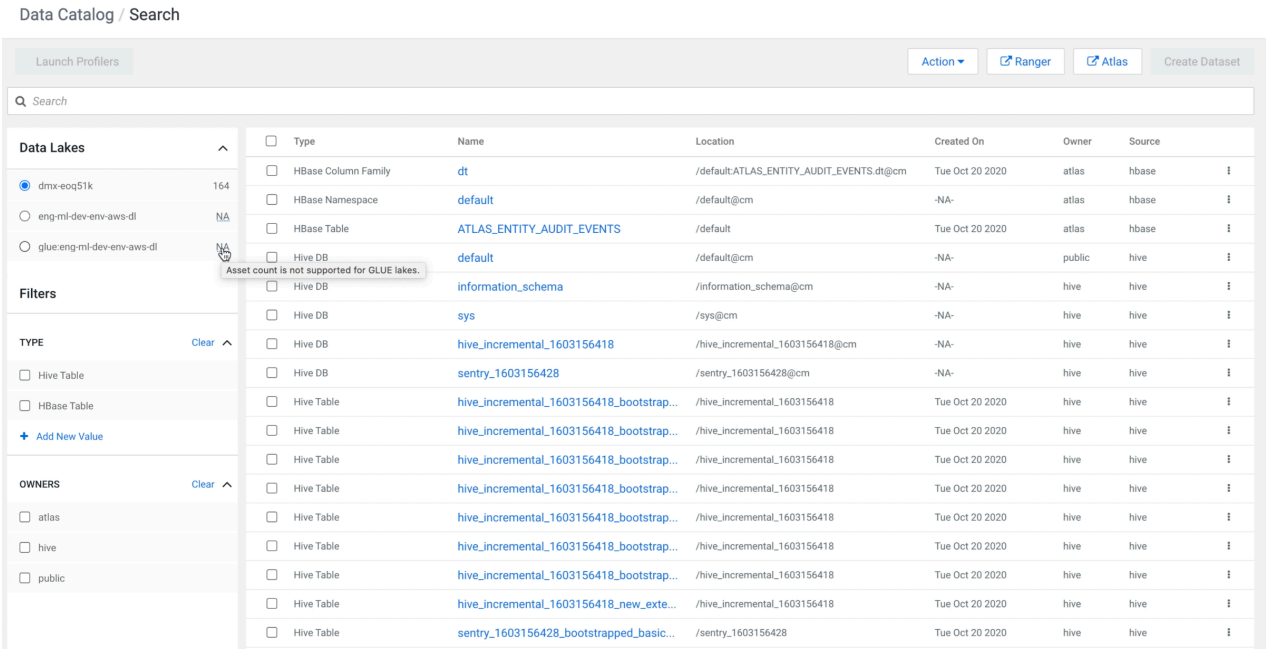
The number of assets that are visible against each data lake indicate that they are applicable to the search query that match. You can select a specific data lake and select one or more search query types to retrieve the total list of available assets in the selected data lake. The total count of the selected data lake can change based on the type of filter that is applied.

You can get the count of all the assets complying with the set search criteria from all the data lakes and display the same for each lake. Using the asset count details, you can optionally change the data lake and with which the result count is obtained from the selected lake. When you select a data lake, the search query gets updated by default. The previous query that was triggered on a previously selected data lake is not carried forward to the currently selected data lake.

For each selected data lake, you can set up different queries and the total asset count varies. Also, in certain scenarios, when a search query is triggered, the data lake count (when hovered on NA) displays the message Asset count for data lakes with Runtime version below 7.2.1 is not supported. The asset count for each selected data lake appears only if the Runtime version is above 7.2.1.



In the event of a Glue lake being selected, the data lake count displays the message Asset count is not supported for GLUE lakes or Asset count is not supported when a GLUE lake is selected.



Searching for assets using Glossary

Use 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.

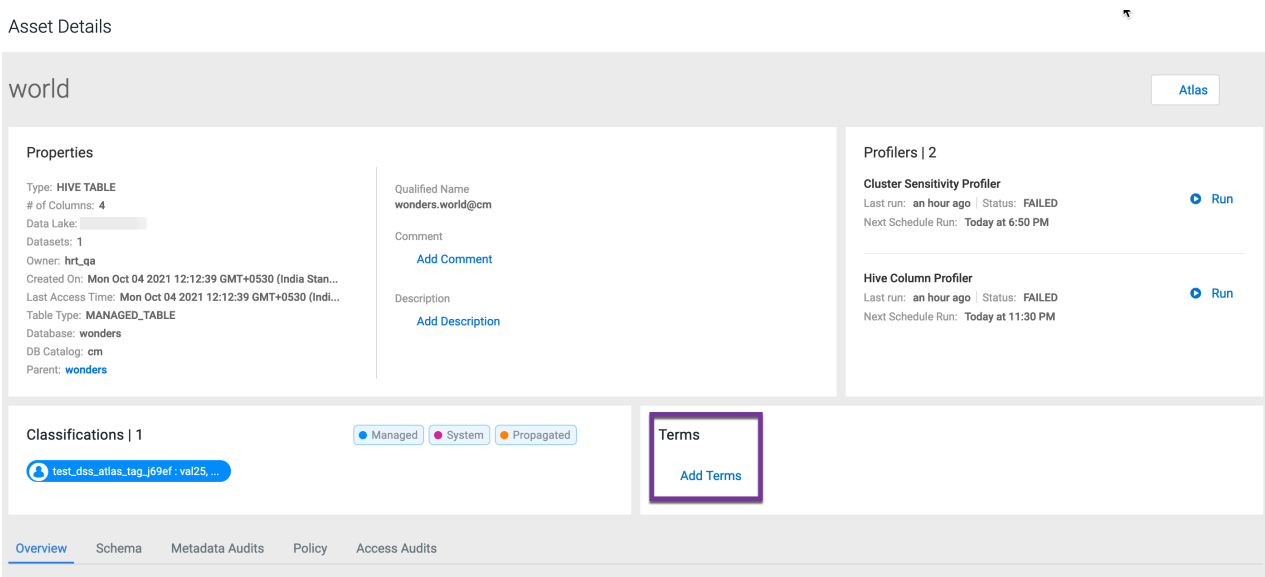
Using Terms in Data Catalog

You can use the Asset Details page in Data Catalog to add or modify "terms" for your selected assets.

A new widget called "Terms" is available in the Asset Details page. You can define rich glossary vocabularies using the natural terminology (technical terms and/or business terms). To semantically relate the term(s) to each other. And finally to map assets to glossary terms(s).

You can assign terms with entities, search for entities, filter entities by glossary term(s), and also search for entities by using associated term(s).

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



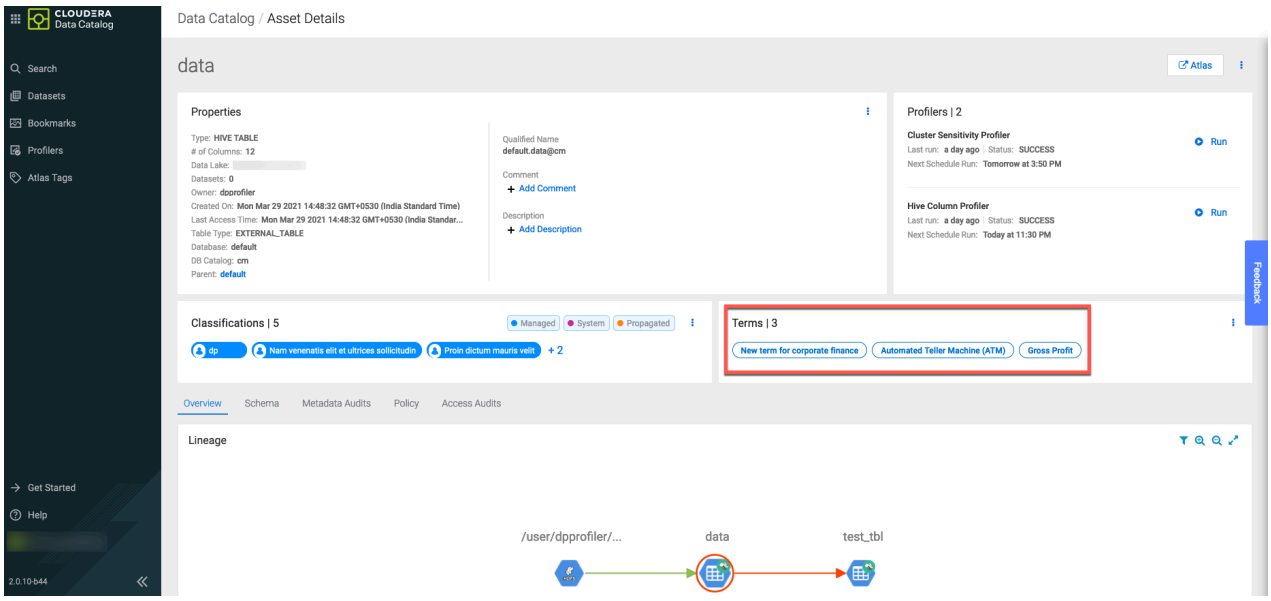
Mapping glossary terms

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

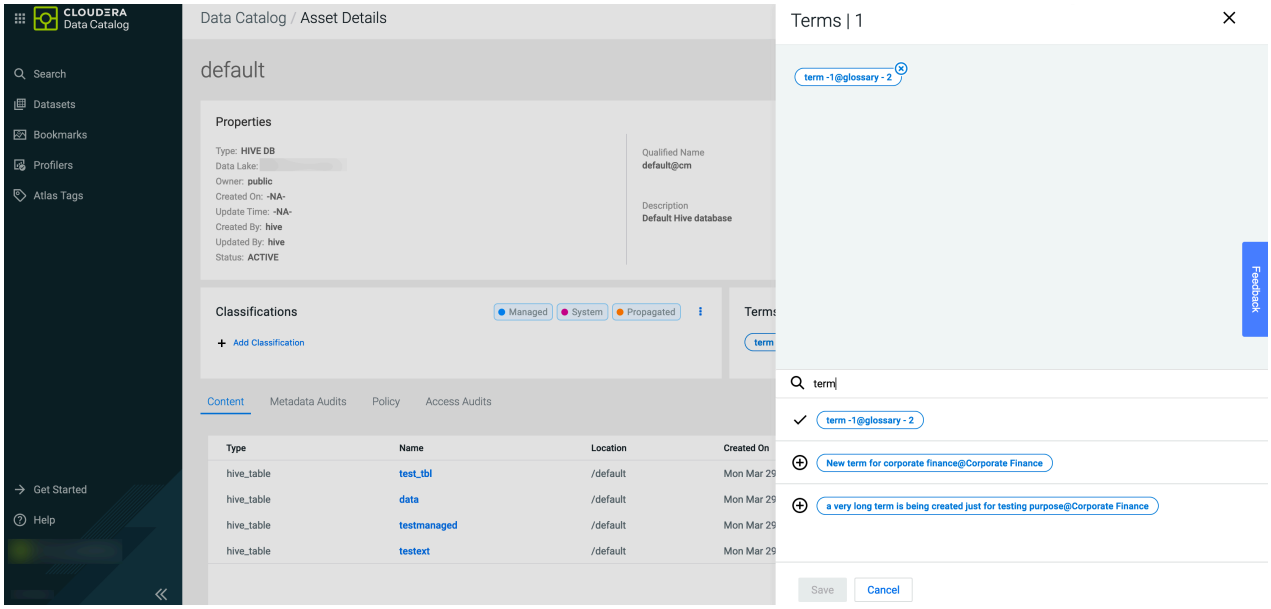
You can search for those terms in Data Catalog and map specific terms with Data assets. You can search for terms in Data Catalog to either add and delete them from the selected data asset. The selected asset displays the total number of terms associated or mapped accordingly.

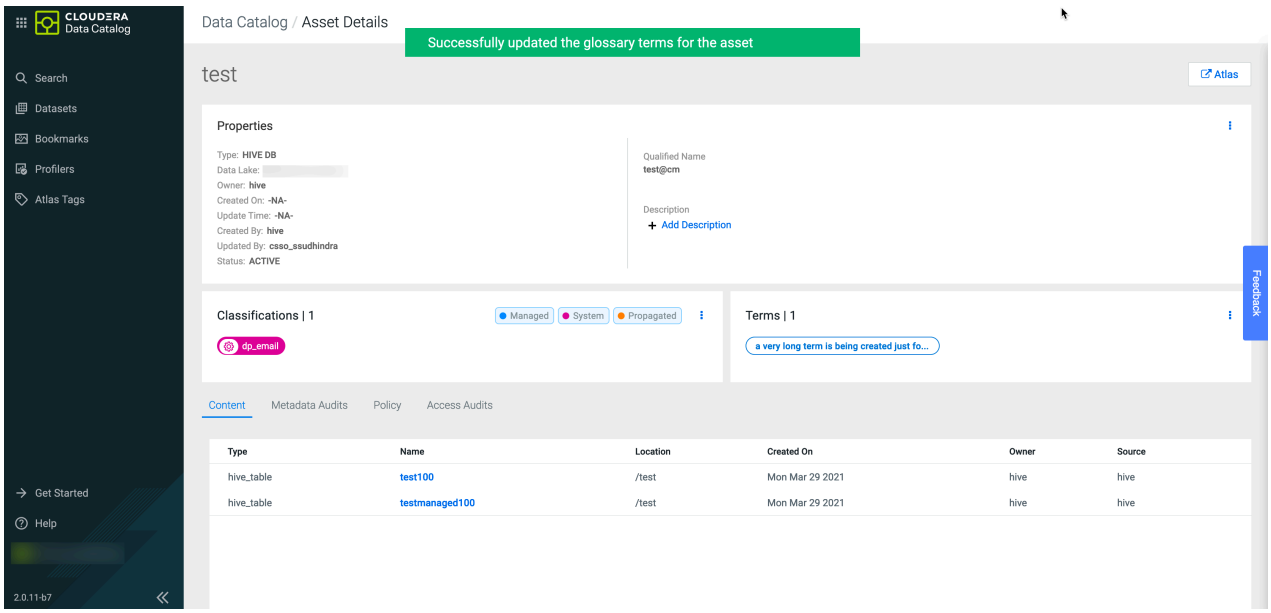
When you map a specific term for your dataset, the term is displayed in the following format:

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

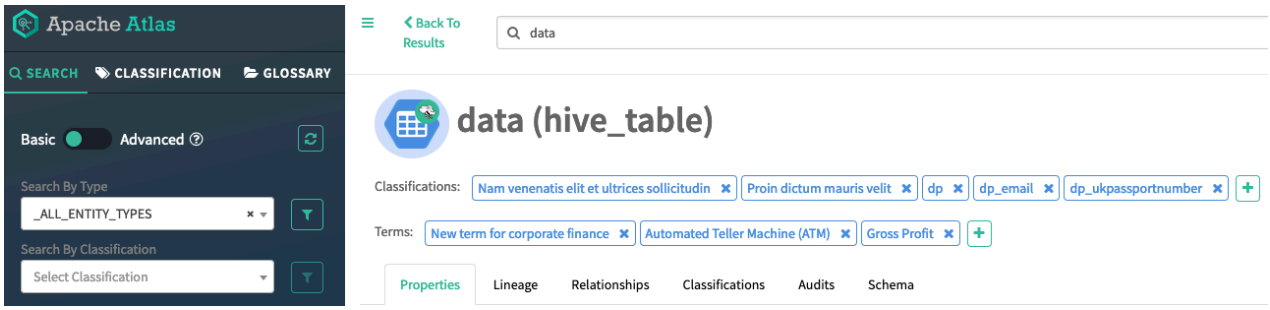


You can use the icon in the Terms widget on the Asset Details page to add new terms for your data assset. Click Save to save the changes.

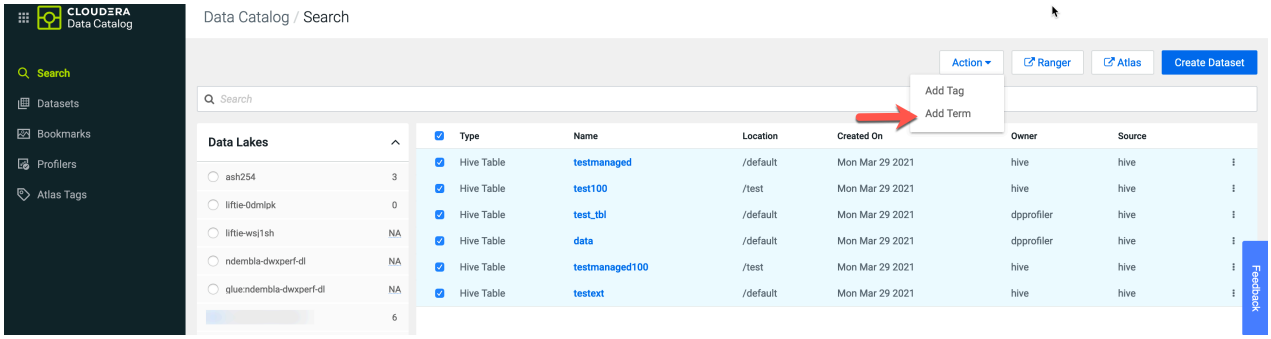


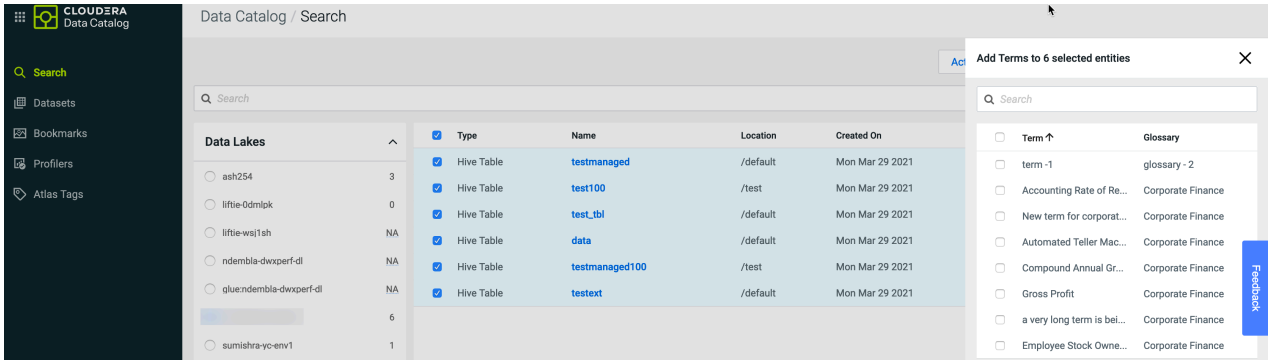


You can search for the same asset in the corresponding Atlas environment as shown in the example image.

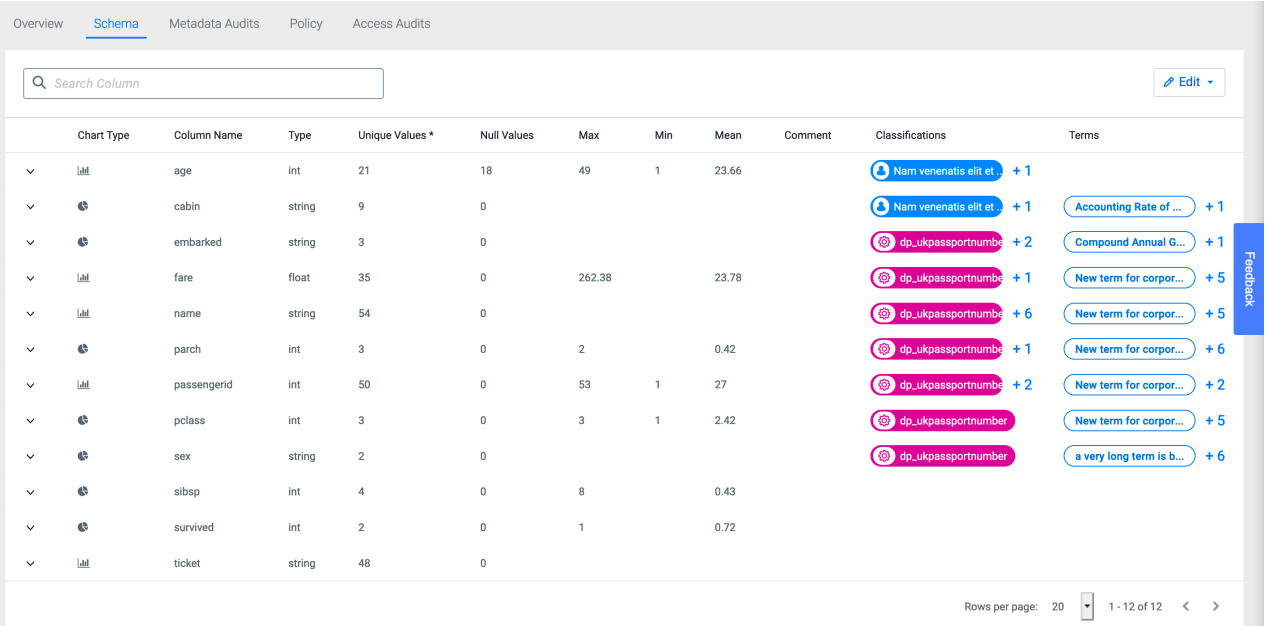


Additionally, you can also associate terms to your datasets by selecting one or more assets on the Data Catalog search page. You can associate terms with multiple datasets at a time.

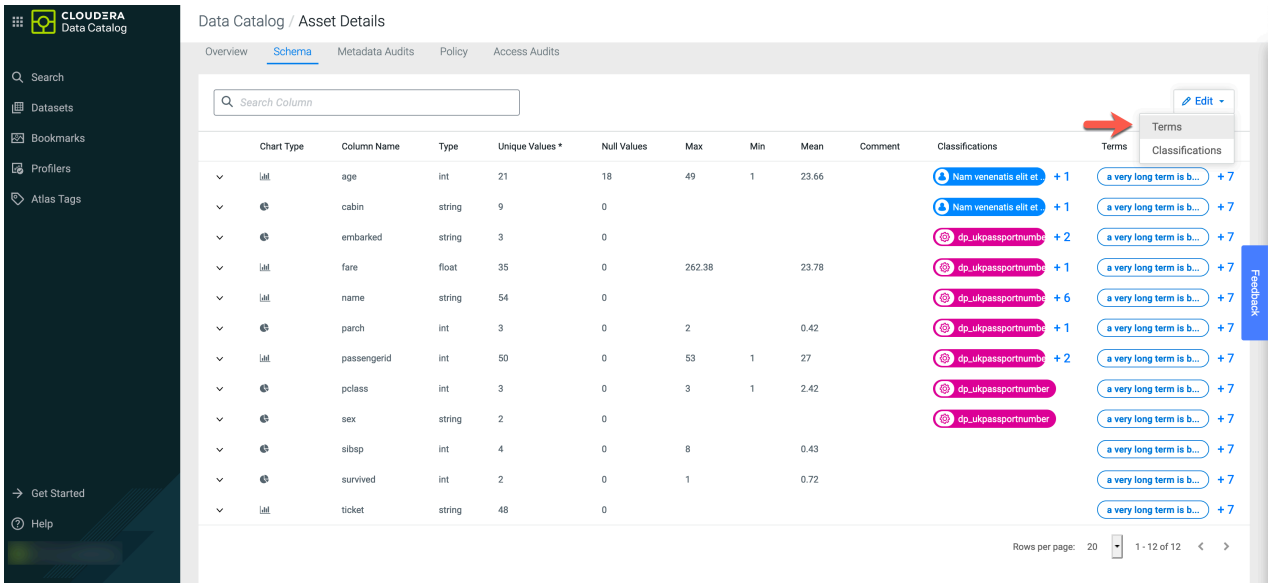




When you select a Hive table asset and navigate to the Asset Details page, under the Schema tab, you can view the list of terms associated with the asset.

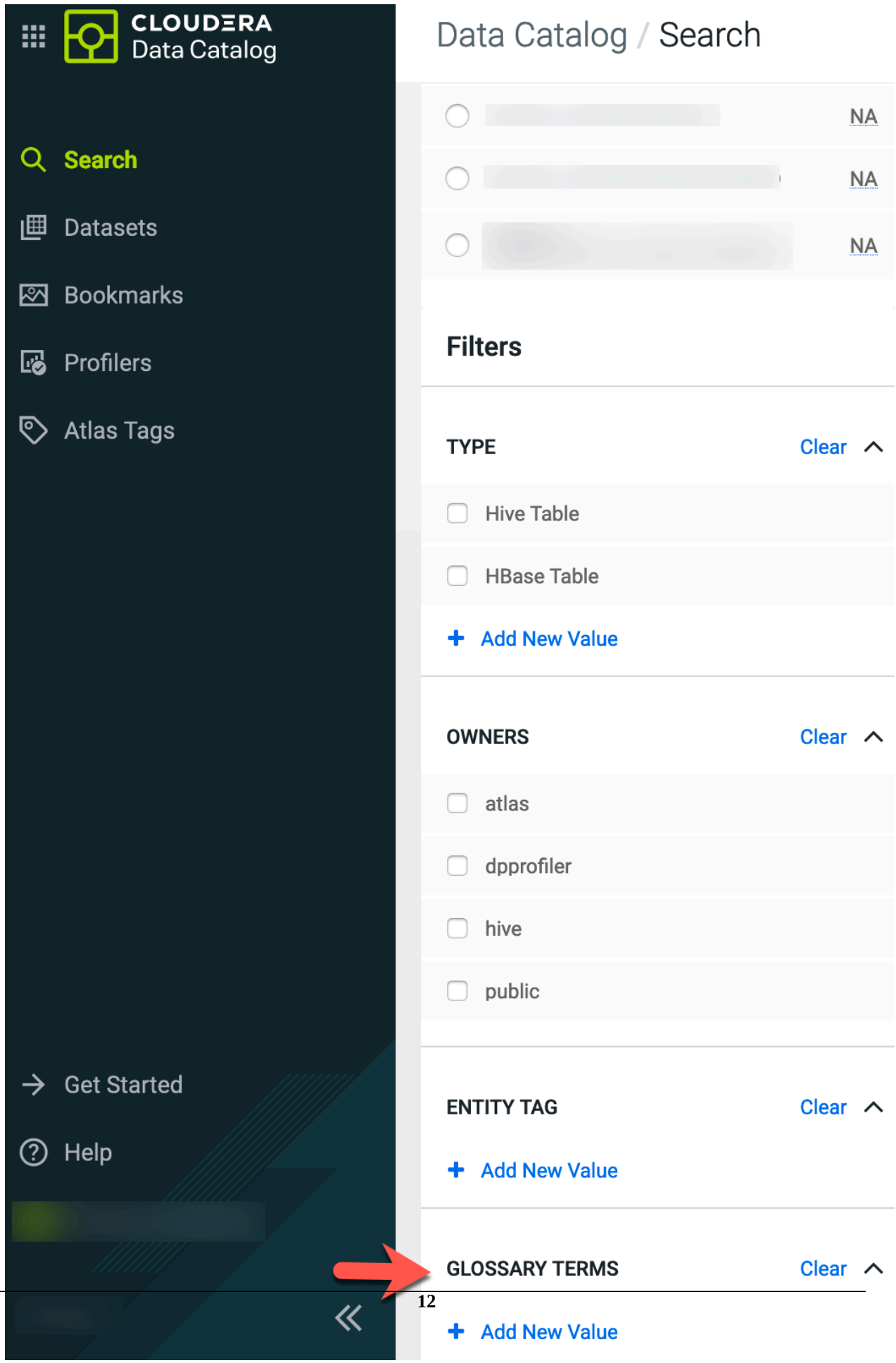


You can add or update the terms for the associated datasets by clicking the Edit button.



Searching for assets using glossary terms

You can search for the datasets using the Glossary terms filter available on the Data Catalog search page.


CLOUDERA
Data Catalog
Search
Datasets
Bookmarks
Profilers
Atlas Tags
Get Started
Help
Data Catalog / Search
NA
NA
NA
Filters
TYPE
Clear
Hive Table
HBase Table
Add New Value
OWNERS
Clear
atlas
dpprofiler
hive
public
ENTITY TAG
Clear
Add New Value
GLOSSARY TERMS
Clear
Add New Value

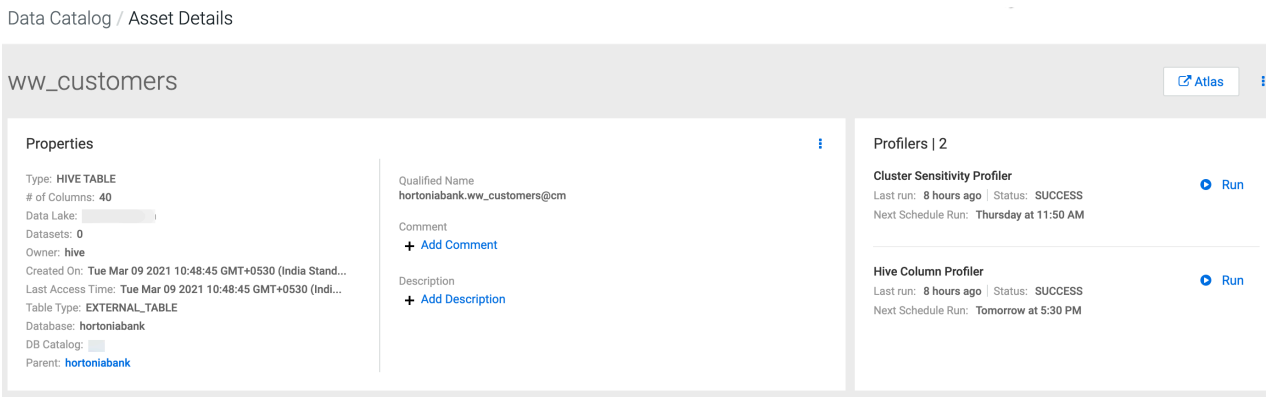
Additional search options for asset types

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

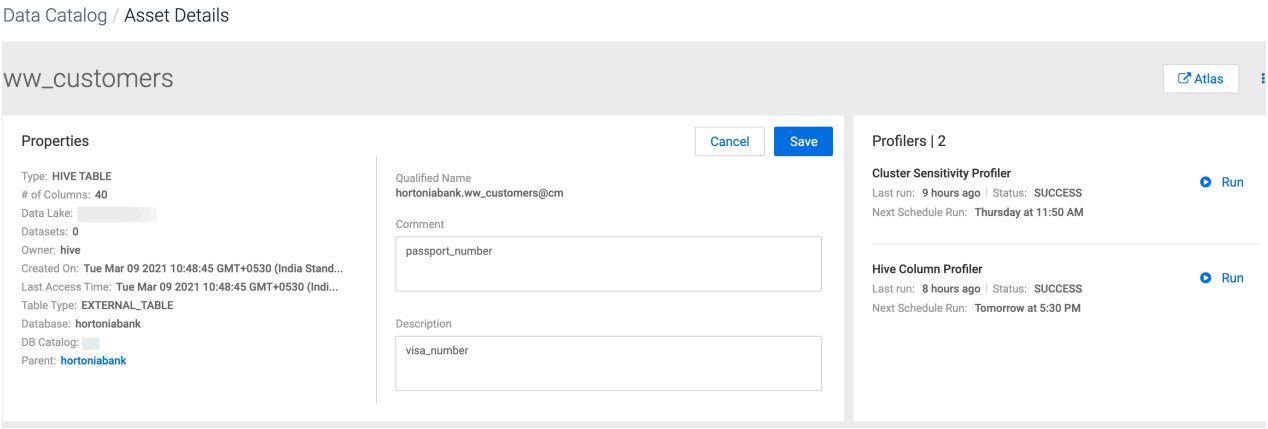
In the Asset Details page for each asset type that you select, you can add or edit comment and description fields. For each asset type in Data Catalog, you can add or edit comments or include a description. Including these values for the selected asset helps you to identify your chosen asset when you perform the search operation.

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

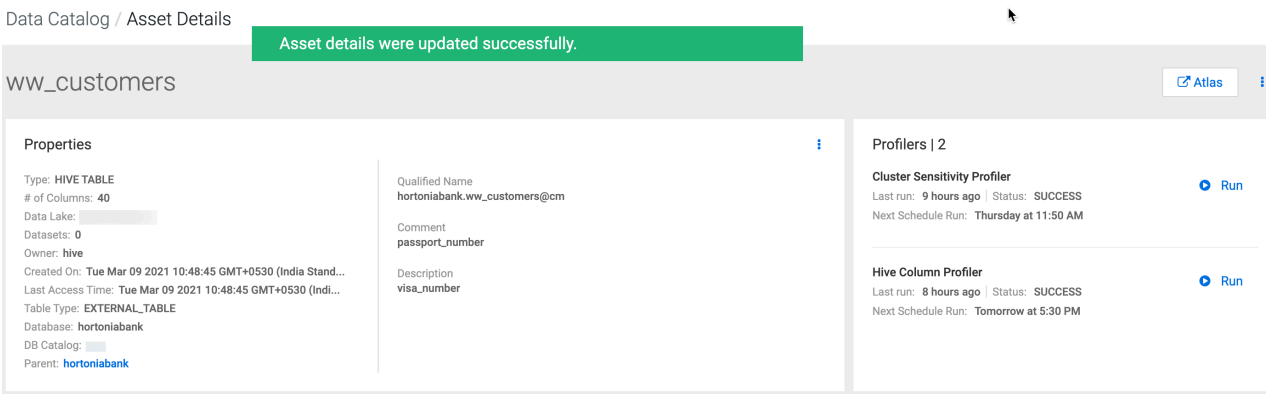
Note: The comment and description options are supported only for Hive table and Hive Column assets. For other asset types, only the description option is supported.



Click + besides Comment and Description to include the respective values.

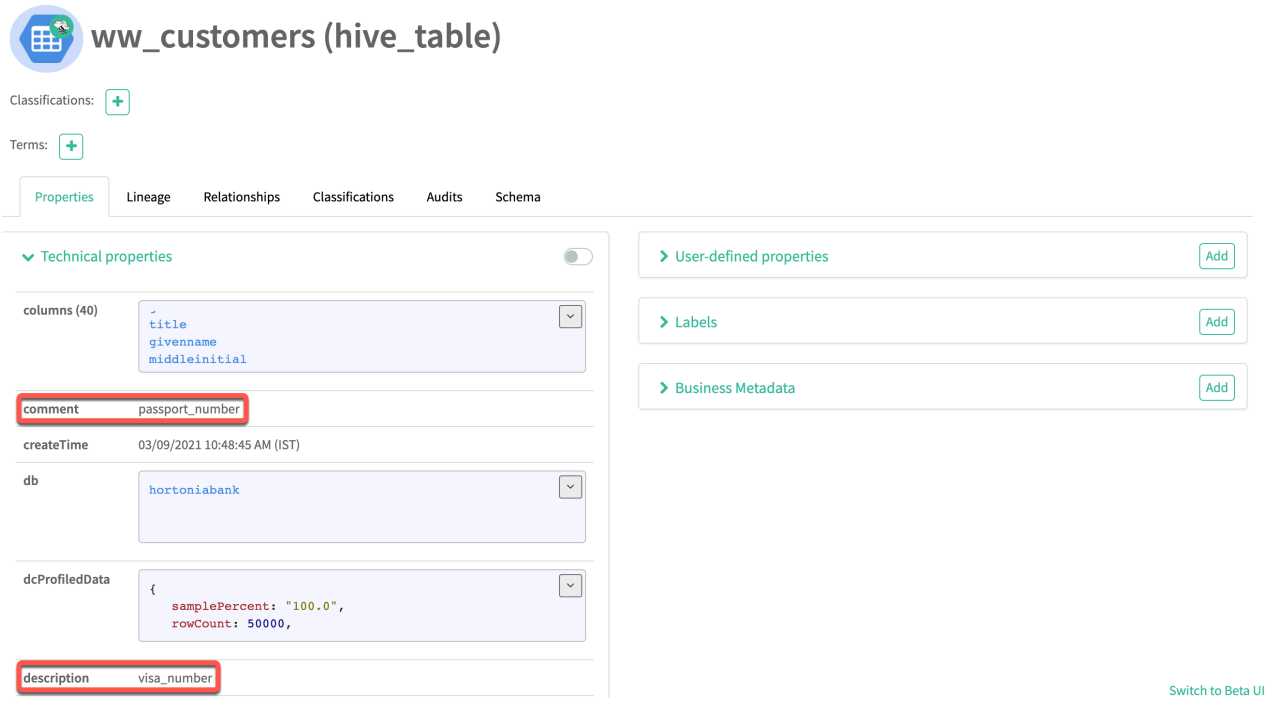


Click Save to save the changes.



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 displayed.

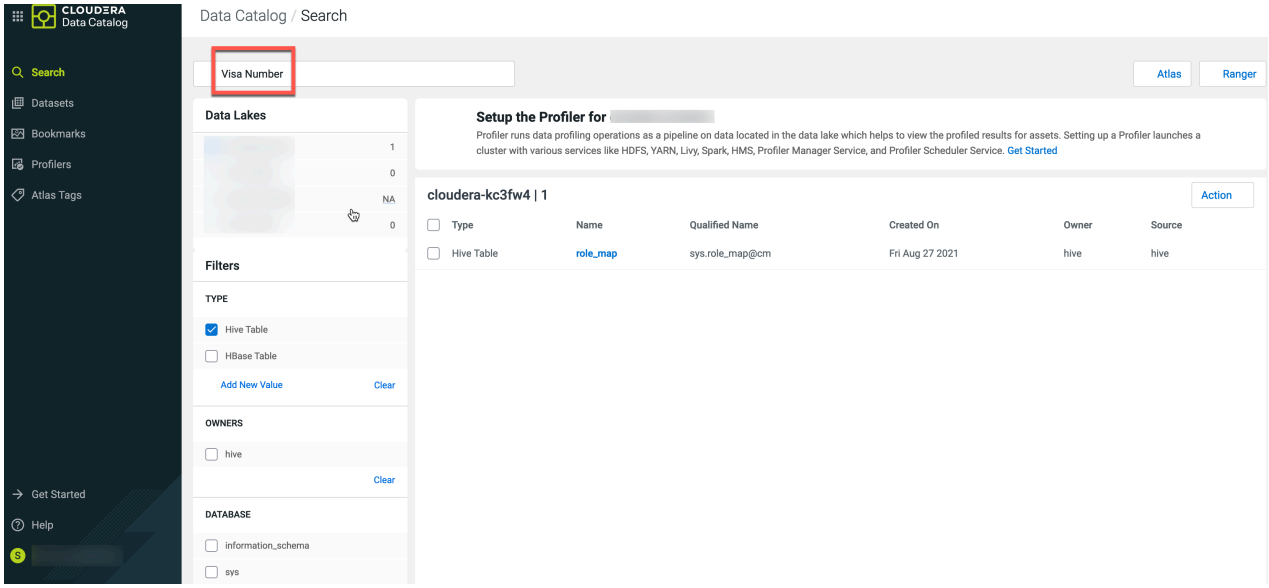


Searching for assets in Data Catalog using additional search options

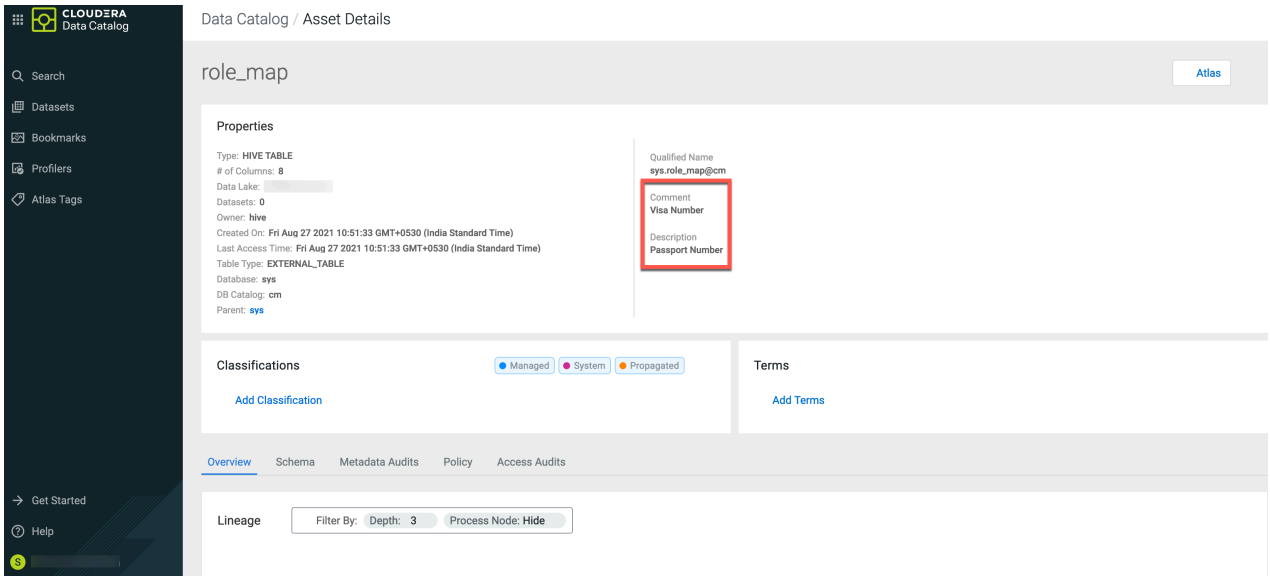
Consider a scenario in Data Catalog, where you select a data asset type and under the Asset Details page, you insert a comment and provide the description for the selected asset.

Navigate to the Data Catalog search query pane and enter the Comment and Description value(s) that you saved for the selected asset type in Data Catalog. The result page displays the asset type that you added for the Comment and Description fields in Data Catalog.

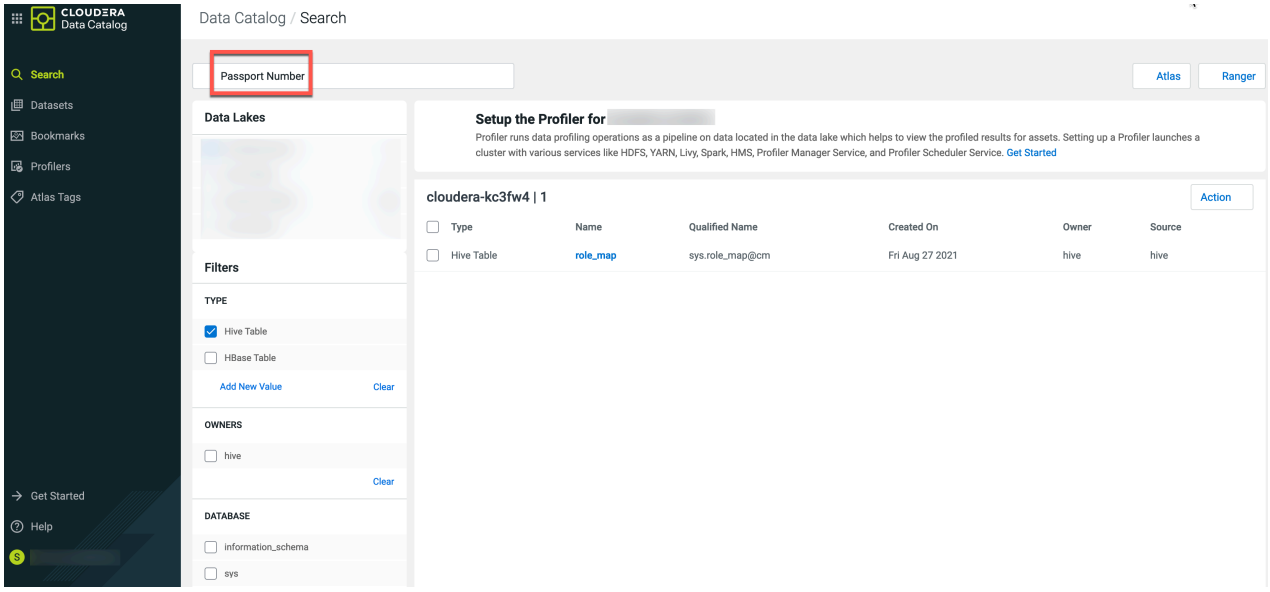
When you query for the entered Comment value for the selected asset type in Data Catalog, the relevant asset type is displayed in the search result page.



Clicking on the asset type in Data Catalog displays the comment and description values as it was assigned in Data Catalog.



When you query for the entered Description value for the selected asset type in Data Catalog, the relevant asset type is displayed in the search result page.



Your search query displays the results.

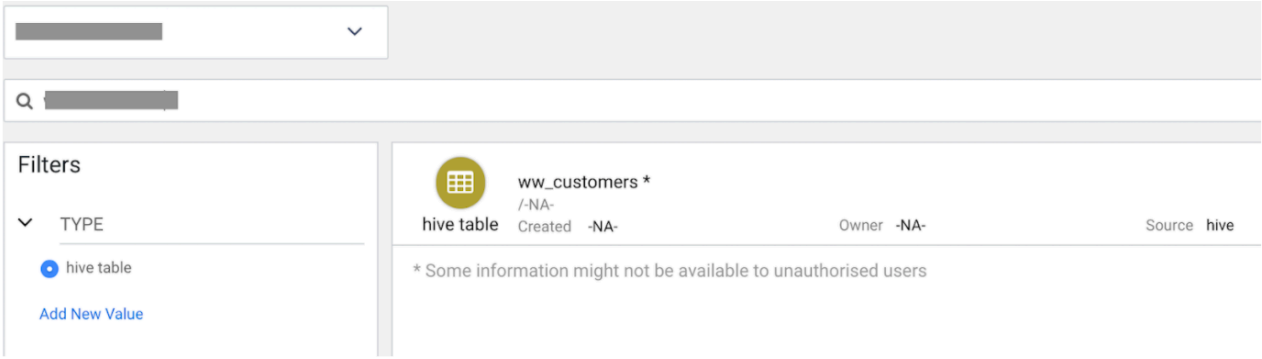
Accessing Tables based on Ranger policies

In Data Catalog service, when a table (in blue color link) is clicked, the Asset Details view page is displayed.

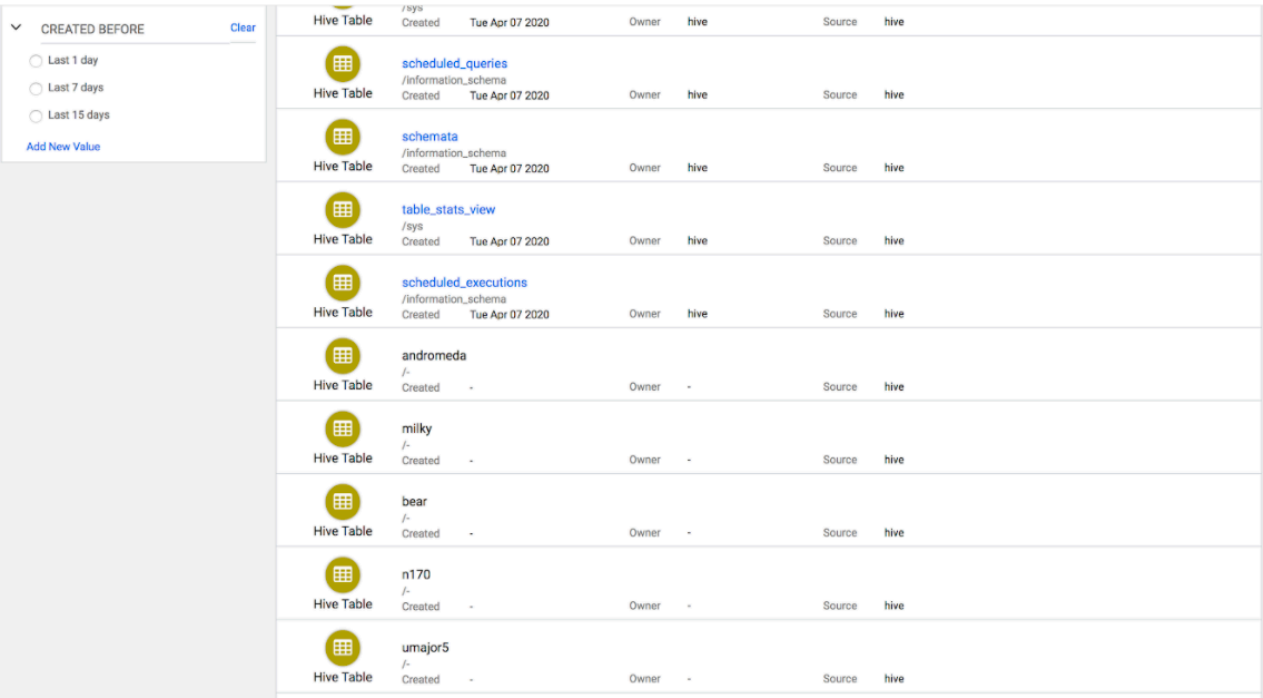
If a user is not authorized to click or view table details, it implies that the user permissions have not been set-up in the Ranger.

Note: The user permissions to view table details are configured in 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 embedded in blue color link. And the ones that do not have read permissions are visible in grey.



Creating Classification for selected assets

You can create a classification that can be associated with an asset.

1. From Data Catalog > navigate to the search page.
2. You can perform one or more of the following:
   - Select Add Classifications on action button in search page
   - Select Add classification in classification widget on Asset Details page.
3. On the Add Classification slider, click Create button.
4. Enter the necessary values in the fields and click the Create button.

Adding Classifications / Terms for selected assets

You can add classification or terms that can be associated with an asset.

Procedure

1. From Data Catalog > navigate to the search page.

2. You can perform one or more of the following:
   a) Select Add Classifications / Terms on action button in the search page.
   b) Select Add Classifications / Terms in classification widget on Asset Details page.
3. On the Add Classifications / terms slider, click on the Add icon against classification / term.
4. Enter other values in the fields, if required and click Save.

Additional Entity type selection for searching Assets

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

Data Catalog users can search and discover assets of more types. Users can search assets of types just like they do for Hive Table with some restrictions.

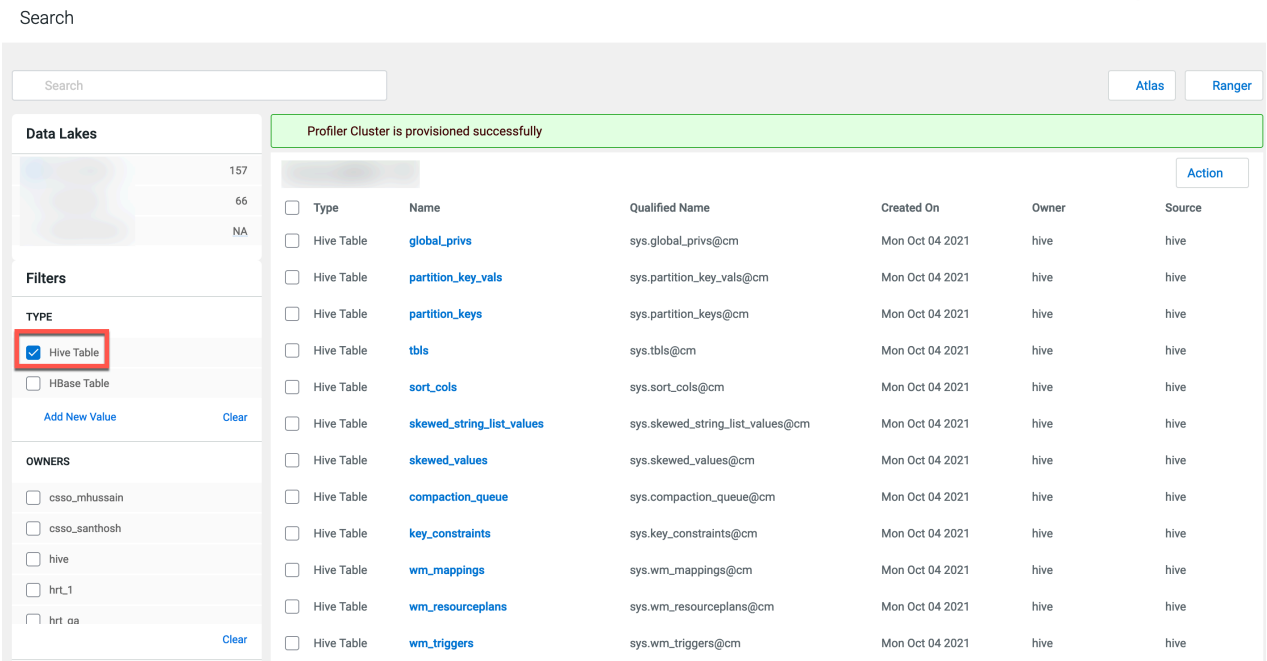
Supported entity types include:

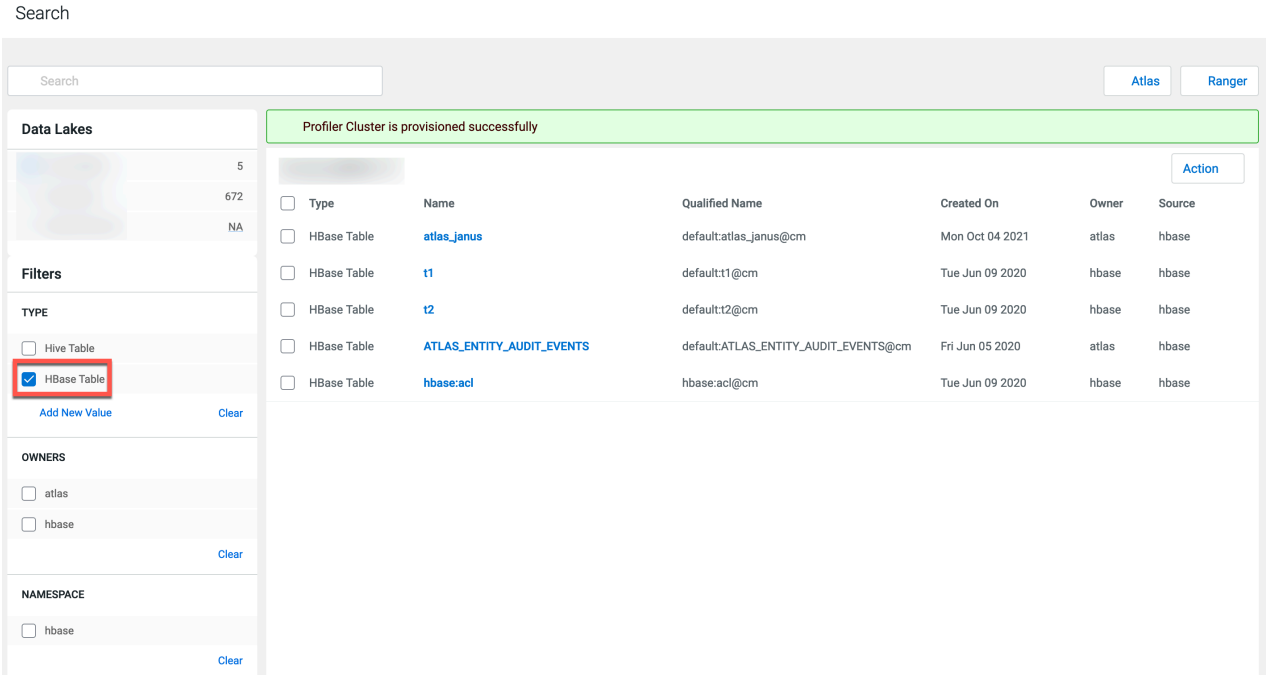
- AWS S3 Object
- AWS S3 Bucket
- AWS S3 Pseudo Dir
- HBase Table
- HBase Column Family
- HBase Namespace
- HDFS Path
- Hive DB
- Hive Table
- Hive Column
- ML Project
- ML Model Build
- ML Model Deployment
- NiFi Flow
- NiFi Data
- Impala Process
- Impala Column Lineage
- Impala Process Execution
- 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
- Spark ML Directory
- Spark ML Model
- Spark ML Pipeline
- Spark Process Execution
- Spark Table

Selecting a type triggers a search query for that type. Currently two types of entities are supported but totally about twelve types of generic entities can be selected to search for assets depending on the data lake.

Owners data is derived from the response received from type based queries.

The following example diagrams depict the entity type selection search results.





Managing Profilers

The Data Catalog profiler engine runs data profiling operations as a pipeline 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.

Data Catalog profiler data testing

You must note the important information about profiler services.

The Data Catalog profilers are not tested at par with the Hive scale, The following dataset has been validated and works as expected

- 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

Launch profiler Cluster

You must launch the Profiler cluster to view the profiler results for your assets and datasets. You must be a Power User to launch Profiler cluster.

About this task

A new user interface which is introduced to launch profilers in Data Catalog. The Profiler Services is now supported by 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 CDP environment.

Attention: By default when you launch a Profiler cluster, the instance type of the Master node will be:

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

Note: This is applicable for the latest build of Data Catalog version 2.0.17: 2.0.17-b26

There are two types of Profiler Services:

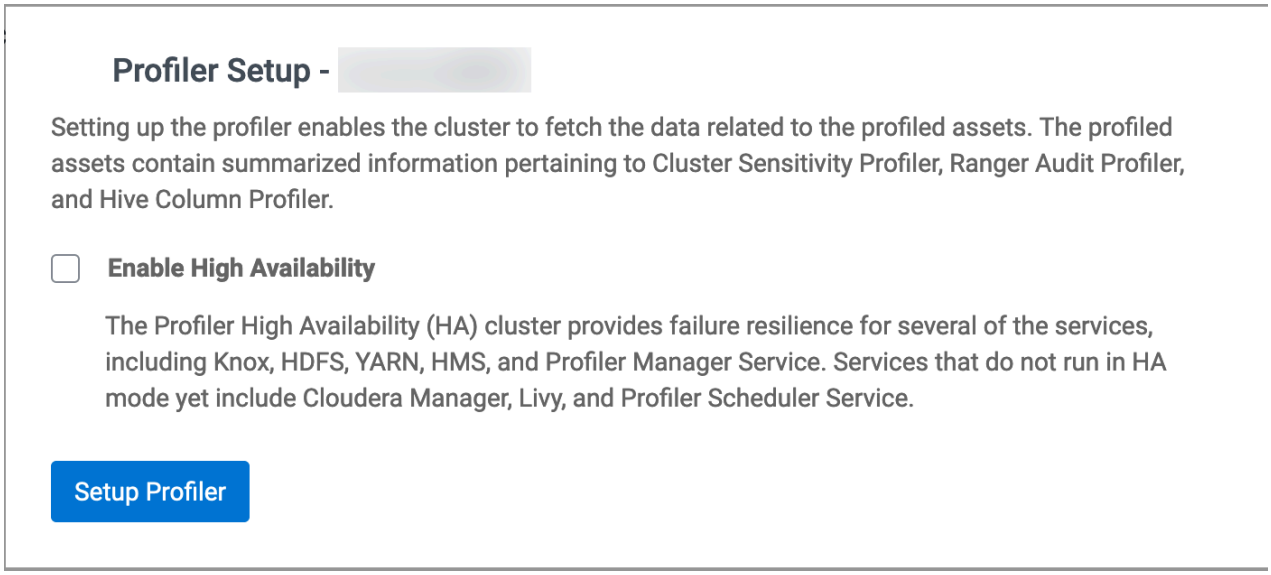
- Profiler Manager
- Profiler Scheduler

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.

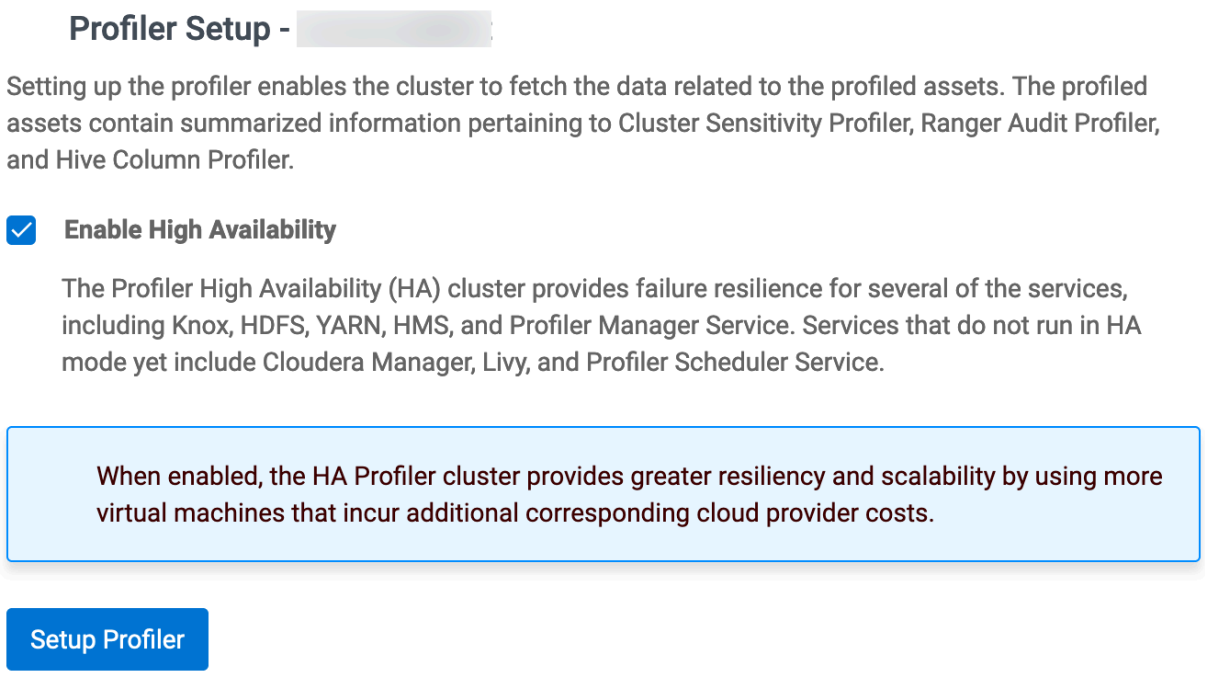
How to launch the cluster profiler

On the Data Catalog search page, select the data lake from which you want to launch the profiler cluster. On the right-hand side of the window, the application displays the page to set up the profiler for the selected data lake. Click the Get Started link to proceed.

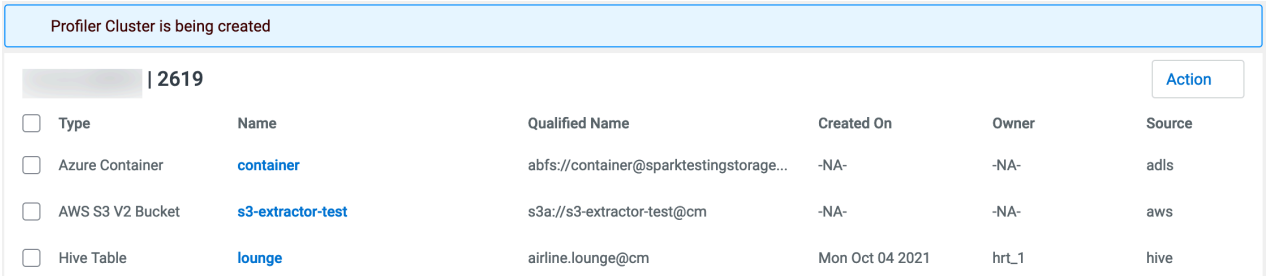


For setting up the profiler, you have the option to enable or disable the HA.

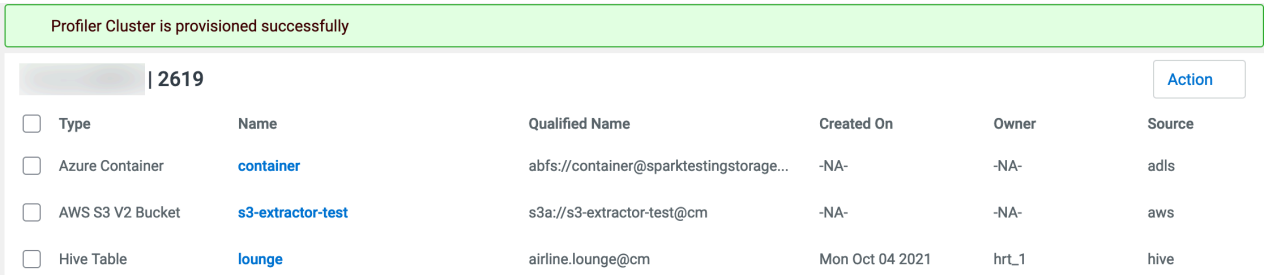
Note that the HA functionality is being supported only from Cloudera Runtime 7.2.10 release onwards. If you are using the Cloudera Runtime version below 7.2.10, you shall not be able to use the HA feature for launching the profiler services.



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

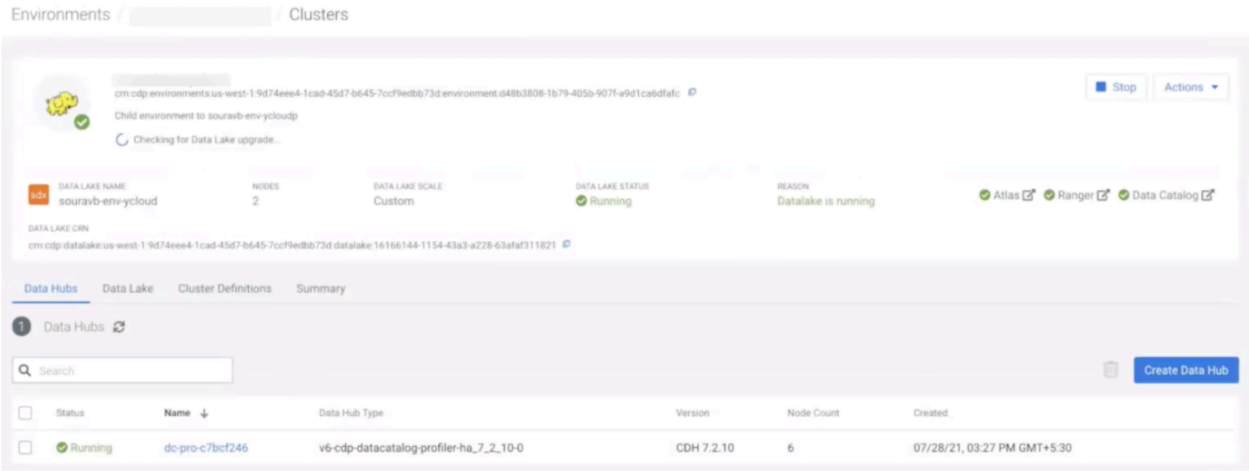


Later, a confirmation message appears that the profiler cluster is created.



Next, you can verify the profiler cluster creation under CDP Management Console > Environments > DataHubs pane.

Note that the newly created profiler cluster has some unique representations under the following categories:



- Data Hub Type - The term “ha” is appended to the type of cluster that is newly created.
- Version - 7.2.10
- Node Count - (Which is 6)

Your Profiler cluster with HA is set up successfully.

Launching profilers using Command-line

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

This, apart from launching the profilers using the Data Catalog UI. The CLI will be one executable and will not have any external dependencies. You can execute some operations in the Data Catalog service using the CDP CLI commands.

Users must have valid permission(s) to launch profilers on a data lake.

For more information about the access details, see Prerequisites to access Data Catalog service.

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

DATA_CATALOG_ENABLE_API_SERVICE

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

In your CDP 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 Data Catalog.

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 Name or CRN of the Datalake.

```
--cli-input-json  (string) Performs service operation based on the JSON stri
ng provided. The JSON string follows the format provided by --generate-cli-s
keleton. 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 out
put. 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.

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

cdp datacatalog launch-profilers --datalake <datalake name or datalake CRN>

Example

cdp datacatalog launch-profilers --datalake test-env-ycloud

{

"datahubCluster": {

"clusterName": "cdp-dc-profilers-24835599",

```
        "crn":
           "crn:cdp:datahub:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:
cluster:dfaa7646-d77f-4099-a3ac-6628e1576160",
```

"creationDate": "2021-06-04T11:31:23.735000+00:00",

"clusterStatus": "REQUESTED",

"nodeCount": 3,

"workloadType": "v6-cdp-datacatalog-profiler_7_2_8-1",

"cloudPlatform": "YARN",

"imageDetails": {

```
            "name":
           "docker-sandbox.infra.cloudera.com/cloudbreak/centos-76:2020-05
-18-17-16-16",
```

"id": "d558405b-b8ba-4425-94cc-a8baff9ffb2c",

```
            "catalogUrl":
           "https://cloudbreak-imagecatalog.s3.amazonaws.com/v3-test-cb-imag
e-catalog.json",
```

"catalogName": "cdp-default"

},

```
        "environmentCrn":
           "crn:cdp:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb
73d:environment:bf795226-b57c-4c4d-8520-82249e57a54f",
```

```
        "credentialCrn":
           "crn:altus:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edb
b73d:credential:3adc8ddf-9ff9-44c9-bc47-1587db19f539",
```

```
        "datalakeCrn":
           "crn:cdp:datalake:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d
:datalake:5e6471cf-7cb8-42cf-bda4-61d419cfbc53",
```

```
        "clusterTemplateCrn":
```

```
            "crn:cdp:datahub:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:c
lustertemplate:16a5d8bd-66d3-42ea-8e8d-bd8765873572"
```

}

}

Deleting profiler cluster

Deleting profiler cluster 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

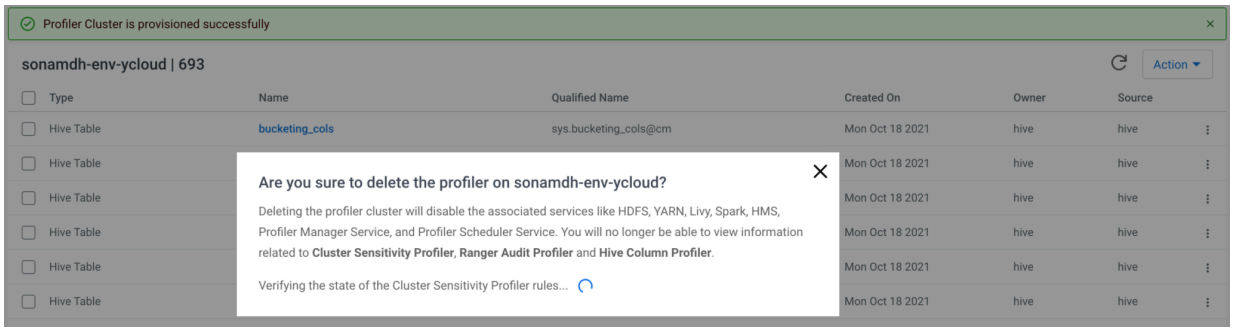
To overcome this situation, when you decide to delete the profiler cluster, there is a provision to retain the status of the Custom Sensitivity Profiler rules. If your profiler cluster has rules that are not changed or updated, you can directly delete the profiler cluster. If the rules were modified 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 cluster.

Procedure

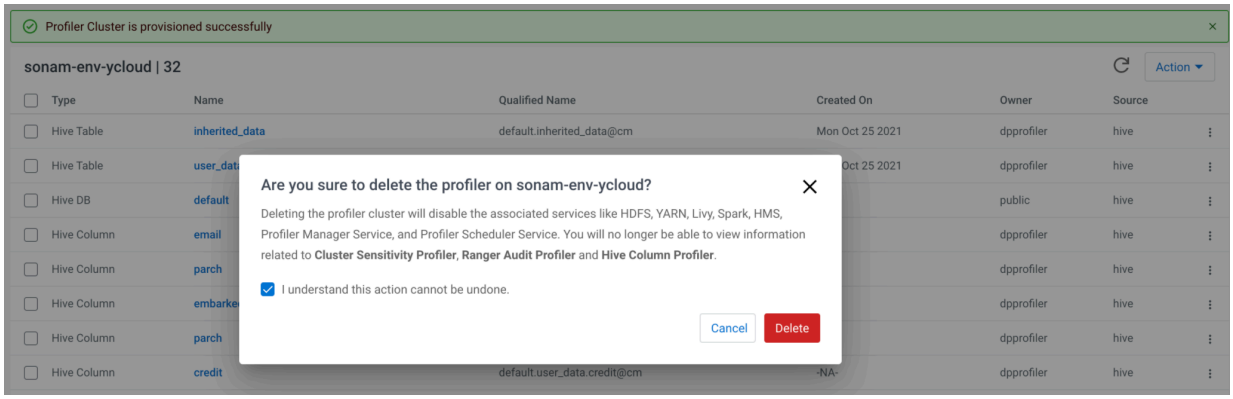
1. On the search page, select the Data Lake from the list.
2. Click the Actions drop-down menu and select Delete Cluster.

3. Click Yes to proceed.

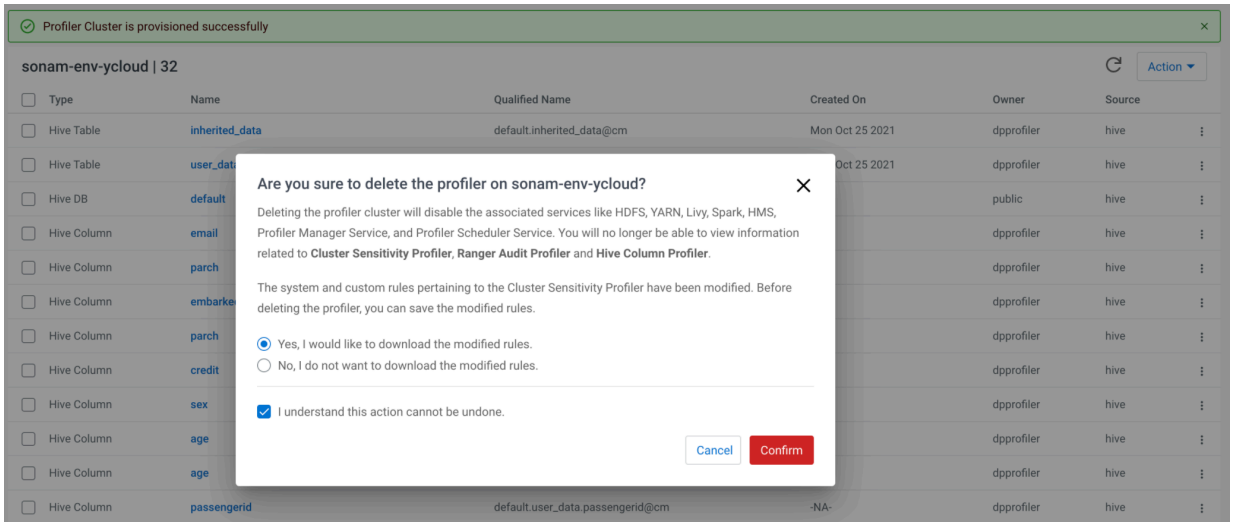
   The application displays the following message.



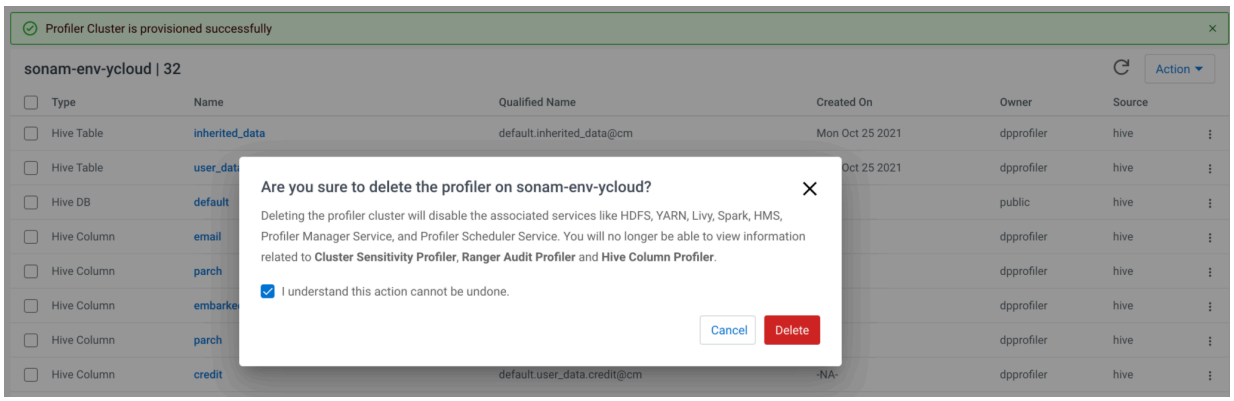
When you upgrade / launch Data Catalog service in Cloudera Runtime version 7.2.14, and later if the profiler cluster is deleted, the following message is displayed.



Note: Using Data Catalog with Cloudera Runtime version 7.2.12 and below, and later when you delete a profiler cluster that has modified Custom Sensitivity Profiler rules, the following message is displayed.



While using Data Catalog with Cloudera Runtime version 7.2.12 and below, and if the profiler cluster does not have any modified Custom Sensitivity Profiler rules, the following message is displayed.



The profiler cluster is deleted successfully.

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

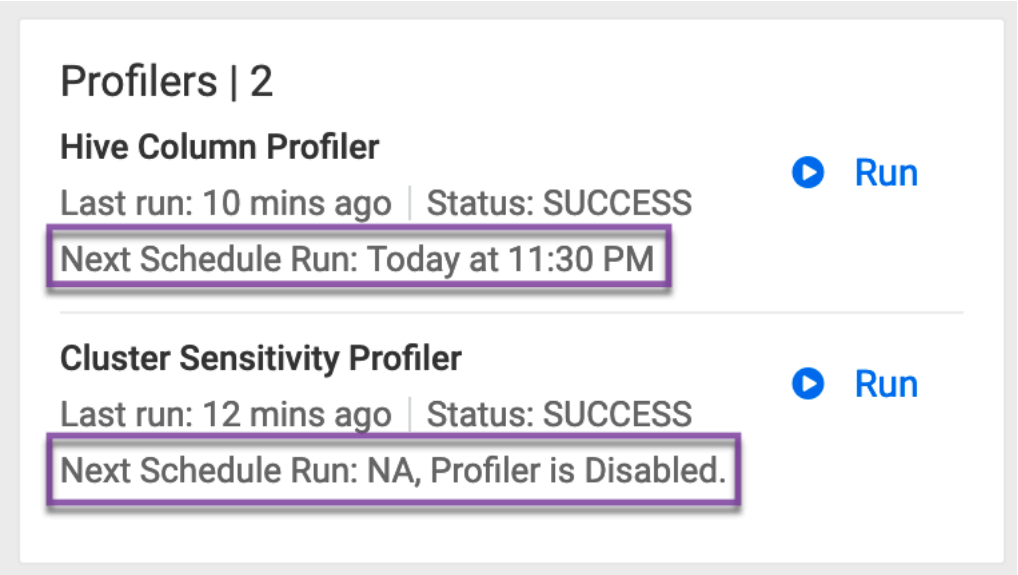
Note: If the profiler cluster is not registered with the data lake, Data Catalog cannot locate or trace the profiler cluster. Users have to delete the profiler cluster from the DataHub page (Cloudera Management Console).

On-Demand Profilers

You can use on-demand profilers to profile specific assets without depending on the cron-based scheduling of profilers jobs. On-demand profiler option is available on the asset details page of the selected asset.

For example, the diagram displays the Asset Details page of an asset. Run On-Demand profiler for Hive Column Statistics and Custom Sensitivity Profiler by clicking on the appropriate Run button. The next scheduled run provides details about the next scheduled profiling for the respective profilers.

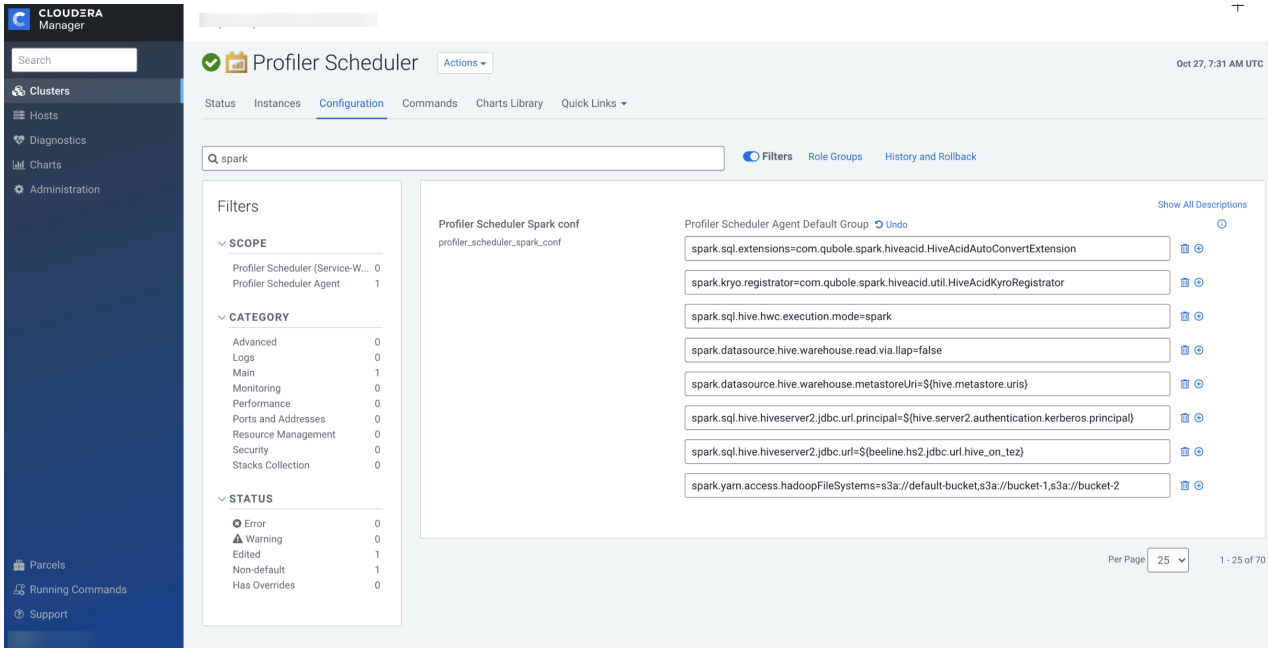
Note: You can use the On-Demand Profiler feature to profile both External and Managed tables.



Profiling table data in non-default buckets

You must configure a parameter in Profiler Scheduler in your Cloudera Manager instance, to profile table data in non-default buckets.

Configuration > Search for "spark" in the filters field > Profiler Scheduler Spark conf > Add spark.yarn.access.hadoopFileSystems=s3a://default-bucket,s3a://bucket-1,s3a://bucket-2



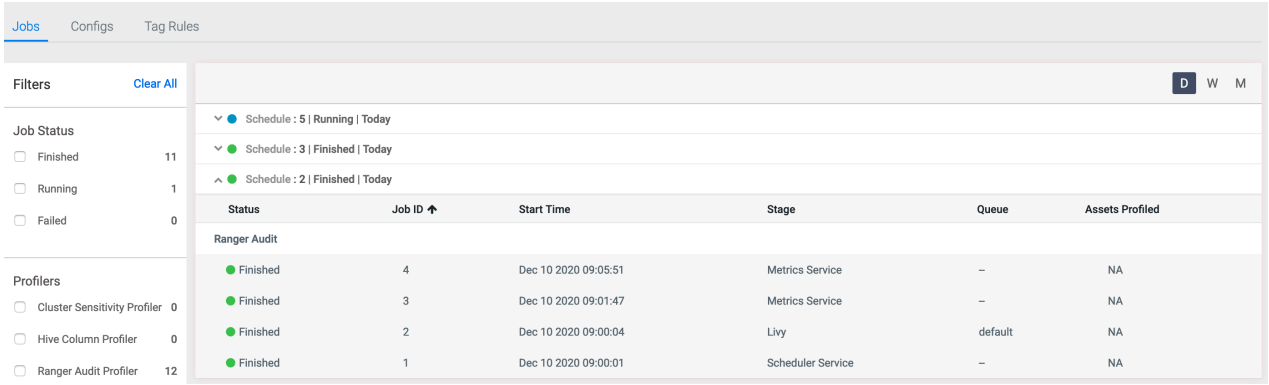
Note: bucket-1 and bucket-2 are non-default buckets.

Attention: For more information, see Accessing data stored in Amazon S3 through Spark.

Tracking Profiler Jobs

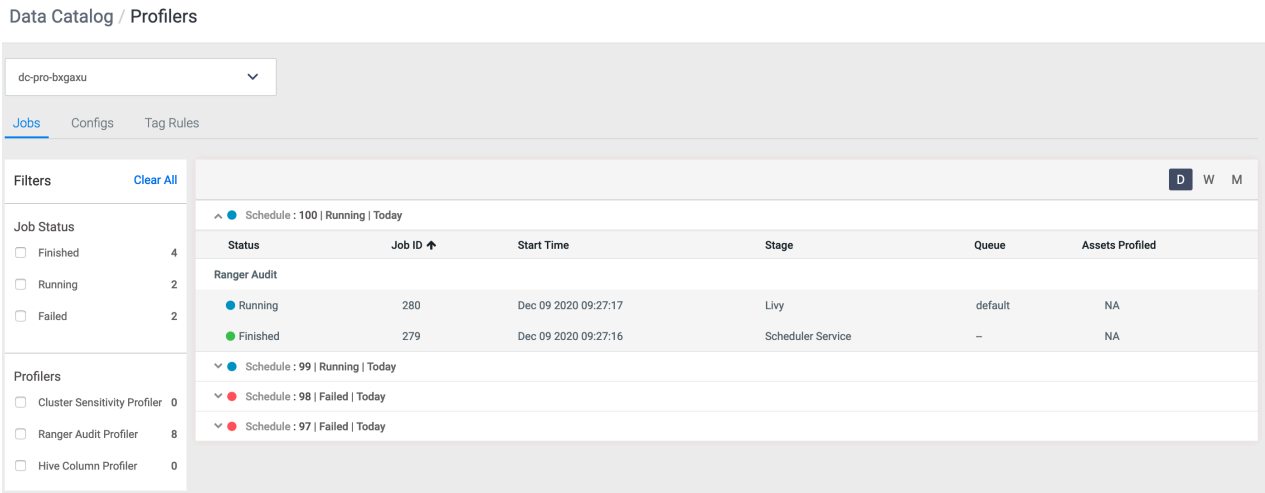
The Data Catalog profiler page is updated to provide a better user experience for tracking respective profiler jobs.

A new placeholder named “Schedule” is introduced under the Profilers section to provide tracking information of each profiler job. Under Schedule, you can find the type of profiler job that has run or in progress or has completed profiling data assets.



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

- Job Status
- Type
- Job ID
- Start Time
- Stage
- Job Queue
- Total assets profiled



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. Whenever the Schedule status appears in green, it indicates that the profiler job has run successfully. When the color appears in blue and red, it indicates that the profiler job is running or has failed.

Profiler job runs in three phases:

- Scheduler Service - 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 a 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.

Clicking on each profiled asset would navigate to the profiled asset details page. The asset profiled page provides information about the profiled asset, profiled status, the profiled job id, and other relevant details.

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.

Important: Currently, the On-Demand schedule is not supported for this version of the profiler. The job schedule is either grayed out or disabled in such a scenario.

Viewing Profiler Jobs

You can monitor the overall health of your profiler jobs by viewing their status on the Profiler Jobs .

Each profiler runs a Spark job on a user-defined schedule defined via the profiler configuration. You can view the status of each of those jobs for all your clusters.

Monitoring the profiler jobs has the following uses:

- By seeing long-term trends in job execution, you can determine the overall health of your profilers.
- If you do a data ingest, you can find out if the profiling has completed.
- Knowing when jobs first failed can help when troubleshooting problems with profilers.

You can take the following actions:

1. Filter by cluster, job status, or profiler.
2. Sort by jobs ID, status, start time, cluster, queue, or profilers.
3. Expand or narrow to show a day, week, or month of jobs.

Viewing Profiler Configurations

You can monitor the overall health of individual profilers by viewing their status on Profiler Configs .

Monitoring the profiler configurations has the following uses:

- See which profilers are active and inactive.
- View asset coverage for a particular profiler over time- for instance, if you change a configuration for a profiler, you can see if new assets become covered.

You can take the following actions:

1. Filter by cluster.
2. Expand the execution status of an individual profiler.
3. Edit the profiler configuration.
4. Toggle each profiler on/off.

Edit Profiler Configuration

In addition to turning on and off the profiler configurations, the individual profilers can be run with their own execution parameters. These parameters are for submission of the profiler job onto Spark. You can edit the configuration of profilers and update these parameters to run profiler jobs.

Procedure

1. Click Profilers in the main navigation menu on the left.
2. Click Configs to view all of the configured profilers.
3. Select the cluster for which you need to edit profiler configuration.

   The list of profilers for the selected clusters is displayed.
4. Click the name of the profiler whose configuration you wish to edit.

   The Profiler Configuration tab is displayed in the right panel.
5. Select a schedule to run the profiler. This is implemented as a quartz cron expression.

   For more information, see Understanding Cron Expression generator on page 32.
6. Select Last Run Check.

Last Run Check configuration enables profilers like Hive Column Statistics and Cluster Sensitivity Profiler to avoid profiling the same asset on each scheduled run. If you have scheduled a cron job, say for about an hour, and have enabled the Last Run Check configuration for two days, this set-up 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, if any.

Caution: This configuration is not applicable to the Ranger Audit Profiler.

7. Update the advanced options.
   - 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.

   For more information, see Configuring SPARK on YARN Applications and Tuning Resource Allocation.
8. Toggle the state of the profiler from Active to Inactive as needed.
9. Click Save to apply the configuration changes to the selected profiler. The changes should appear in the profiler description.

Additional Configuration for Cluster Sensitivity Profiler

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

Procedure

1. Click Profilers in the main navigation menu on the left.
2. Click Configs to view all of the configured profilers.
3. Select the cluster for which you need to edit profiler configuration.

   The list of profilers for the selected clusters is displayed.
4. Click the Cluster Sensitivity Profiler to edit.

   The Profiler Configuration tab is displayed in the right panel.
5. Toggle the Enable button to enable Cluster Sensitivity Profiler. Select Disable if you do not want to run the Cluster Sensitivity Profiler.
6. Select the Sample Data Size.
   a) From the drop down, select the type of sample data size.
   b) Enter the value based on the previously selected type.
7. Select the queue, schedule, and advanced configuration details as specified in Edit Profiler Configuration.
8. Add Asset Filter Rules as needed to customize the selection and deselection of assets which the profiler profiles. For more information, see Setting Asset filter rules on page 32.
9. Toggle the state of the profiler from Active to Inactive as needed.
10. Click Save to apply the configuration changes to the selected profiler. The changes should appear in the profiler description.

Additional Configuration for Hive Column Profiler

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

Procedure

1. Click Profilers in the main navigation menu on the left.
2. Click Configs to view all of the configured profilers.
3. Select the cluster for which you need to edit profiler configuration.

   The list of profilers for the selected clusters is displayed.
4. Click the Hive Column Profiler to edit.

   The Profiler Configuration tab is displayed in the right panel.
5. Select the queue and schedule details as specified in Edit Profiler Configuration on page 30.

6. Add Asset Filter Rules as needed to customize the selection and deselection of assets which the profiler profiles. For more information, see Setting Asset filter rules on page 32.

   Note: The schedule for Hive Column Profiler is set to run once every six hours. After installation, you will be able to see the output of Hive Column Profiler after six hours. If you want to view the output in advance, update the cron expression accordingly.

7. Select the Sample Data Size.
   a. From the drop down, select the type of sample data size.
   b. Enter the value based on the previously selected type.
8. Click Save to apply the configuration changes to the selected profiler. The changes should appear in the profiler description.

Understanding Cron Expression generator

A cron expression details about when the schedule executes and visualizes the next execution dates of your cron expression. The cron expression utilizes the quartz engine.

The cron expression uses a typical format as seen in the table:

```
Cron Expression: * * * *        * ? *
```

Each * in the cron represents a unique value.

For example, consider a cron with the following values:

```
Cron Expression: 1 2 3 2 5 ?       2 0 2 1
```

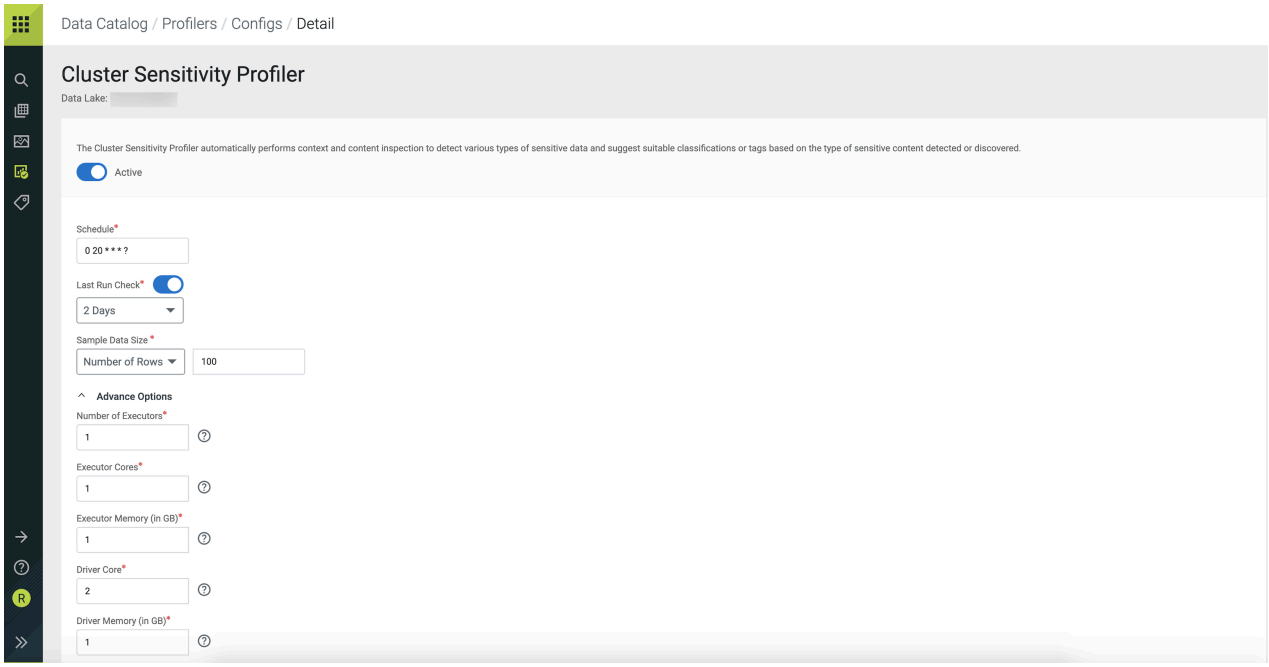
The cron is scheduled to run the profiler job at: 03:02:01am, on the 2nd day, in May, in 2021.

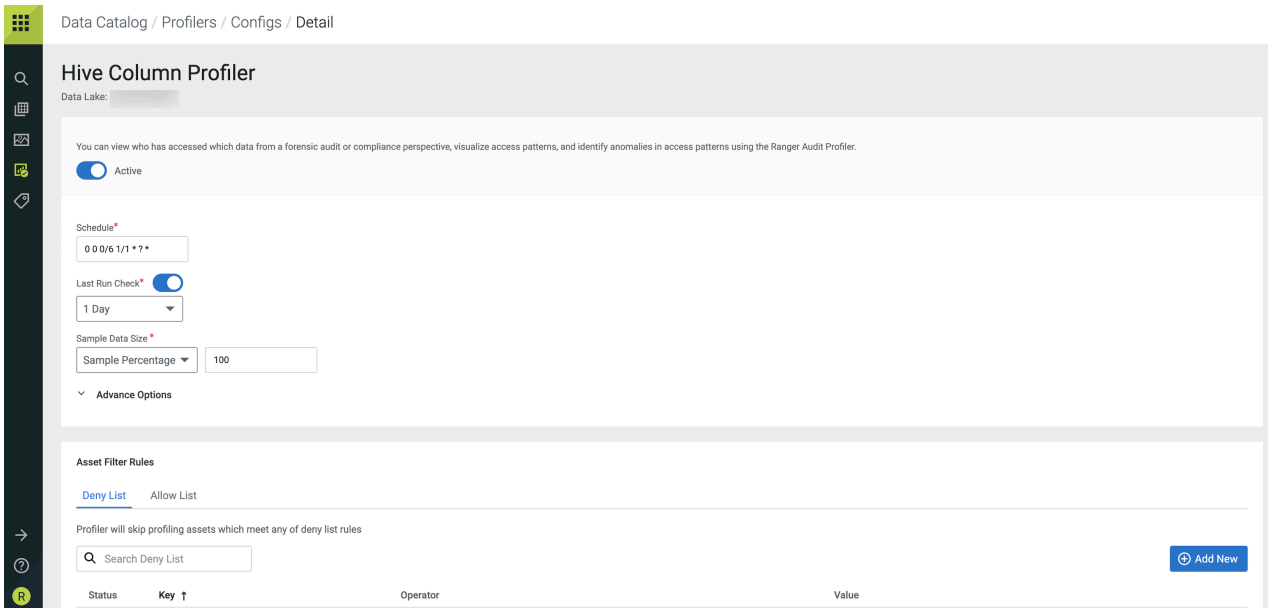
You can change the value of cron as and when it is required depending on how you want to schedule your profiler job.

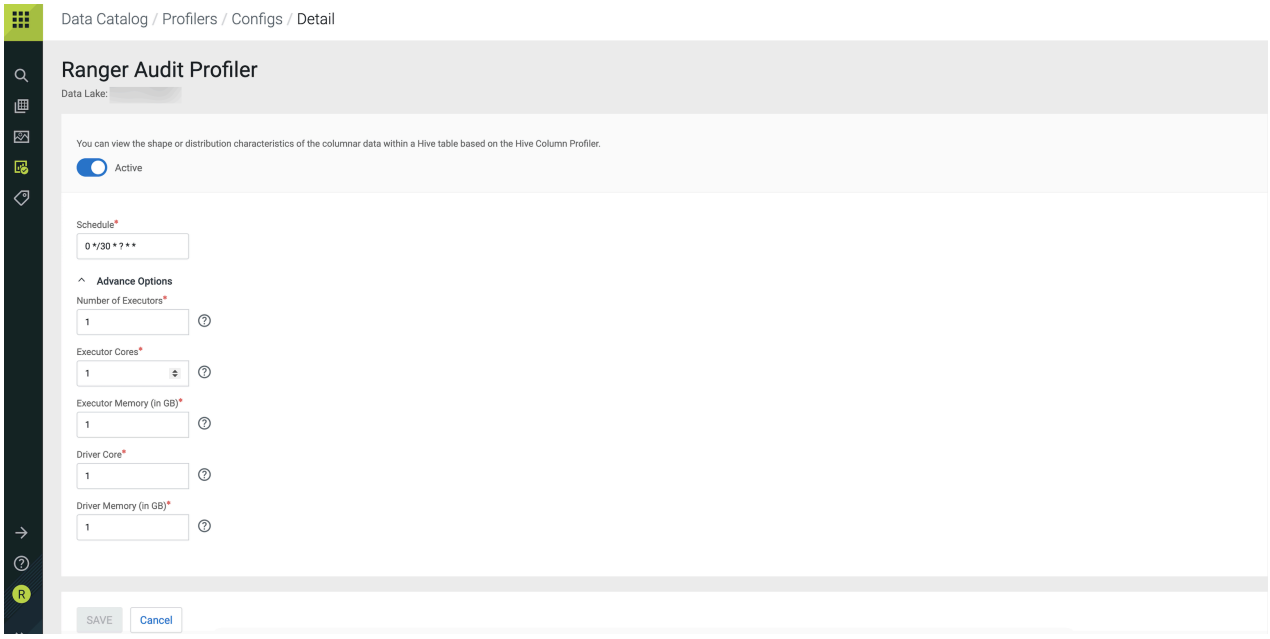
Setting Asset filter rules

Add Asset filter rules as needed to customize the selection and deselection of assets which the profiler profiles.

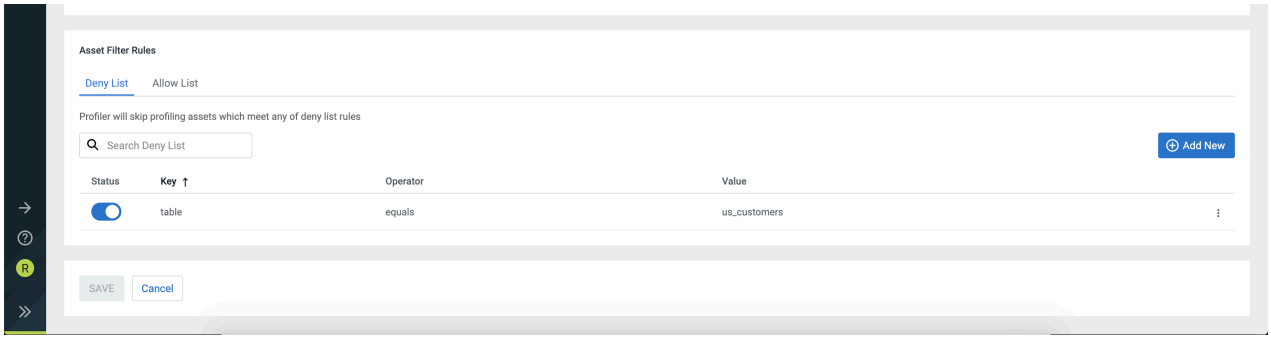
Note: You can configure the Deny-list and Allow-list for both Cluster Sensitivity Profiler and Hive Column Profiler. The same filter rules do not apply to Ranger Audit Profiler.

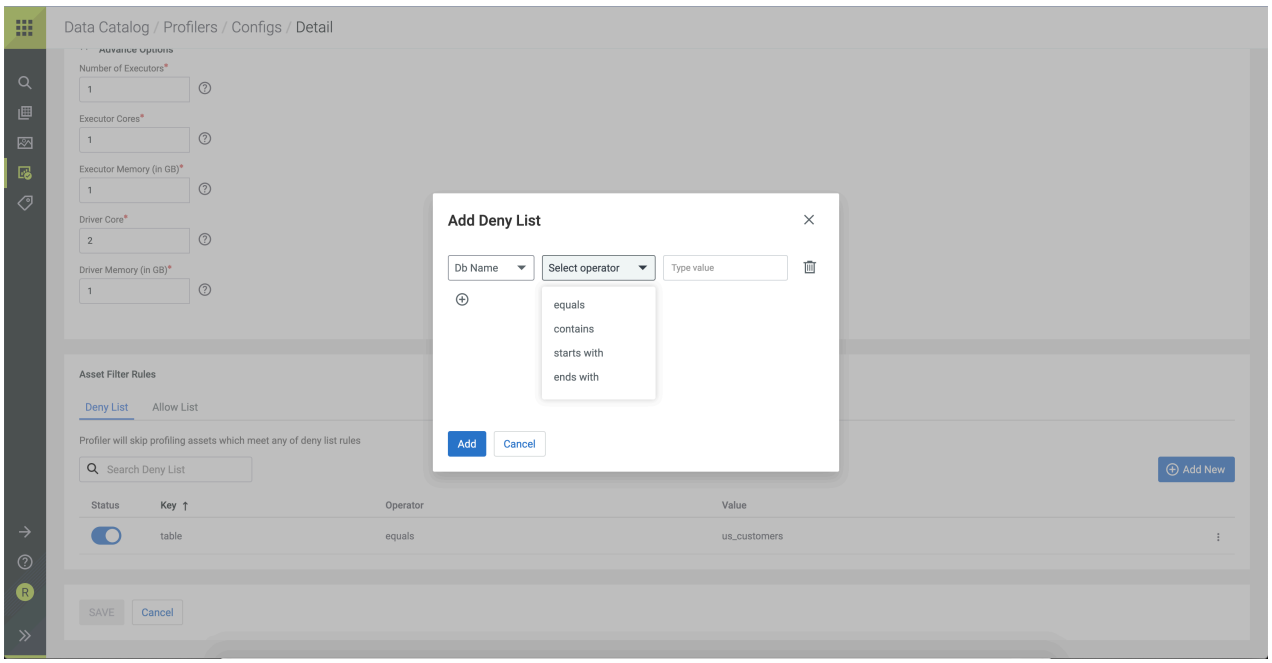

Data Catalog / Profilers / Configs / Detail
Cluster Sensitivity Profiler
Data Lake:
The Cluster Sensitivity Profiler automatically performs context and content inspection to detect various types of sensitive data and suggest suitable classifications or tags based on the type of sensitive content detected or discovered.
Active
Schedule*
0 20 * * * ?
Last Run Check*
2 Days
Sample Data Size *
Number of Rows
100
Advance Options
Number of Executors*
1
Executor Cores*
1
Executor Memory (in GB)*
1
Driver Core*
2
Driver Memory (in GB)*
1


Data Catalog / Profilers / Configs / Detail
Hive Column Profiler
Data Lake:
You can view who has accessed which data from a forensic audit or compliance perspective, visualize access patterns, and identify anomalies in access patterns using the Ranger Audit Profiler.
Active
Schedule*
0 0 0/6 1/1 * ? *
Last Run Check*
1 Day
Sample Data Size *
Sample Percentage
100
Advance Options
Asset Filter Rules
Deny List
Allow List
Profiler will skip profiling assets which meet any of deny list rules
Search Deny List
Add New
Status
Key
Operator
Value

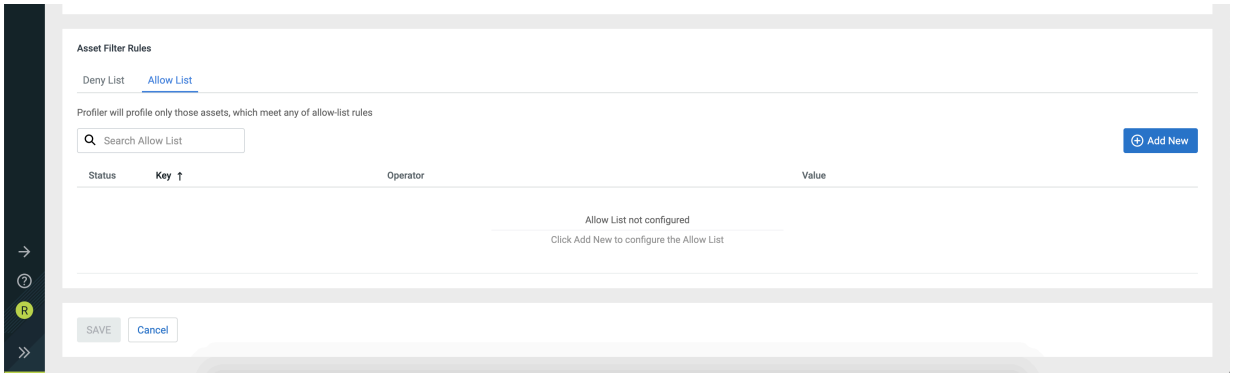


- Deny-list - The profiler will skip profiling assets that meet any defined Deny-list criteria.
  - Select the Deny-list tab.
  - Click Add New to include rules for Deny-list.
  - Select the key from the drop down list. You can select a database name, name of the asset, name of the owner of the asset, path to the assets, or created date.
  - 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.
  - Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example here.
  - Click Done. Once it is added, you can toggle the state of the new rule to enable it or disable it as needed.





- Allow-list - The profiler will include only assets that satisfy any defined Allow-list criteria. If no Allow-list is defined, the profiler will profile all the assets.
  - Select the Allow-list tab.
  - Click Add New to include rules for the Allow-list.
  - Select the key from the drop down list. You can select a database name, name of the asset, name of the owner of the asset, path to the assets, or created date.
  - 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.
  - Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
  - Click Done. Once it is added, you can toggle the state of the new rule to enable or disable it as needed.



Note: If an asset meets both Allow-list and Deny-list rules, the Deny-list rule overrides the Allow-list.

Backing up and Restoring Profiler Database

Using certain scripts that can be executed by the root users, you can take the backup of the Profiler databases. Later, if you want to delete the existing DataHub cluster and launch a new DataHub cluster, you must have an option to restore the old data.

Data Catalog includes Profiler services that run data profiling operations on data that is located in multiple Data Lakes. As of the latest Cloudera Runtime release, the Profiler services run on a DataHub cluster. When you delete the DataHub cluster, the profiled data and the user configuration information stored in the local databases are lost.

Profiler clusters run on the DataHub cluster using a couple of embedded databases - profiler_agent and profiler_metrics.

Note: If you download the modified Custom 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 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 preserve Profiler data in the DataHub cluster. You might want to delete the DataHub cluster but preserve the Profiler data.
- When you want to re-launch the Profiler and access the older processed data. You might want to delete a DataHub cluster but preserve the Profiler data of the DataHub cluster.
- **Note:** For users using Data Catalog on Cloudera Runtime 7.2.14 version, note the following
  - No user action or manual intervention needed after the upgrading DataHub cluster to 7.2.14 version is completed.
  - 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 belowCloudera Runtime 7.2.11 version to 7.2.11:

Navigate 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 below or equal to Cloudera Runtime 7.2.11 version to 7.2.12:

Navigate 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 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 script

When you run the script, note that there are a couple of phases through which you can accomplish your task.

Firstly you backup your Profiler database and next you can restore the Profiler database.

To backup the Profiler database:

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

   Attention: For users of Cloudera Runtime below 7.2.8 version, contact Cloudera Support.

   Note: If the profiler cluster having the Cloudera Runtime version 7.2.11 or below, you must run the following command:

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

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

   ```
   bash /opt/cloudera/parcels/CDH/lib/profiler_manager/profiler
   admin/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.

To restore the Profiler database:

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

   Attention: For users of Cloudera Runtime below 7.2.8 version, contact Cloudera Support.

   Note: If the profiler cluster having the Cloudera Runtime version 7.2.11 or below, you must run the following command:

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

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

   ```
   bash /opt/cloudera/parcels/CDH/lib/profiler_manager/profiler
   admin/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 detailed:

- 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

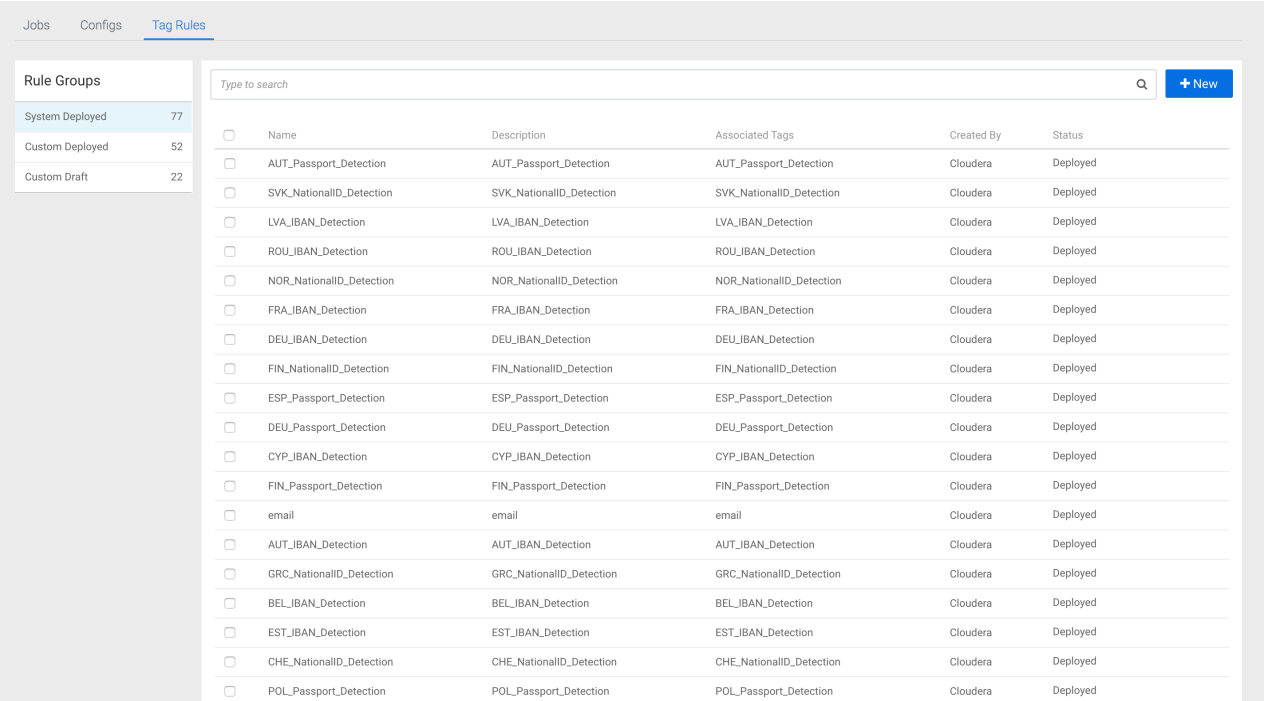
1. From Profiler Configs
2. Select the profiler to proceed further.

Profiler Tag Rules

You can use preconfigured tag rules or create new rules based on regular expressions and allow or deny files on specific columns in your tables.

Rules are categorized into three groups:

- System Deployed : These are in-built 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.



Tag Management

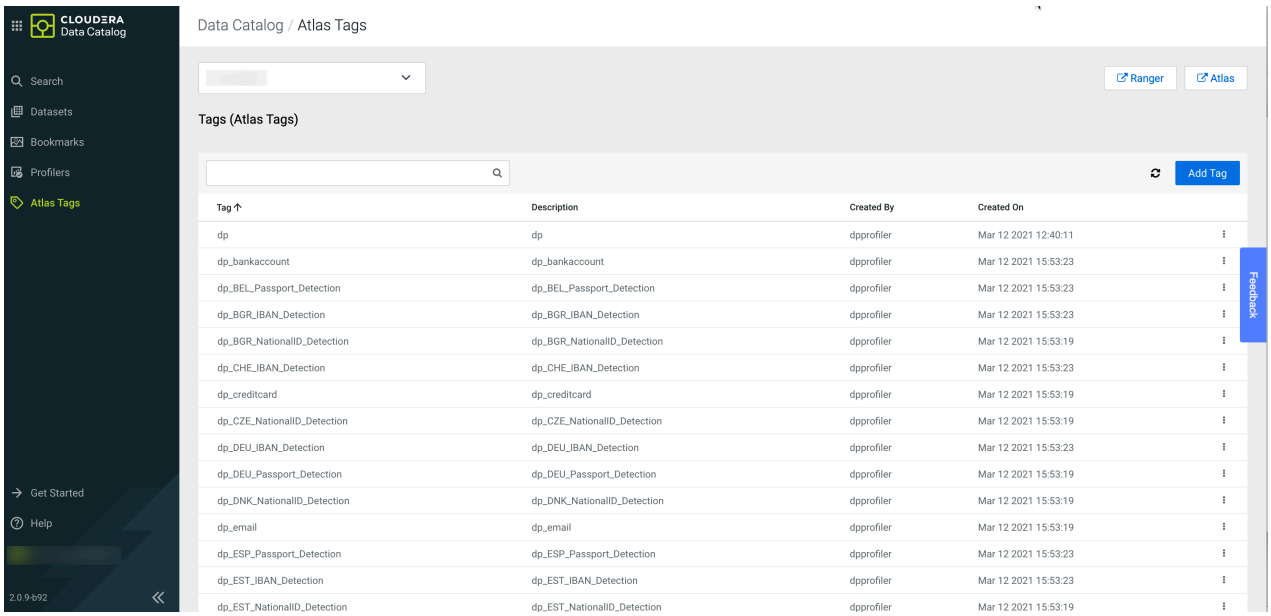
From Atlas tags UI in Data Catalog, you can create, modify, and delete any of the Atlas tags in a Data Catalog instance.

You can access the Atlas link by logging into Data Catalog > Atlas Tags .

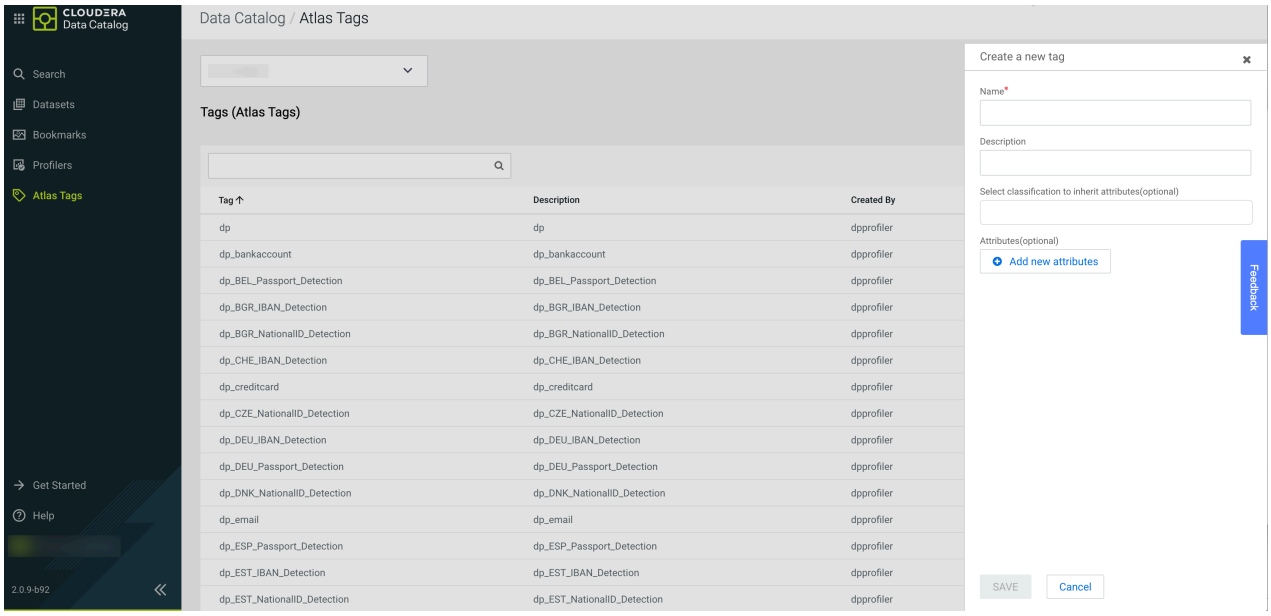
Atlas Tags allows the user to perform the following activities with a selected Data Lake for tag management:

- Selecting a Data Lake
- Searching for a tag
- Adding a tag
- Editing a tag
- Deleting a tag

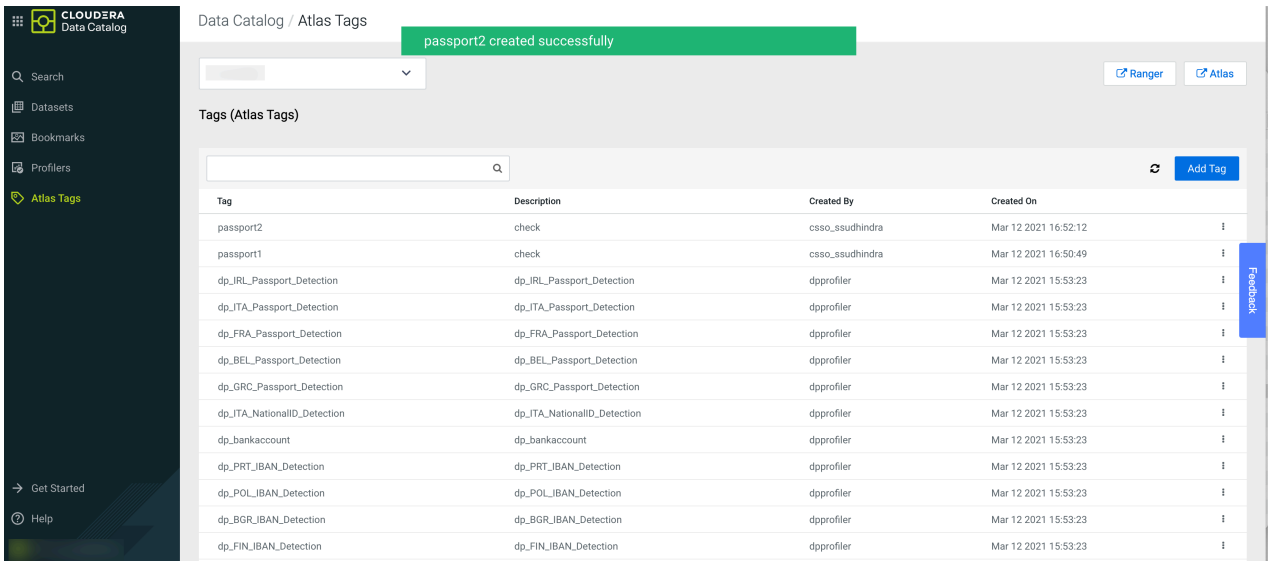
The new Atlas tags UI is displayed as seen in the diagram.



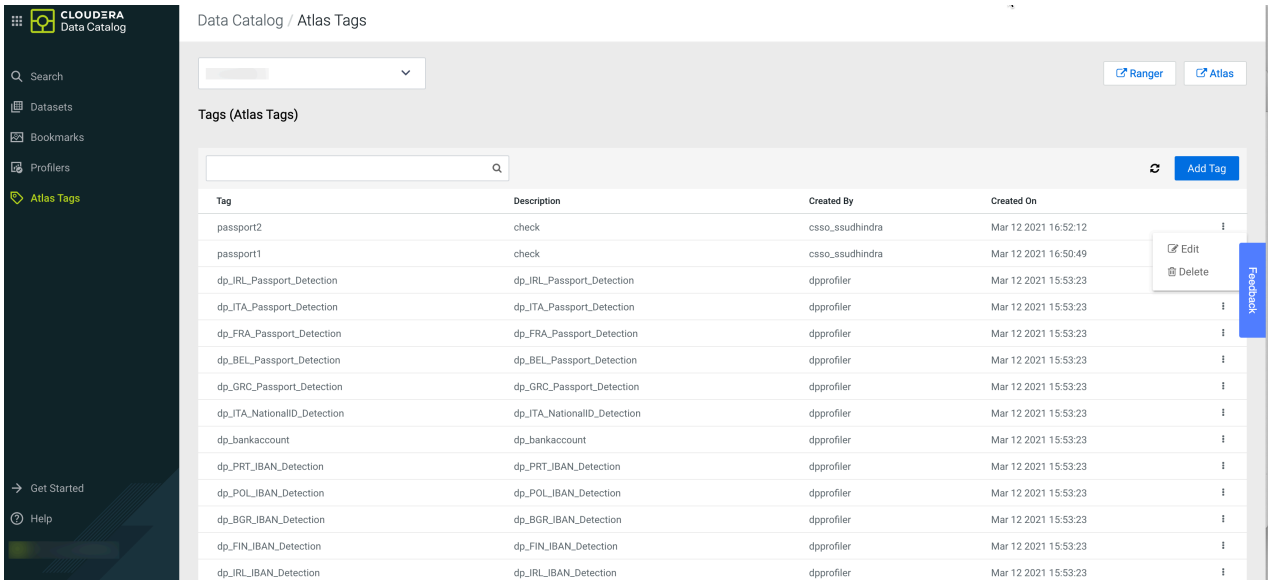
You can create a new tag in the Atlas tags UI. The following diagram provides an overview about the Create a new tag page.



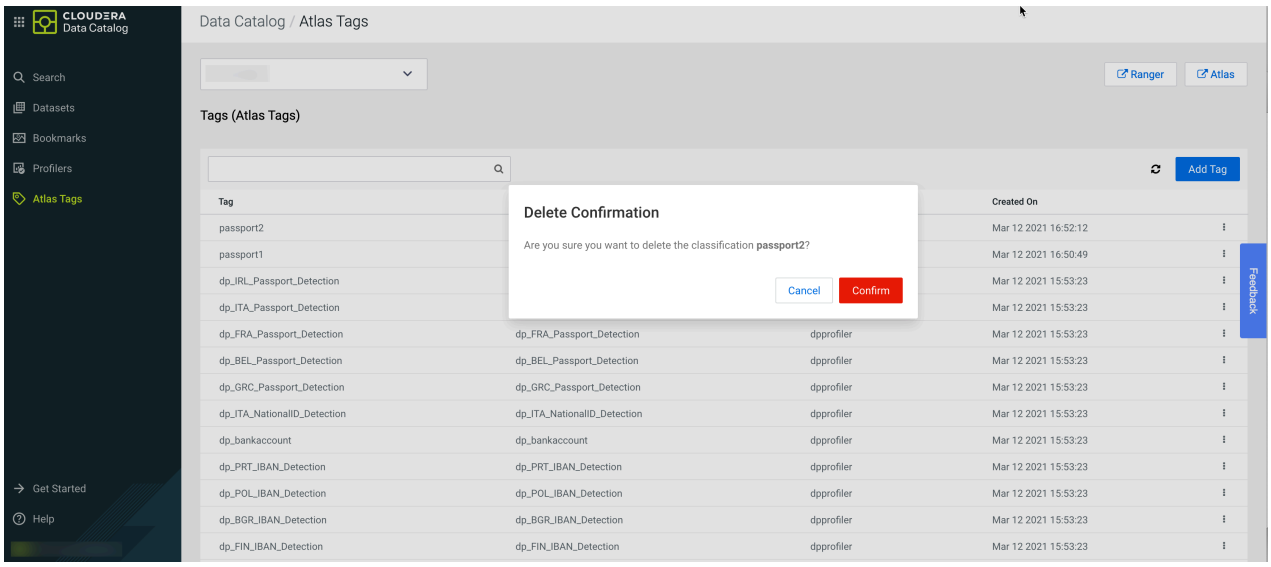
You can add or update Atlas tags. The created or updated tag is highlighted in the tag list as seen in the following diagram.



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.


CLOUDERA
Data Catalog
Search
Datasets
Bookmarks
Profilers
Atlas Tags
Get Started
Help
Data Catalog / Atlas Tags
Ranger
Atlas
Tags (Atlas Tags)
Add Tag
Tag
Created On
Delete Confirmation
Are you sure you want to delete the classification passport2?
Cancel
Confirm
passport2 Mar 12 2021 16:52:12
passport1 Mar 12 2021 16:50:49
dp_IRL_Passport_Detection Mar 12 2021 15:53:23
dp_ITA_Passport_Detection Mar 12 2021 15:53:23
dp_FRA_Passport_Detection dp_FRA_Passport_Detection dpprofiler Mar 12 2021 15:53:23
dp_BEL_Passport_Detection dp_BEL_Passport_Detection dpprofiler Mar 12 2021 15:53:23
dp_GRC_Passport_Detection dp_GRC_Passport_Detection dpprofiler Mar 12 2021 15:53:23
dp_ITA_NationalID_Detection dp_ITA_NationalID_Detection dpprofiler Mar 12 2021 15:53:23
dp_bankaccount dp_bankaccount dpprofiler Mar 12 2021 15:53:23
dp_PRT_IBAN_Detection dp_PRT_IBAN_Detection dpprofiler Mar 12 2021 15:53:23
dp_POL_IBAN_Detection dp_POL_IBAN_Detection dpprofiler Mar 12 2021 15:53:23
dp_BGR_IBAN_Detection dp_BGR_IBAN_Detection dpprofiler Mar 12 2021 15:53:23
dp_FIN_IBAN_Detection dp_FIN_IBAN_Detection dpprofiler Mar 12 2021 15:53:23
Feedback