CDP Public Cloud / Data Catalog

Data Catalog Operations

Date published: 2019-11-14
Date modified: 2024-06-03

https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2024. 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

Managing Datasets.....5
Creating Datasets.....5
Edit Datasets.....6
Delete Datasets.....6

Collaborate with other users.....7

Search for Assets.....8
Filters.....8
Viewing Ranger and Atlas applications.....9
Prepopulating Asset Owners.....9
Accessing Data Lakes.....10
Navigating to tables and databases in Hue.....10
Integrating Data Catalog with AWS Glue Data Catalog.....10
Setting up AWS Glue Catalog with CDP Data Catalog.....11
Working with AWS Glue.....15
Accessing AWS Glue.....15
Prerequisites for accessing Hue tables and databases.....19
Accessing Hue tables and databases.....19
Accessing Hue tables and databases from the search page.....20
Accessing Hue tables and databases from the Asset Details page.....21
Accessing Hue assets.....22
Searching for assets across multiple data lakes.....24
Download CSV option.....25
Searching for assets using Glossary.....26
Using Terms in Data Catalog.....27
Mapping glossary terms.....27
Searching for assets using glossary terms.....30
Additional search options for asset types.....32
Searching for assets in Data Catalog using additional search options.....33
Accessing Tables based on Ranger policies.....34
Creating Classification for selected assets.....35
Adding Classifications / Terms for selected assets.....36
Additional Entity type selection for searching Assets.....36

Viewing Data Asset Details.....38
Viewing Data Assets.....38
View Data Asset Schema.....40
Navigating from the container asset to the parent asset from Asset Details page.....40
View Authorization Policies on a Data Asset.....41
View Data Asset Audit Logs.....42
Navigation Support for Hive entity within Lineage.....42
Adding Hive asset to one or more datasets on Asset Details screen.....43
Viewing Atlas Entity Audits.....45

Managing Profilers **47**
Data Catalog profiler data testing 48
Launch profiler Cluster 48
Launching profilers using Command-line 51
Deleting profiler clusters 54
On-Demand Profilers 57
Profiling table data in non-default buckets 58
High Availability support for Profiler services 58
Tracking Profiler Jobs 61
Viewing Profiler Jobs 62
Viewing Profiler Configurations 62
Edit Profiler Configuration 63
Additional Configuration for Cluster Sensitivity Profiler 63
Additional Configuration for Hive Column Profiler 64
Understanding Cron Expression generator 65
Setting Asset filter rules 65
Backing up and Restoring Profiler Database 68
About the script 68
Running the script 69

Enable or Disable Profilers **71**

Profiler Tag Rules **71**

Tag Management **72**

Tagging Multiple Assets **75**

Creating Custom Profiler Rules **78**
Adding Custom Regular Expressions 78
Adding Lookup Files 79
Using Behaviors 79
Regular expressions 80
File based denylist and allowlist checks 80
Luhn algorithm 80
Using DSL Grammar 80

Managing Datasets

You can view, create, edit, and delete Datasets.

In Cloudera Data Catalog, click Datasets to view all the datasets.

Search for Datasets

On the Datasets page, enter a search string in the search box to view all asset collections with names that contain the search string.

Filter Datasets by Tags

You can view and filter datasets with tags. Select the tag from the drop down list or enter the tag in the filter box. Any dataset with the filter tag assigned to a column will appear in the results.

Related Information

Understanding datasets

Creating Datasets

You can group data assets into datasets. This enables you to organize data based on business classifications, purpose, protection requirements, or more. Examples of datasets are: customer profiles, sales assets, financials, PII, and HR data.

Procedure

1. From the **Datasets** page, click Add Datasets.

   The **Add** page appears.
2. Enter the following information.

Field Name	Description	Example Values
Name	Enter an appropriate dataset name. This name cannot be duplicated across the system. (Mandatory)	Customer Profiles, Sales Assets, Financials
Description	Describe the purpose or intent of the dataset. (Mandatory)	Contains customer profiles: data assets for US and WW.
Data Lake	Assign the dataset to one data lake. Choose from a list of available Data lakes. (Mandatory)	dss_bbsh_clust3
Tags	Add tags to your dataset for context and subsequent lookup. Tags enable your to quickly catalog, search and retrieve asset collections as well as share such information with others in the future. (Optional)	se, pii, geo, finance
Public/Private	Select Public if you want other users to have access to this dataset. Select Private if only you want to have access to this dataset. **Note:** You can change the status of the asset collection later. Click the lock icon on the **Dataset Details** page to change the access state of the dataset.	Public/Private

3. Click Next.

   The **Dataset Details** page appears for the new dataset.
4. Click Add Assets to add related data assets into your dataset.

   The Asset Search page appears.
5. Search for assets using Basic Search.
   a) Search using the name of the asset by entering the name in the search bar.
   b) Use filters to search for specific assets based on the attributes of assets. Click Filter to display the filters available.
      - Created Time: From the dropdown list, select the time to refine the search on the basis of when the asset has been created.
      - Owner: Enter the name of the owner to refine the search on the basis of the owners of the assets.
      - DB Name: Enter the name of the database.
      - Tag: Enter the names of the tags.
   c) Select one more than one filter if needed.
   d) Click Search to view the assets. The Results appear.
   e) Click Reset to reset the filters and search again.
   f) From the list, click to select the assets that you like to add to your dataset.
6. Search for assets using Advanced Search, if needed. Advanced search uses facets of technical and business metadata about the assets, such as those captured in Apache Atlas, to help users define and build collections of interest. Advanced search conditions are a subset of attributes for the Apache Atlas type hive_table.
7. Click Done.

   The assets are added to the dataset and the Search page is refreshed.
8. Close the Search tab.

   The Datasets Details page appears.
9. Click Save.

Edit Datasets

You can edit Datasets by adding or removing assets and changing the access state of the Datasets.

Procedure

1. Click a Dataset in the list to edit it. The Details page of that Dataset appears.
2. On the Assets tab, click Edit to edit the content of this Dataset. The Dataset appears in edit mode. If another user is editing this Dataset, an error message will appear saying that this Dataset is being edited by another user and you cannot edit it.
3. Add or remove assets in the Dataset.
   a) Click Add to add new assets to this Dataset.
   b) Select one or more assets and click Remove to remove assets from this Dataset.
4. Click Save to save the changes that you made to the Dataset.
5. Click Cancel to undo any changes that you made to this Dataset.

Delete Datasets

You might want to delete an Datasets if you no longer need to track those Datasets, or if you want to reassign those assets to another Dataset. You can delete Datasets at any time. Deleting an Datasets does not delete the assets contained therein, it only disassembles the Datasets. You can re-create Datasets or reassign assets to new Datasets.

Procedure

1. From Data Catalog Datasets page, click the More Options icon beside the name of the Dataset you want to delete.
2. Click Delete.
3. Click Confirm.
   You are returned to the Datasets home page.

Collaborate with other users

You can collaborate and share insights with other users in the enterprise regarding various datasets.

You can rate datasets and view the average rating of a dataset. This can help other users to find datasets with higher ratings easily. You can also add your knowledge and insights about the asset collection by adding comments. Other users can respond to your comments or add their comments about each data asset collection.

On the right hand side of each asset collection page, you can see additional details about the dataset. The collaboration details are also displayed in this tab. The tab displays the following details - average rating for the asset collection, the number of likes, the number of comments, and the bookmark icon indicating if the dataset is bookmarked by the current user or not.

You can perform the following collaboration actions for each dataset.

Like a Dataset

You can let other users know that you like a Dataset. The like icon on the Dataset page displays the total number of likes received by this Dataset.

Click the like icon to add the Dataset to your list of liked collections.

Comment and discuss about a Dataset

You might want to share your knowledge or insights about this Dataset with other users. Data Catalog allows you to collaborate with other users by adding comments.

Click the comment icon to add a comment about this Dataset. The Collaborate tab expands. Click Actions menu to reply to an existing comment. You can continue to add comments for each Dataset.

Bookmark the Dataset

In addition to sharing with other users, you can also bookmark Datasets for easy access in the future.

Click the bookmark icon to add the Dataset to your list of bookmarks. This Dataset will appear in the list of bookmarks when you click the Bookmarks link on the left navigation menu.

Rate the Dataset

You can also rate the datasets on a scale of one to five. Click the star icon to rate the open Dataset. The Collaborate tab expands.

Click the stars to provide your own rating. The rating on the Datasets page shows the average of the rating provided by various users. The Rating section also displays the number of votes given for this Dataset.

View the tags of an Dataset

You can add tags while creating the Dataset. You can also click on the tags to search for Datasets with similar tags. There are two types of tags. System tags are automatically generated based on the details of the assets in the Datasets. You can add more tags that appear in the list of user generated tags.

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.

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 a database to view all the assets stored in that type of database. By default, only the following options are visible:
  - Hive Table
  - HBase Table

  Additional Type filters can be added by clicking + Add New Value.

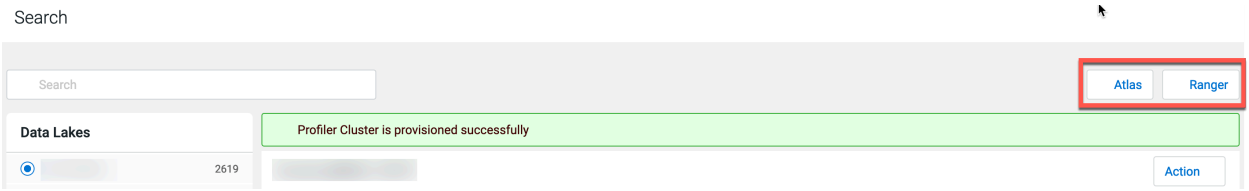
  Note: For information purposes, the Database filter is displayed as Namespace in case of HBase tables.
- Entity Tag - Use entity tags to refine your search results. You can add business metadata as entity tags in Atlas and use these tags to refine your search results and view the details of the required 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 the following type of table assets by tags that have been applied on their children entities, that is, columns or column families using the column tags filter:
  - Hive
  - HBase
  - Iceberg
- 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.

Viewing Ranger and Atlas applications

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



The Atlas and Ranger buttons seen on the search page of Data Catalog allows you to navigate to the specific SDX component. Clicking on Atlas navigates to the Atlas landing page in a new browser tab and the same process takes place to land you on the Ranger landing page when you click on Ranger. This allows finer access and control over the assets and the surrounding policies.

Clicking on Atlas and Ranger links enables you to sign into the respective services and proceed further.

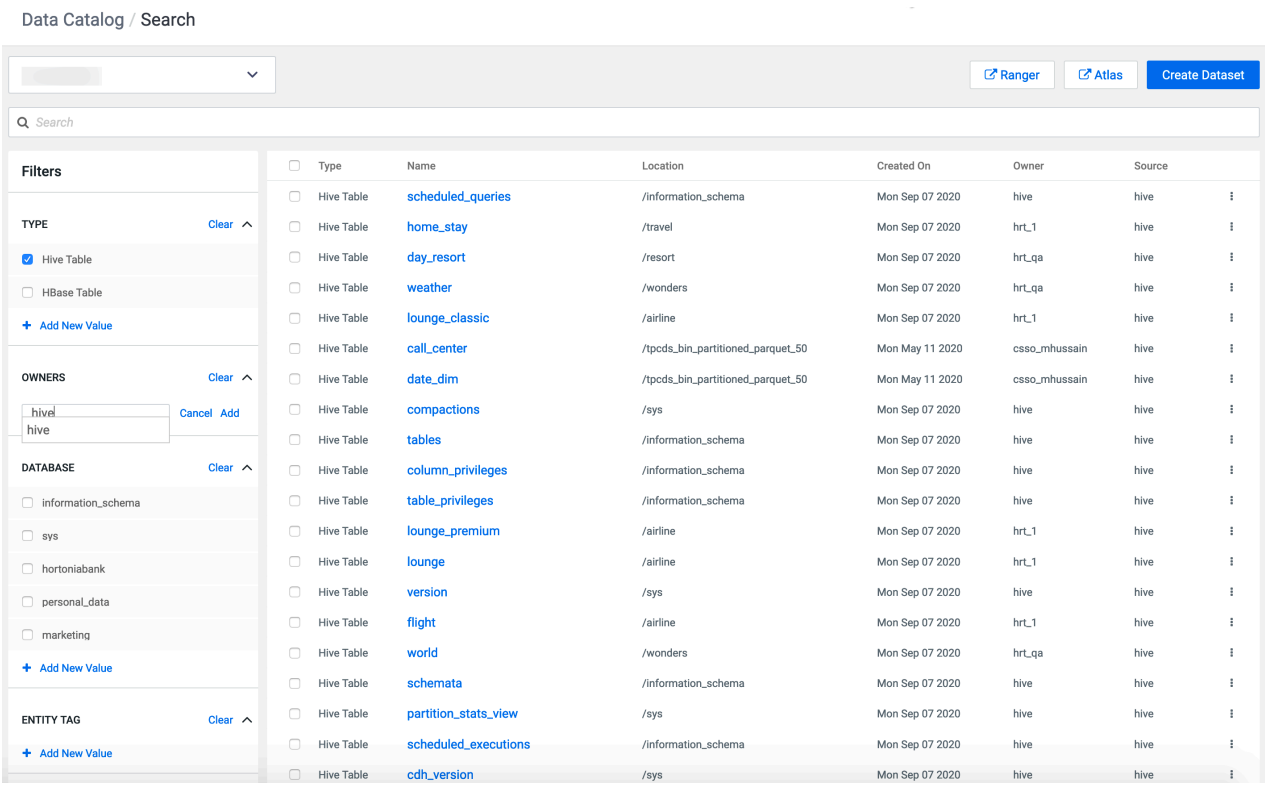
Prepopulating Asset Owners

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.

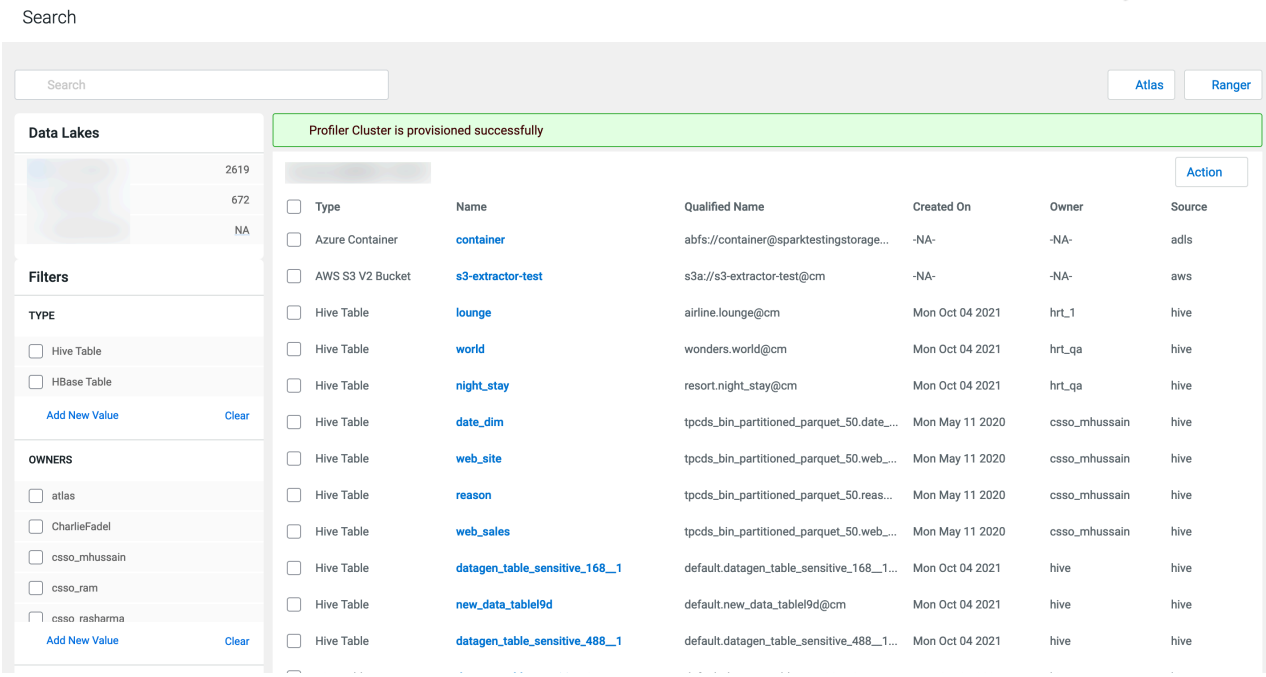


Accessing Data Lakes

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



Related Information

Introduction to data lakes

Understanding data lake details

Navigating to tables and databases in Hue

Data Catalog helps you to explore assets within your Data Lake. Hue allows you to run analytics on Hive assets. The integration between Data Catalog and Cloudera Data Warehouse (CDW) service provides a direct web link to the Hue instance from the Data Catalog web UI, making it easy to navigate across services.

Integrating Data Catalog with AWS Glue Data Catalog

Integrating Cloudera Data Catalog with AWS Glue Catalog enables the users to browse and discover data as well as register data into SDX (through metadata translation or copy), so that it can be used with Cloudera Data Hub and other relevant experiences.

While using AWS Glue in Cloudera Data Catalog, you will be able to experience a complete snapshot metadata view, along with other visible attributes that can power your data governance capabilities.

How integration works

Assuming that the SDX is running in the users' AWS account (that contains the same AWS account which has Glue DataCatalog and the data that has to be discovered), the credentials with the ExternalDataDiscoveryService (which is hosted in SDX) must be shared, so that these two entities can interact with each other. These credentials are used to launch SDX and other workload clusters on the users' AWS account.

Prerequisites:

- You must have full access to AWS Glue Catalog and also have access to the EMR cluster's Hive Metastore instance.
- You must set up the CDP.
- You must have access to your AWS IT Admin and CDP Admin user credentials, which is required to enable CDP to access AWS/EMR managed data in CDP.

Note: AWS policies are managed by AWS IAM and the AWS roles are added to the CDP Management Console. Refer to Using Instance Profile in AWS and Using credentials in Management Console. For more information about the AWS access, see AWS Environments.

Setting up AWS Glue Catalog with CDP Data Catalog

You must map your Cloudera Data Catalog instance with AWS Glue Catalog.

Procedure

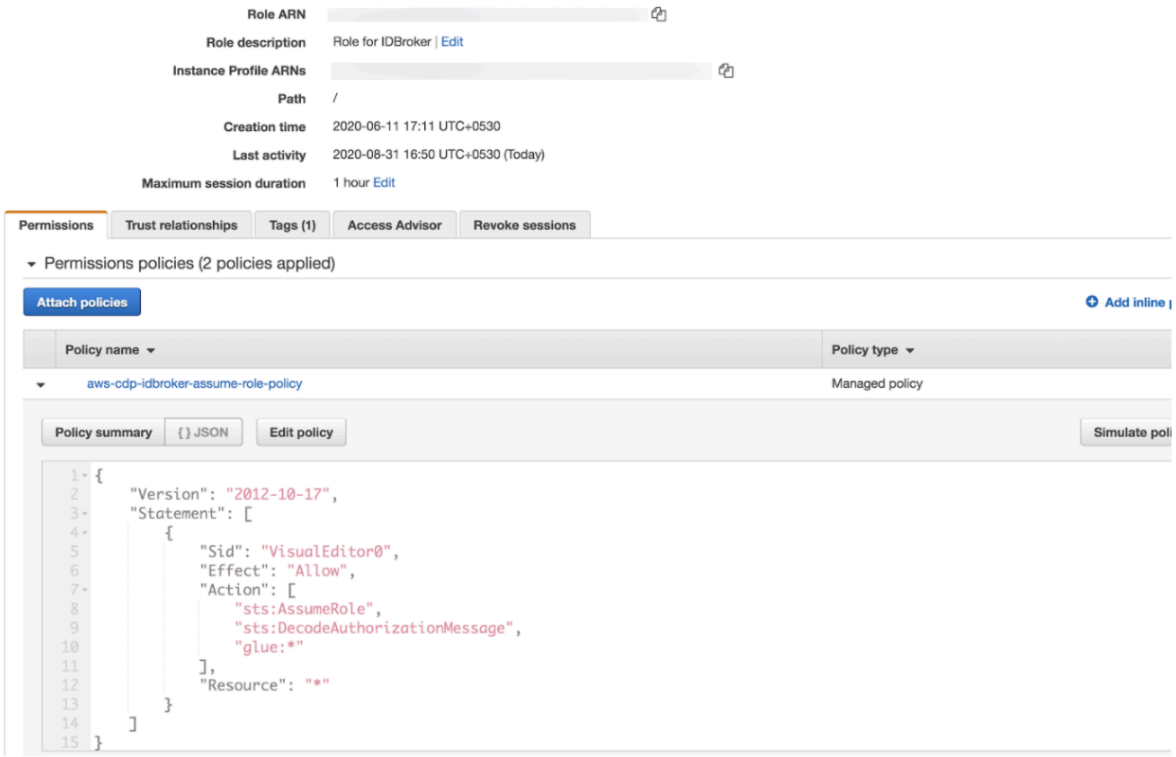
1. Enable the entitlement for your Cloudera Data Catalog instance by running the following command on your Cloudera environment. For example:

```
$ cdp coreadmin grant-entitlement --entitlement-name
DATA_CATALOG_ENABLE_AWS_GLUE --account-id {account_id}
```

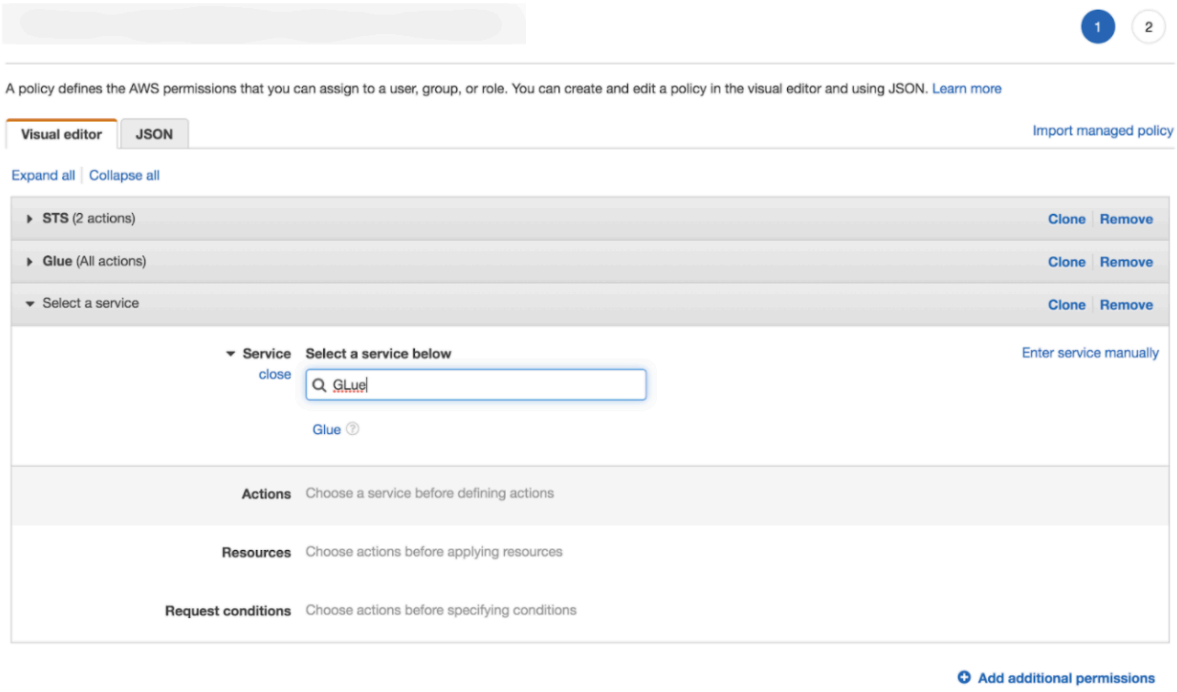
2. You must add relevant permissions in the corresponding AWS account:
   a) Include permission to access Glue Catalog service by editing the policy accordingly.

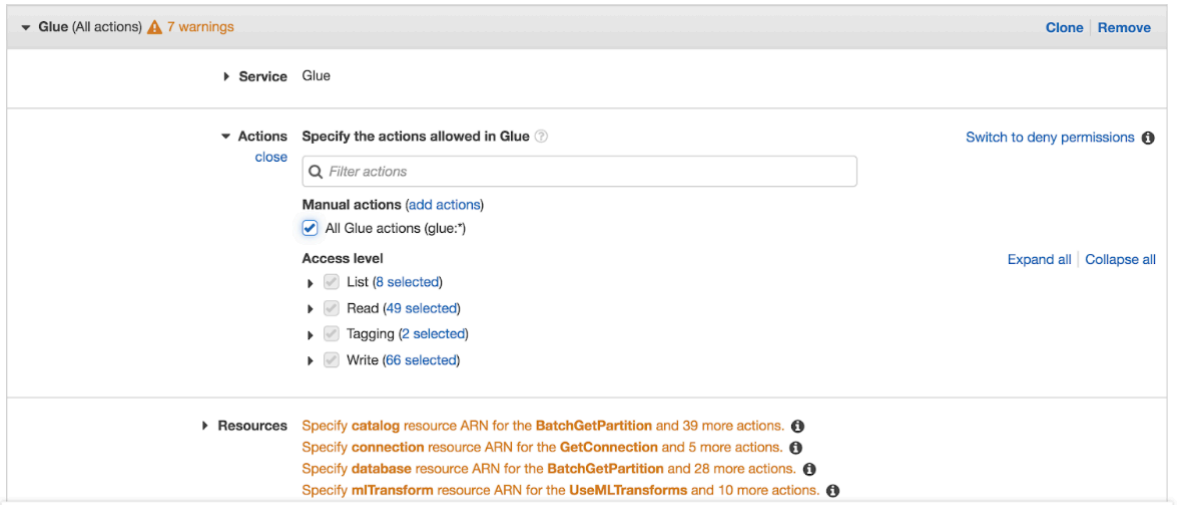
      Make a note of the Assumer Instance Profile role that you intend to use and include full access authorization for AWS Glue.

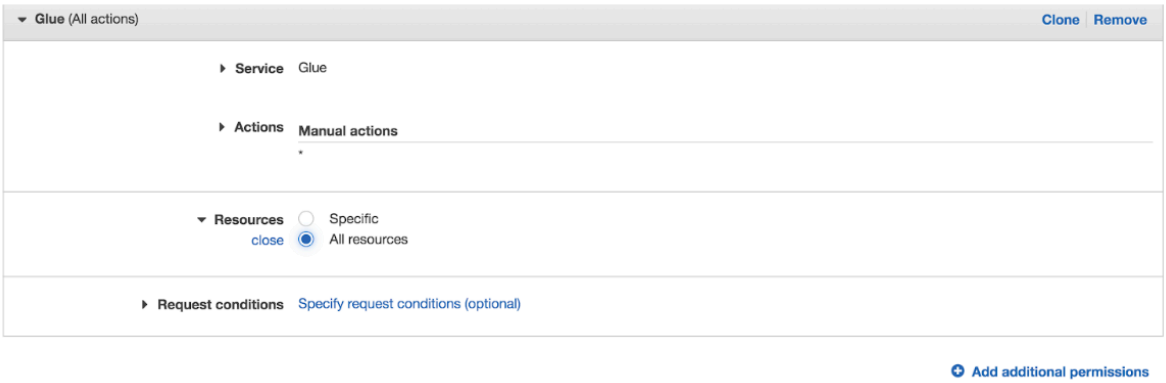
      Refer to the following images as a guidance to complete the set up.



Note: For Role ARN and Instance Profile ARNs, you must include the appropriate account number and role respectively.







b) Search for the role attached to the Instance Profile of the CDP environment. Use the Instance Profile that you have configured above with Glue related policy in your AWS Environment creation command.

Use the following examples to setup AWS environment and AWS data lake as part of the Glue setup:

```
cdp environments create-aws-environment --profile default --cli-input-js
on '
{"environmentName”:”ab-ds-cli-7321”,
 "credentialName”:”cd2d-1234”,
 "Region":"us-region-2”,
 "securityAccess":{-insert the value--"},
 "Authentication":{---insert the value---"},
 "logStorage":{"storageLocationBase":"s3a://demo-e2e-test-state-bucket/
ab-ds-cli-7321/logs","instanceProfile":"arn:aws:iam::<xxxxxxxxxxx>:insta
nce-profile/<role-name>"},
 "vpcId":"vpc-0123456”,
 "subnetIds":["subnet-04fe923b902aa5cf2","subnet-099c7a631f0ebed3c"],
 "s3GuardTableName":"dc-pro-cli-7210",
 "Description":"ab-ds-cli-7321",
"enableTunnel":false,
 "workloadAnalytics":false,
 "freeIpa":{"instanceCountByGroup":1},
 }'

cdp environments set-id-broker-mappings \
--environment-name "ab-ds-cli-7321" \
--profile default \
--set-empty-mappings \
--data-access-role arn:aws:iam::<xxxxxxxxxxxx>:role/add-role \
--ranger-audit-role arn:aws:iam::<xxxxxxxxxxxx>:role/add-role
```

Similarly, while setting up the data lake use the Instance Profile that you configured above with Glue related policy in your data lake creation command:

```
cdp datalake create-aws-datalake --profile default --runtime 7.2.12 --cl
i-input-json '
{"datalakeName":"ab-ds-cli-7321-sdx",
 "environmentName":"ab-ds-cli-7321",
 "cloudProviderConfiguration":{"instanceProfile":"arn:aws:iam::<xxxxxx
xxxxx>:instance-profile/<role-name>","storageBucketLocation":"s3a://demo
-e2e-test-state-bucket/ab-ds-cli-7321"},
 "scale":"LIGHT_DUTY",
 }'
```

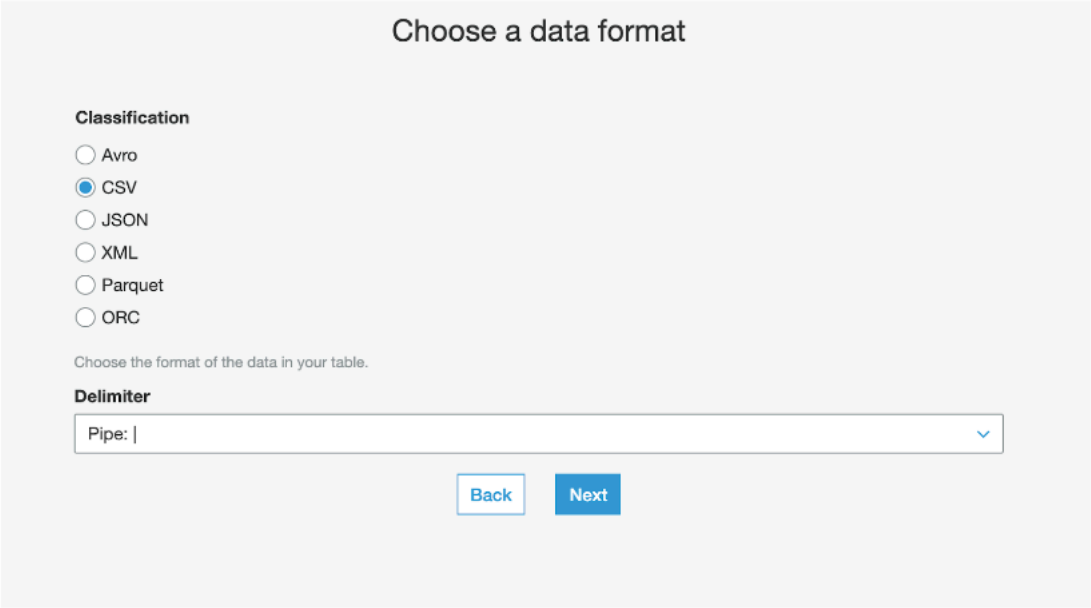
For more information, see Creating an AWS environment with a medium duty data lake using the CLI.

c) Navigate to the attached policy for the role.

d) When you manually create tables in AWS Glue Data Catalog, you must set the **fully qualified path** for the table location.

For example: s3://my-aws-server-node-1/something/something.amazonaws.com/dc-pro-721-storage/glue/

3. 3. You must set up the AWS Glue Data Catalog. For more information, see Populating the Glue Data Catalog. You must select only the CSV format which is currently supported for Cloudera Data Catalog and the delimiter which is used in the data.



4. While creating tables in AWS Glue Data Catalog manually, set the fully qualified path for location. For example: s3://my-aws-server-node-1/something/something/dc-pro-721-storage/glue/

What to do next

AWS Glue metadata must be registered with Cloudera Data Catalog.

Working with AWS Glue

The AWS Glue metadata must be registered with the Cloudera Data Catalog. The Glue contains the metadata that is synchronized with Cloudera Data Catalog. The Glue metadata is accessed using the Data lake option in the Cloudera Data Catalog. After setting up AWS Glue with Cloudera Data Catalog and once the Glue synchronization with CDP is complete, the AWS Glue data lake appears.

Note: For each AWS environment, there would be a separate listing under the data lake option. The Glue data lake name / identifier in Cloudera Data Catalog follows the format: <glue: Data Lake Name>. The Glue assets are of the type: GLUE EXTERNAL TABLE

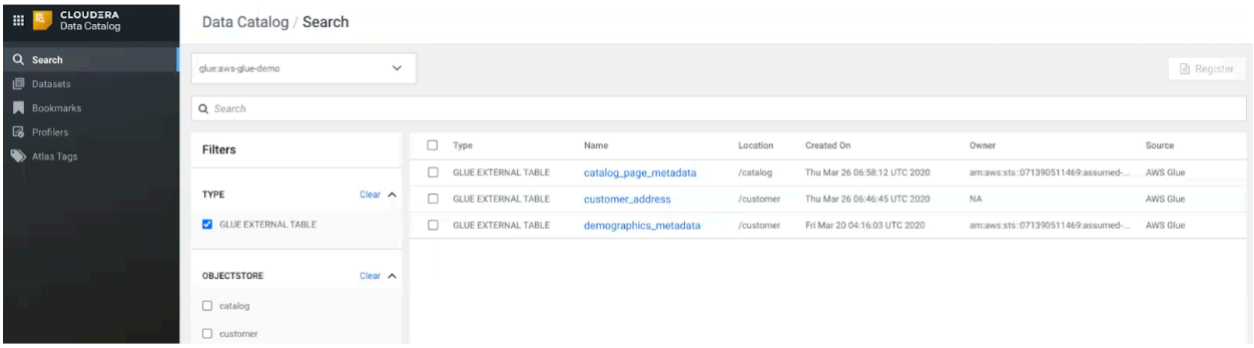
Accessing AWS Glue

To access the Glue metadata in Cloudera Data Catalog, you must note the following in your instance.

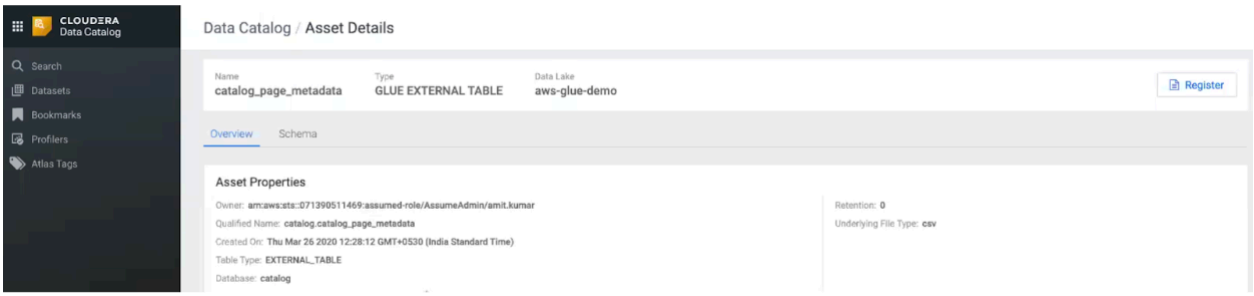
- List the Glue Metadata by selecting the Glue data lake.
- Select one or more Glue assets and register the same with Cloudera platform.
- Verify if the registered Glue assets are listed in the Cloudera Data Catalog owned data lake.
- Select the registered Glue asset and click to open the **Asset Details** page.

Listing the Glue assets

In Cloudera Data Catalog, when you select the AWS Glue data lake, you can view the list of Glue assets. These metadata assets are directly sourced from Glue.



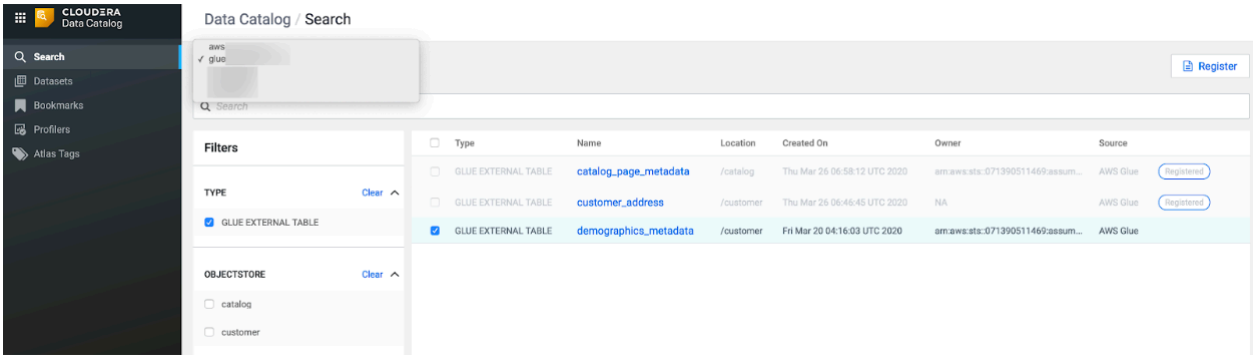
When you click on one of the assets, the Asset Details page is displayed.



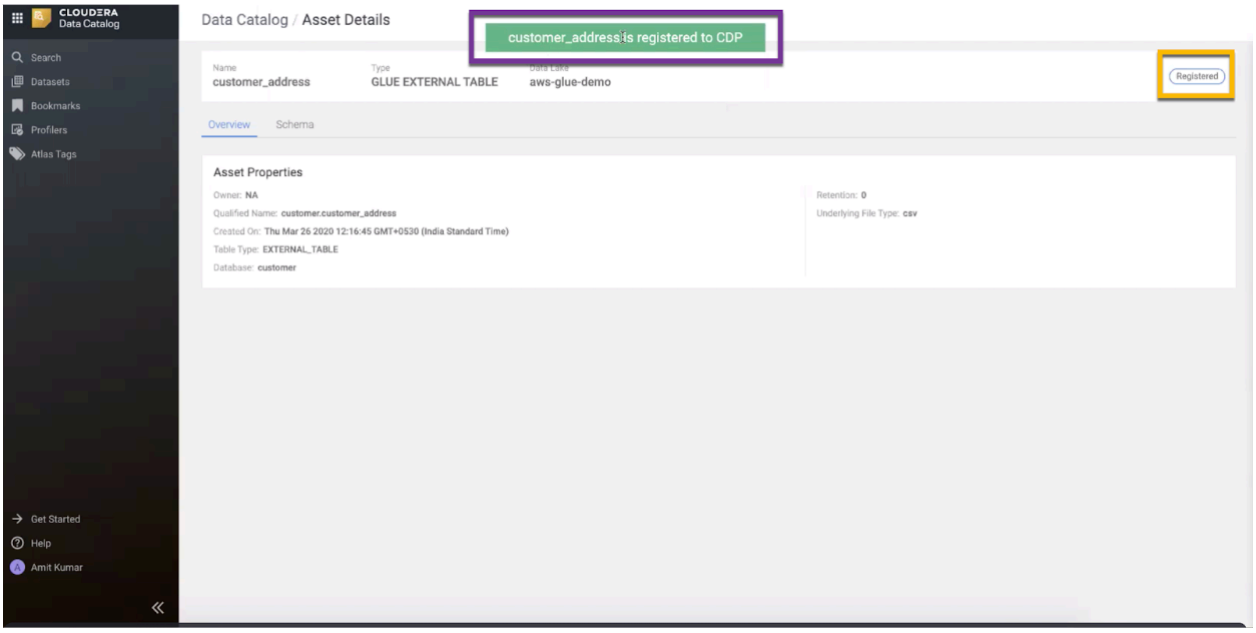
Next, on the main Cloudera Data Catalog page, you must select the Glue data lake and select one of the Glue assets and register the asset to the Cloudera platform. Click Register.

Optionally, you directly click on the Glue asset and register the asset on the **Asset Details** page.

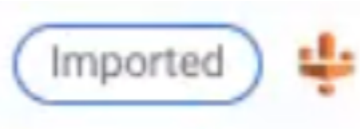
Note: You can select one or more Glue assets on the asset listing page to register them in CDP.

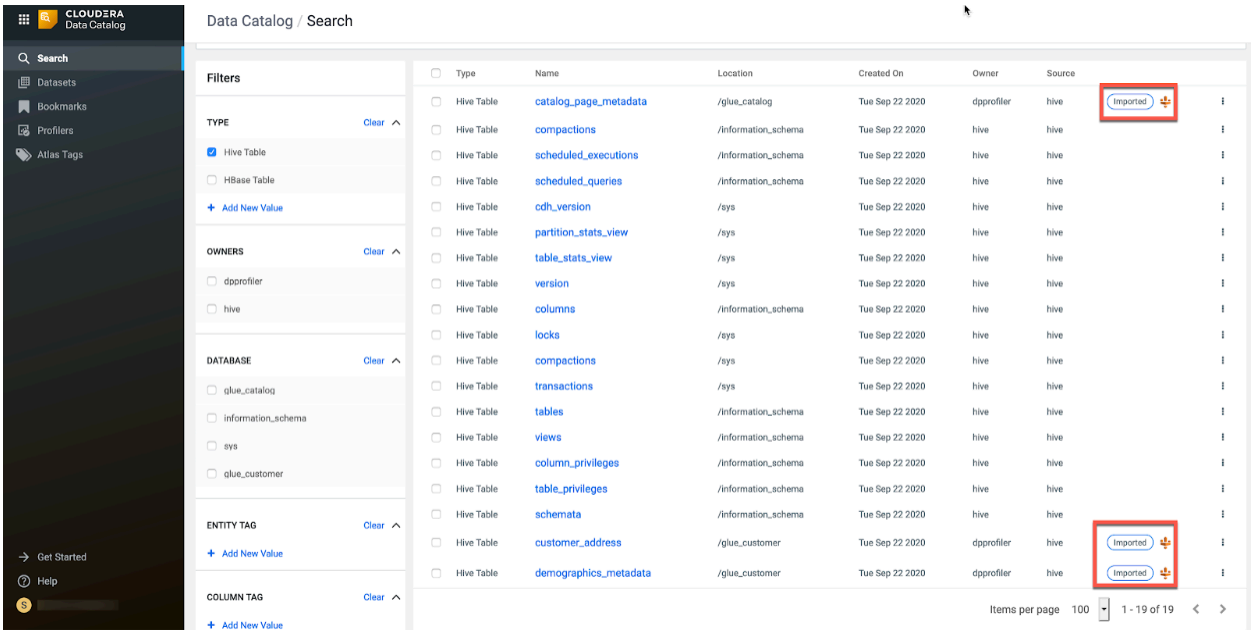


Once the Glue asset is registered, the asset is imported into CDP.



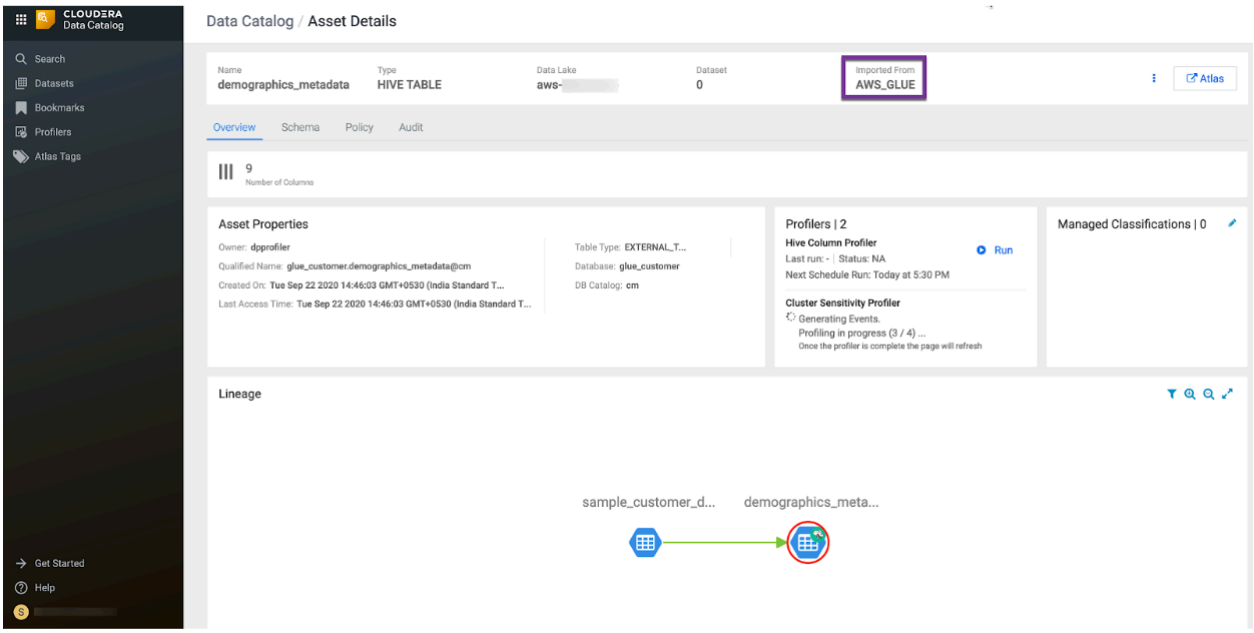
Next, navigate back to the Cloudera Data Catalog main page and select the Cloudera Data Catalog owned data lake and select the type as Hive Table. The search results lists all the Hive table assets and you can view the Glue registered asset(s) as well. The registered Glue asset can be identified using





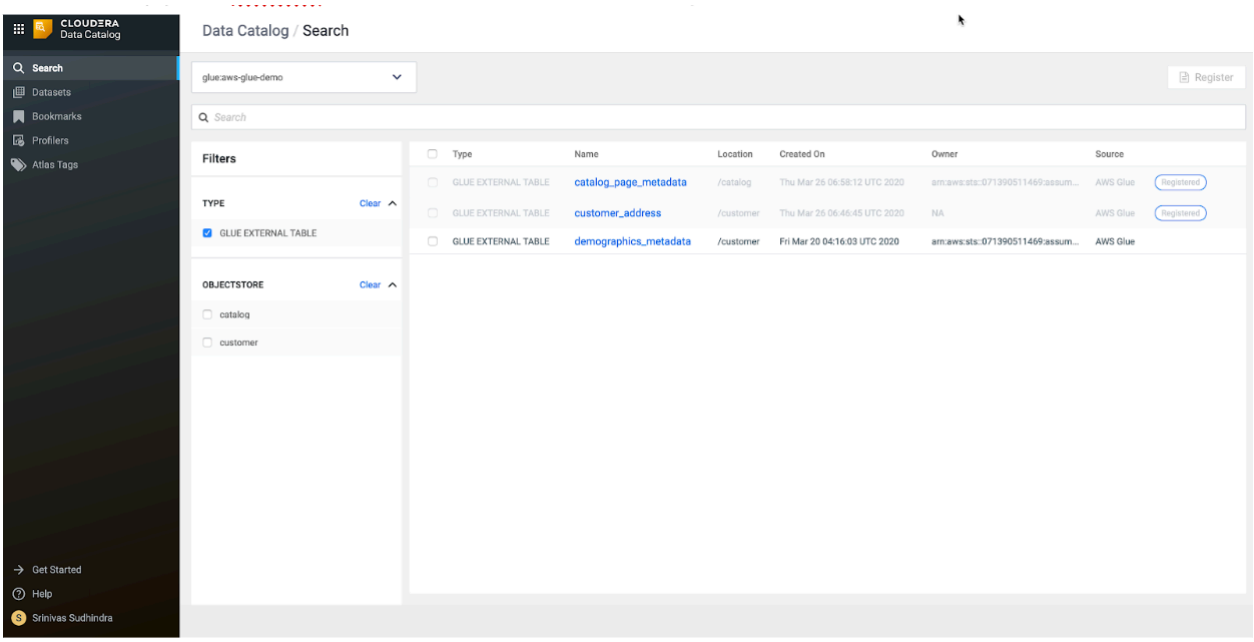
Note: The entries in Cloudera Data Catalog for Glue assets are created in the Hive Metastore.

Click on the Glue registered asset and you can view the **Asset Details** page for the selected Glue asset.



The Asset Details page for the Glue asset is populated by Atlas. While registering the Glue data, the data was written to the Hive Metastore and later Atlas synchronised the metadata.

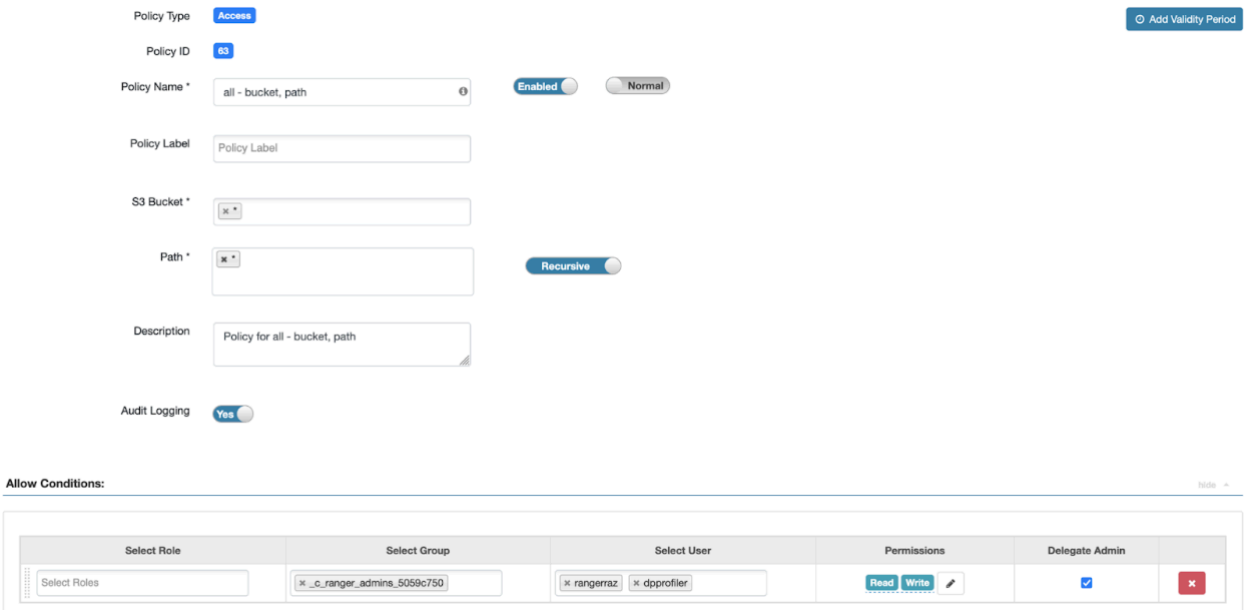
Go back to the main Data Catalog page and select the Glue data lake. Note that the registered Glue asset(s) are greyed out or cannot be selected again.



You can still view the registered Glue assets (powered by Atlas) by clicking on the same and it navigates to the Asset Details page as seen above in the image.

Working with Ranger Authorization Service (RAZ) enabled AWS environment

For RAZ enabled AWS environment, you must employ the following permission settings to work with Cloudera Data Catalog - Glue integration.



Note: By default dpprofiler user is not included in the allowed users list. You must manually add the dpprofiler user in the allowed users list.

Prerequisites for accessing Hue tables and databases

Hue is a part of the CDW service. Before you can access Hue from Data Catalog, you must set up the CDW service.

Before you begin

You must have the DWAdmin role to perform these tasks.

Procedure

1. An AWS or Azure environment in CDW that has been activated and registered in the CDP Management Console.

   Activating an AWS or Azure environment in CDW enables the CDW service to use the existing Data Lakes that are present in the CDP environment. This also creates a default Database Catalog that is used by a Virtual Warehouse.

   For more information, see Activating AWS environments or Activating Azure environments. Currently, CDW only supports AWS and Azure Cloud environments
2. A Hive or an Impala Virtual Warehouse created using CDW. Depending on your data set type, create a Hive or an Impala Virtual Warehouse using CDW.

   For more information, see Adding a new Virtual Warehouse.
3. Select the asset type in the Data Catalog UI.

   The asset type must be either hive_table or hive_db.

   After these conditions are met, Hue links for the selected assets are displayed in the Data Catalog UI.

What to do next

Accessing Hue tables and databases

When you log into Data Catalog, you can access web links to the Hue instances from the Data Catalog Search page or the Asset Details page.

Before you begin

You can access the Hue tables and databases by using one of the following options in Data Catalog,

Procedure

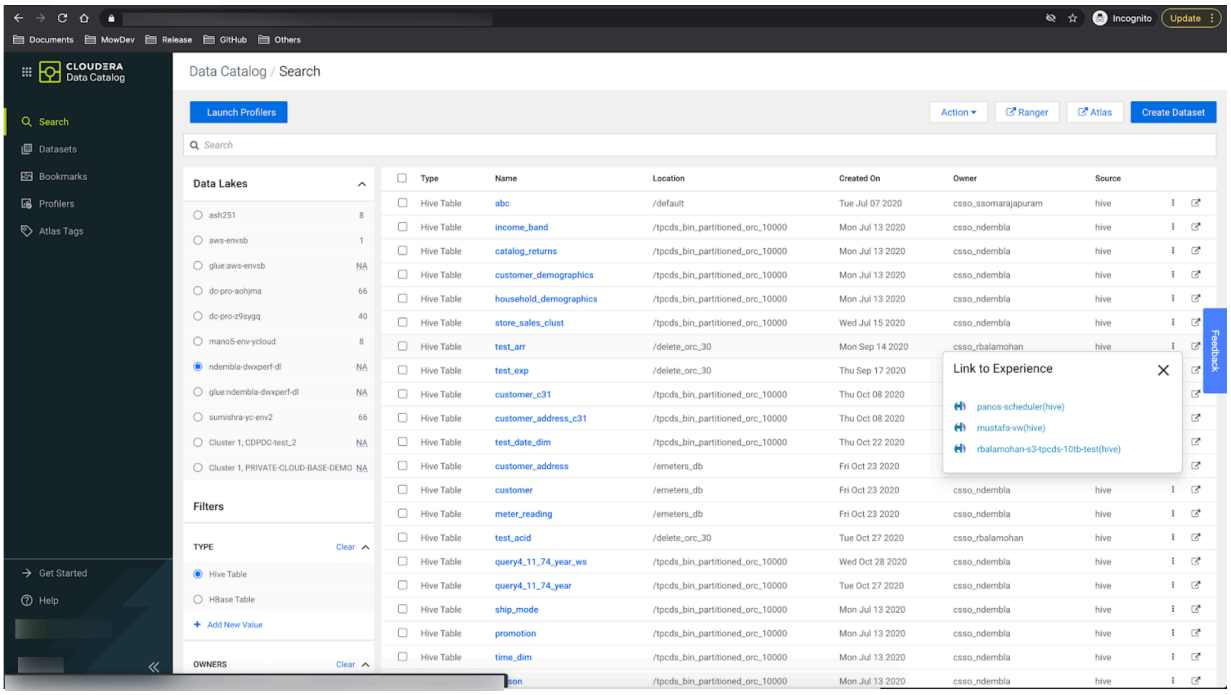
1. Log into Cloudera Data Catalog UI.
2. When you navigate to the entity search page, you can click on the icon present at the end of the row which opens the Link to Experiences dialog.
3. Directly on the selected asset's Asset Details page, you can click Link to Experiences.

Accessing Hue tables and databases from the search page

When you log into Data Catalog, you can access web links to the Hue instances from the Data Catalog Search page.

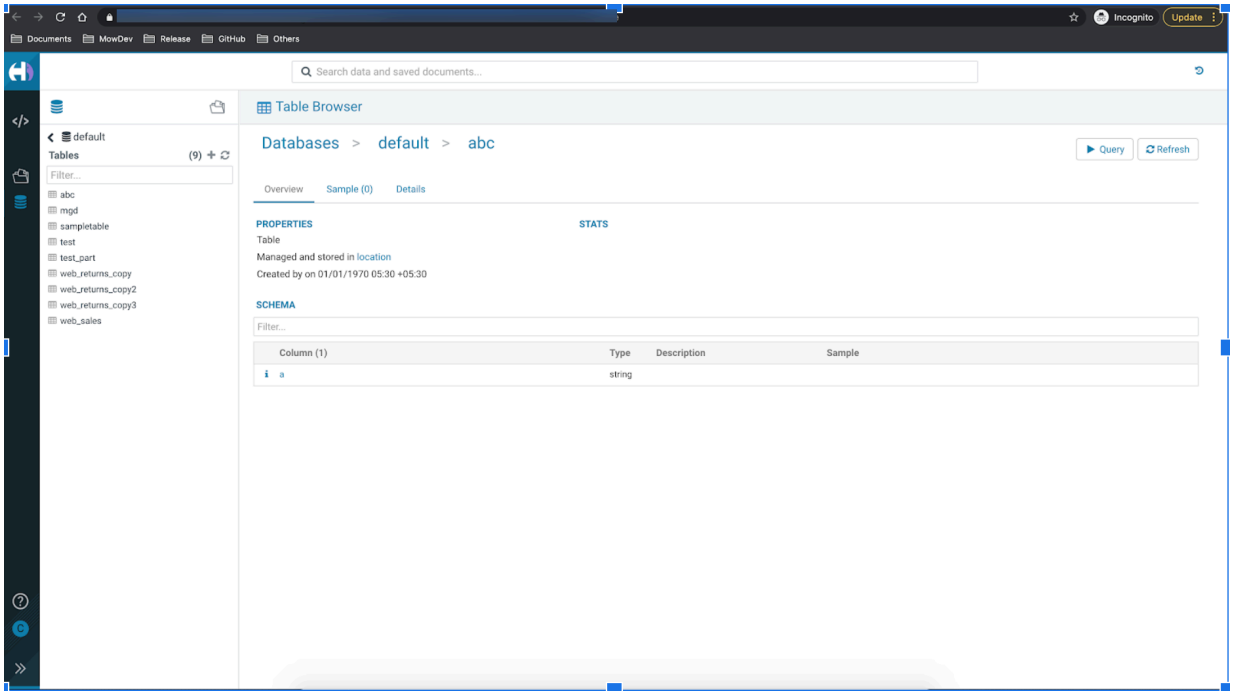
Procedure

1. Log into Cloudera Data Catalog UI.
2. Select the Data Lake associated with your environment.
3. Select the asset type under Filters, for example Hive Table.
   Hive tables present within the Data Lake are displayed.
4. Click the icon at the end of the row for a table that you want to display in Hue.



A Link to Experiences pop-up appears. It contains a list of Hue instances within the virtual warehouses that are associated with this particular table.

5. Click the link to open Hue web interface.
   The Hive table is displayed on the Hue Table Browser.



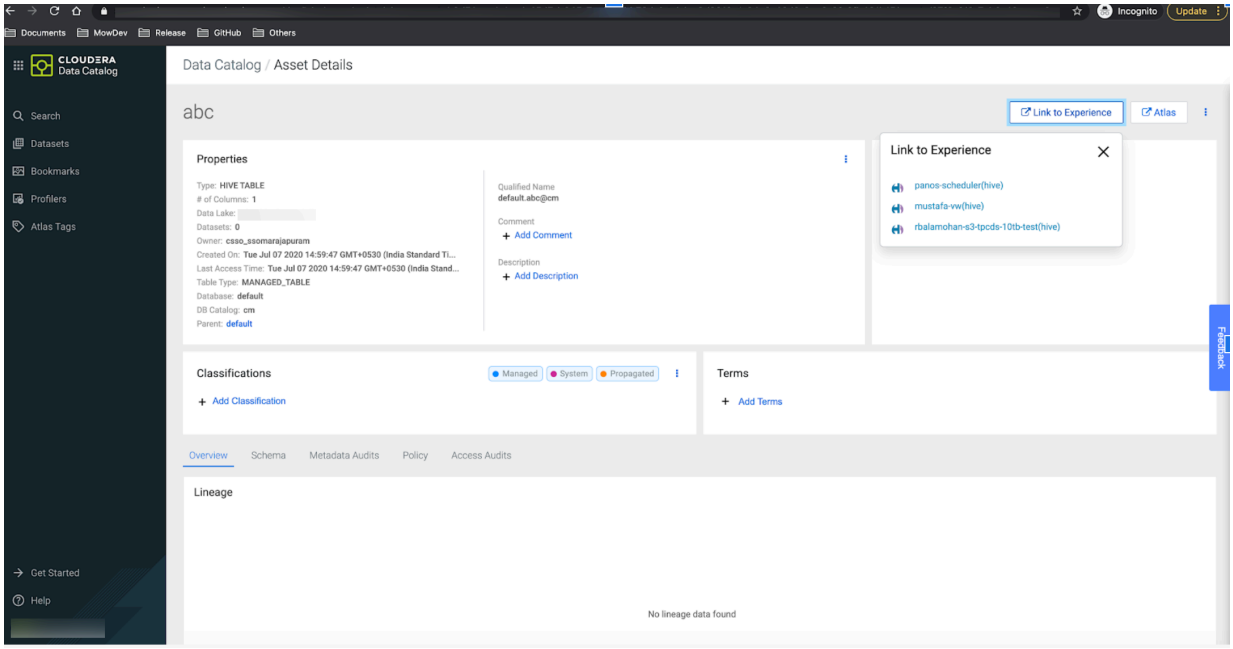
Accessing Hue tables and databases from the Asset Details page

When you log into Data Catalog, you can access web links to the Hue instances from the Data Catalog Search Asset Details page.Log into Cloudera Data Catalog

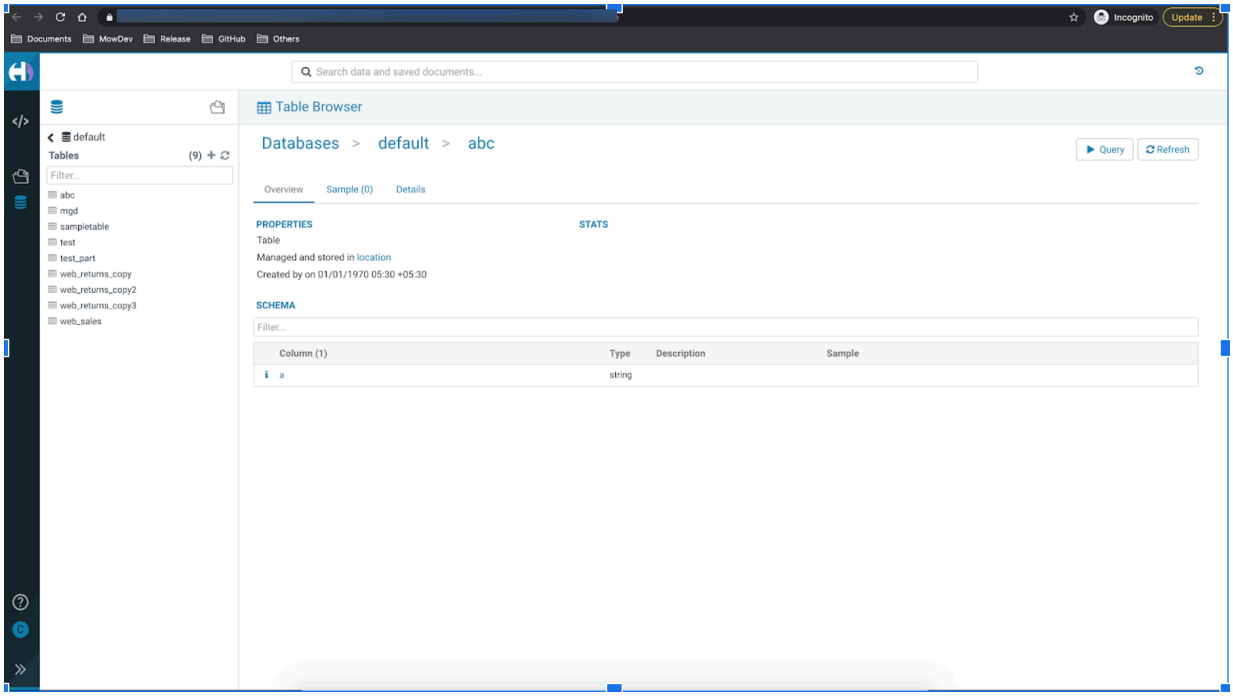
Procedure

1. Log into Cloudera Data Catalog UI.
2. Select the Data Lake associated with your environment.
3. Select the asset type under Filters, for example Hive Table. All the Hive tables present within the Data Lake are displayed

4. Click on the asset name to open the Asset Details page. The Asset Details page contains the Link to Experience button as shown in the following image:



5. Click Link to Experience.
   The pop-up contains a list of Hue instances within the Virtual Warehouses that are associated with this particular table.
6. Click the link to open Hue web interface.
   The Hive table is displayed on the Hue Table Browser.



Accessing Hue assets

When you are on the Data Catalog search page or the Asset Details page, to access Hue assets, an external link button is used.

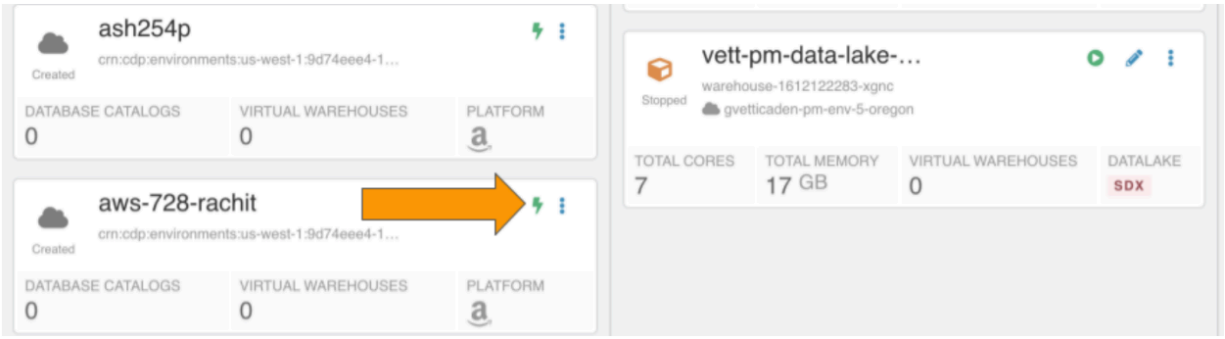
Before you begin

You must fulfill all the three conditions as listed in Prerequisites for accessing Hue tables and databases on page 19 for viewing the external link button.

You must activate Cloudera Data Warehouse on your selected environment.

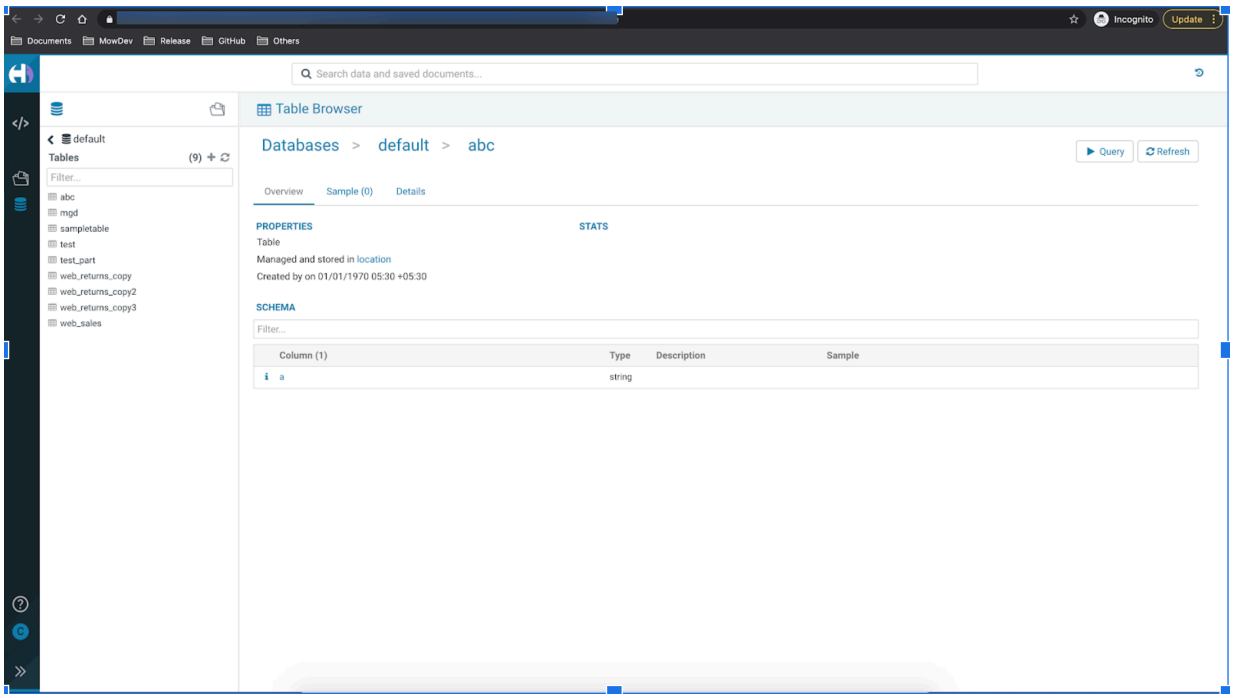
Procedure

1. Login to Cloudera Data Platform.
2. Under Management Console > select Data Warehouse
3. On the Overview page, click on the left pane property displaying Environments.
   AWS/Azure environments are listed.
4. Locate the enabled AWS / Azure environment and activate CDW by clicking on the bolt icon.
   This action also creates a default Database Catalogs.



5. Once the environment is activated and a DB Catalog is created and running, you can create a virtual warehouse as per the requirement.

   For more information about AWS and Azure Environment, see AWS environments overview and Azure environments overview.



You will now be able to navigate to the Hue assets on these virtual warehouses using the links provided by Data Catalog.

Searching for assets across multiple data lakes

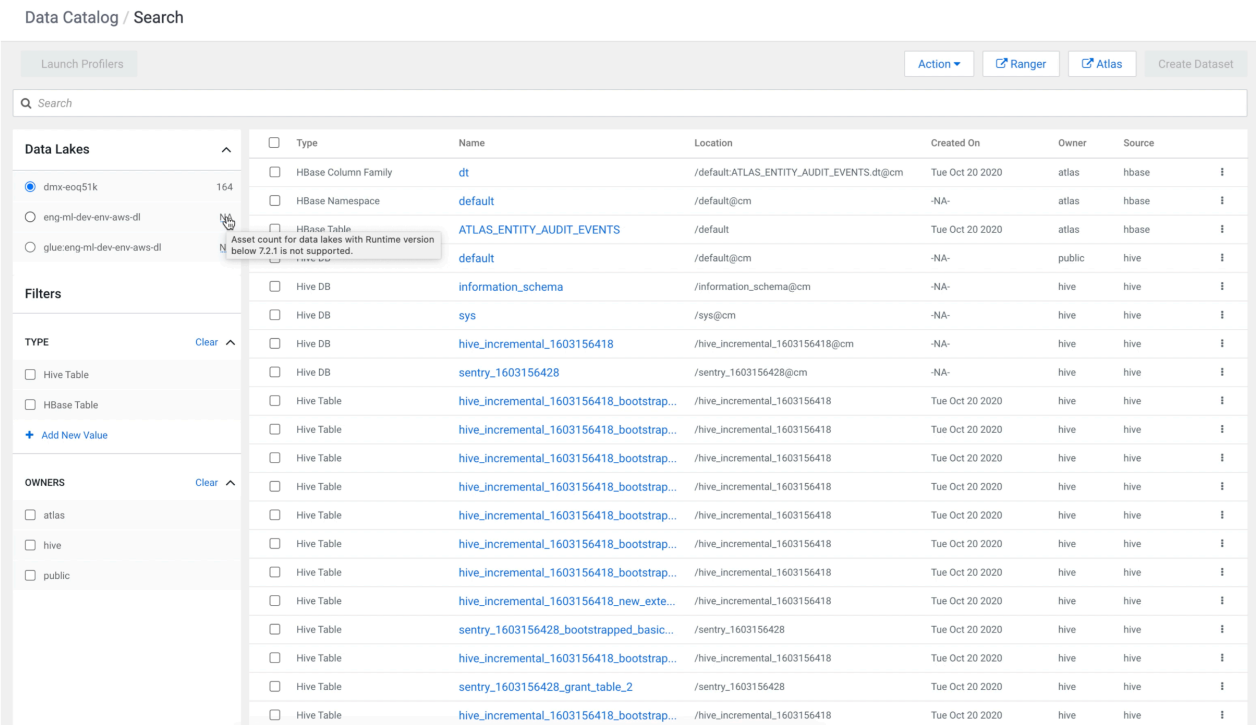
In Data Catalog, the data lake search capabilities has been enhanced with the way you can search for the assets. You can now view the complete list of data lakes that are available in a specific Data Catalog instance.

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.

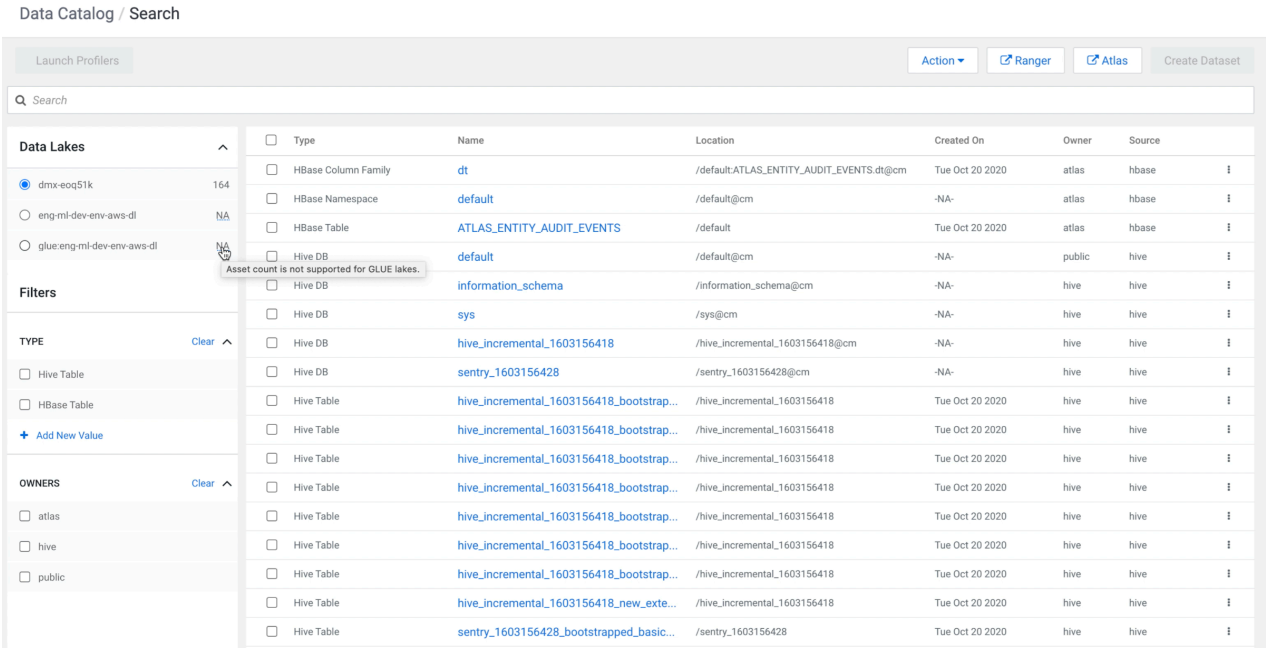
The number of assets that are visible against each data lake indicate that they are applicable to the search query that match. You can select a specific data lake and select one or more search query types to retrieve the total list of available assets in the selected data lake. The total count of the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

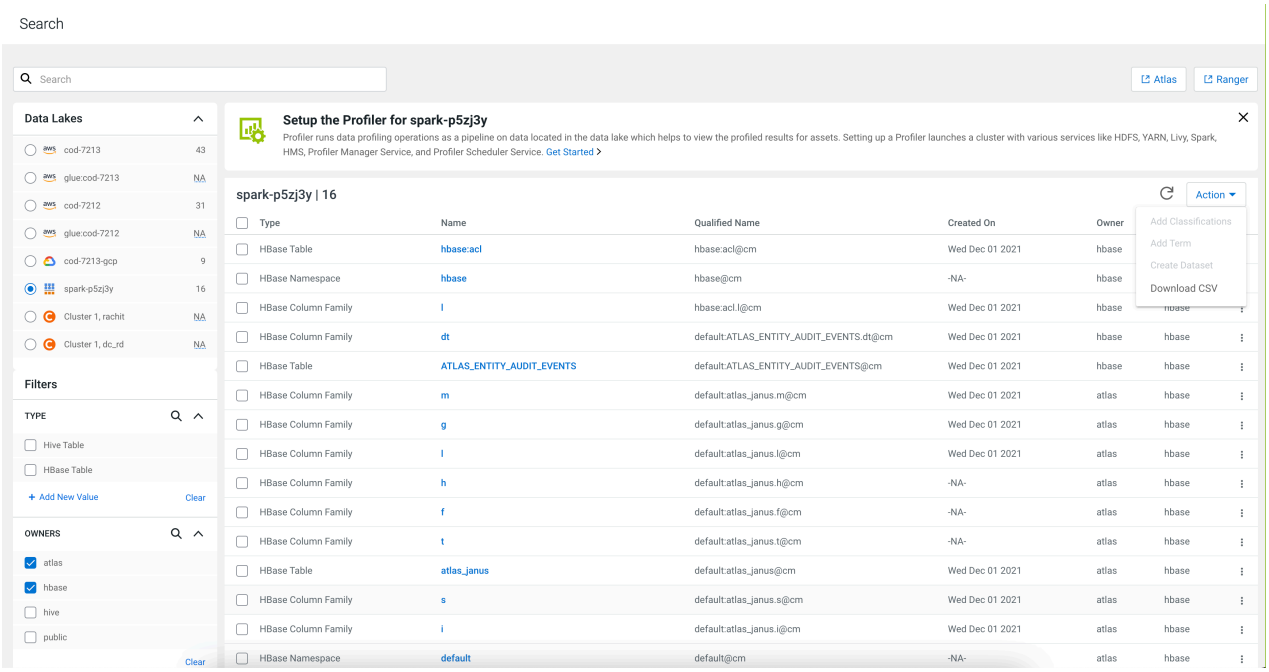


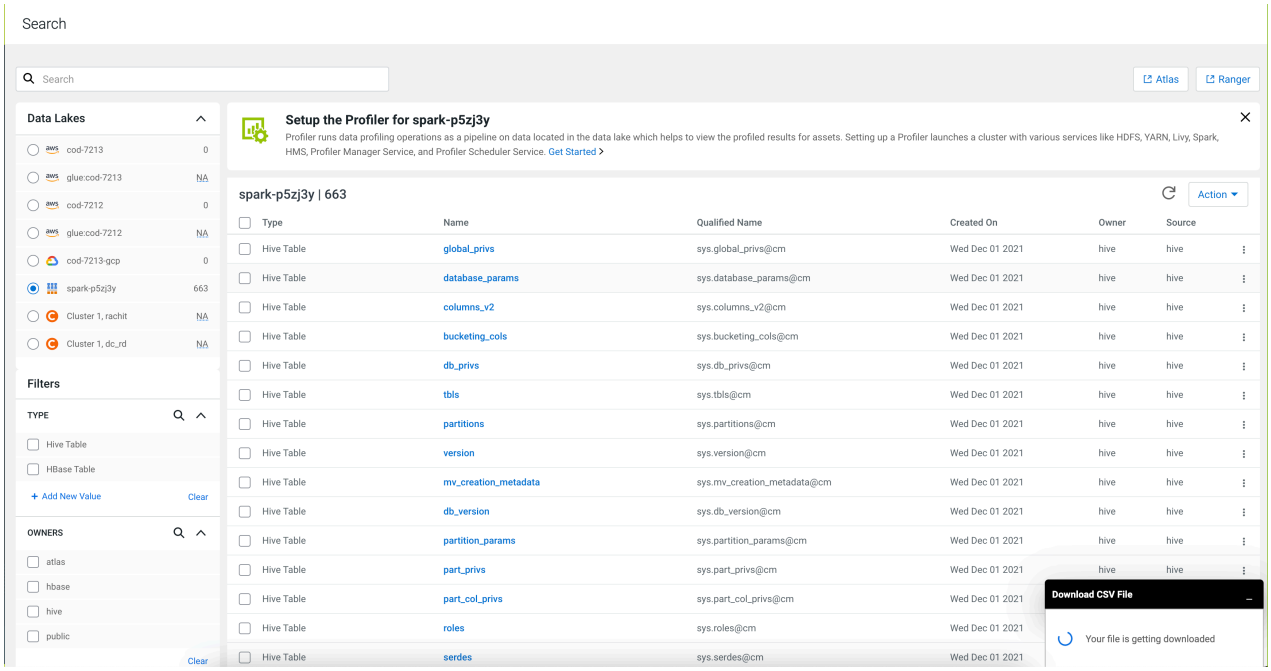
Download CSV option

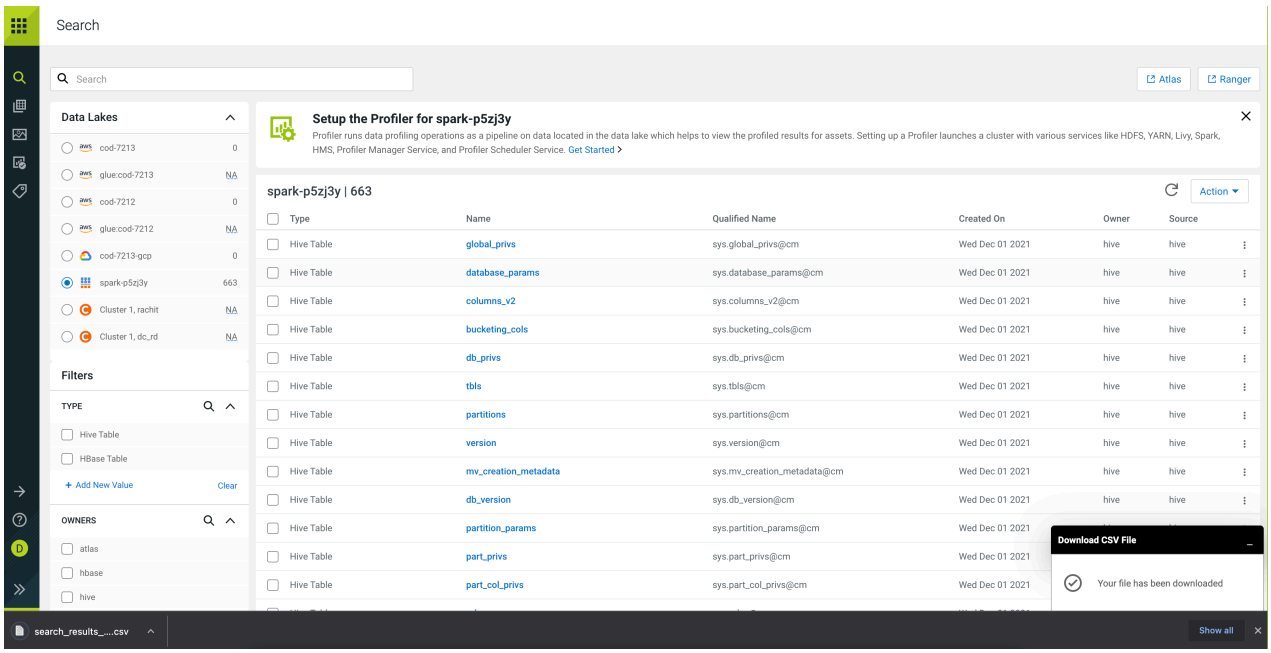
The download CSV feature in Data Catalog allows you to download the search result for the current / specified query with the selected data lake. The feature allows you to download upto 10000 rows for the current search query.

The CSV format does not confirm with any specific order or continuation in the downloaded results. For example, a user downloads 10000 assets once and later downloads the CSV again with about 10000 assets. The CSV may not contain the search results in the same order as it was downloaded earlier or previously.

The following example images provides a sample download flow.







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.

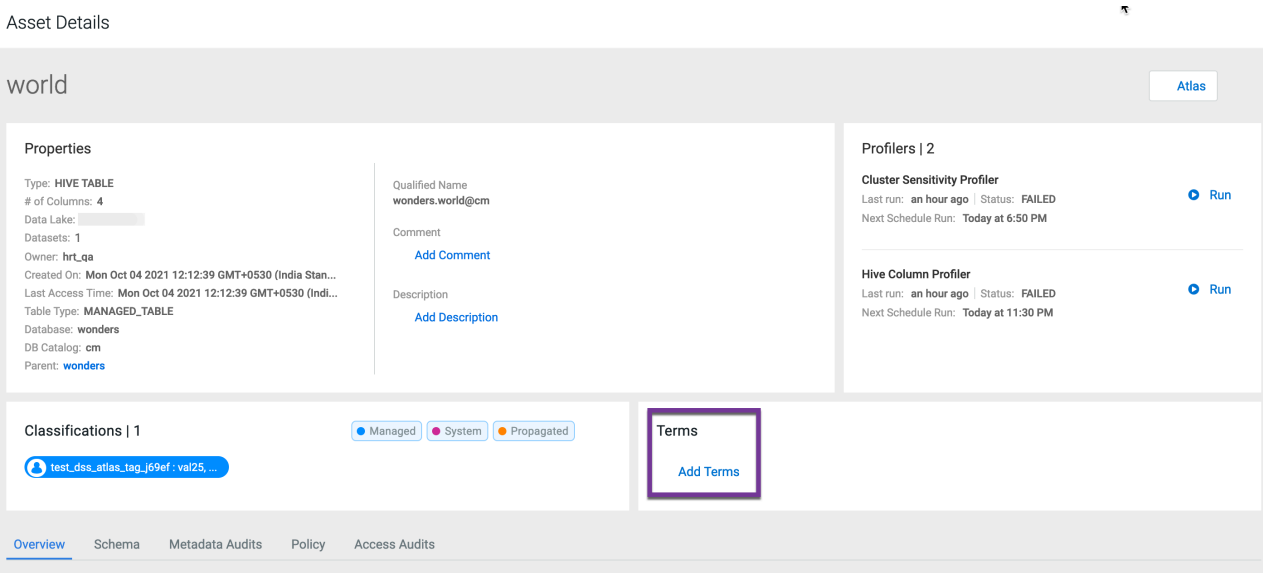
Using Terms in Data Catalog

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.



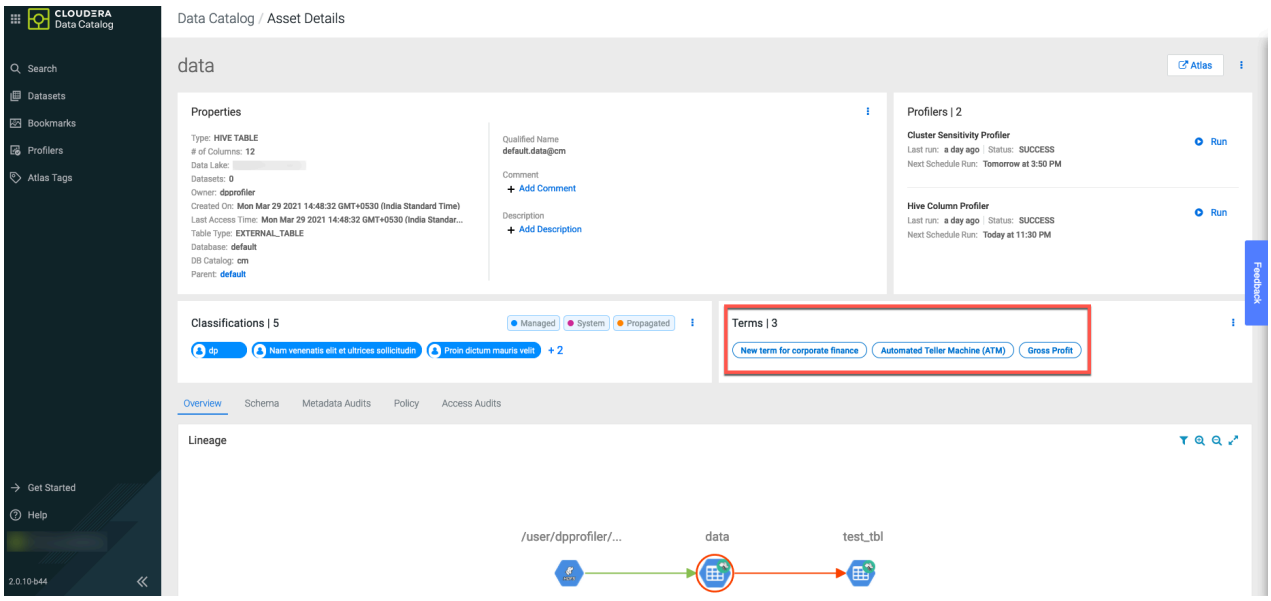
Mapping glossary terms

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

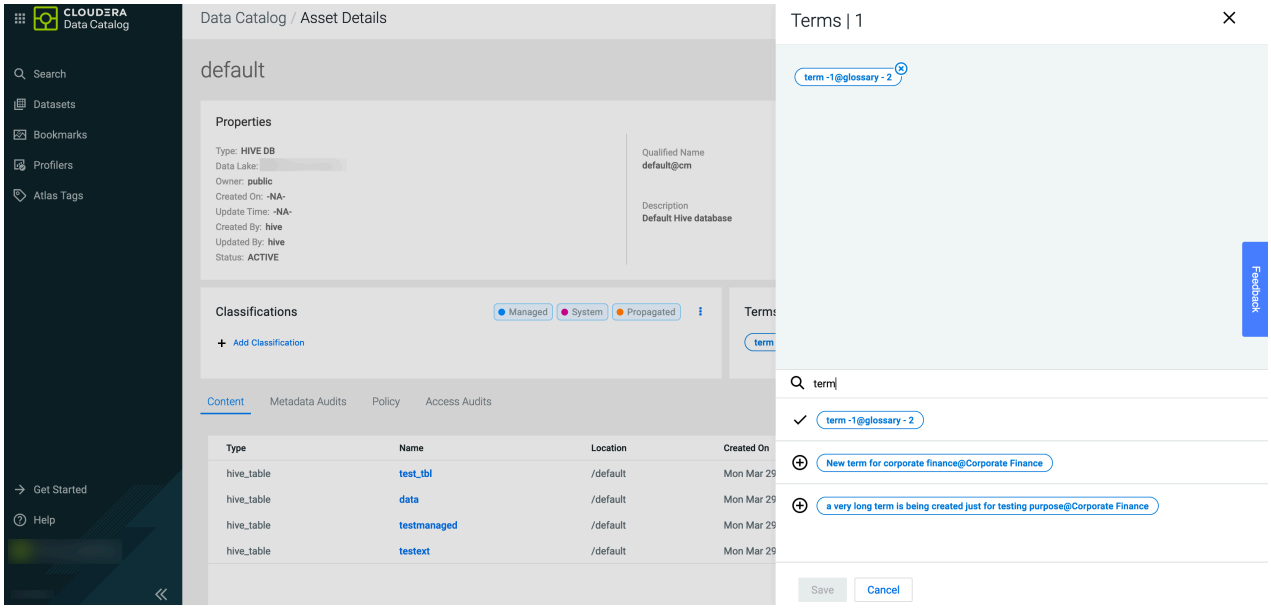
You can search for those terms in Data Catalog and map specific terms with Data assets. You can search for terms in Data Catalog to either add and delete them from the selected data asset. The selected asset displays the total number of terms associated or mapped accordingly.

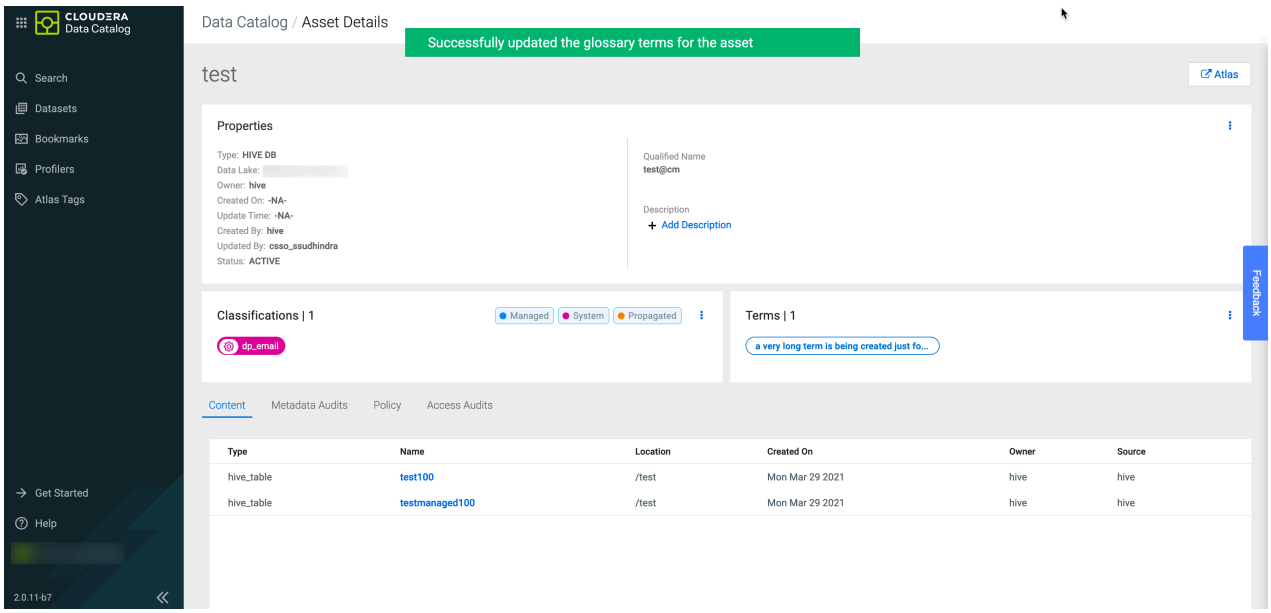
When you map a specific term for your dataset, the term is displayed in the following format:

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

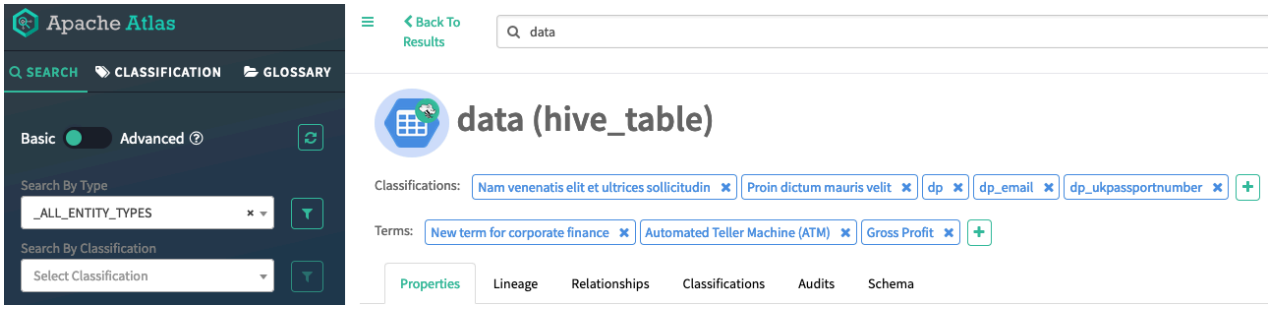


You can use the icon in the Terms widget on the Asset Details page to add new terms for your data assset. Click Save to save the changes.

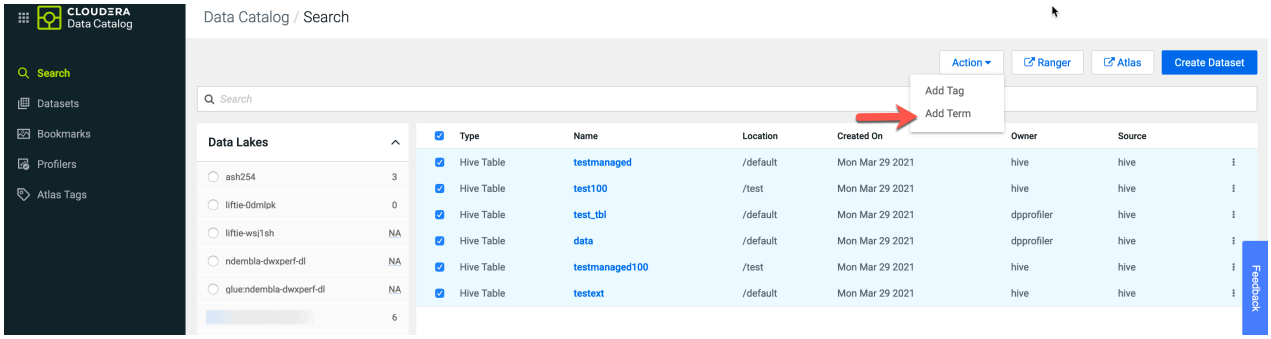


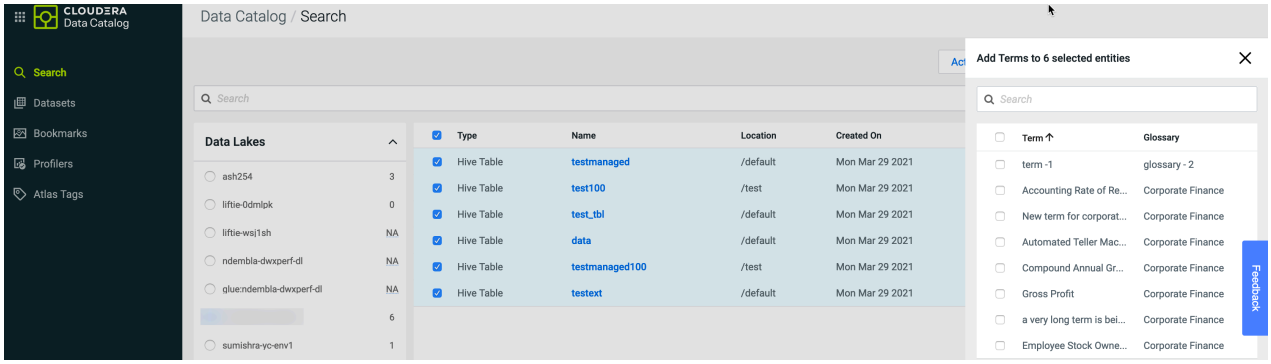


You can search for the same asset in the corresponding Atlas environment as shown in the example image.

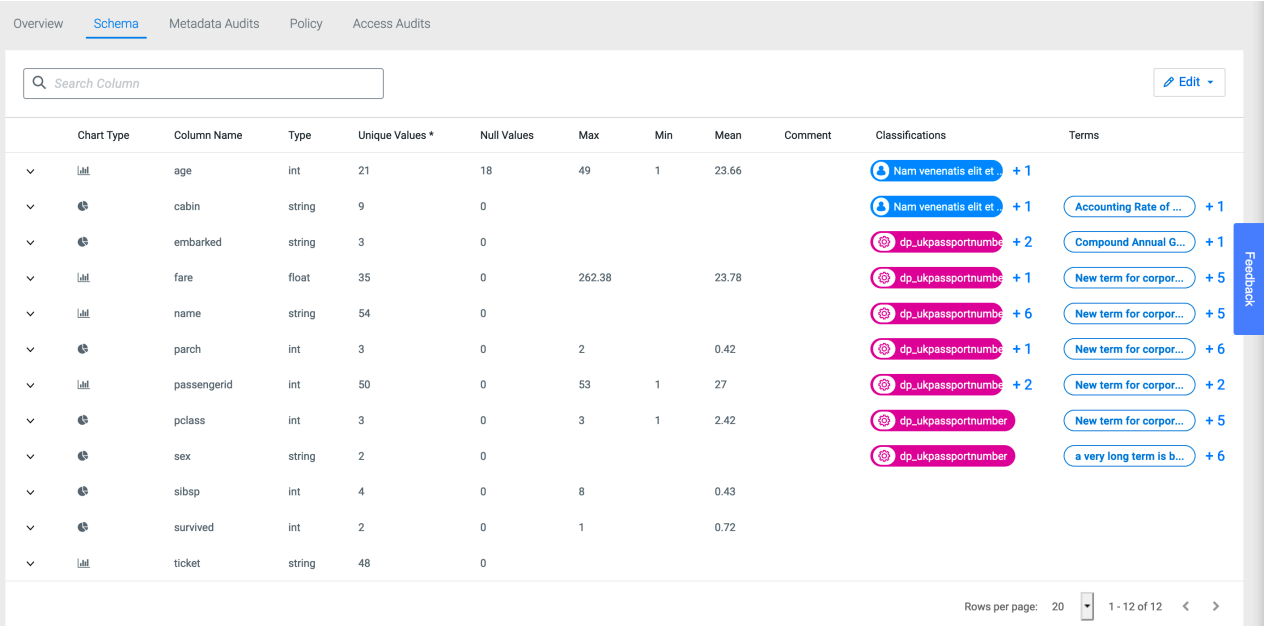


Additionally, you can also associate terms to your datasets by selecting one or more assets on the Data Catalog search page. You can associate terms with multiple datasets at a time.

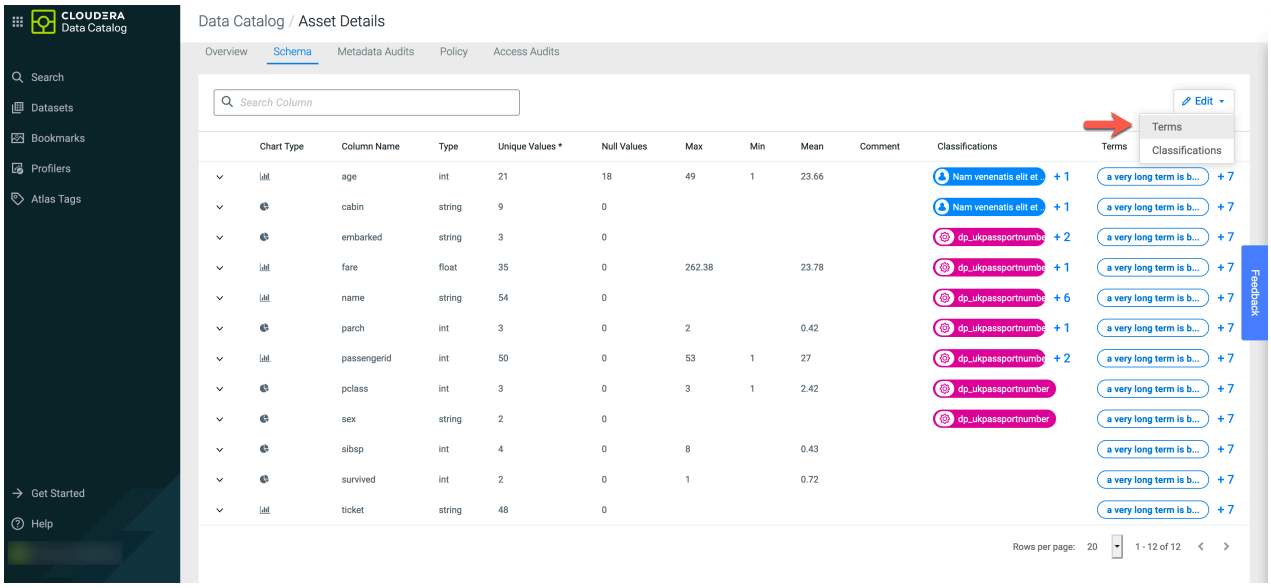




When you select a Hive table asset and navigate to the Asset Details page, under the Schema tab, you can view the list of terms associated with the asset.

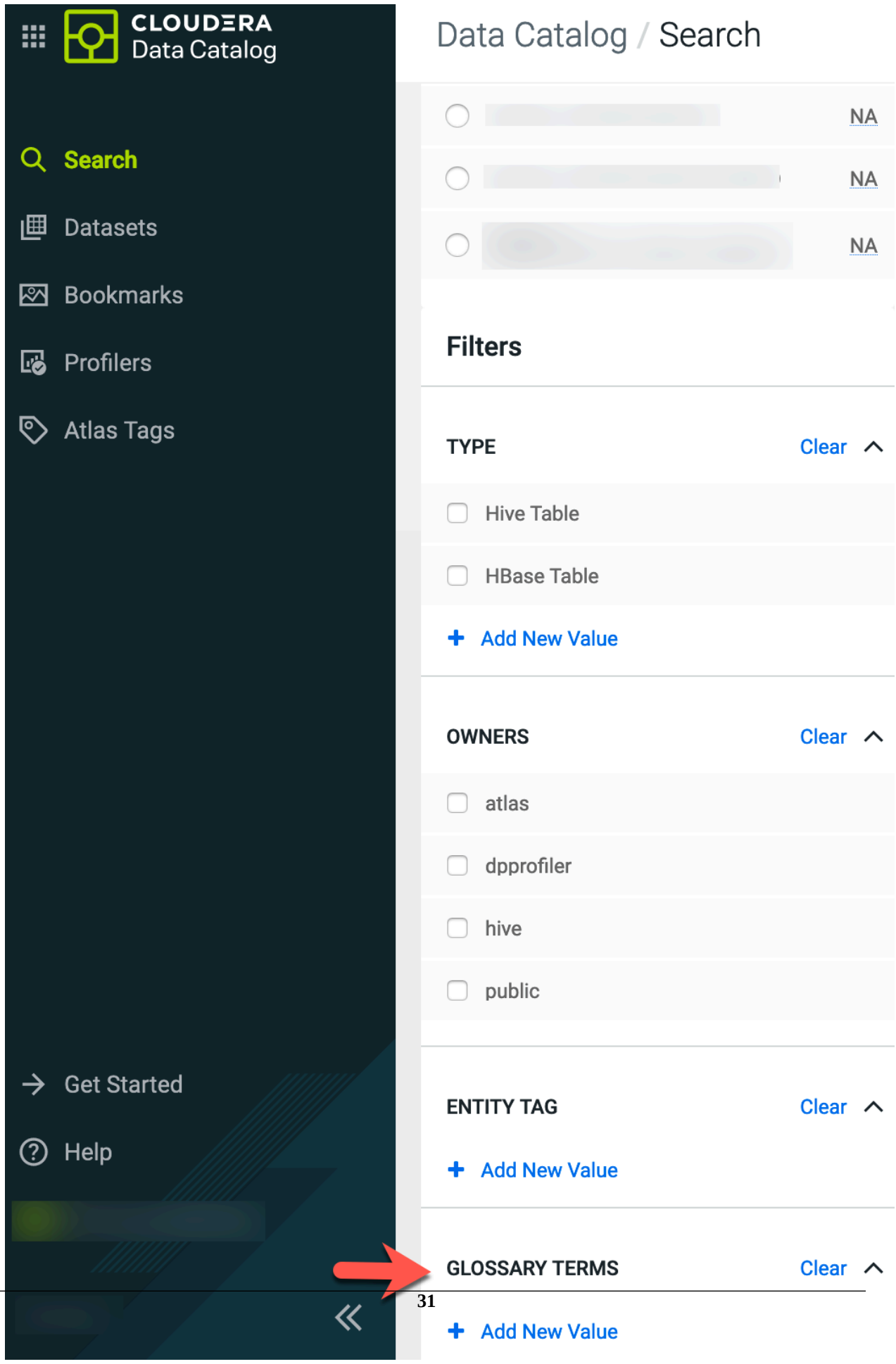


You can add or update the terms for the associated datasets by clicking the Edit button.



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

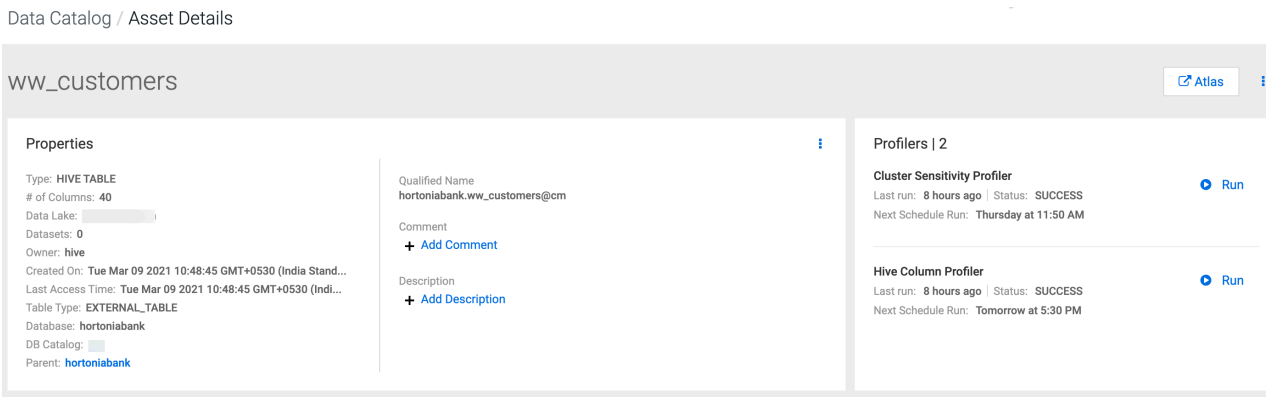
Additional search options for asset types

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.

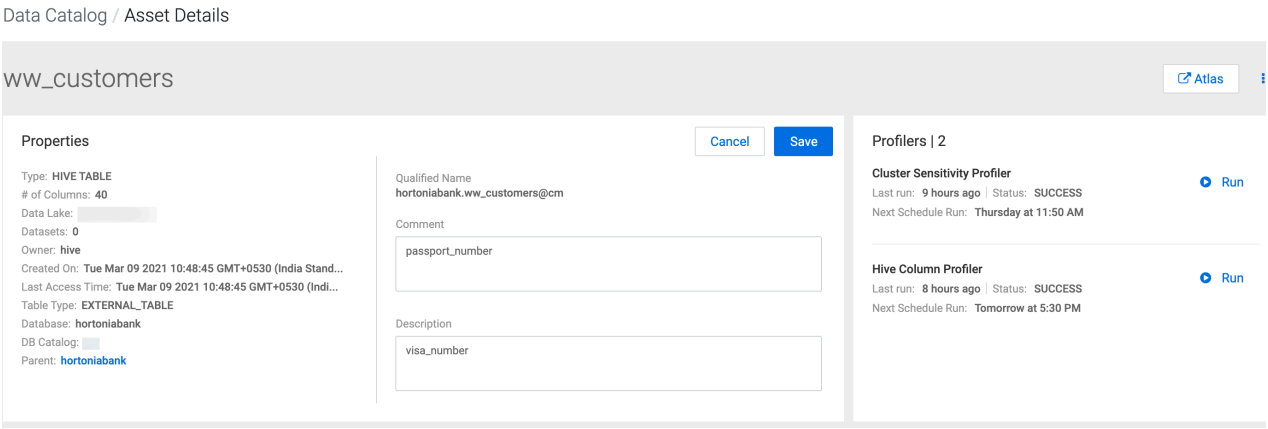
In the Asset Details page for each asset type that you select, you can add or edit comment and description fields. For each asset type in Data Catalog, you can add or edit comments or include a description. Including these values for the selected asset helps you to identify your chosen asset when you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

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

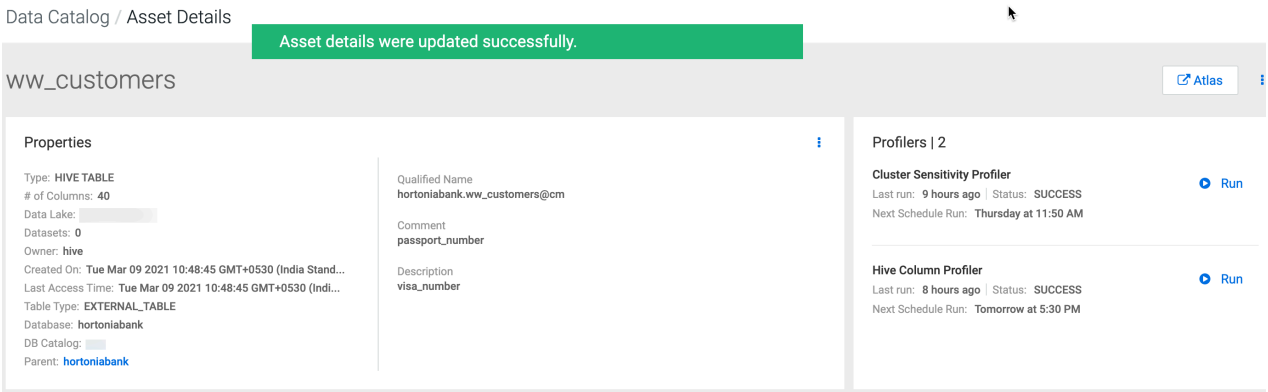
Note: The comment and description options are supported only for Hive table and Hive Column assets. For other asset types, only the description option is supported.



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

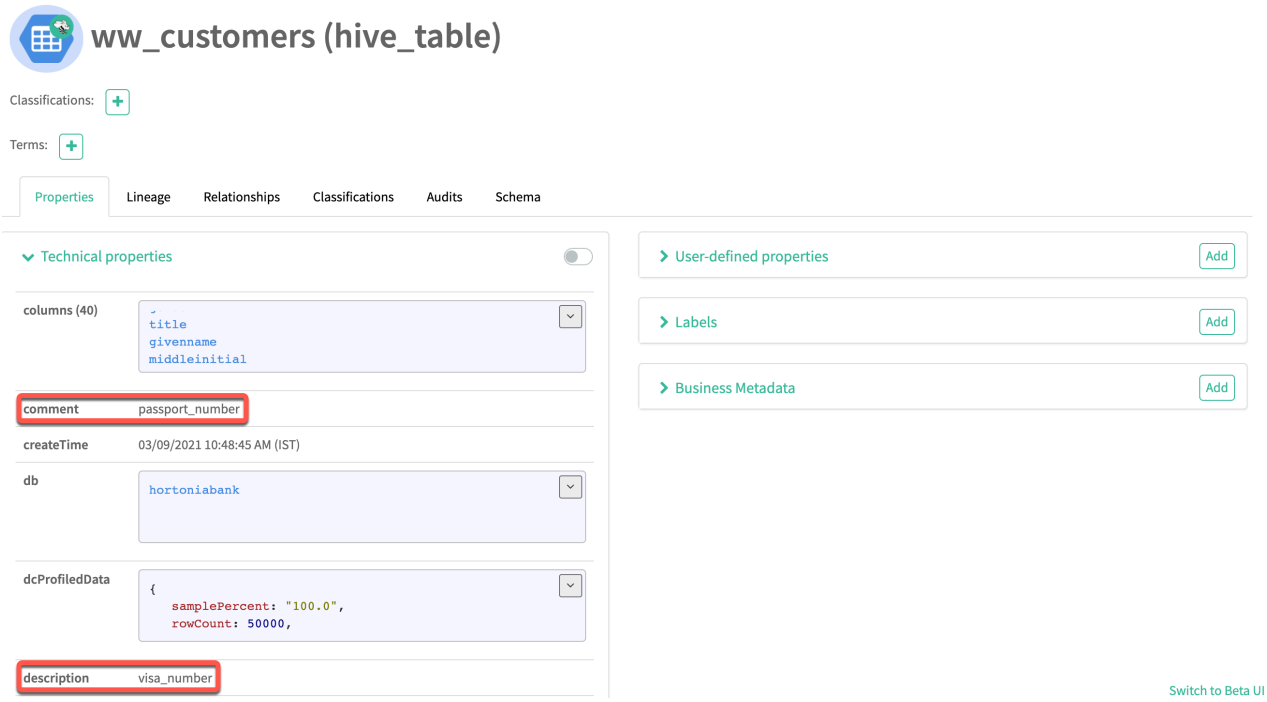


Click Save to save the changes.



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.

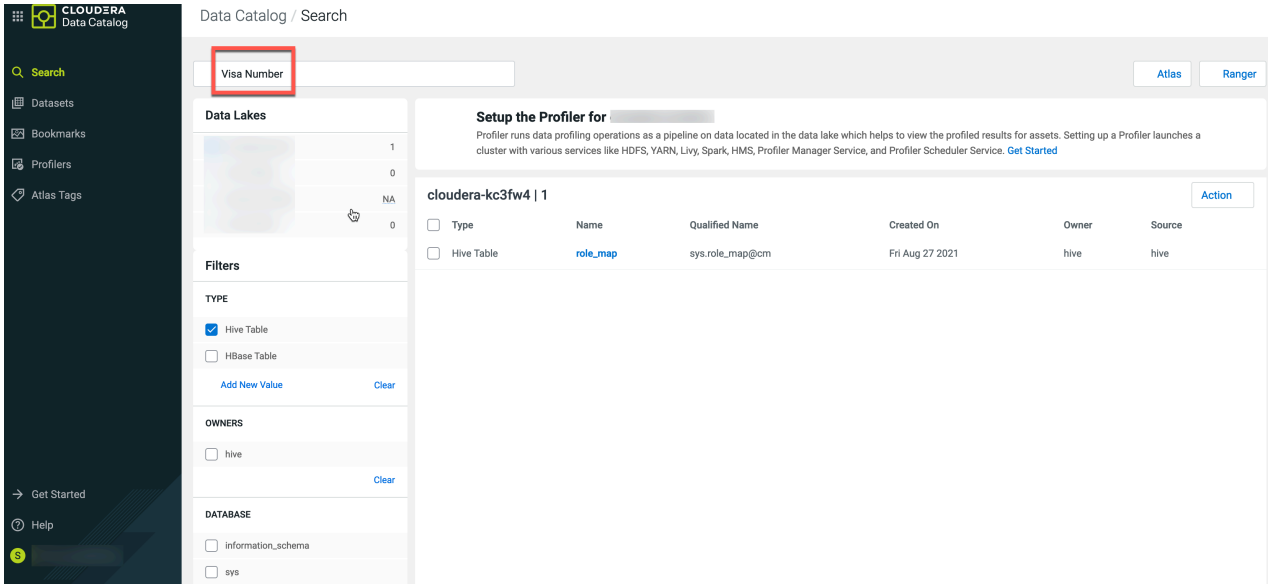


Searching for assets in Data Catalog using additional search options

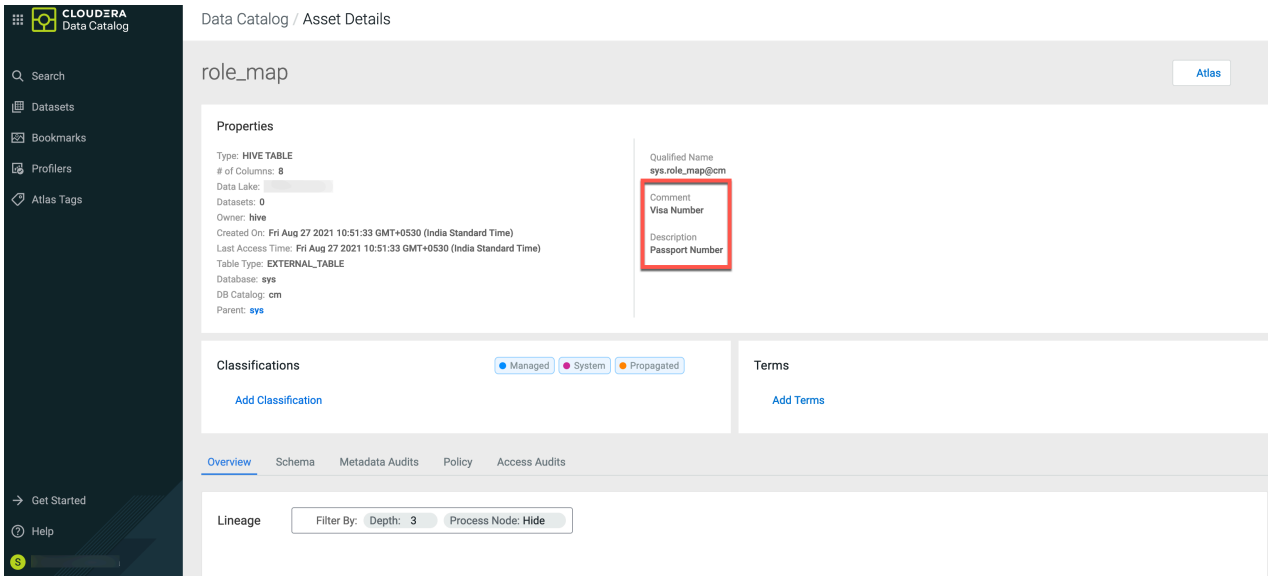
Consider a scenario in Data Catalog, where you select a data asset type and under the Asset Details page, you insert a comment and provide the description for the selected asset.

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.

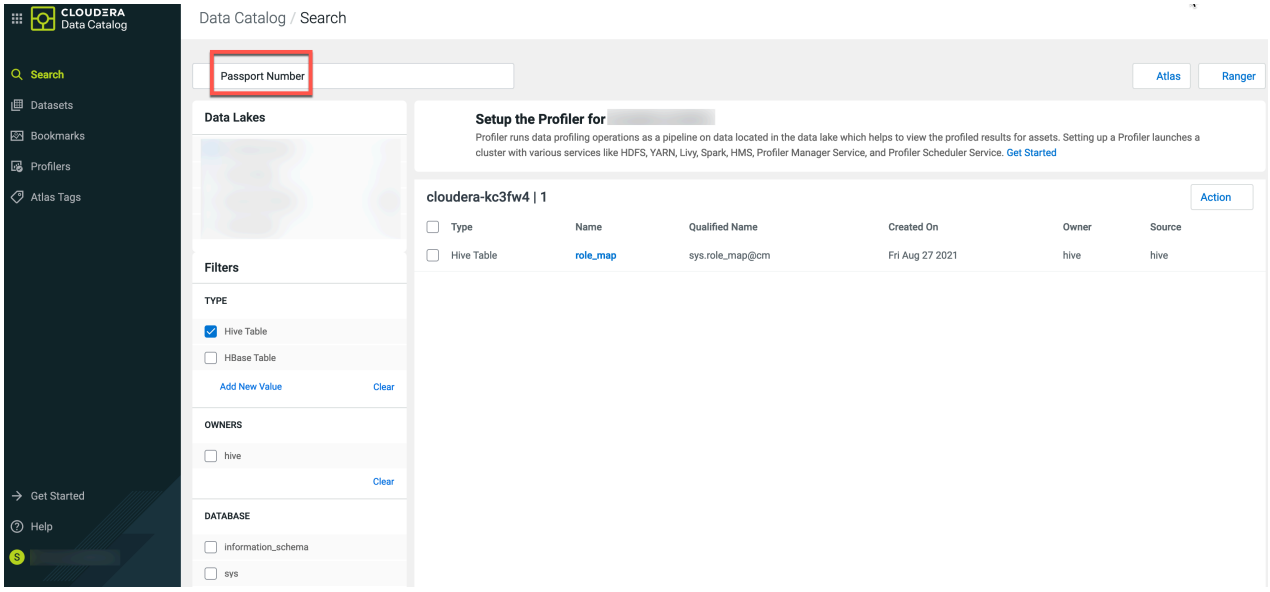
When you query for the entered Comment value for the selected asset type in Data Catalog, the relevant asset type is displayed in the search result page.



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.

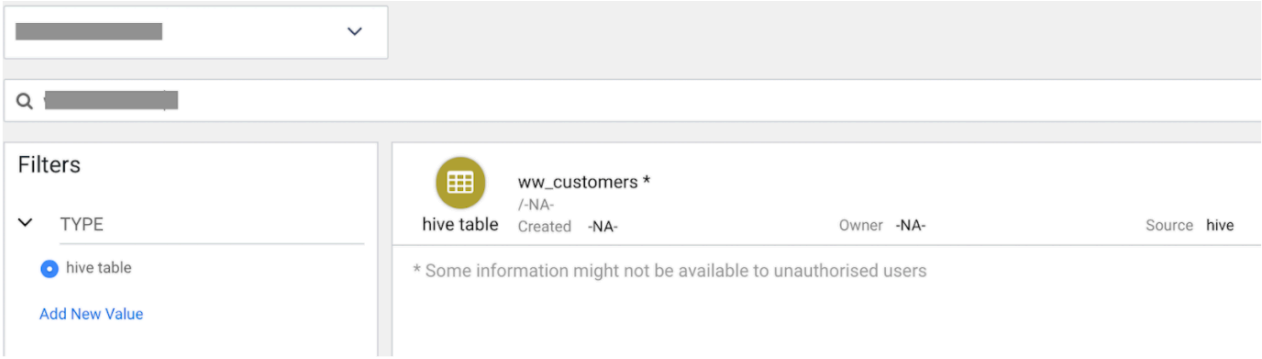
Accessing Tables based on Ranger policies

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

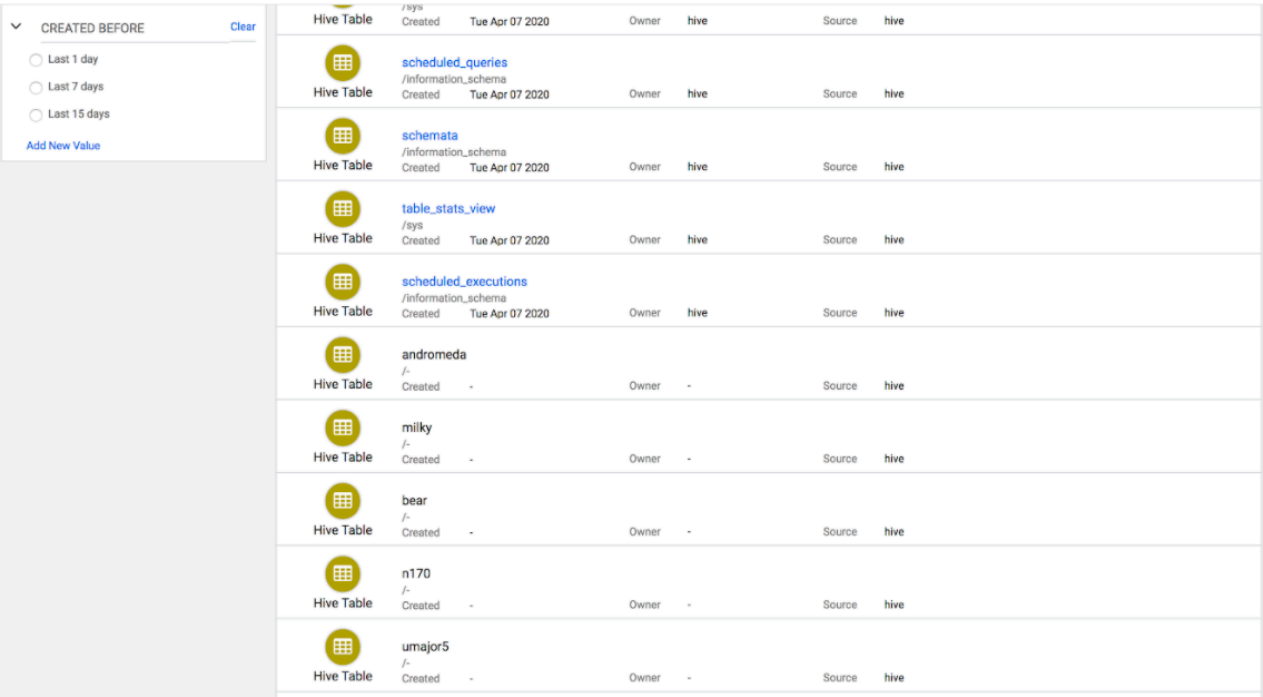
If a user is not authorized to click or view table details, it implies that the user permissions have not been set-up in the Ranger.

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:

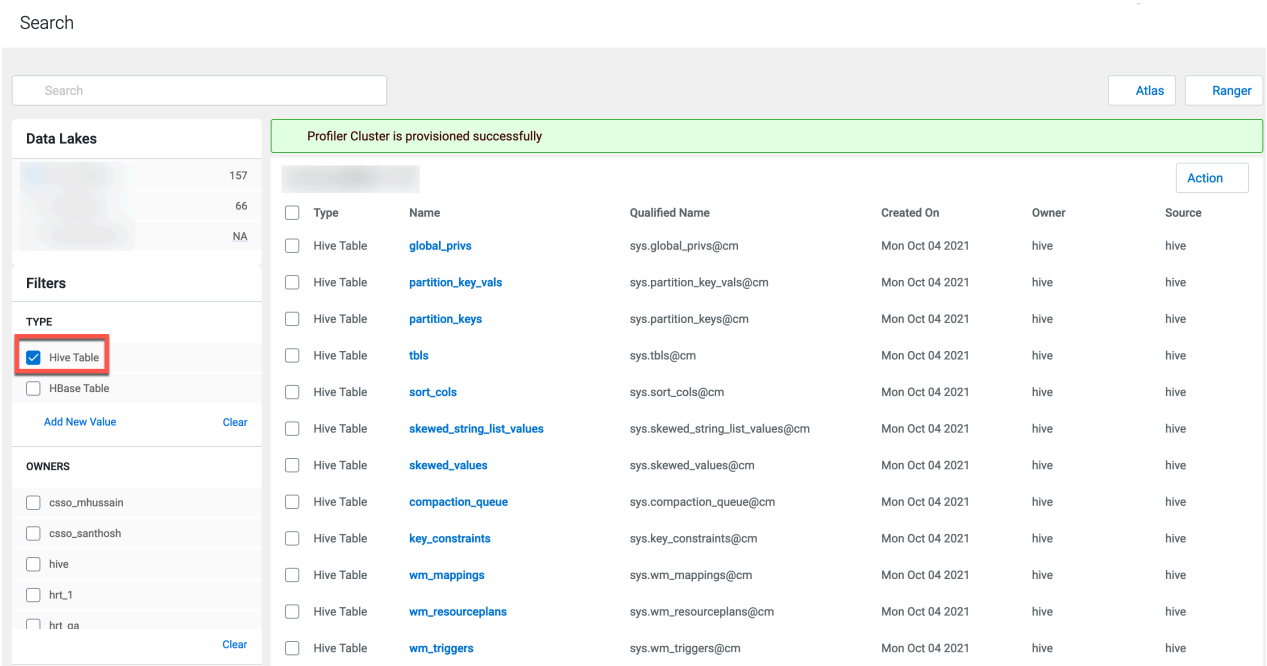
- Azure BLOB
- Azure Container
- Azure Directory
- AWS S3 Object
- AWS S3 V2 Object
- AWS S3 Bucket
- AWS S3 V2 Bucket
- AWS S3 Pseudo Dir
- AWS S3 V2 Directory
- HBase Table
- HBase Column Family
- HBase Namespace
- HDFS Path
- Hive DB
- Hive Table
- Hive Column
- ML Project
- ML Model Build
- ML Model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

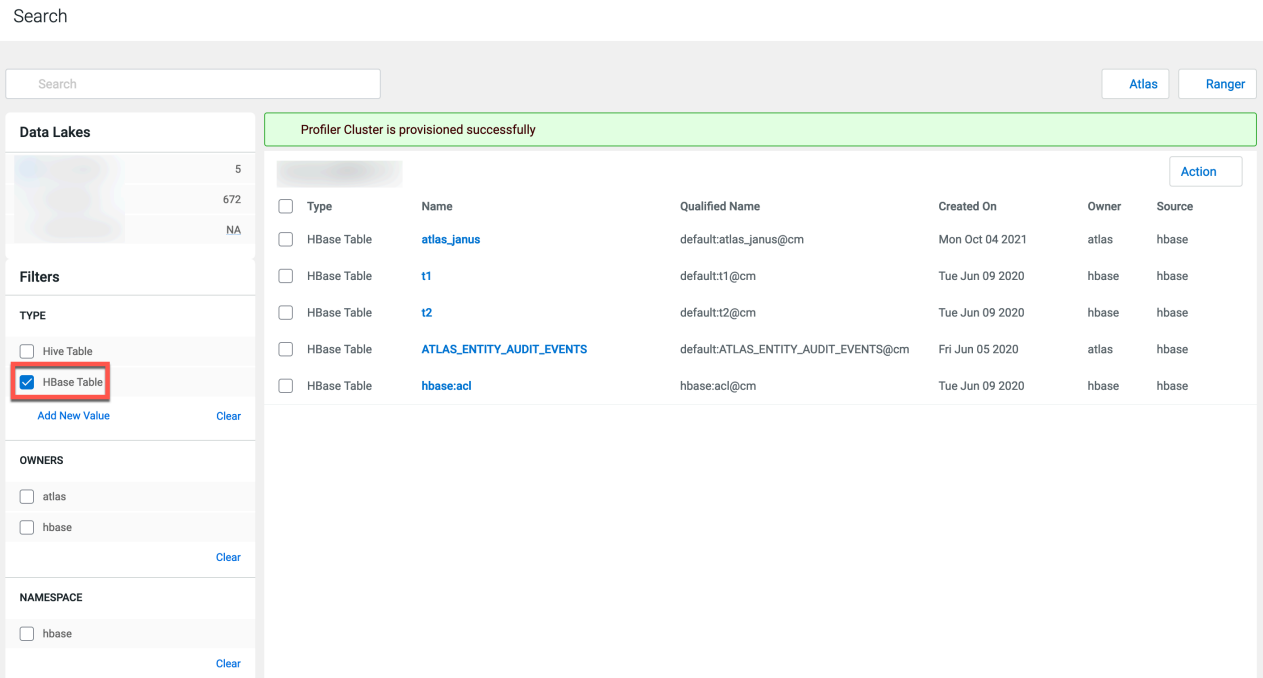
- RDBMS Foreign Key
- RDBMS Index
- RDBMS Instance
- RDBMS Table
- Spark Process
- Spark Application
- Spark Column
- Spark Column Lineage
- Spark DB
- Spark ML Directory
- Spark ML Model
- Spark ML Pipeline
- Spark Process Execution
- Spark Table

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.





Viewing Data Asset Details

The Asset Details page comprises four tabs (Overview, Schema, Policy, and Audit).

To access the Asset Details page, click an asset in the Data Catalog Search page. This brings you to the Overview tab, the first of the four tabs that form the Asset Details page.

- Overview: Displays an overview for the data asset.
  - Table properties: Number of rows, number of columns, number of partitions, owner, and tags
  - Lineage: Shows the chain of custody for the data from relevant metadata repositories such as Apache Atlas. Lineage shows both upstream paths (lineage) into and downstream paths (impact) out of a given asset.
- Schema: Displays the schema of the data asset for structured data (such as Hive tables) from the relevant metadata repositories (such as Atlas).
- Policy: The policy view shows security (authorization) policies defined on assets such as those present in Apache Ranger. It includes both resource (physical asset based) as well as classification based policies
- Audit: The data asset audit logs page shows the most recent access audits from Apache Ranger.

Viewing Data Assets

The Data Asset Overview page displays all the Apache Atlas metadata associated with a particular data asset.

About this task

The Data Asset Overview page displays:

Asset properties: Displays properties information relevant to asset type, like in case of Hive table - Number of rows, number of columns, number of partitions, and the owner.

From the Data Catalog search page, click to select a data asset.

The Asset Overview window opens.

The following matrix captures the supported fields for different asset types:

Asset Type	Lineage	Tagging	Access Metrics	Schema	Policy	Audit	Atlas Punch out
aws_S3_bucket	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
aws_S3_Object	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
aws_S3_pseudo_dir	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
aws_s3_v2_object	Yes	Yes	Not Supported	Not Supported	Not Supported	Yes	Yes
aws_s3_v2_directory	Yes	Yes	Not Supported	Not Supported	Not Supported	Yes	Yes
aws_s3_v2_bucket	Yes	Yes	Not Supported	Not Supported	Not Supported	Yes	Yes
adls_gen2_directory	Yes	Yes	Not Supported	Not Supported	Not Supported	Yes	Yes
adls_gen2_blob	Yes	Yes	Not Supported	Not Supported	Not Supported	Yes	Yes
adls_gen2_container	Yes	Yes	Not Supported	Not Supported	Not Supported	Yes	Yes
Hive DB	Not Supported	Yes	Not Supported	Not Supported	Yes	Yes	Yes
Hive Table	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Hive Column	Yes	Yes	Not Supported	Not Supported	Yes	Yes	Yes
Hbase Namespace	Yes	Yes	Not Supported	Not Supported	Yes	Yes	Yes
Hbase Table	Yes	Yes	Not Supported	Yes	Yes	Yes	Yes
Hbase Column Family	Yes	Yes	Not Supported	Not Supported	Yes	Yes	Yes
Iceberg Table	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Iceberg Column	Yes	Yes	Yes	Not supported	Yes	Yes	Yes
impala_process	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
impala_column_lineage	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
impala_process_execution	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
ML_Project	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
ML_Model_Build	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
ML_Model_Deploy	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_db	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_column	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_foreign_key	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes

Asset Type	Lineage	Tagging	Access Metrics	Schema	Policy	Audit	Atlas Punch out
rdbms_index	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_instance	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
rdbms_table	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_process	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_application	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_column	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_column_lineage	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_db	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_ml_directory	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_ml_model	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_ml_pipeline	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_process_execution	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes
spark_table	Yes	Yes	Not Supported	Not Supported	Not Supported	Not Supported	Yes

View Data Asset Schema

From the Asset Details Schema page, you can view the schema of the data asset for structured data (such as Hive tables) from the relevant metadata repositories (such as Atlas).

Procedure

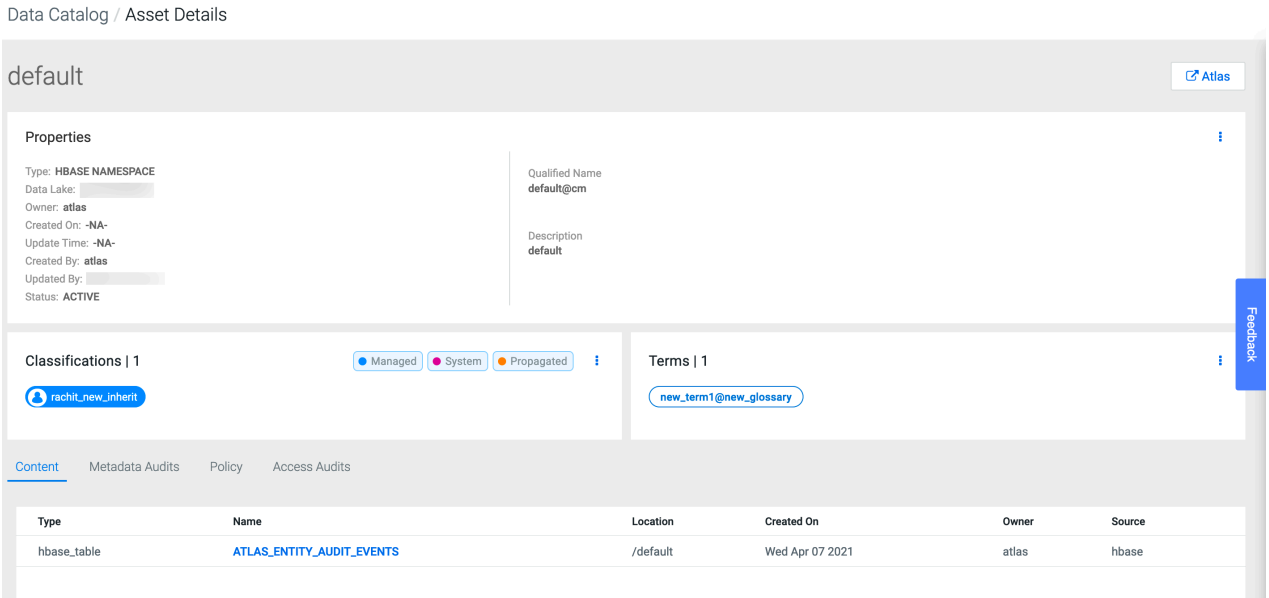
1. From the Data Catalog search page, select an asset.
   The Asset Overview window opens.
2. Click Schema.
   The Schema table shows the data asset schema as retrieved from Apache Atlas.
3. (Optional) To edit tags:
   a) Click Edit Tags.
   b) Click the (+) icon.
   c) Select or deselect the tags you choose, then click Save.

      You can now manage and edit tags at the table level.

Navigating from the container asset to the parent asset from Asset Details page

A generic Assets Details page is available for container data types like buckets and databases.

The Contents tab (similar to the Schema tab) lists all the contents of the selected entity. Clicking on any element available in the selected entity list navigates you to the Asset Details page.



For example, for a database entity having a list of tables, clicking on any listed table navigates to the Asset Details page of the same table. This page helps you understand the parent-child relationship as far as asset management is concerned. The Contents tab displays entities that are contained within assets of container entity types. The entities in the table of Contents tab are clickable, which will allow you to navigate to the Asset Details page of these contained assets.

The following table lists the entity types, their parent, and contents.

Type	Parent	Content
Hive DB	-	Hive Table, Iceberg Table
HBase Namespace	-	HBase Table
ML Project	-	ML Model Build
ML Model Build	ML Project	ML Model Deployment
AWS S3 Bucket	-	AWS S3 Pseudo Dir
AWS S3 Pseudo Dir	AWS S3 Bucket	AWS S3 Object
RDBMS Instance	-	RDBMS DB
RDBMS DB	RDBMS Instance	RDBMS Table

View Authorization Policies on a Data Asset

The Asset Details Policy page displays all the Apache Ranger policy details associated with a particular data asset. This helps you understand how data access is secured and protected: what users can see what data (or metadata) under what conditions (security policies, data protection, and anonymization).

Procedure

1. From the Data Catalog search page, select a data asset.
   The Asset Overview window opens.
2. Click the Policy tab.
   The Policy table shows the data asset policies as retrieved from Apache Ranger.

View Data Asset Audit Logs

The Asset Details Audit page displays all the Apache Ranger audit events associated with a particular data asset. This helps you to view who has accessed what data from a forensic audit or compliance perspective, and to visualize access patterns and identify anomalies.

Procedure

1. From the Data Catalog search page, select a data asset.
   The Asset Overview window opens.
2. Click the Audit tab.
   The Audit table shows the most recent raw audit event data by type of access and access outcome (authorized/unauthorized).
3. (Optional) You can filter the audit results by Access Type or Result.

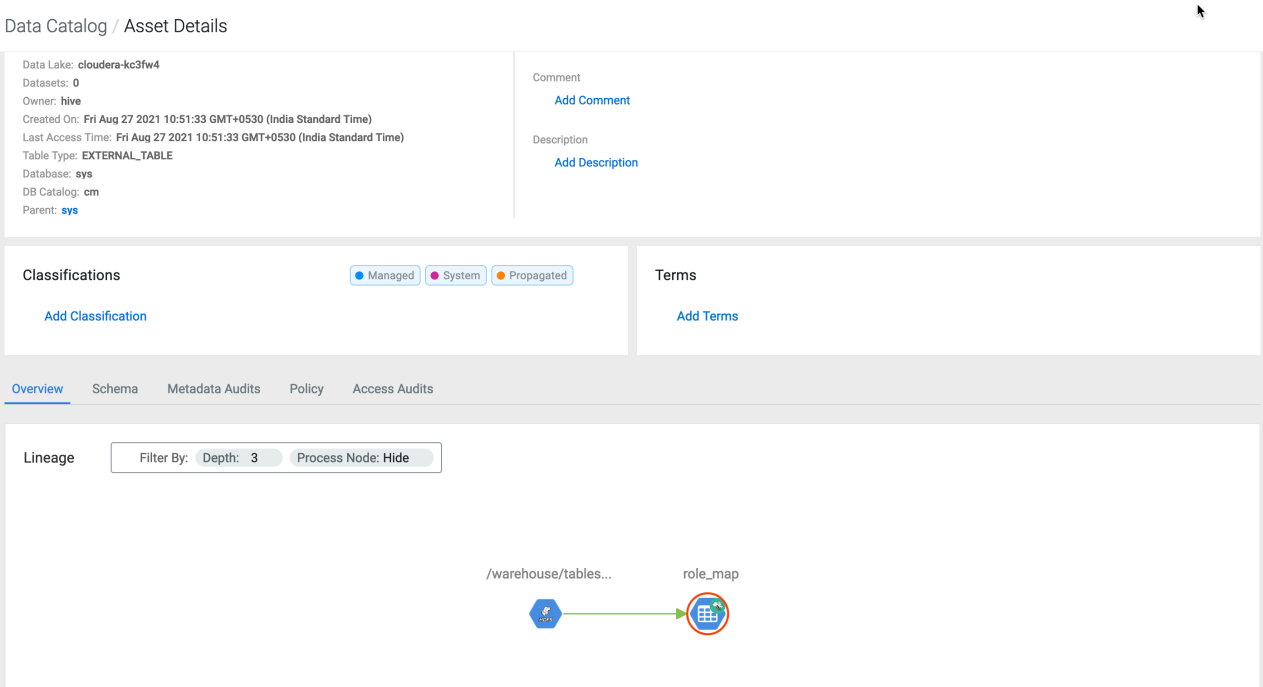
   Access type: SELECT, UPDATE, CREATE, DROP, ALTER, INDEX, READ, WRITE.

   Result: ALLOWED, DENIED.

Navigation Support for Hive entity within Lineage

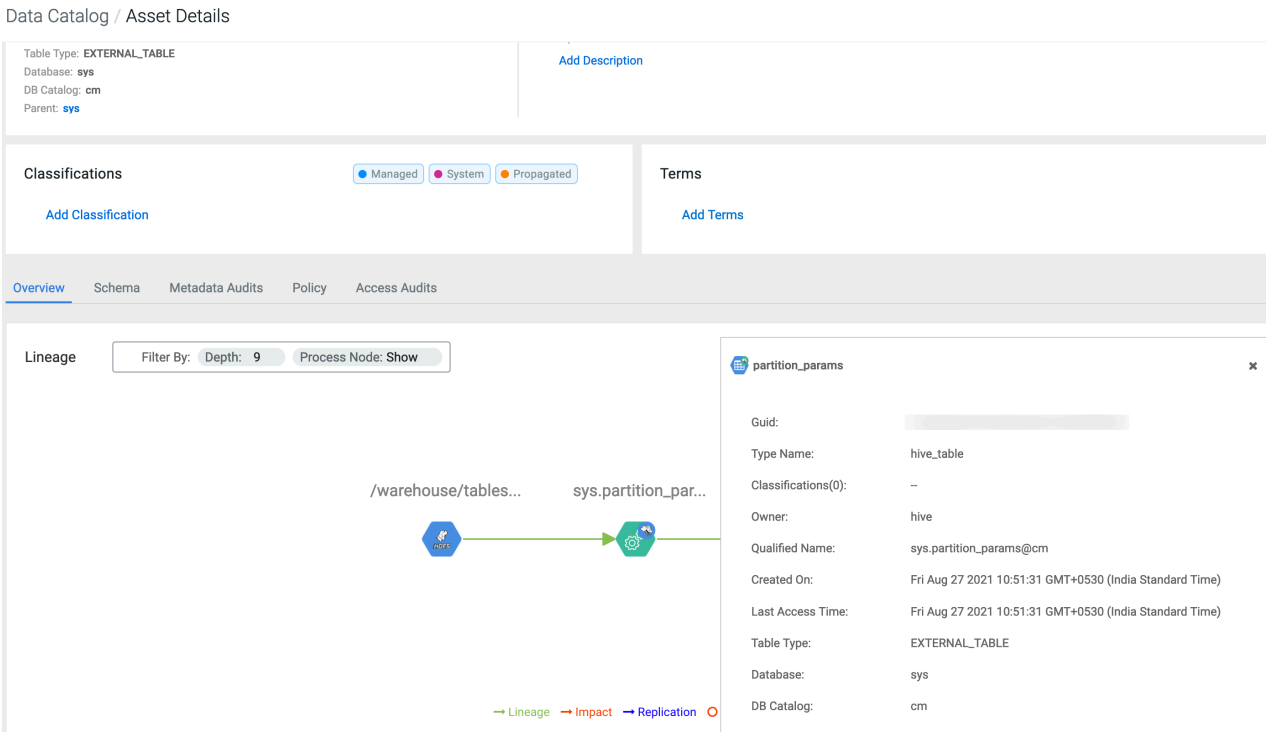
When you click a Hive entity within lineage, the Asset Details page of the selected Hive entity is displayed.

Previously, when you clicked on any entity for which slider information was available, a slider would display the entity details. As of now, as seen in the corresponding images, the Asset Details page of the Hive entity is displayed. The option selected in Depth drop-down and Show Process nodes are now displayed on the upper-left corner of the Lineage module.



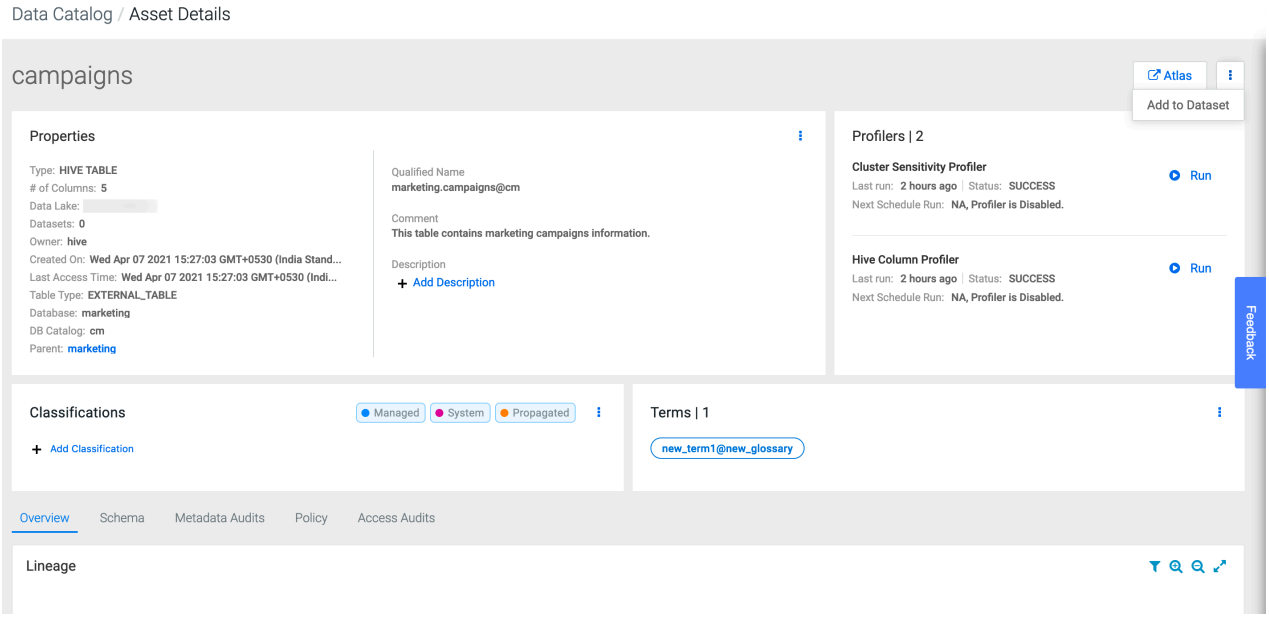
Alternatively, if you do not want to navigate away from the current page and want to view the information with respect to any entity, hover on the entity and click the information icon to view the details.

The screenshot depicts the slider information for the clicked entity:



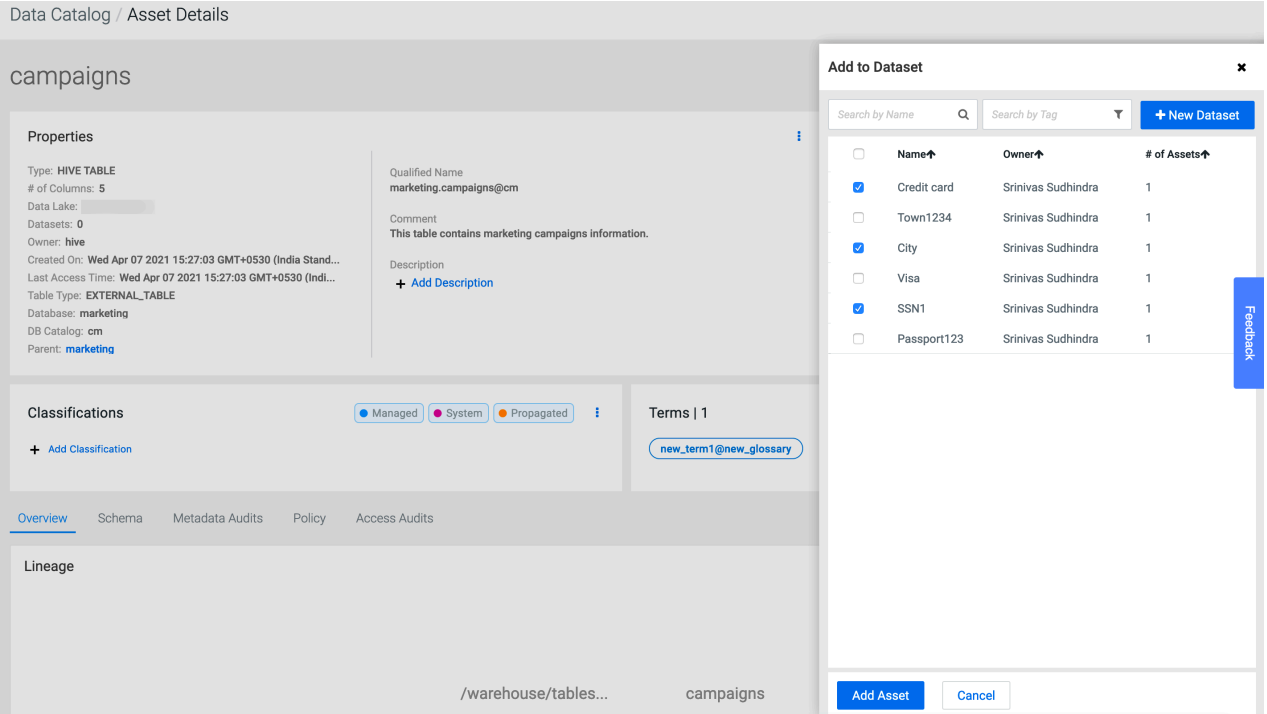
Adding Hive asset to one or more datasets on Asset Details screen

On the Asset Details screen, users are provided with an option to add the asset to the dataset as shown in the diagram.

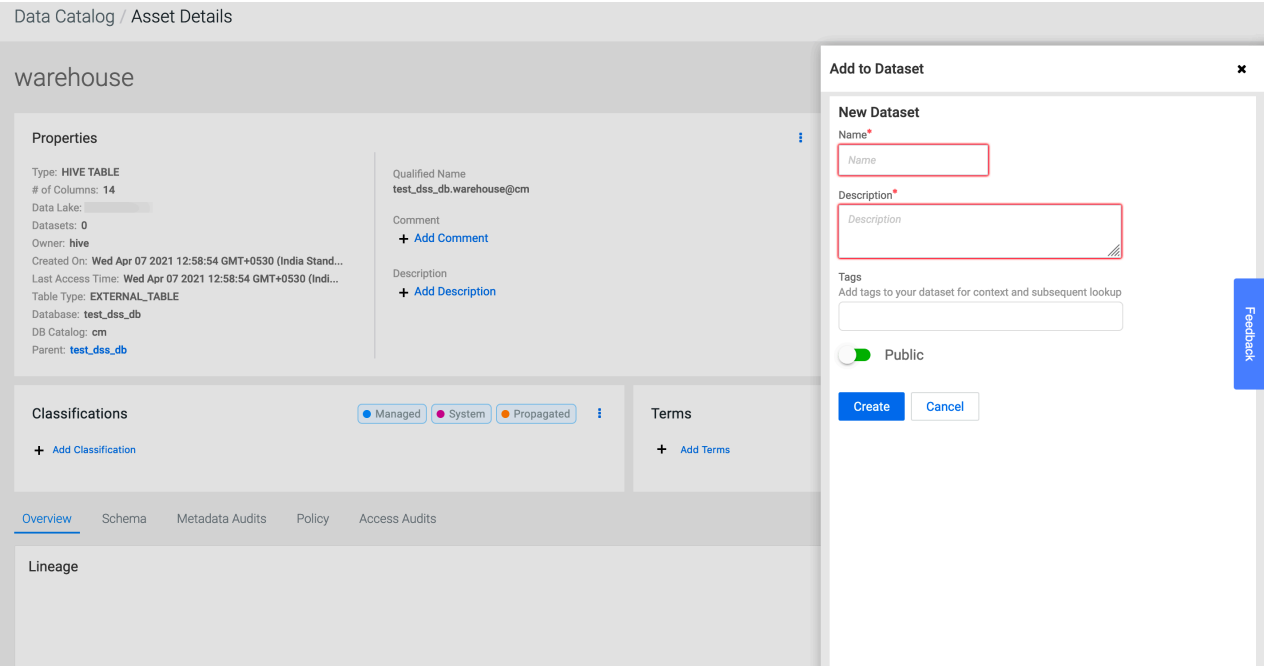


The Add to Dataset window provides an option to add the asset into one or more existing datasets or even create a new one.

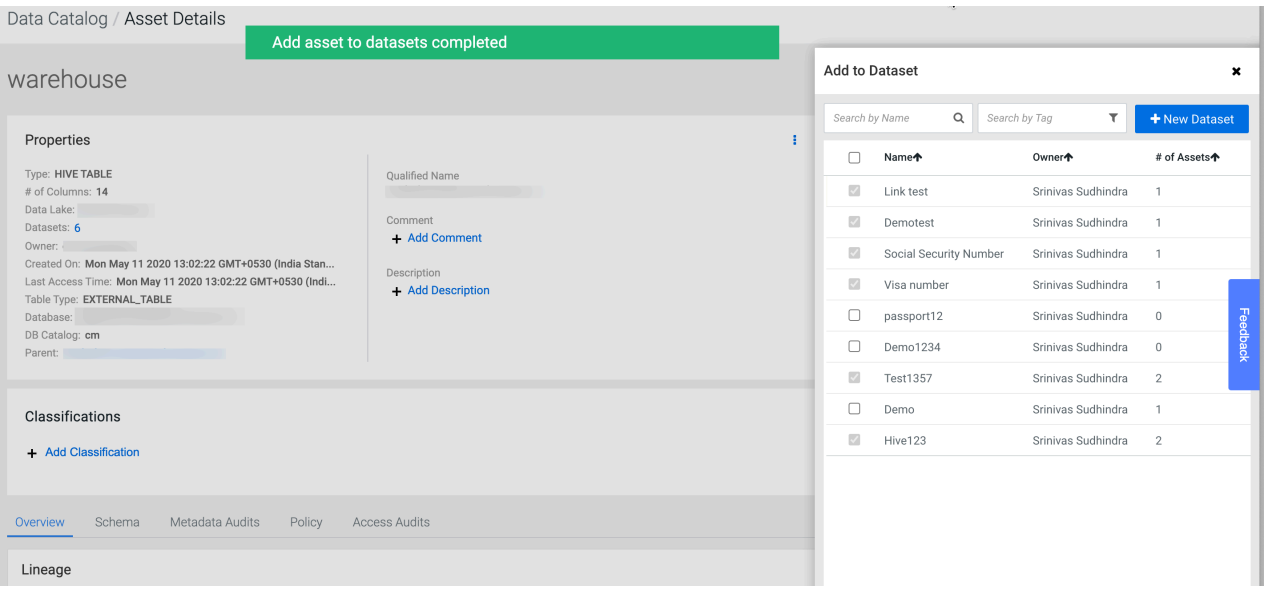
Datasets that already contain the asset are disabled and marked as checked. Datasets which are currently in edit state are disabled and marked with a *.



Users can search for an existing dataset by name or by tags applied on the dataset. Users can select one or more datasets from the list and then click on the Add Asset button which adds the asset to these dataset(s).



There are instances, where there are no datasets present or the user just wants to create a new dataset to add the asset. In that case, the user can click on the New Dataset button which opens up a new dataset form. Once the user fills in the form and clicks on the Create button, a new dataset with the given properties is created and the asset is added to it automatically. This is reflected in the datasets list where the newly added dataset is highlighted.

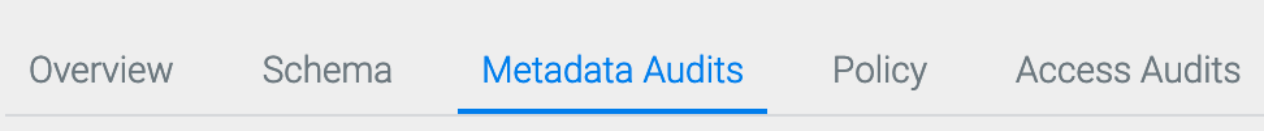


Viewing Atlas Entity Audits

In Data Catalog, Atlas audits help Data Stewards to identify and track the entity changes or modifications that are performed over a period of time.

Information about the Atlas entity audit events are displayed for each entity in the Asset Details page in Data Catalog. Using this information, Data Stewards can distinguish between entity audits and data audits that emanate from Ranger.

On the Asset Details page, a new tab called Metadata Audits displays information related to the selected entity type and about the events that occurred based on the user activities.



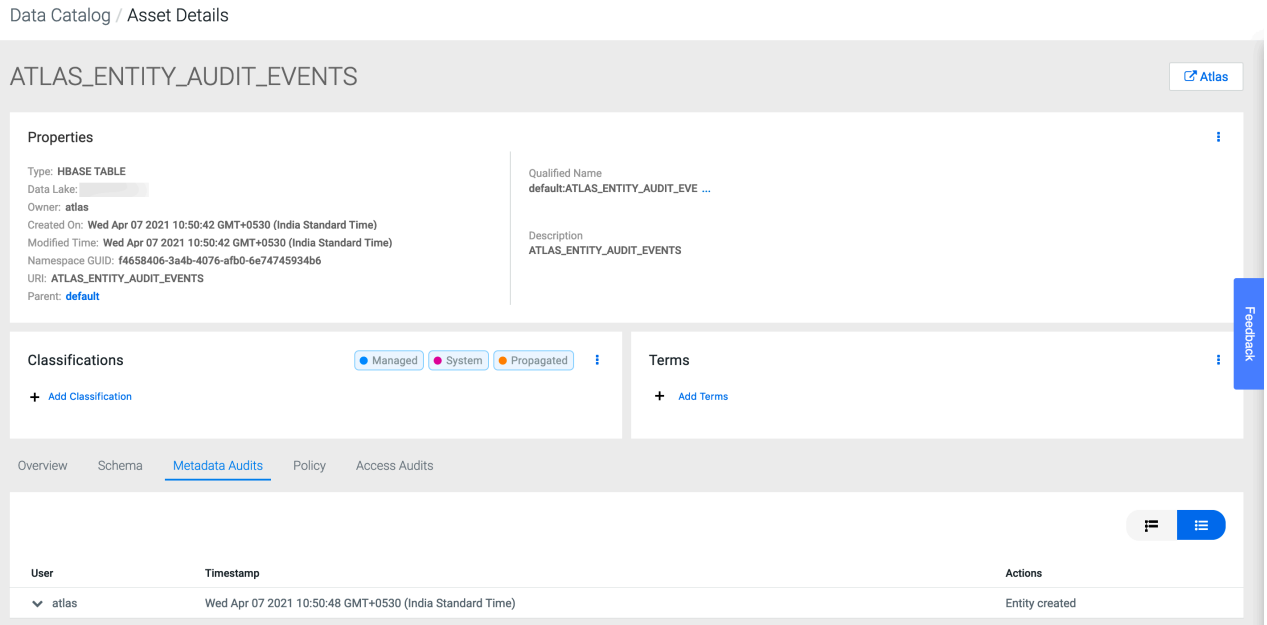
Clicking on Metadata Audits, tab, you can view manage information about:

- The user who made the changes to the specific entity
- The time when the entity was changed
- The kind of change that was made to the entity
- Any other relevant changes pertaining to the audit entries

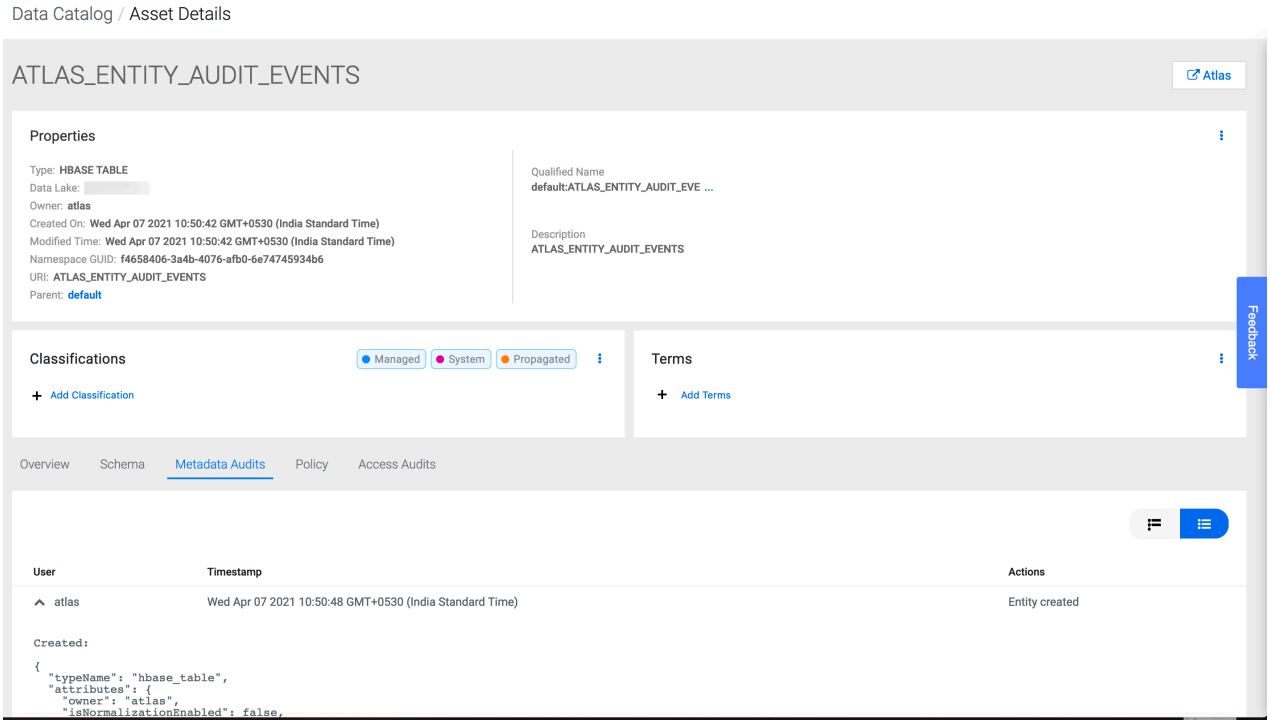
The changes that can be identified for:

- Created entities and related updates
- Tagged entities
- Labeled entities
- Export and Import operations

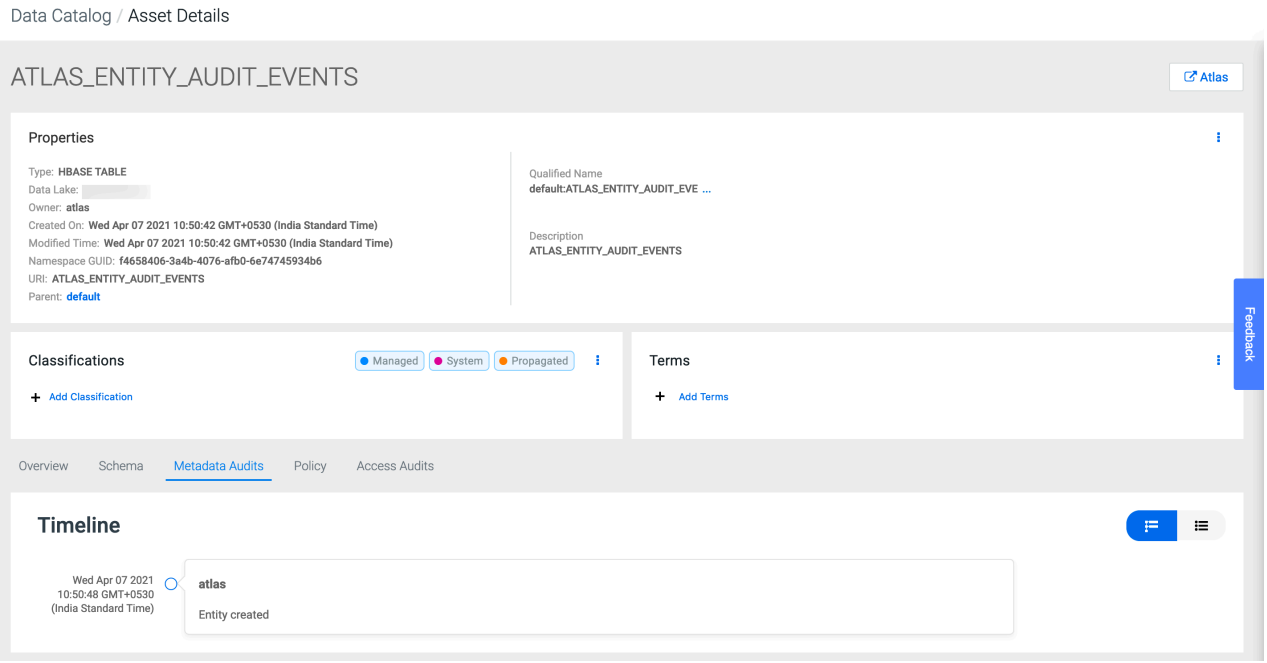
For example, the following image displays information about the Atlas audit events that are performed by each Atlas user that is displayed in the Asset Details page in Data Catalog.



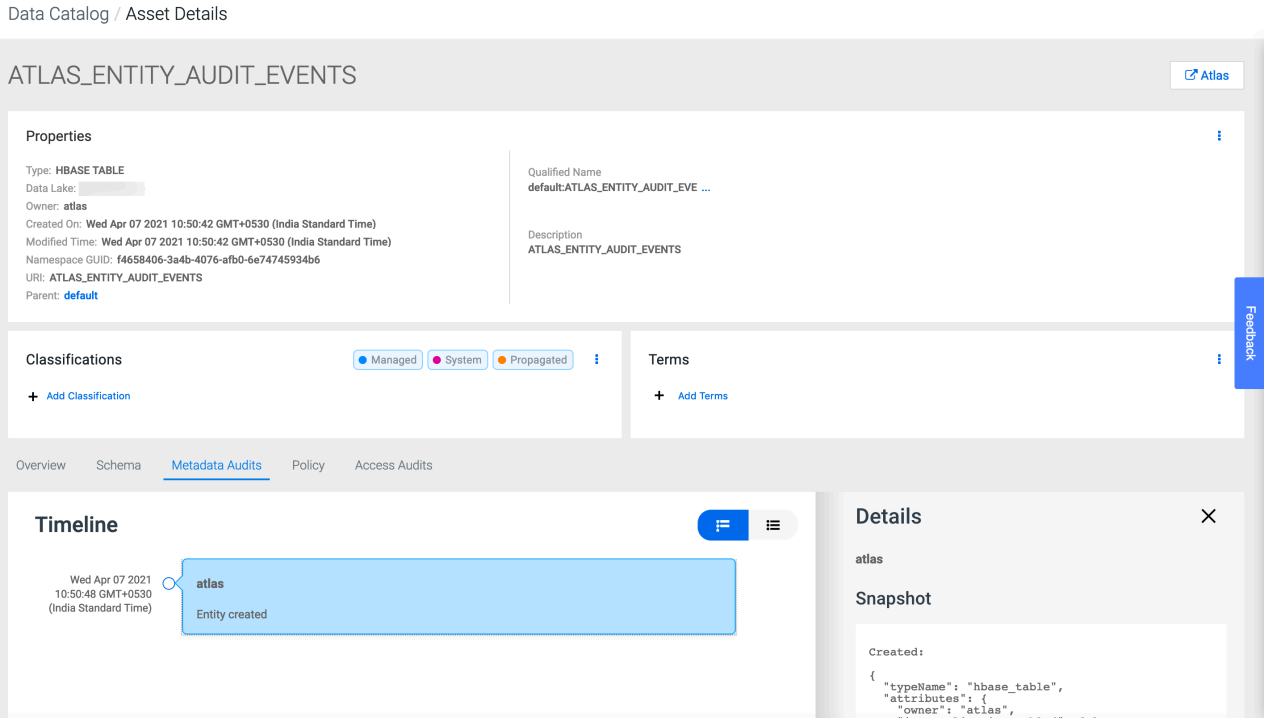
Clicking on any line item displays the JSON format, which is directly derived from Atlas, in other words the source of data available in Atlas.



Use the toggle icon (on the top-right corner) for viewing Atlas Audits in different formats. By default, you can view Metadata Audits in tabular format in the Asset Details page and when you toggle the view icon, you can view the Timeline format. The events are listed as timelines in this format.



Clicking on a user in the Timeline format displays the JSON data, which is again derived from Atlas.



Managing Profilers

The Cloudera 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.

Important: Profilers do not support Iceberg tables in this release.

Related Information

Understanding the Data Catalog Profiler
Understanding the Cluster Sensitivity Profiler
Understanding the Ranger Audit Profiler

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:

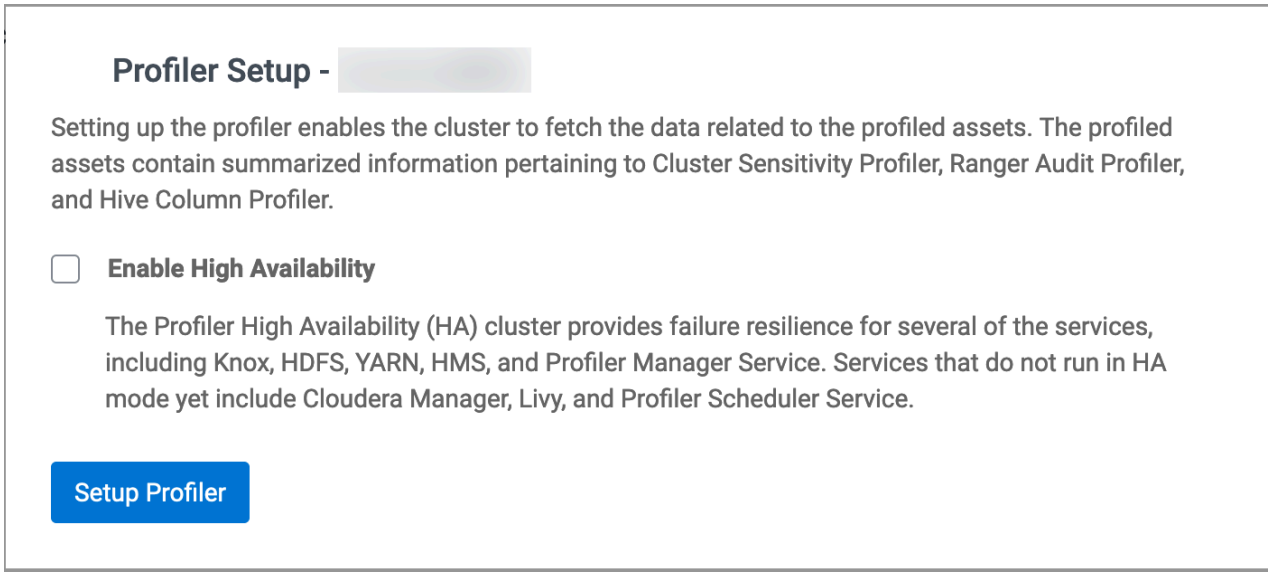
- Profiler Manager
- Profiler Scheduler

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.

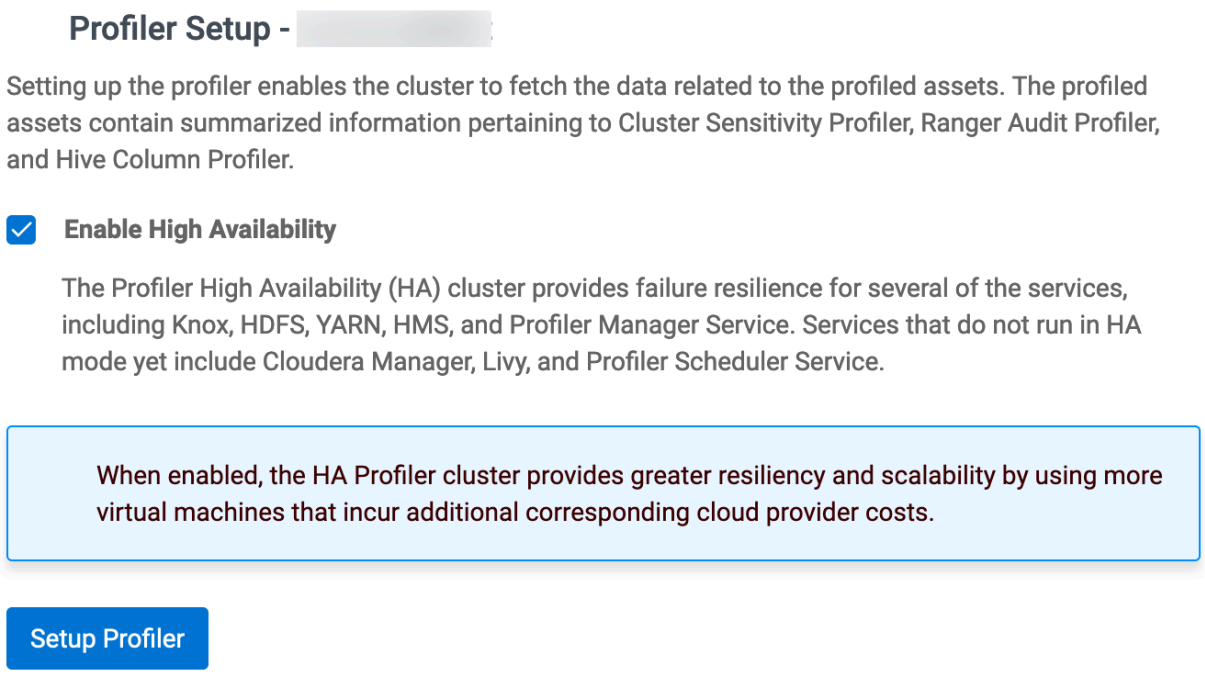
How to launch the cluster profiler

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.

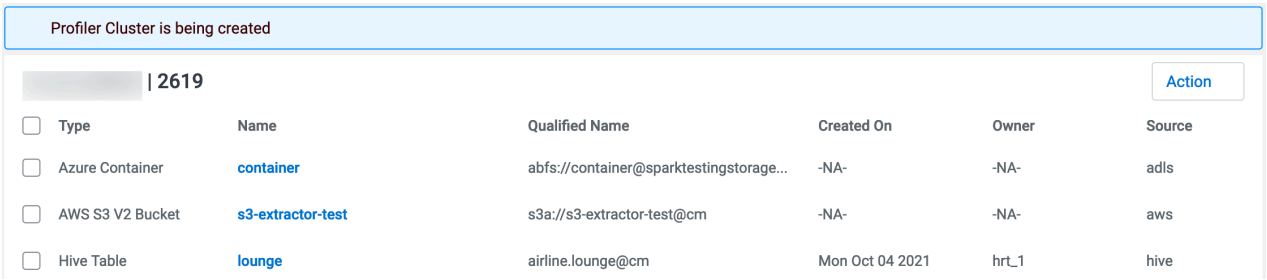


For setting up the profiler, you have the option to enable or disable the HA.

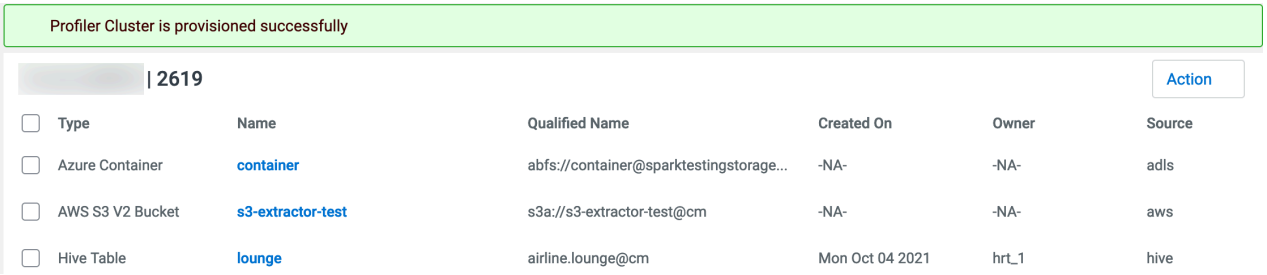
Note that the HA functionality is being supported only from Cloudera Runtime 7.2.10 release onwards. If you are using the Cloudera Runtime version below 7.2.10, you shall not be able to use the HA feature for launching the profiler services.



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

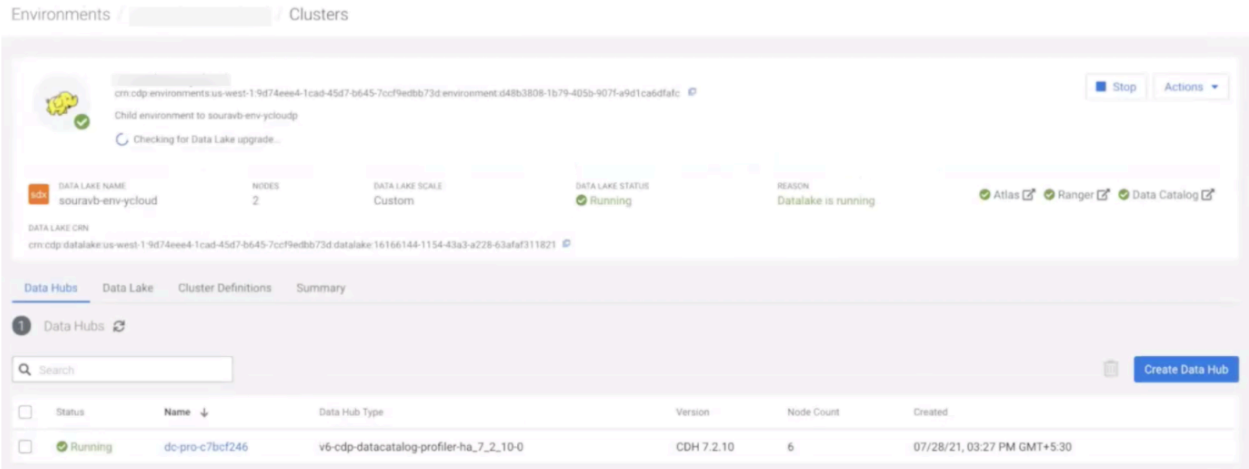


Later, a confirmation message appears that the profiler cluster is created.



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.

Related Information

Understanding the Data Catalog Profiler
Understanding the Cluster Sensitivity Profiler
Understanding the Ranger Audit Profiler

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 clusters

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. removes all the Custom Sensitivity Profiler rules and other updates to the specified profiler.

About this task

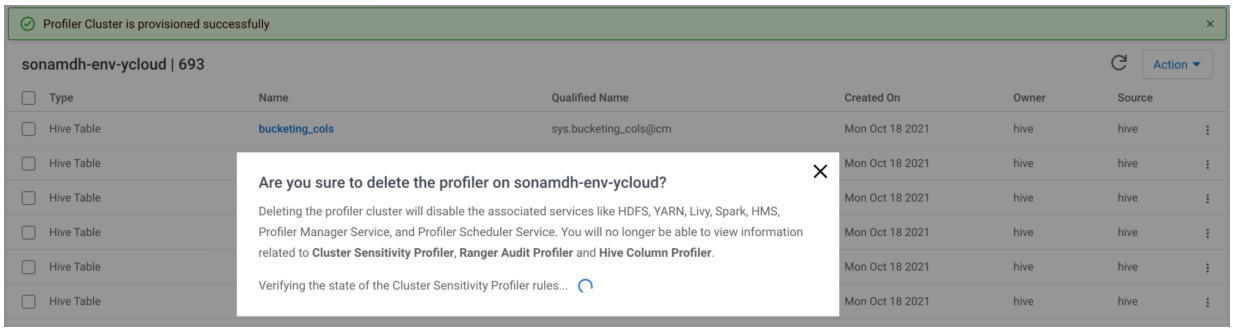
To overcome this situation, when you decide to delete the profiler cluster, there is a provision to retain the status of the Custom Sensitivity Profiler rules. If your profiler cluster has rules that are not changed or updated, you can directly delete the profiler cluster. If the rules were modified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

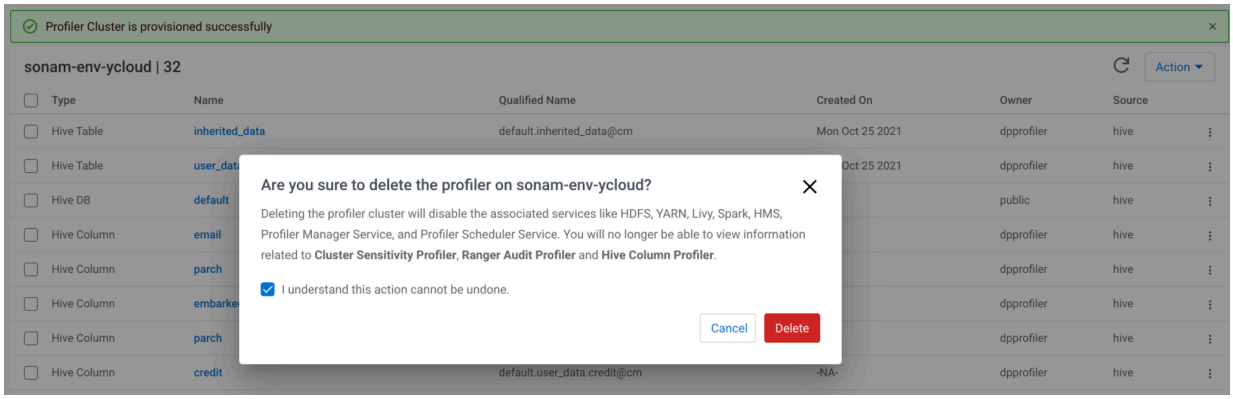
1. On the search page, select the Data Lake from the list.
2. Click the Actions drop-down menu and select Delete Cluster.
3. Click Yes to proceed.
4.

5. Click Delete.

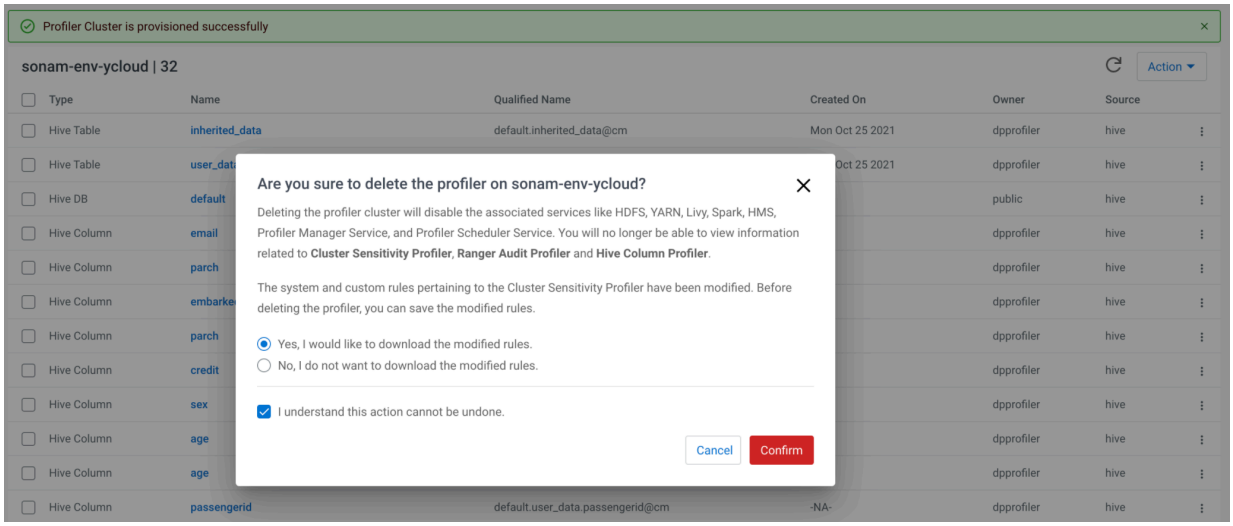
   The application displays the following message.



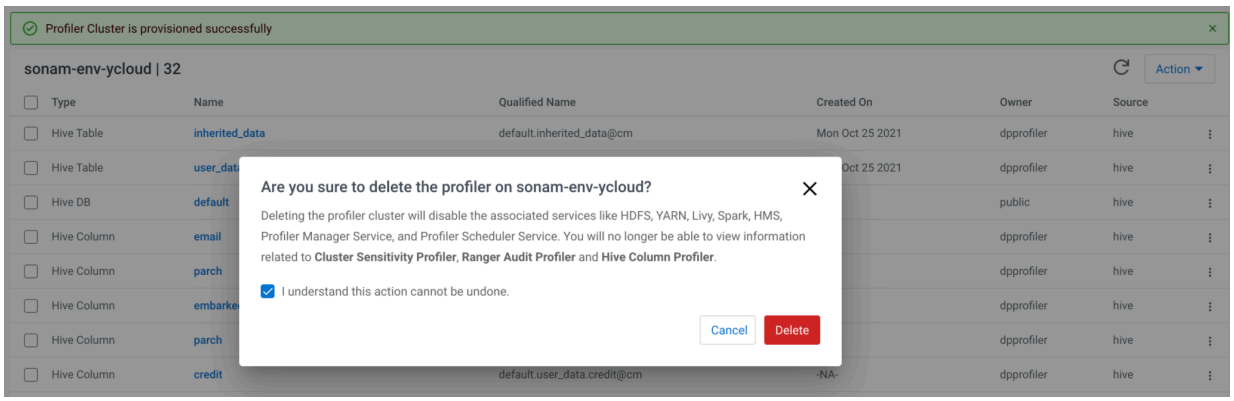
When you upgrade / launch Data Catalog service in Cloudera Runtime version 7.2.14, and later if the profiler cluster is deleted, the following message is displayed.



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

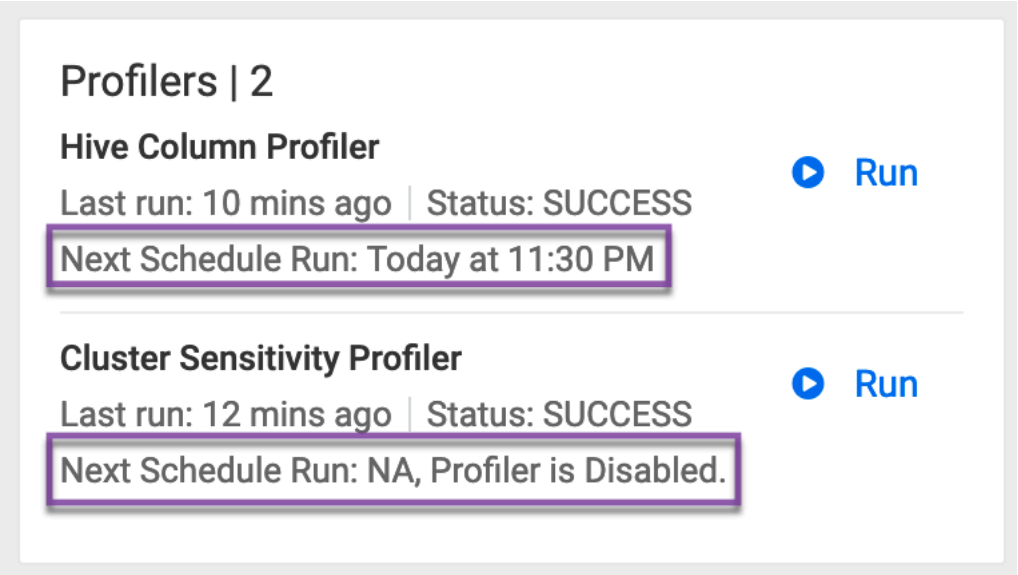
Note: If the profiler cluster is not registered with the data lake, Data Catalog cannot locate or trace the profiler cluster. Users have to delete the profiler cluster from the DataHub page (Cloudera Management Console).

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. 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.

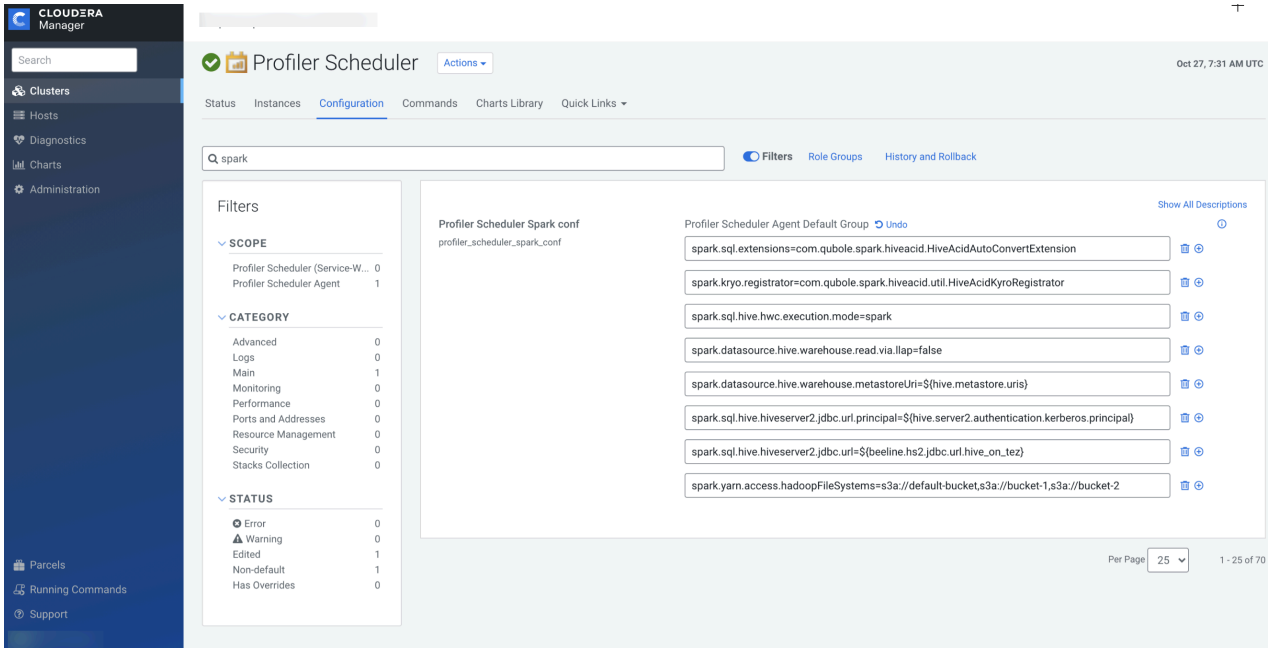
Note: You can use the On-Demand Profiler feature to profile both external and managed tables.



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.

High Availability support for Profiler services

The Profiler Services is now supported by enabling the High Availability (HA) feature.

There are two types of Profiler Services:

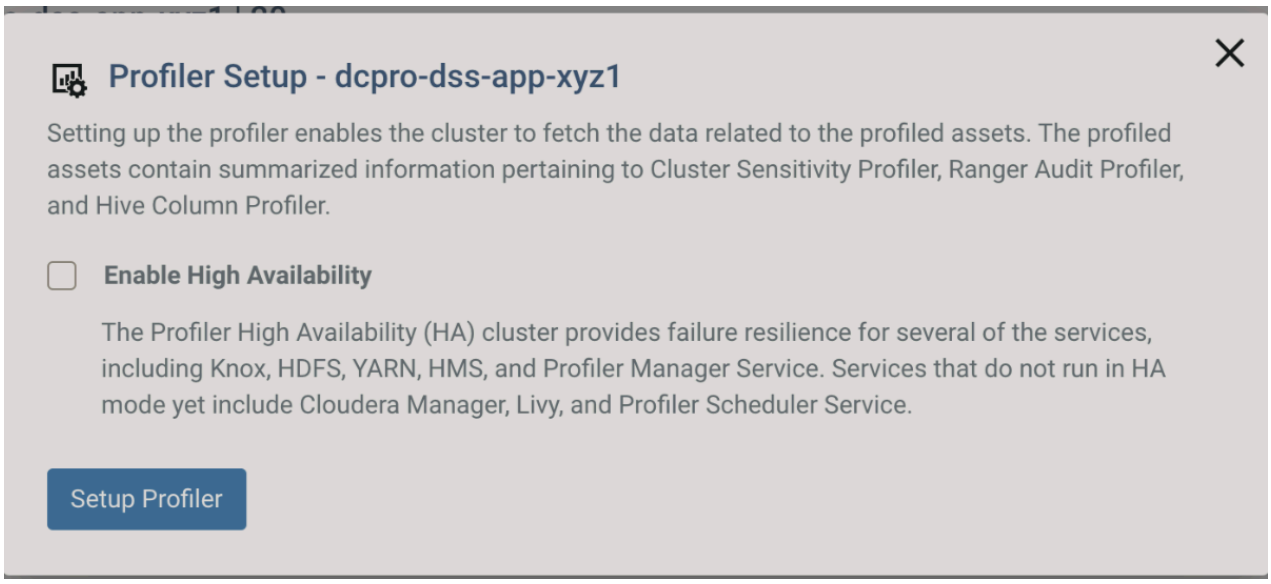
- Profiler Manager
- Profiler Scheduler

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.

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

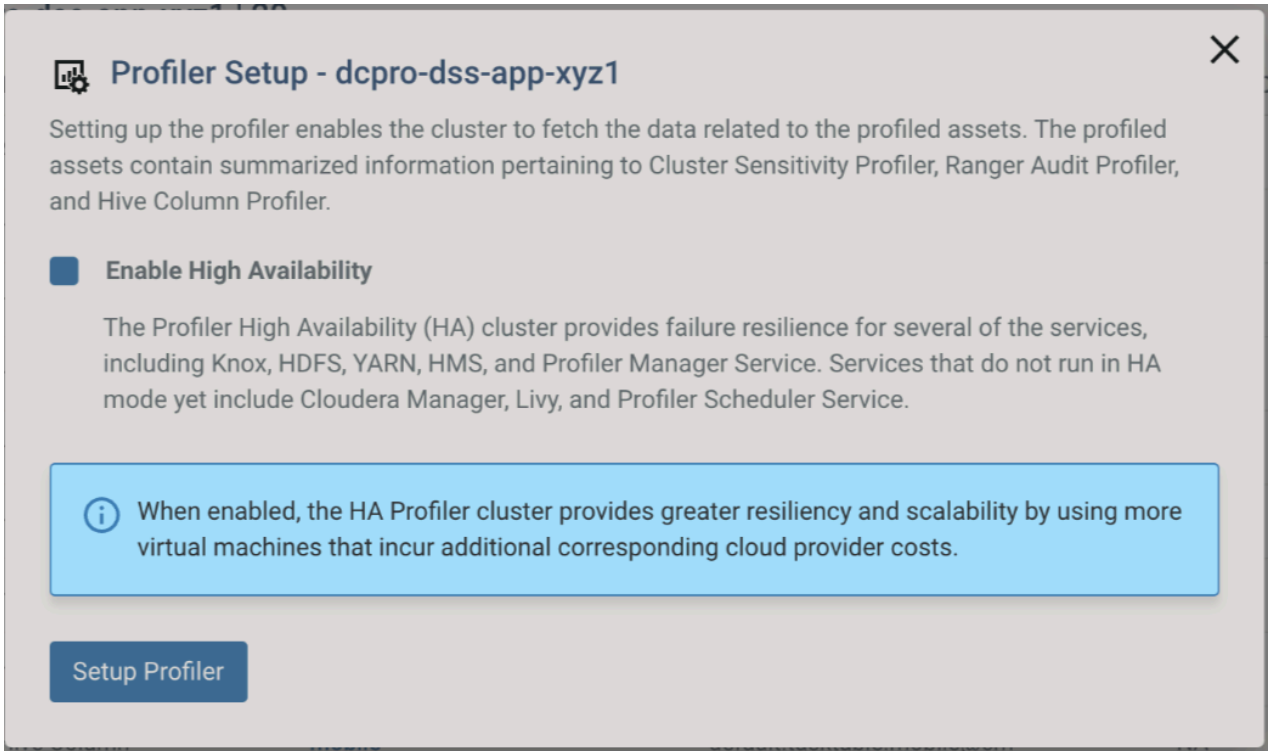
As of this update, there is a new user interface which is introduced to launch profilers in Data Catalog.

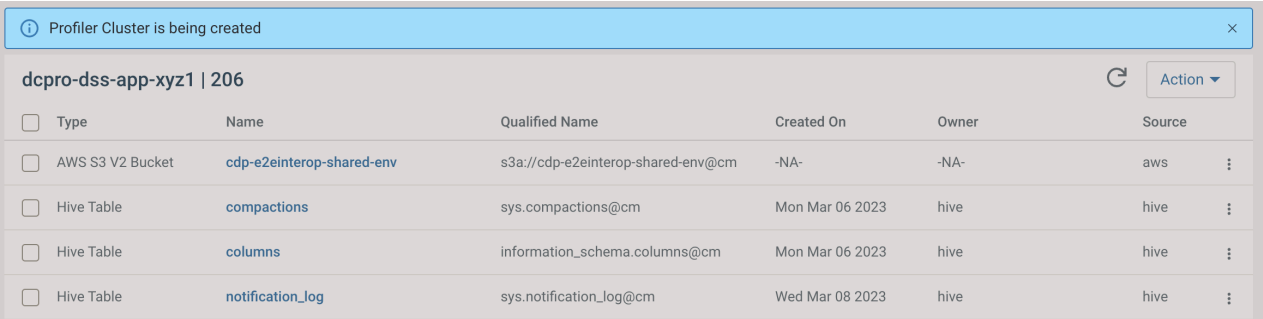
On the Data Catalog search page, select the data lake from which you want to launch the profiler cluster. On the right-hand side of the window, the application displays the page to set up the profiler for the selected data lake. Click the Get Started link to proceed.



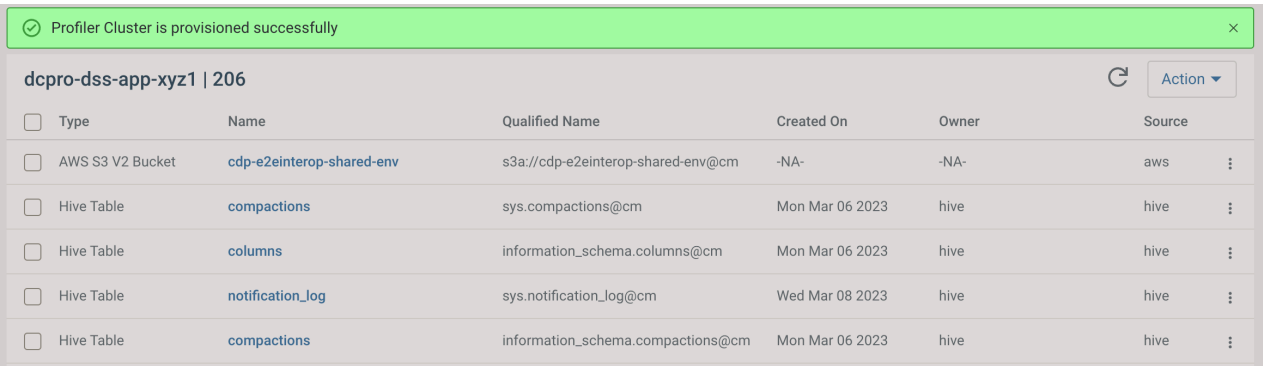
For setting up the profiler, you have the option to enable or disable the HA. Note that the HA functionality is being supported only from Cloudera Runtime 7.2.10 release onwards. If you are using the Cloudera Runtime version below 7.2.10, you shall not be able to use the HA feature for launching the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

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

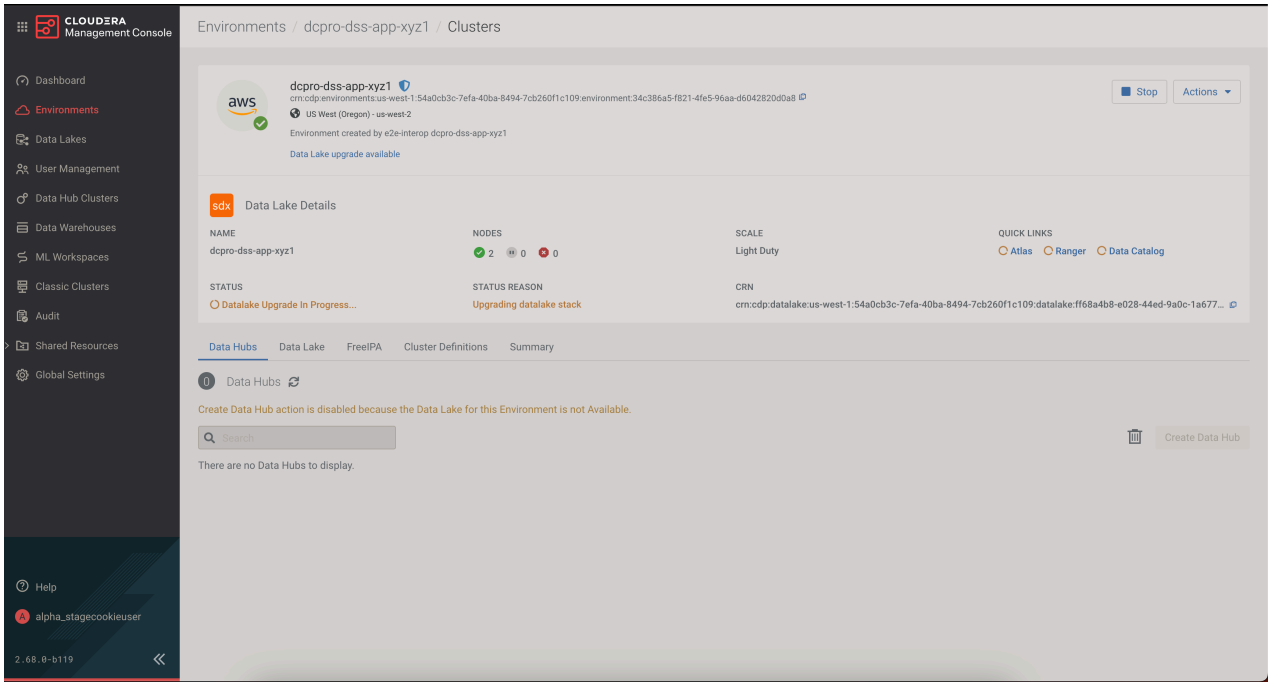




Later, a confirmation message appears that the profiler cluster is created.



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:



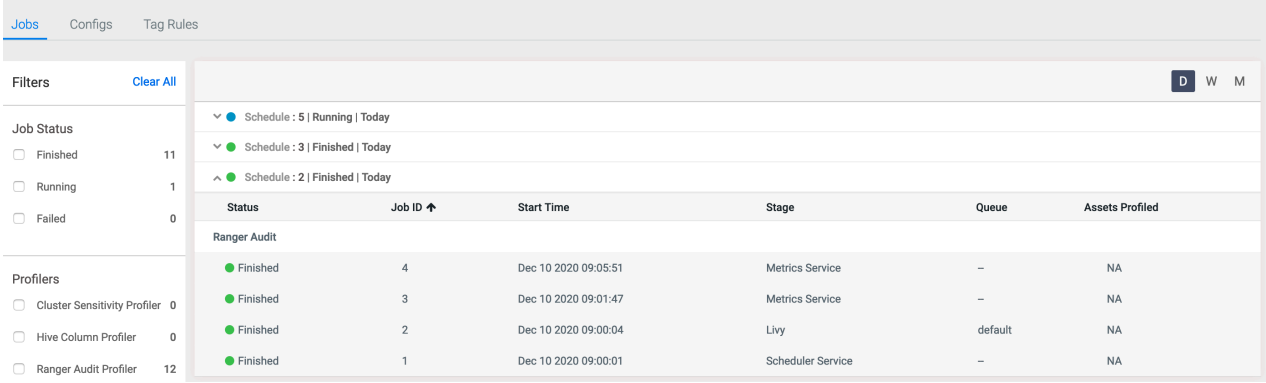
- Data Hub Type - The term “ha” is appended to the type of cluster that is newly created.
- Version - 7.2.10
- Node Count - (Which is 6)

  Your Profiler cluster with HA is set up successfully.

Tracking Profiler Jobs

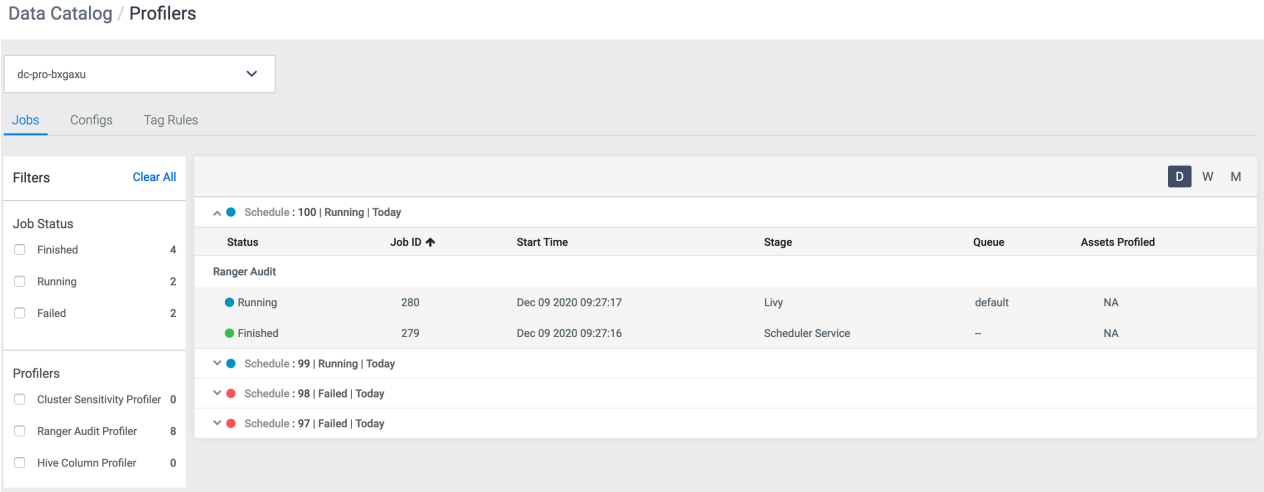
The Data Catalog profiler page is updated to provide a better user experience for tracking respective profiler jobs.

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.

Related Information

Understanding the Data Catalog Profiler
Understanding the Cluster Sensitivity Profiler
Understanding the Ranger Audit Profiler

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:

- Verify 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.

Related Information

Understanding the Data Catalog Profiler
Understanding the Cluster Sensitivity Profiler
Understanding the Ranger Audit Profiler

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 65.
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 65.
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.

Related Information

Understanding the Data Catalog Profiler

Understanding the Cluster Sensitivity Profiler

Associating Custom Sensitivity Profiler tags to Hive assets

Associating Custom Sensitivity Profiler tags to Hive assets fails on the 7.2.16 Cloudera Runtime cluster.

The issue can be resolved by adding the Spark configuration:

spark.sensitive.profiler.tableSamplePercentage=50 to profiler scheduler conf by adding the configuration under the key - profiler_scheduler_spark_conf.

The value passed in the configuration must be an integer value ranging from 1 to 100.

This value indicates the percentage of rows to be picked up as a sample for the profiler job.

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 63.
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 65.

   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 the 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:

Each * in the cron represents a unique value.

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

For example, consider a cron with the following values:

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

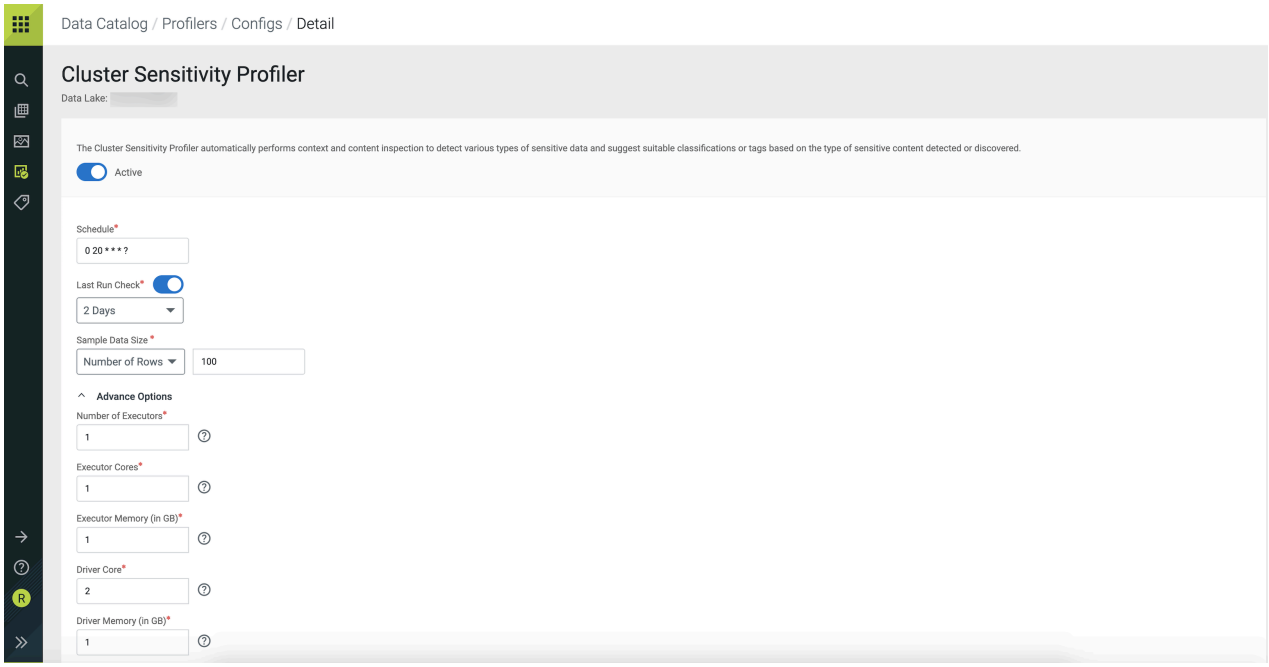
The cron is scheduled to run the profiler job at: 03:02:01am, on the 2nd day, in May, in 2021.

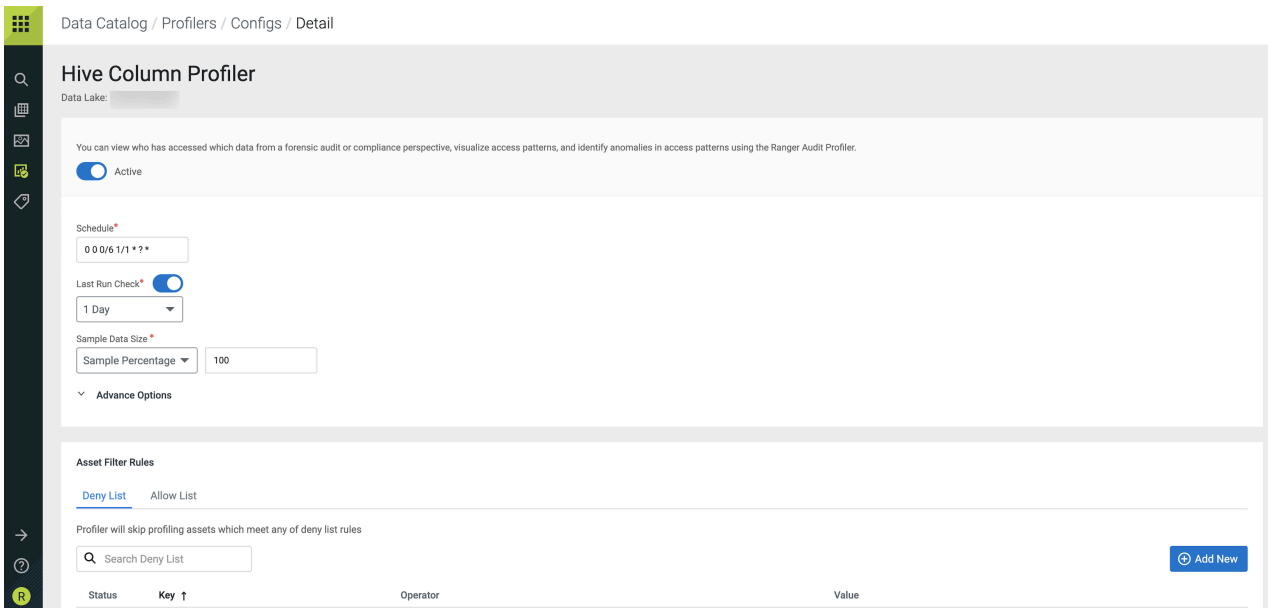
You can change the value of cron as and when it is required depending on how you want to schedule your profiler job.

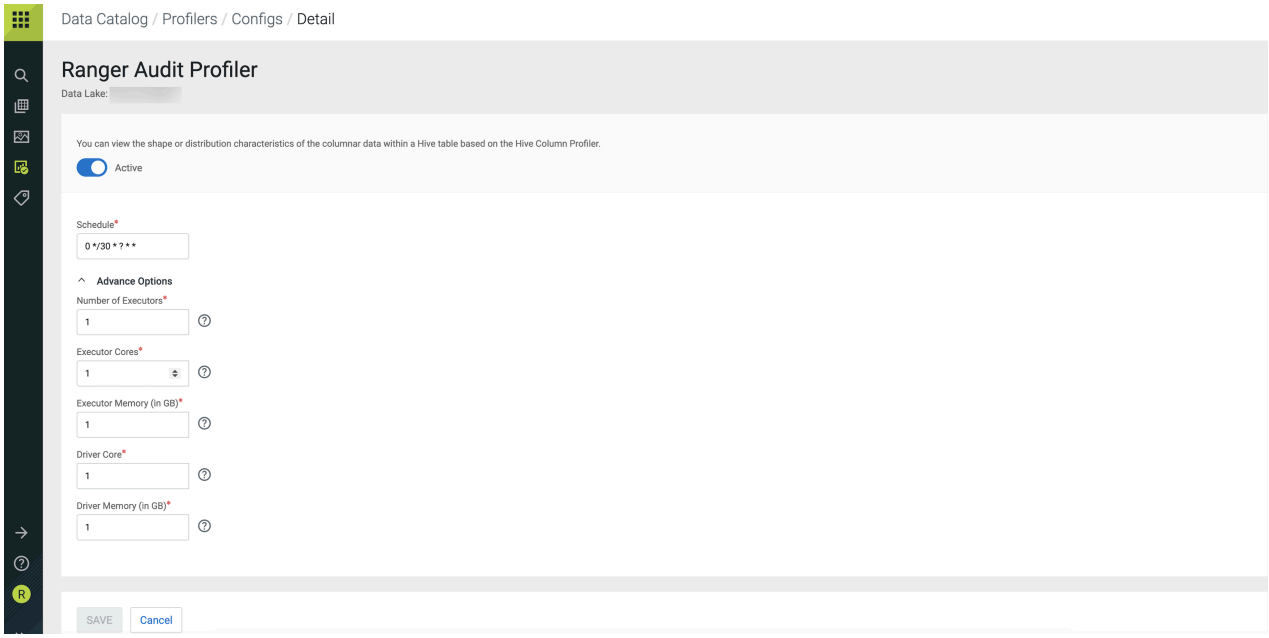
Setting Asset filter rules

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

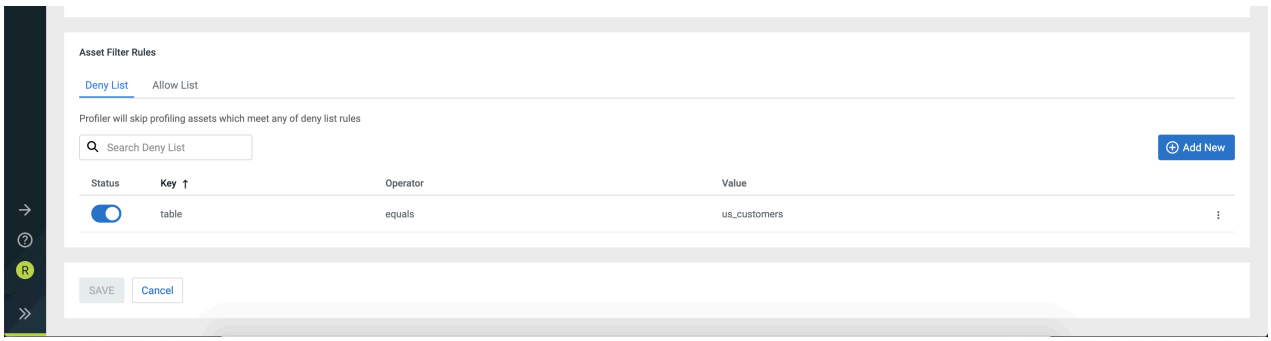
Note: You can configure the Deny-list and Allow-list for both Cluster Sensitivity Profiler and Hive Column Profiler. The same filter rules do not apply to the Ranger Audit Profiler.

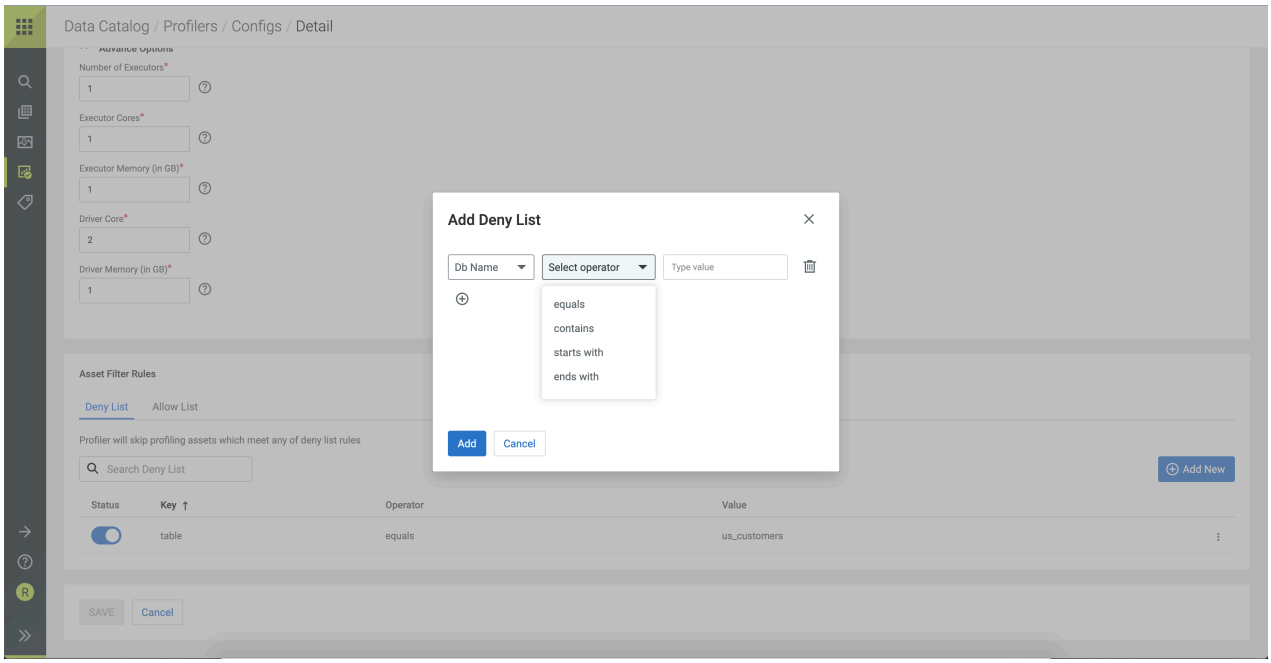




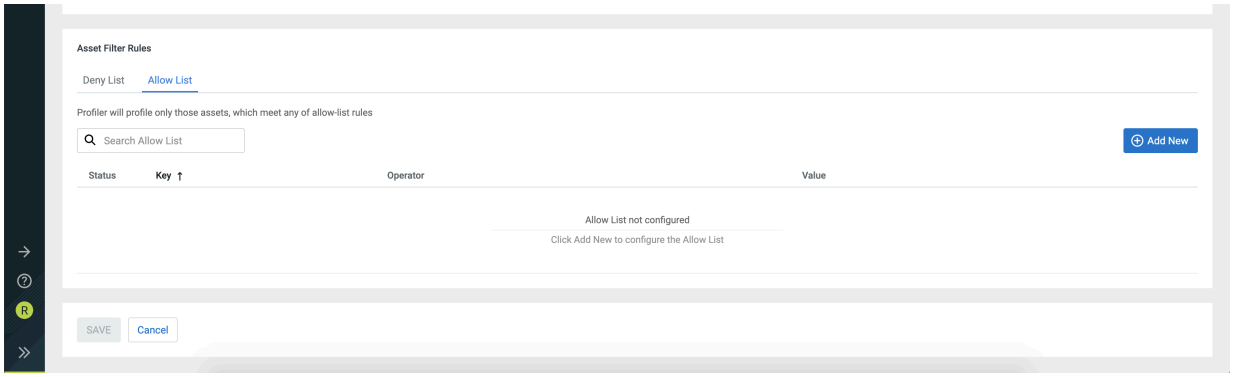


- Deny-list - The profiler will skip profiling assets that meet any defined Deny-list criteria.
  - Select the Deny-list tab.
  - Click Add New to include rules for Deny-list.
  - Select the key from the drop down list. You can select a database name, name of the asset, name of the owner of the asset, 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.


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
table
equals
us_customers
SAVE
Cancel


Data Catalog / Profilers / Configs / Detail
Number of Executors*
Executor Cores*
Executor Memory (in GB)*
Driver Core*
Driver Memory (in GB)*
Add Deny List
Db Name
Select operator
Type value
equals
contains
starts with
ends with
Add
Cancel
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
table
equals
us_customers
SAVE
Cancel

- 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

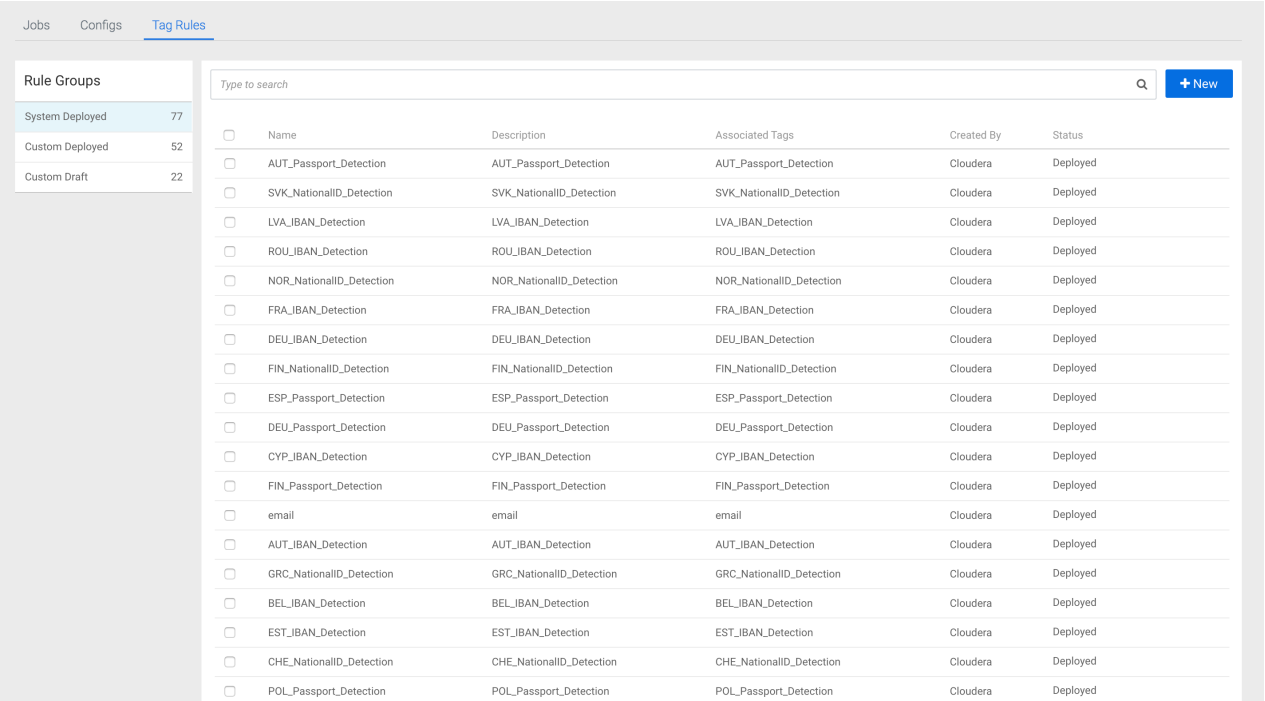
1. From Profiler Configs
2. Select the profiler to proceed further.

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

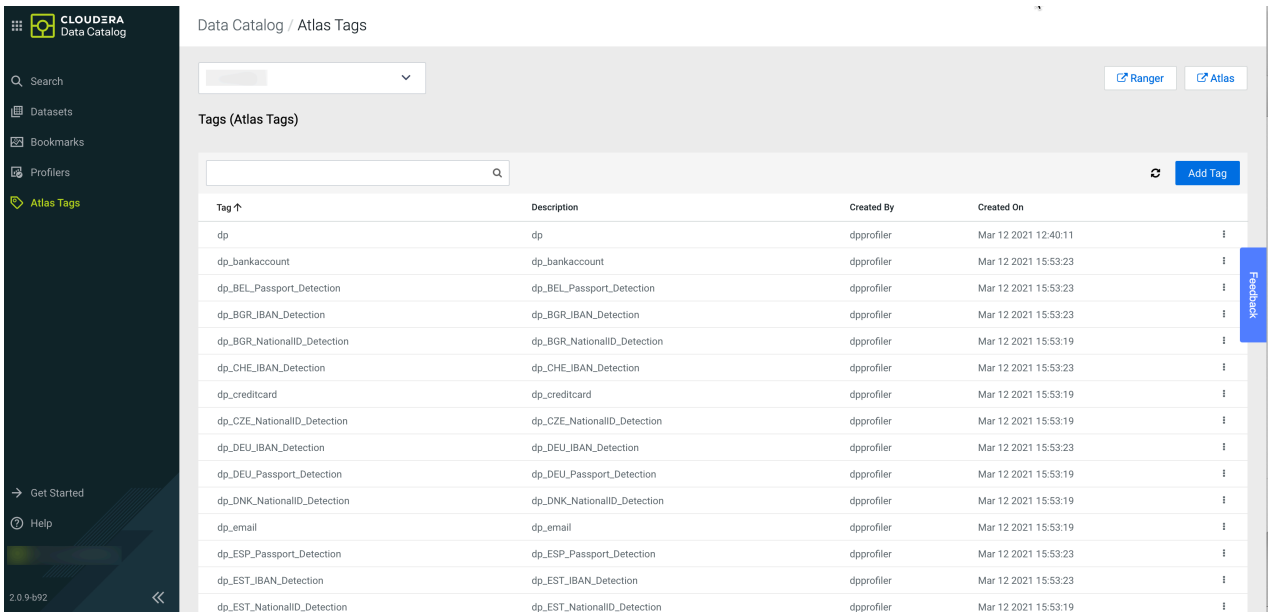
From Atlas tags UI in Data Catalog, you can create, modify, and delete any of the Atlas tags in a Data Catalog instance.

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

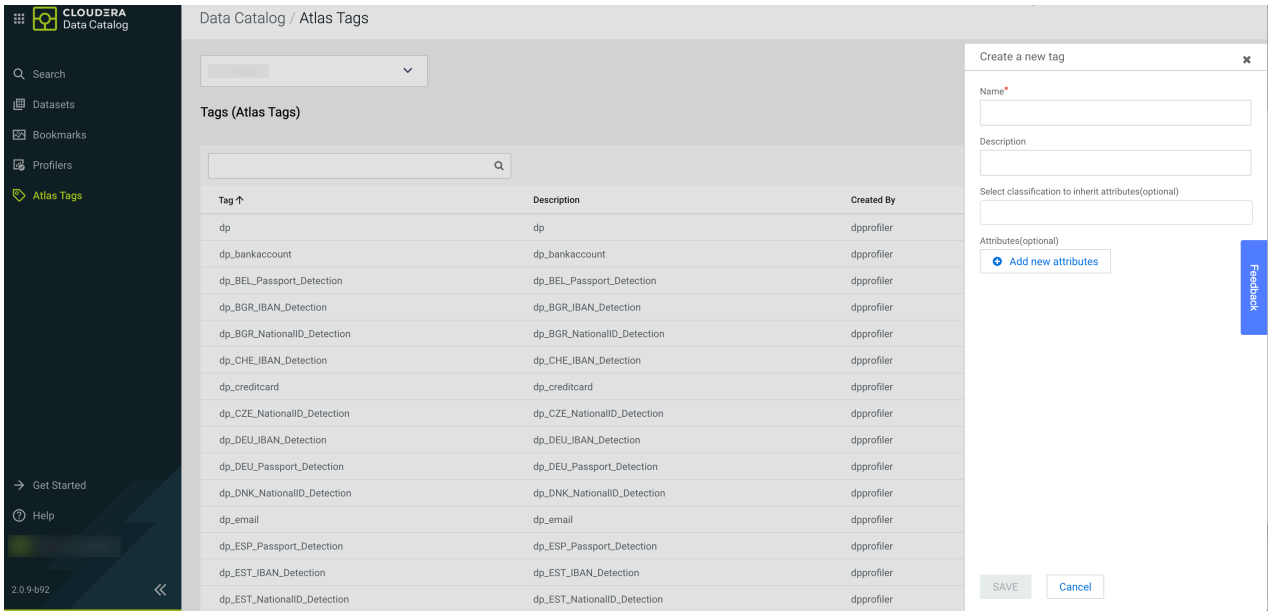
Atlas Tags allows the user to perform the following activities with a selected Data Lake for tag management:

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

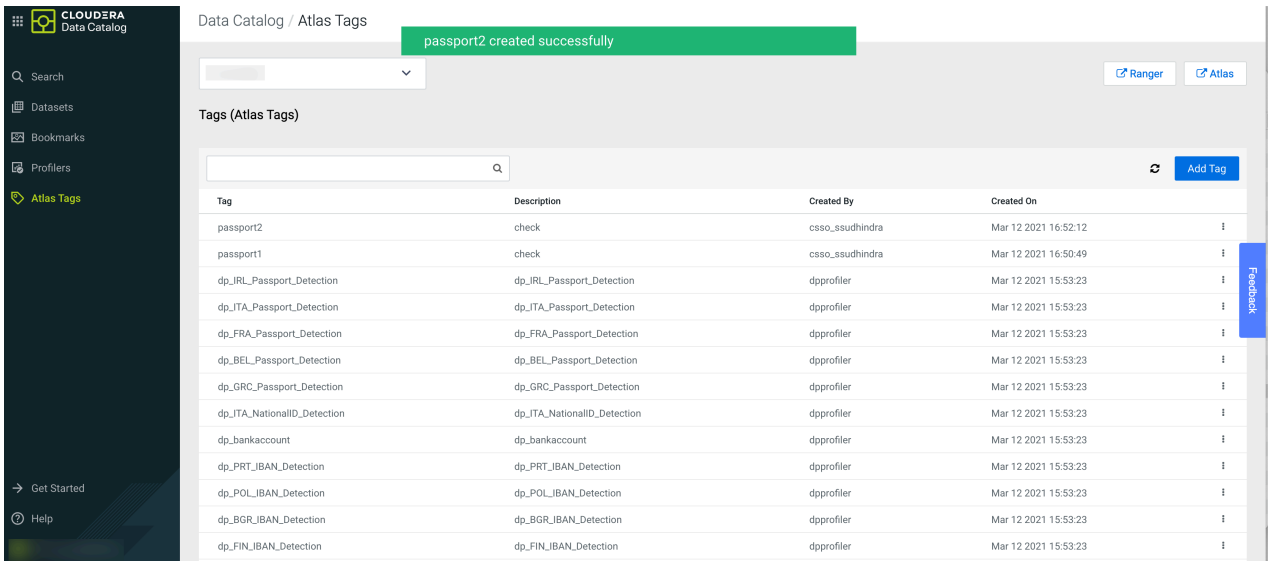
The new Atlas tags UI is displayed as seen in the diagram.



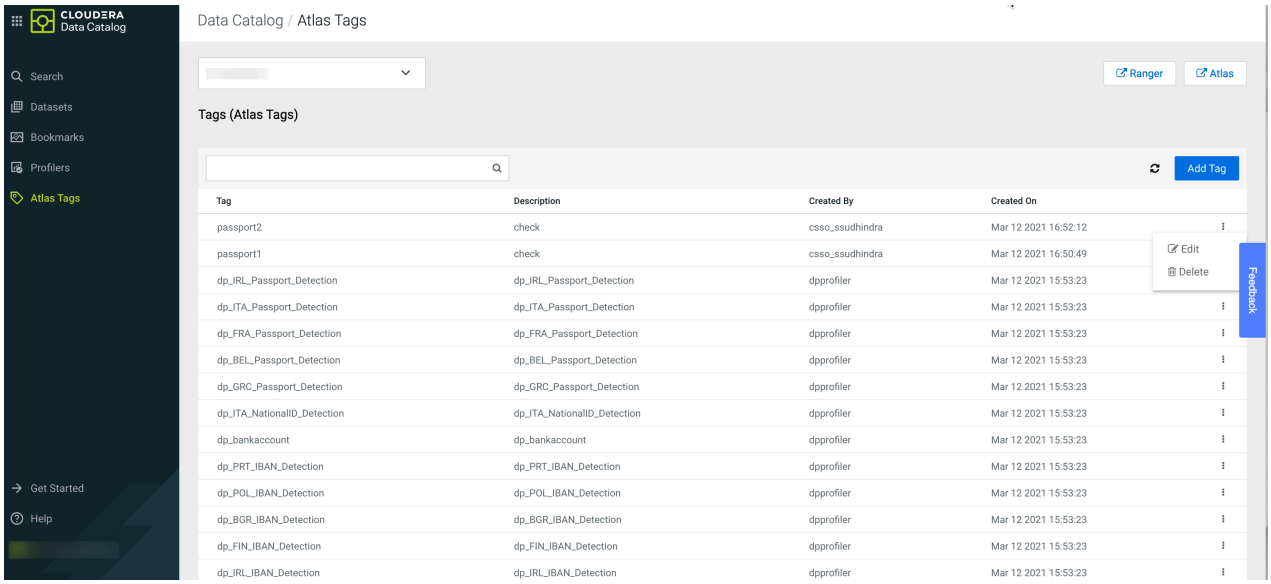
You can create a new tag in the Atlas tags UI. The following diagram provides an overview about the Create a new tag page.



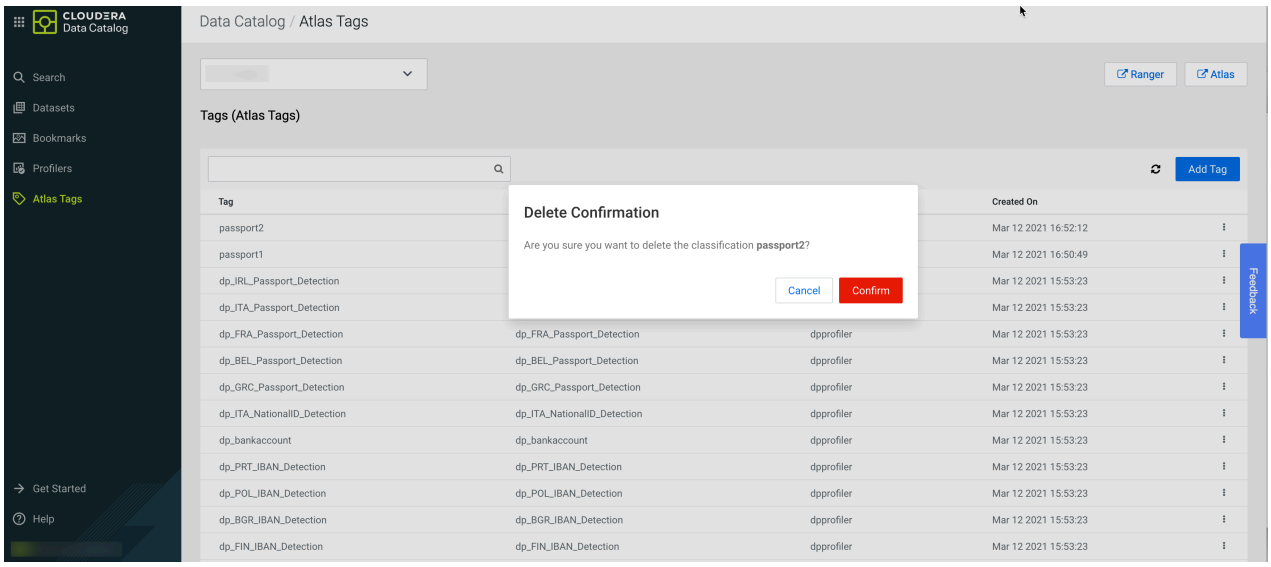
You can add or update Atlas tags. The created or updated tag is highlighted in the tag list as seen in the following diagram.



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.



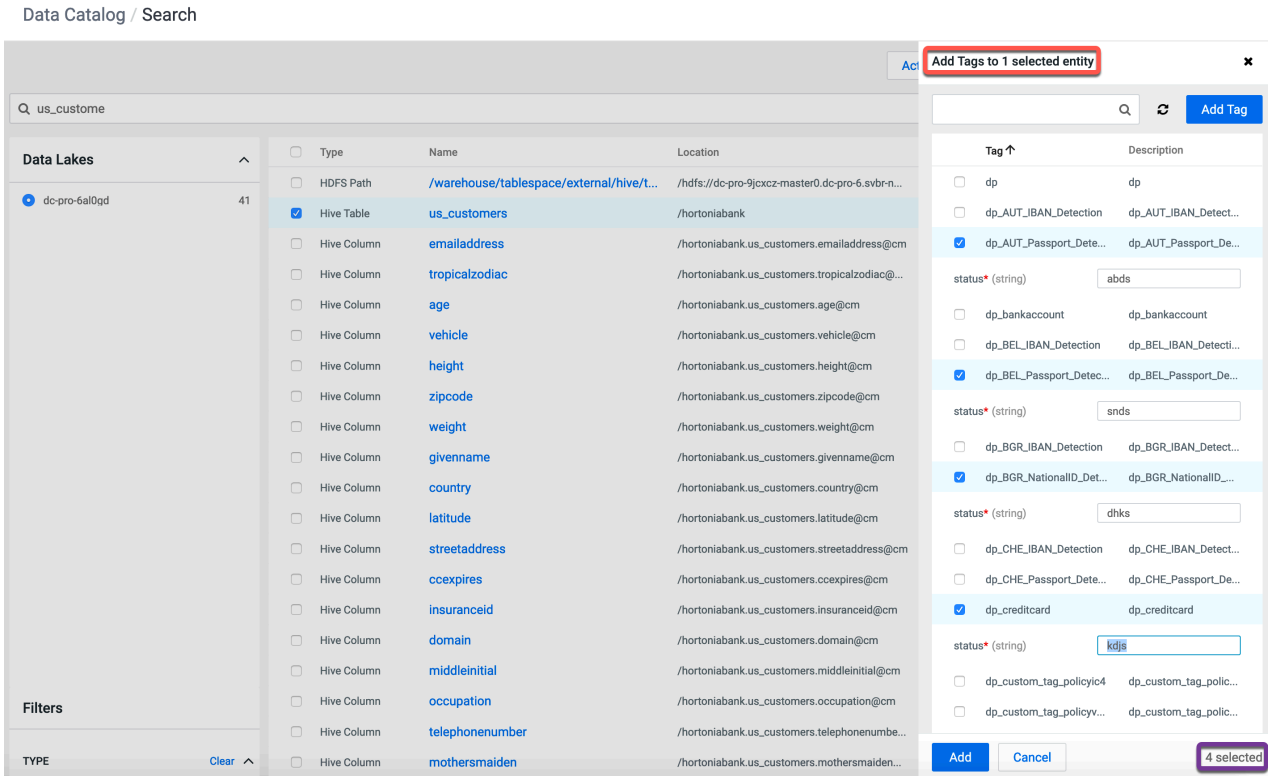
You can delete one Atlas tag at a time. A separate confirmation message appears.



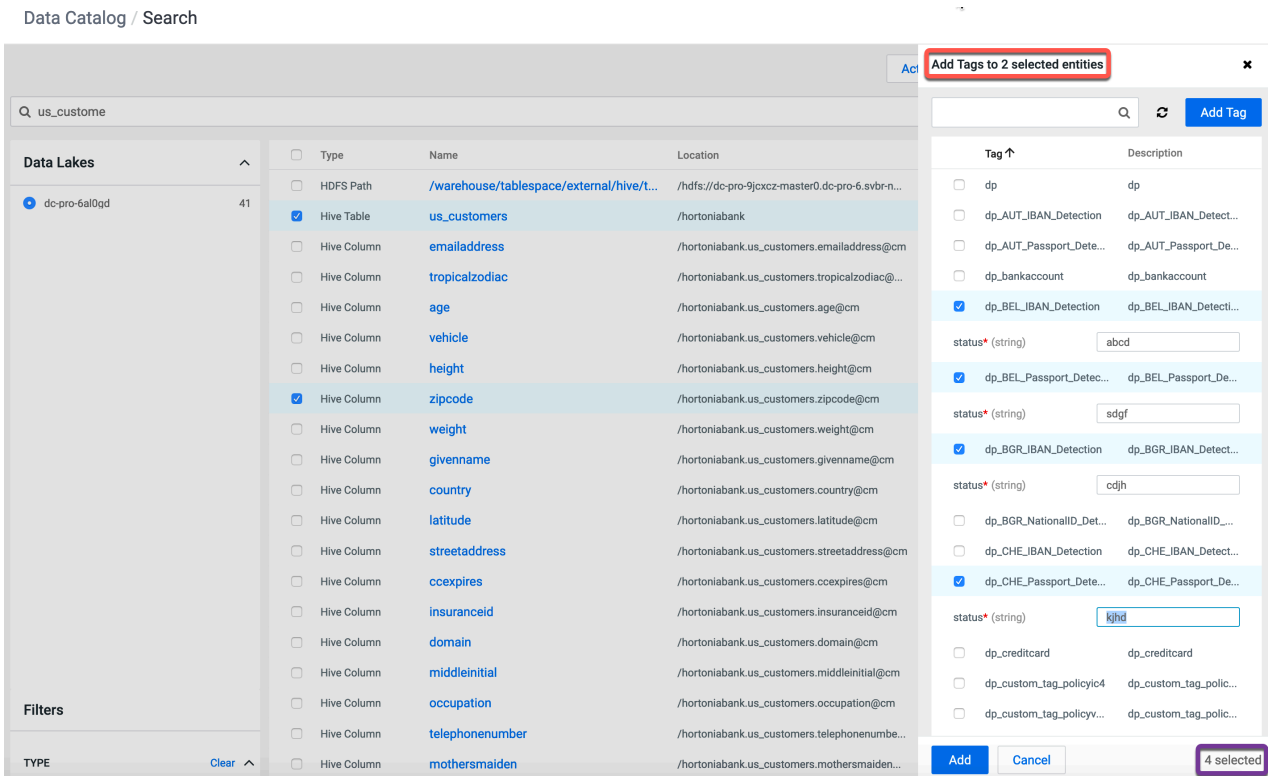
Tagging Multiple Assets

On the Data Catalog search page, you can add tags to multiple assets based on the asset type that you select based on the search result.

When you select an asset, you can add one or more available tags to the selected asset. You can also create one or more new tags and associate the newly created tags to the selected asset. The number of selected assets that you plan to tag is displayed. As you add the number of tags to one or more selected assets, The Add Tag panel displays the number of tags assigned.



Note: If you do not save your changes without clicking the Add button in Add Tags panel, the changes are not retained in the Data Catalog instance. You have to retag the assets and later click the Add button.

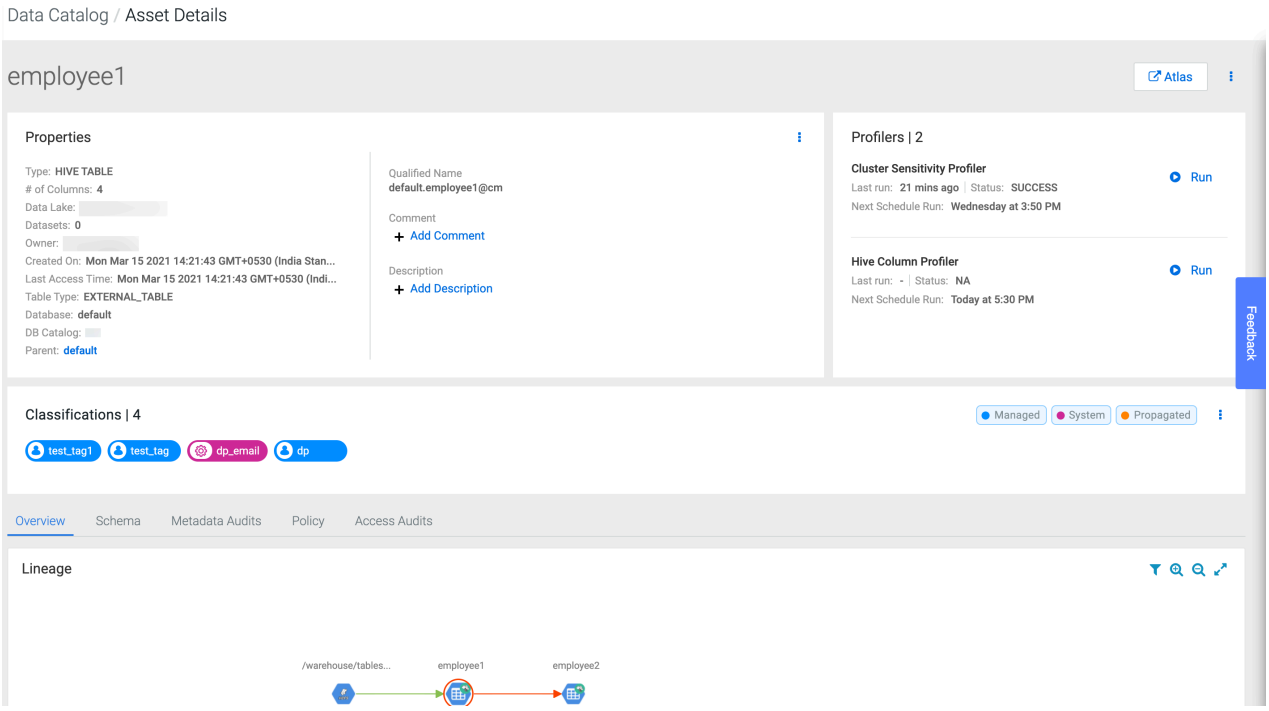


When you add one or more tags to the selected entities, the assigned tags are displayed having been tagged to the number of selected entities. Another scenario could throw a message saying that the selected asset is already tagged.

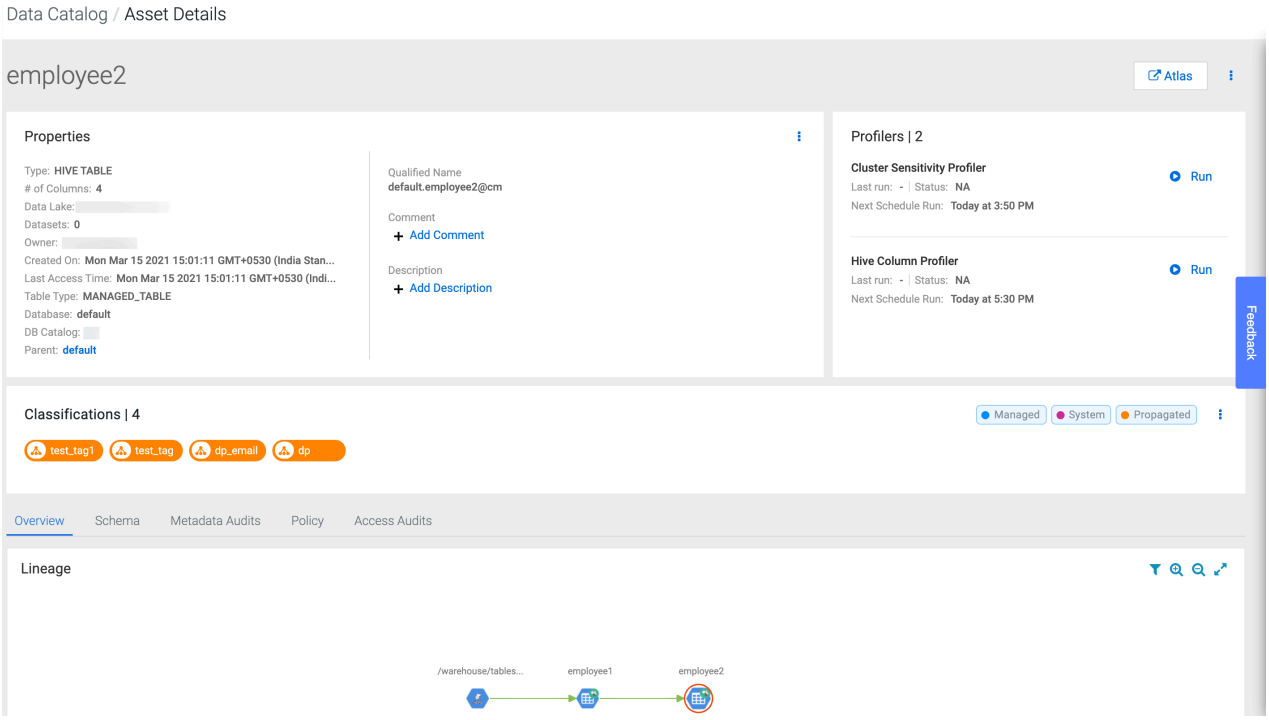
Propagated Asset tagging

Data Catalog supports the concept of propagated tags. This feature is derived from Apache Atlas.

Whenever you add a new tag, you can mark them as propagated and use those tags accordingly while tagging assets.



For example, consider table1 as a parent asset and table2 as a child asset. Create a tag and mark that tag as propagated, and later apply the same tag to table1. The tag gets applied to table2 as well. Propagated tag works on the basis of parent -> child tagging relationship.



Note: When you delete or remove the propagated tag from the parent asset, the same tag is removed from all the child assets.

Attention: The propagated tag concept is not supported with child -> parent relationships.

Creating Custom Profiler Rules

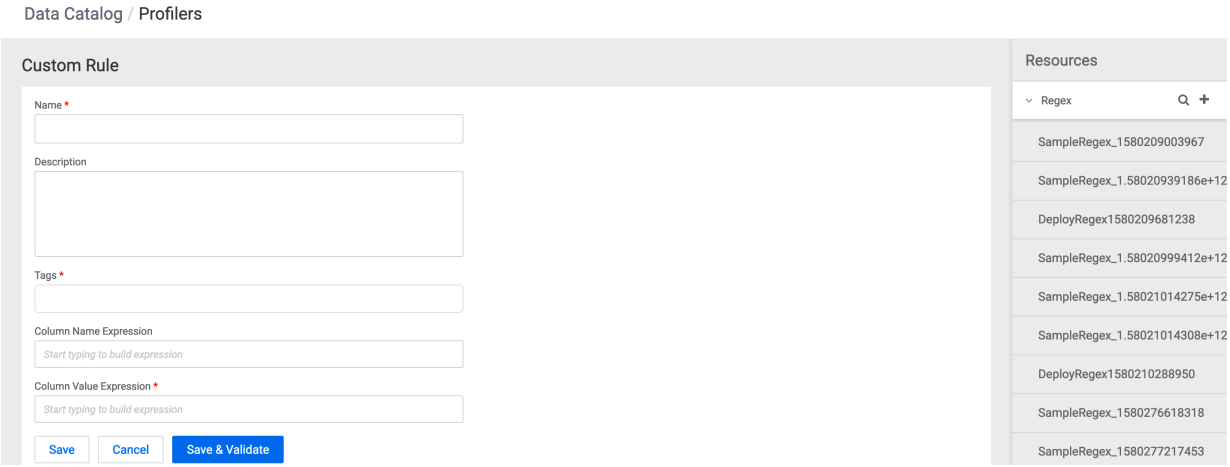
You can create a custom profiler by adding the required tags, regex entries, and attaching whitelist or blacklist files to specific columns within your tables.

Procedure

1. On the Profilers page, click Tag Rules.
2. On the Tag Rules tab, click New to create a new profiler tag rule.
3. Enter the name of the new custom profiler tag rule.
4. Enter the description for the custom tag rule.
5. Select the Tags. You can select tags from the drop down list and or enter a new value to create a new tag.

   New tags that you create here are added with a dp_ prefix in the list of Atlas tags. For example, if you add a new tag called credit_card, this tag will be added as dp_credit_card in Atlas.
6. Enter the rule for the column name. As you enter the values, regex name and resource names are auto populated. Select the column that is needed for your custom profiler.
7. Enter the column value for the DSL.

   Based on your entry, Data Catalog auto populates values from the entries already available in the Resources tab. You can use a combination of regex entries and whitelist or blacklist files and other behaviors. For more information about behaviors, see DSL Grammar.
8. Click Save and Validate.



   In the validation pop up window that appears, enter data to validate your custom profiler tag rule. Make sure you separate each data entry with a new line.
9. Click Save to create a tag rule and validate and deploy it later.

Adding Custom Regular Expressions

To use custom regex entries within your new custom profiler tag rules, you can also add new regex values.

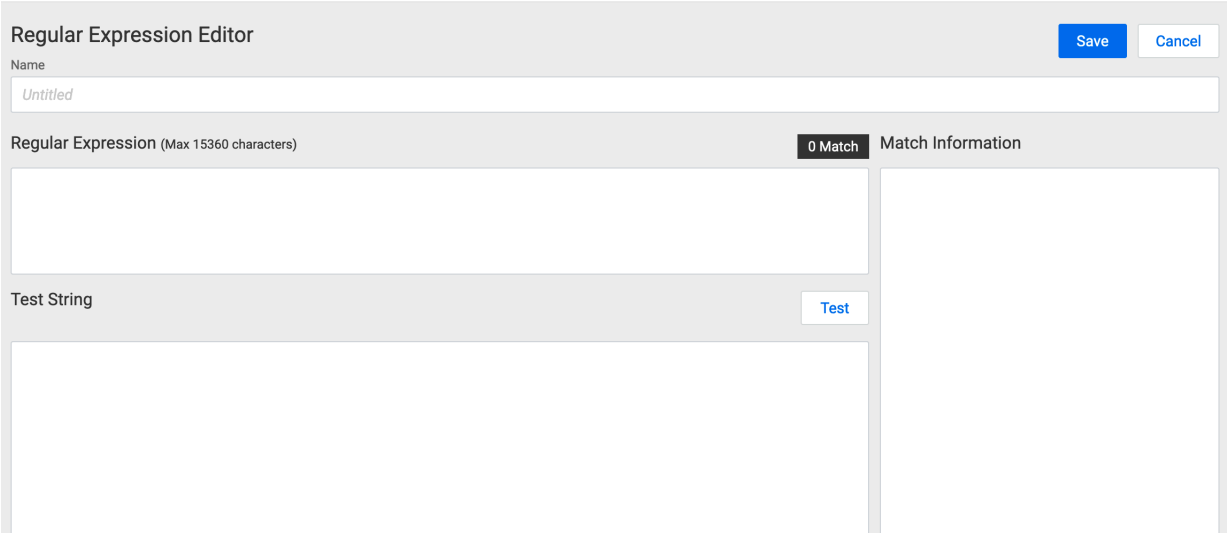
Procedure

1. Click Resources in the right panel on the New Custom Profiler Rules page.
2. Click + icon on the Regex tab. The Regular Expression Editor page appears.
3. Enter the name of the new regular expression.

4. Enter a valid regular expression.

   For example:

```
\b(([a-zA-Z0-9_\-\.]+)@((\[[0-9]{1,3}\.[0-9]{1,3}\.[0-9]{1,3}\.)|(([a-zA-Z0-9\-]+\.)+))([a-zA-Z]{2,4}|[0-9]{1,3})(\]?))\b
```



5. Enter the list of test strings to evaluate the new regular expression.

   If the test string is valid, then the match information gets auto populated in the Match Information box.
6. Click Save to add the new regular expression to the list of Regex Resources.

Adding Lookup Files

When you have too many allowed and denied entries and cannot add them inline, you can create allowed or denied files with one value in each line and add them to your DSL.

Procedure

1. Click Resources in the right panel on the New Custom Profiler Rules page.
2. Click + icon on the Lookups tab. The New Lookup File page appears.
3. Enter the name of the new Lookup file.
4. Click Choose File to upload the file.
5. Click Save.

Using Behaviors

You can use various behaviors to take single inputs of type text and evaluate them to a Boolean value.

The profiler can take column values of any type and pass the values to each behaviour as text. Behaviors include the following:

1. Regular expressions
2. File based allowlist and denylist checks
3. Luhn algorithm

Regular expressions

You can include one or more regular expressions and evaluate to True if one of these matches the provided value.

Keyword: regex

A regex that matches everything can be defined as follows:

```
regex(\"[\\\\s\\\\S]+\")
```

A regex that includes multiple expressions can be defined as follows:

```
regex(\"[\\\\s\\\\S]+\",\"^[0-9]*$\")
```

File based denylist and allowlist checks

When the number of allowlist and denylist entries are many and cannot be defined inline, you can create allowlist or denylist files with one value in each line.

Keywords: allow-list, deny-list

Make sure to place the file in an HDFS location that is accessible to the Profiler Agent user.

You can provide the location of the file as an attribute in the DSL definition.

For example:

```
whitelist(\"/apps/dpprofiler/profilers/sensitive_info_profiler/1.0/lib/krapt
r/meta/whitelist\")
```

```
blacklist(\"/apps/dpprofiler/profilers/sensitive_info_profiler/1.0/lib/krapt
r/meta/blacklist\")
```

Luhn algorithm

You can do a Luhn check on identification numbers in columns.

Keyword: luhn_check

Use the Luhn algorithm or Luhn formula to validate a variety of identification numbers such as credit card number, IMEI numbers, National Provider Identifier numbers in the United States, Canadian Social Insurance numbers, Israel ID Numbers, and Greek Social Security Numbers.

Using DSL Grammar

Using DSL grammar, you can combine different behaviours in intuitive ways to bring out functionality while creating custom profiler rules.

The two behaviors available in this framework are as follows:

1. falseIdentity - Always evaluates to false, regardless of the input.
2. trueIdentity - Always evaluates to true, regardless of the input.

These two behaviors are used in the following examples and descriptions.

Binary AND operator

Keyword: `and`

And works the same way it does other languages. Hence following observations.

```
falseIdentity and trueIdentity == falseIdentity
```

```
falseIdentity and falseIdentity == falseIdentity
```

```
trueIdentity and trueIdentity == trueIdentity
```

```
trueIdentity and falseIdentity == falseIdentity
```

Here we are using == to show their equality.

Binary OR operator

The or operator works the same way it does in other languages.

```
falseIdentity or trueIdentity == trueIdentity
```

```
falseIdentity or falseIdentity == falseIdentity
```

```
trueIdentity or trueIdentity == trueIdentity
```

```
trueIdentity or falseIdentity == trueIdentity
```

Expand DSL to use as follows.

```
val rule1= falseIdentity and trueIdentity and trueIdentity
```

```
val rule2= trueIdentity and trueIdentity and trueIdentity
```

```
val rule3=rule1 and rule2
```

```
rule3 or trueIdentity
```

The above expression evaluates to true.