CDP Private Cloud Data Services / Data Catalog 1.5.5

Profilers

Date published: 2023-12-16
Date modified: 2026-01-30



https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2026. All rights reserved.

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

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

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

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

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

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

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

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

Contents

Launching profilers 4

Deleting profilers 9

On-Demand Profilers 10

Tracking Profiler Jobs 13

Viewing profiler configurations 17

Configuring the Cluster Sensitivity Profiler 20
- Profiler tag rules 21
- Creating custom profiler rules 22
  - Adding custom regular expressions 23
  - Using DSL grammar 24
  - Creating tag rules 26

Configuring the Data Compliance Profiler 29
- Profiler tag rules 35
  - Approving Data Compliance Profiler tags 37
  - Creating tag rules 41

Configuring the Hive Column Profiler 46

Configuring the Statistics Collector profiler 48

Configuring the Ranger Audit Profiler 54

Configuring the Activity Profiler 55

Understanding the Cron Expression generator 57

Enabling or disabling profilers 57

Launching profilers

In on premises environments, the Kubernetes-based Compute Cluster is automatically launched after starting the profilers in the user interface.

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

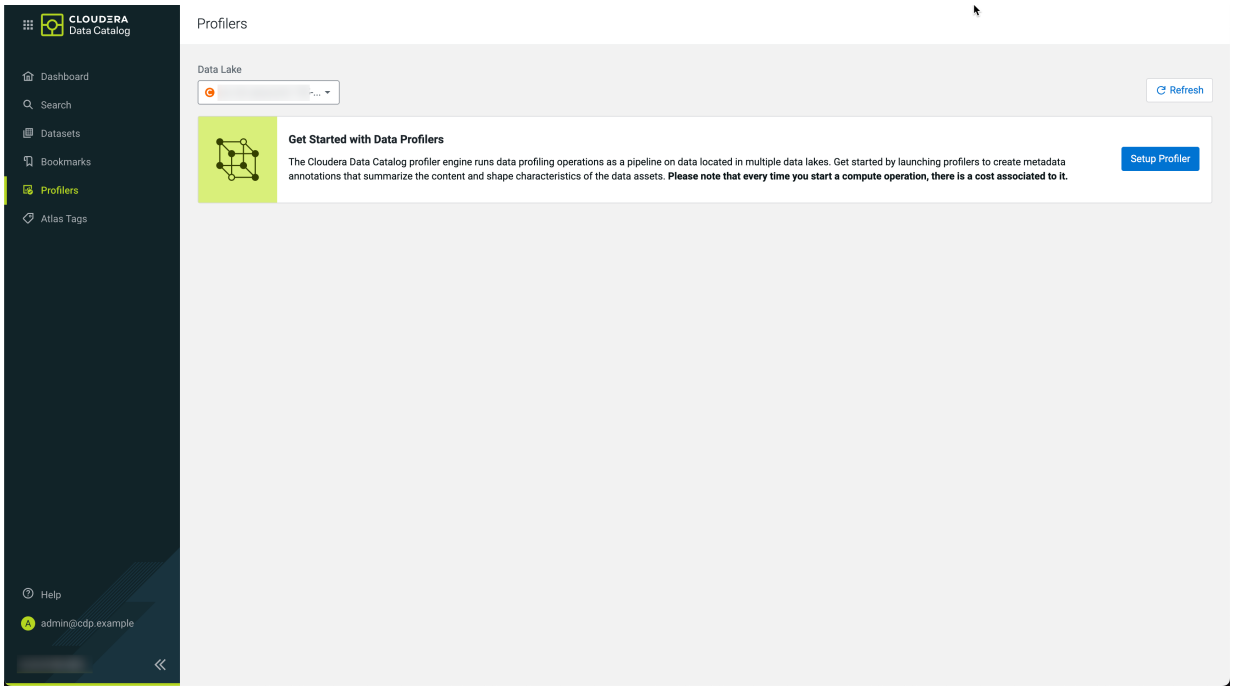
For After Cloudera Data Catalog 1.5.5 SP2

Important: If profilers were launched before upgrading to Cloudera Data Catalog on premises 1.5.5 SP2, delete them and relaunch them to avoid receiving the following message:

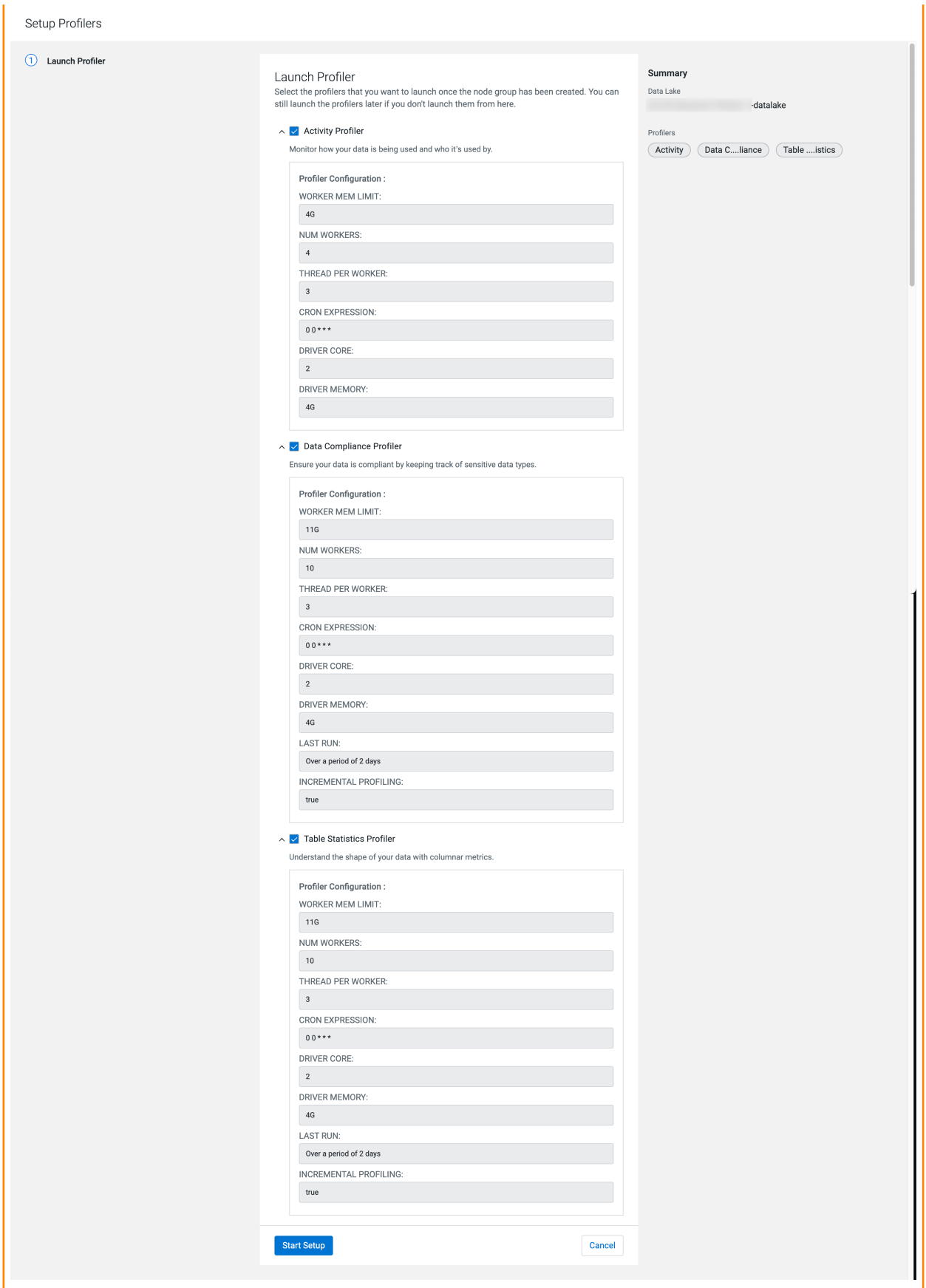
```
The auto-scaling node group is unavailable at the moment but
 profilers seem to have been launched. This has caused an
 inconsistency in Cloudera Data Catalog. Please delete all profilers
 and relaunch them to continue profiling your assets.
```

The safest way is to delete all profilers before upgrading to Cloudera Data Catalog on premises 1.5.5 SP2.

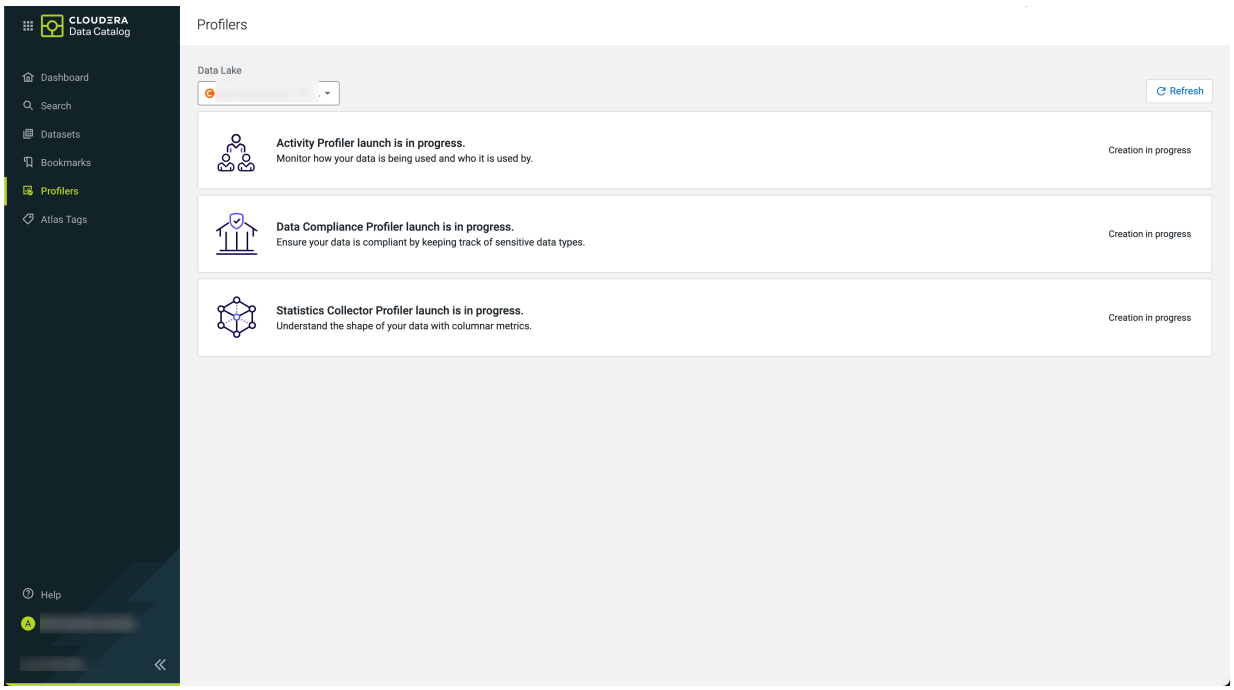
On the **Profilers** page, click the Setup Profiler link to proceed.



Select the required profilers with their default configuration, then click Start Setup.

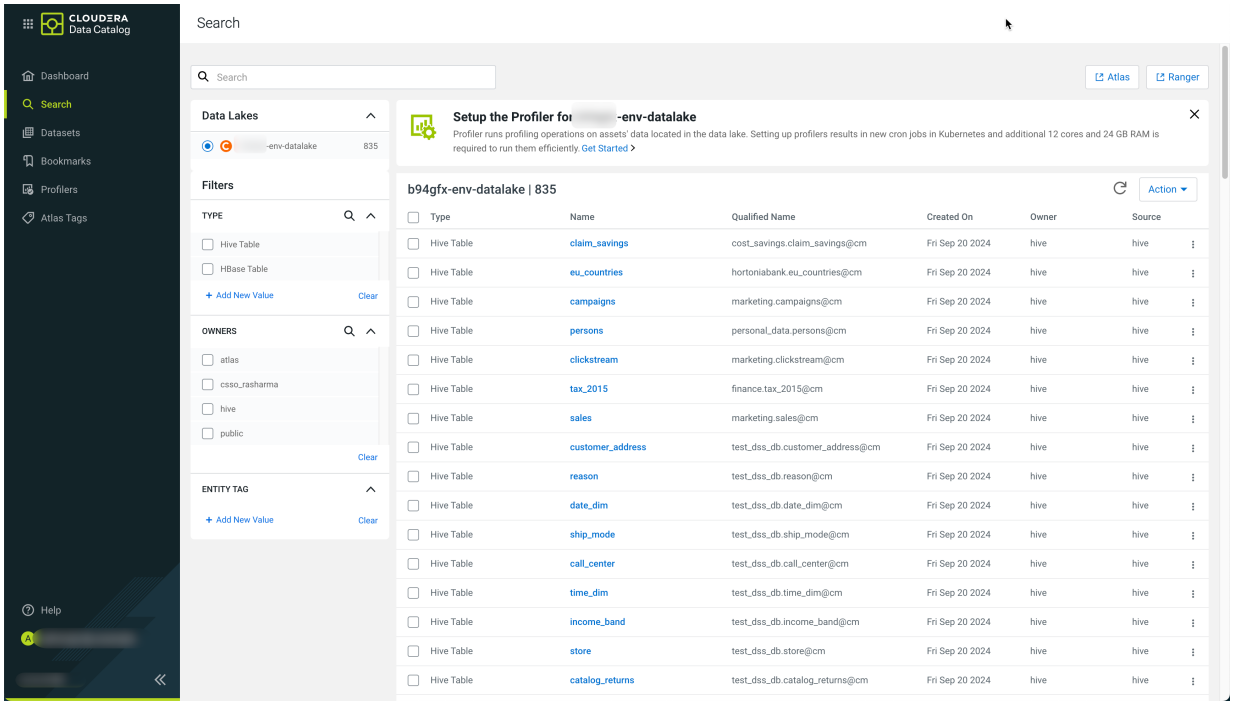


Wait for the profiler worker and driver setup to complete.

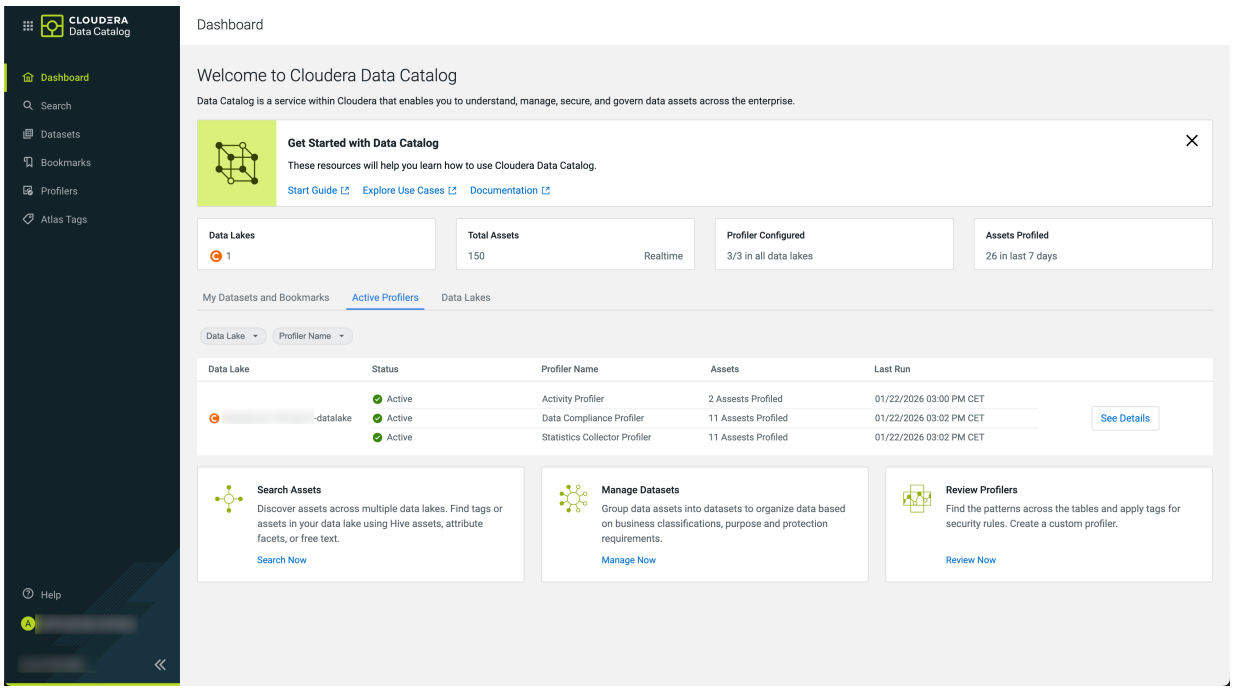


For Before Cloudera Data Catalog 1.5.5 SP2

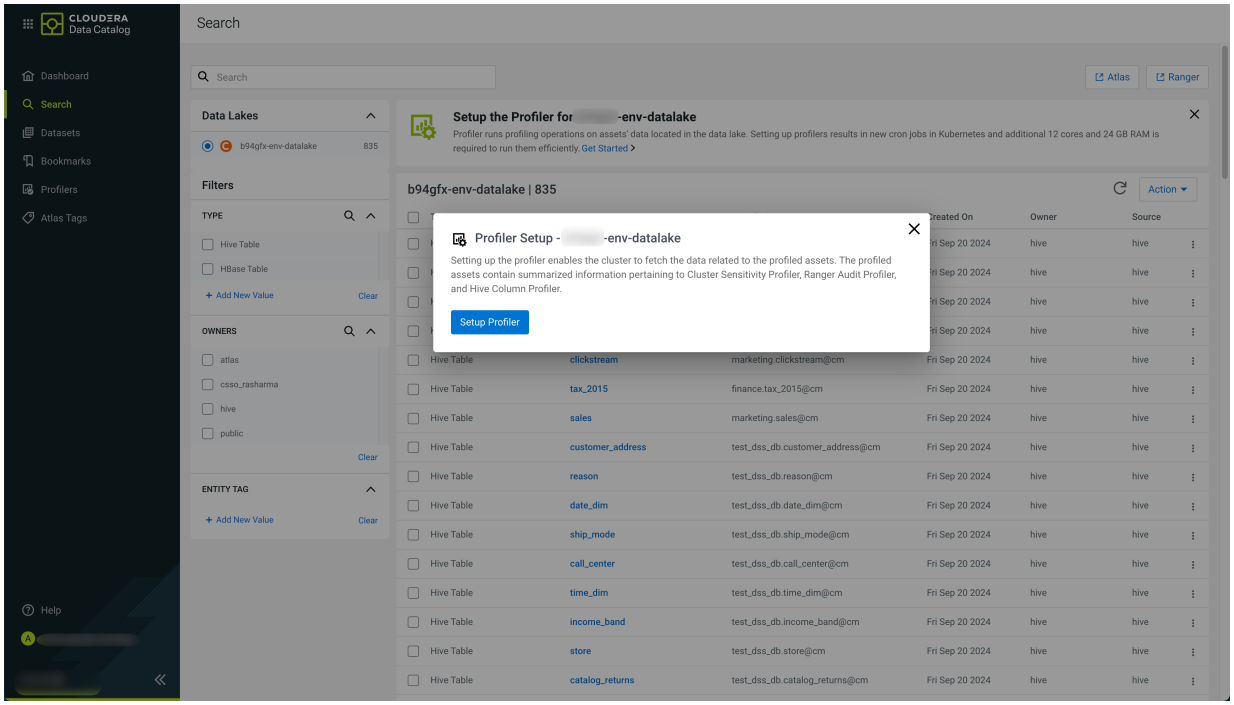
On the **Search** page, click the Get Started link to proceed.



The finished profilers will also be available from the **Dashboard**.

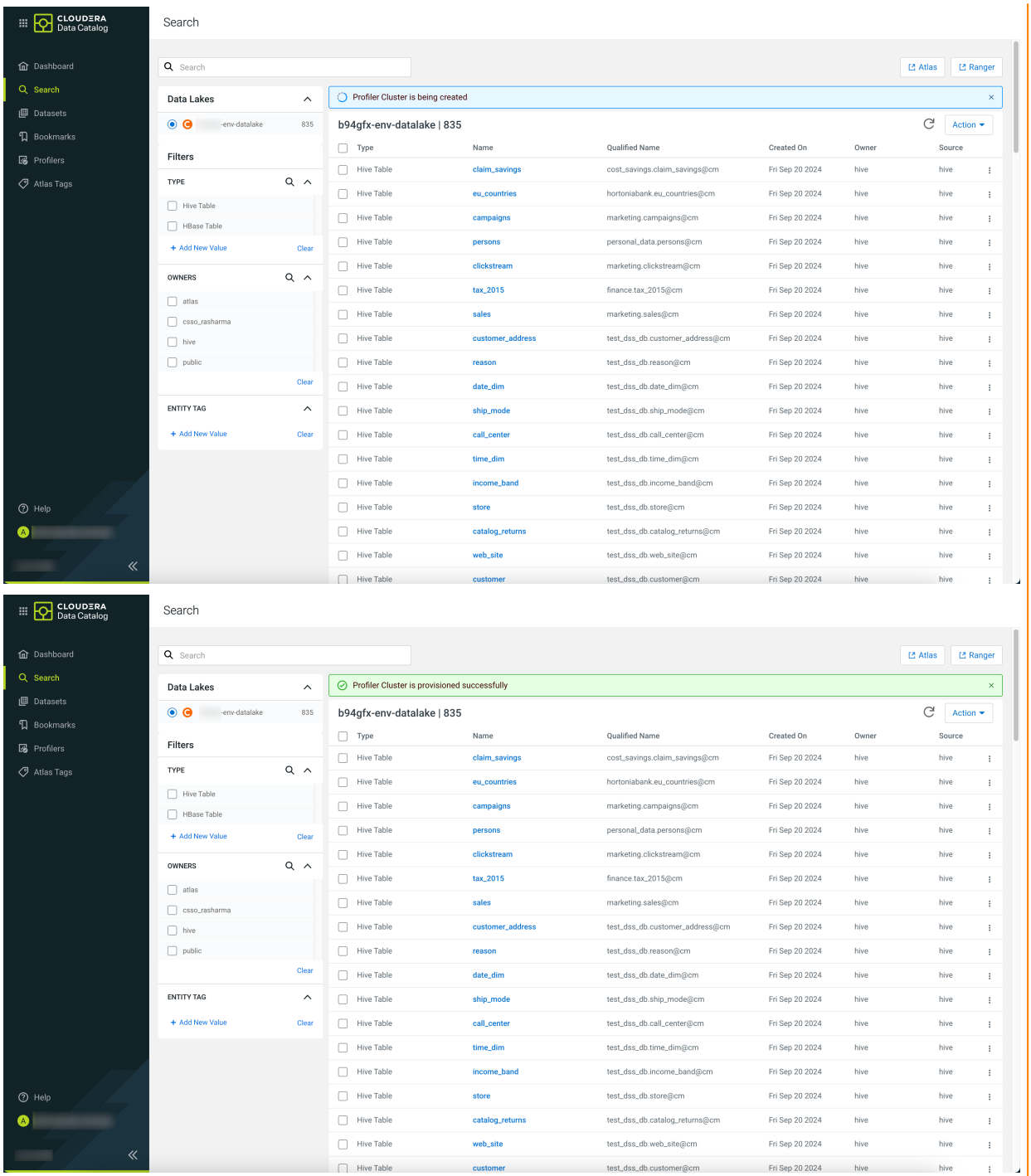


Click Setup Profiler. Cloudera Data Catalog processes the request and the profiler creation will start.



The High Availability (HA) feature for profilers, including launching and managing jobs are supported by default. No separate action is required to enable the HA functionality or its components.

Note: Once you schedule the profiler jobs, navigate to the **Profilers** page to view the status of the respective profiler jobs.



Related Information

Cloudera Data Catalog Profilers before 1.5.5 SP2
Cloudera Data Catalog Profilers after 1.5.5. SP2
The Cluster Sensitivity Profiler
Data Compliance Profiler
The Ranger Audit Profiler
The Activity Profiler
The Hive Column Profiler
The Statistics Collector Profiler

Deleting profilers

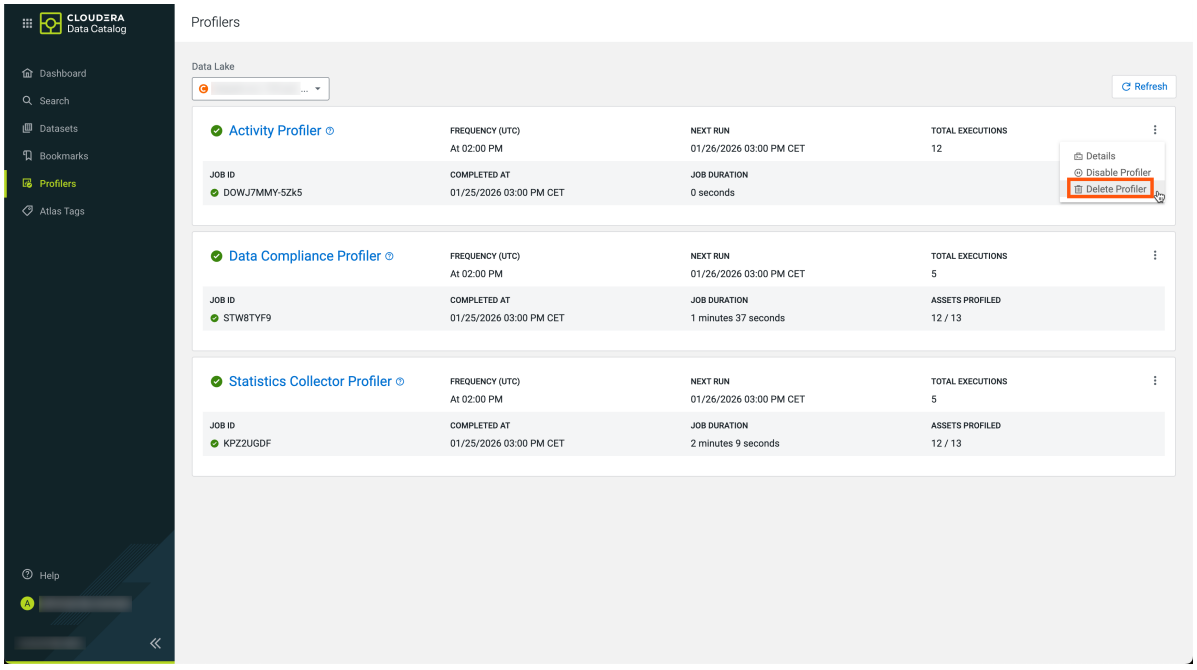
Learn how to delete profiler jobs and clusters, including steps to retain or download modified rules, and understand the impact on associated data and Kubernetes objects.

Warning:

- When you delete the scheduled jobs, the associated Kubernetes cron job object is deleted from the Kubernetes cluster.
- The associated data (cron jobs, keytabs and associated configMaps of the profilers) are also deleted.
- In an OpenShift (OCP) environment, deleting the profiler uninstalls the Helm release, gracefully removing the deployments, config maps, and jobs.
- In a standard Kubernetes environment (such as ECS), the compute namespace is deleted, which cascades to delete all underlying pods, jobs, and config maps.

For After Cloudera Data Catalog 1.5.5 SP2

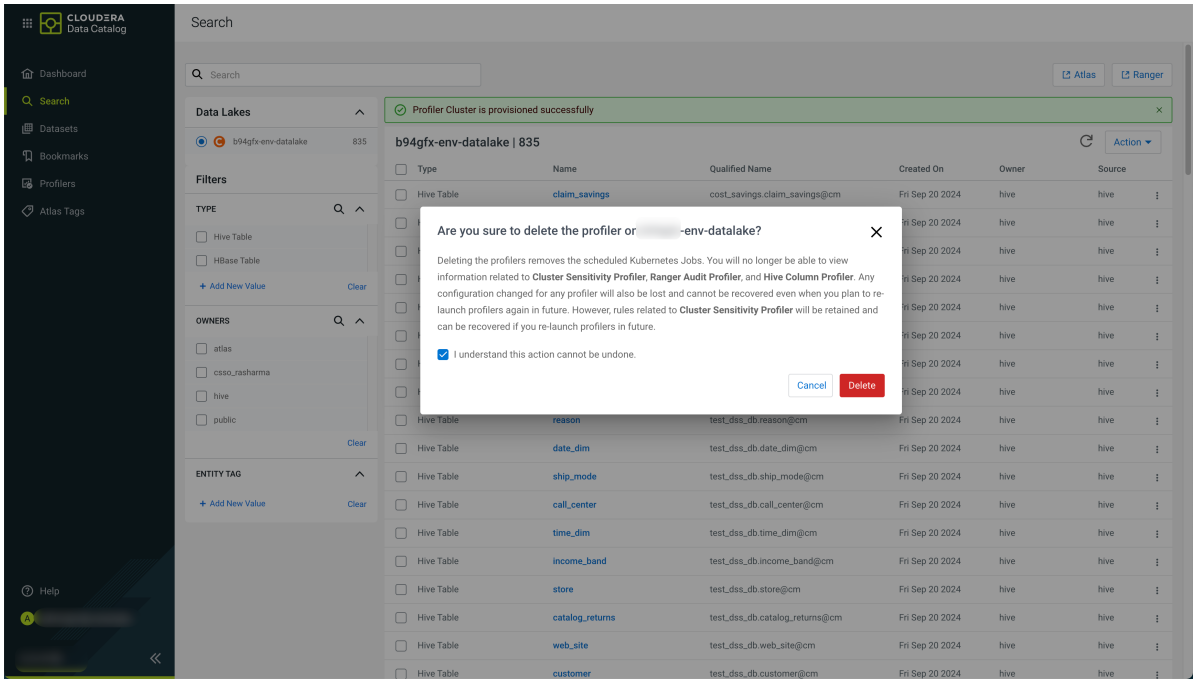
1. Go to **Profilers** and select your Data Lake.
2. Click the ⋮ icon, then Delete Profiler.



3. Confirm your selection.

For Before Cloudera Data Catalog 1.5.5 SP2

1. On the **Search** page, select the Data Lake.
2. Click the Actions drop-down menu and select Delete Cluster.
3. Click Yes to proceed.
4. If you agree, select the warning message I understand this action cannot be undone.



5. Click **Delete**.

   The application displays the following message:

   The profiler cluster is deleted successfully

On-Demand Profilers

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

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

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

For After Cloudera Data Catalog 1.5.5 SP2

Figure 1: Confirming on-demand profilers

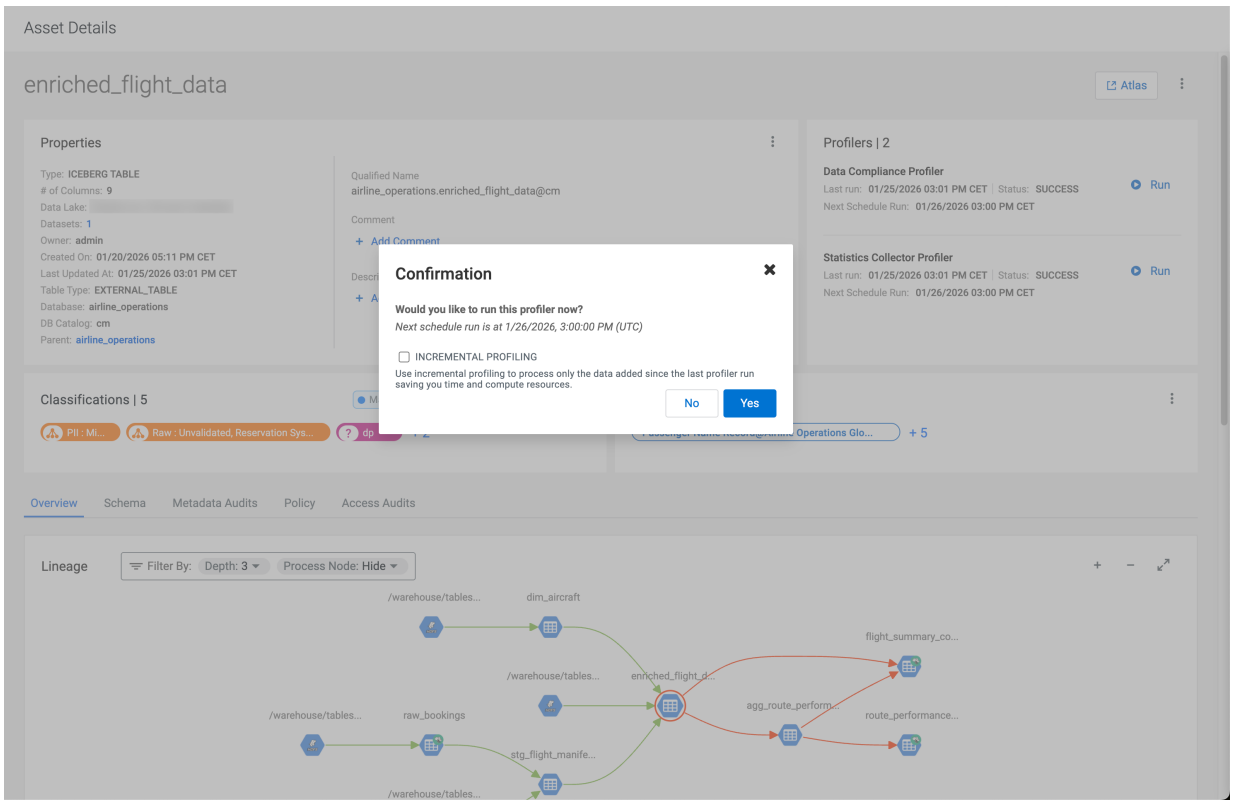


Figure 2: Tracking on-demand profilers

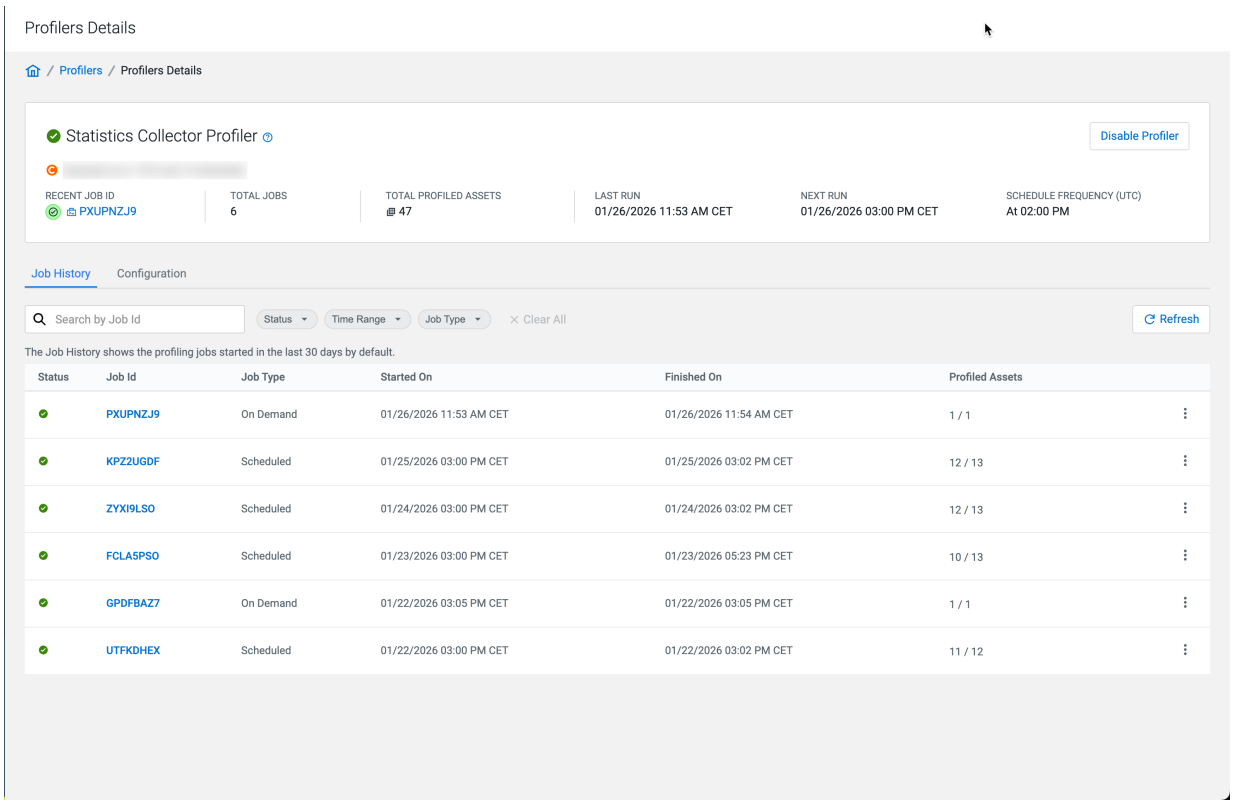
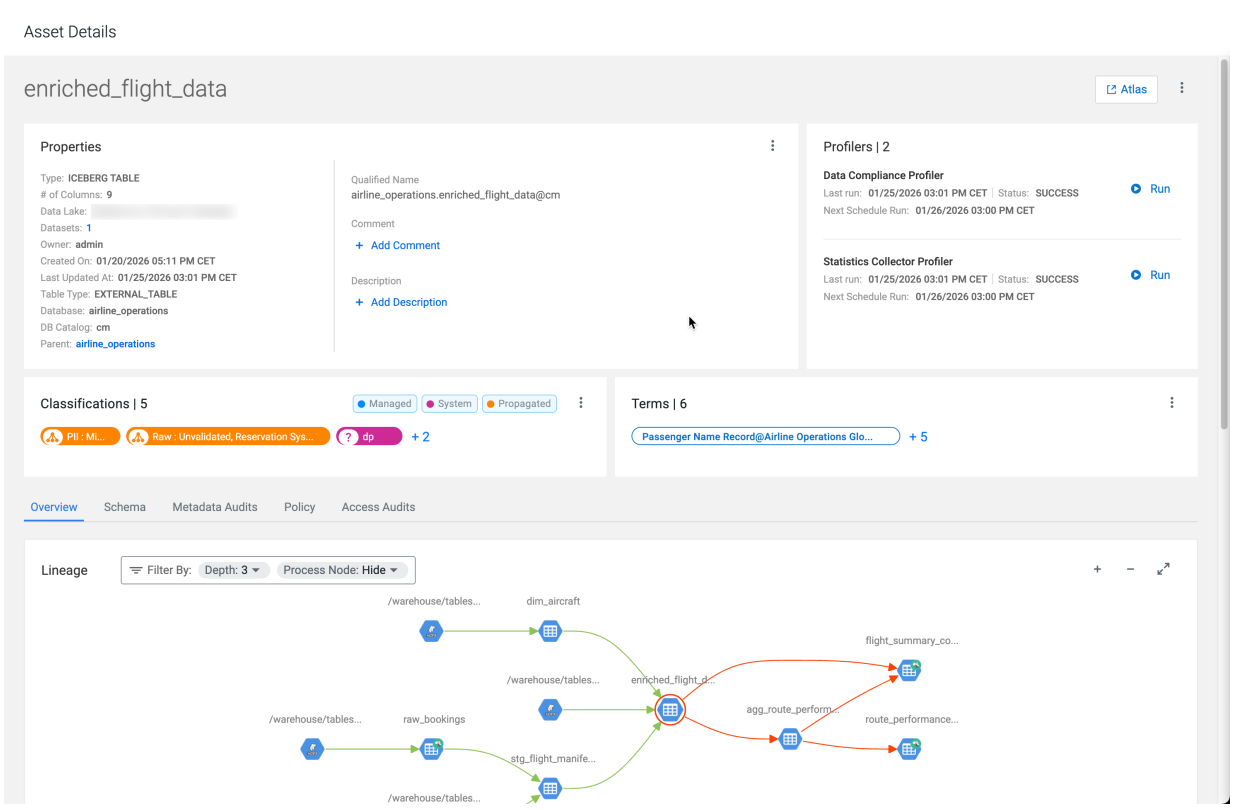


Figure 3: On-demand profilers in progress



For Before Cloudera Data Catalog 1.5.5 SP2

Figure 4: On-demand profilers in progress

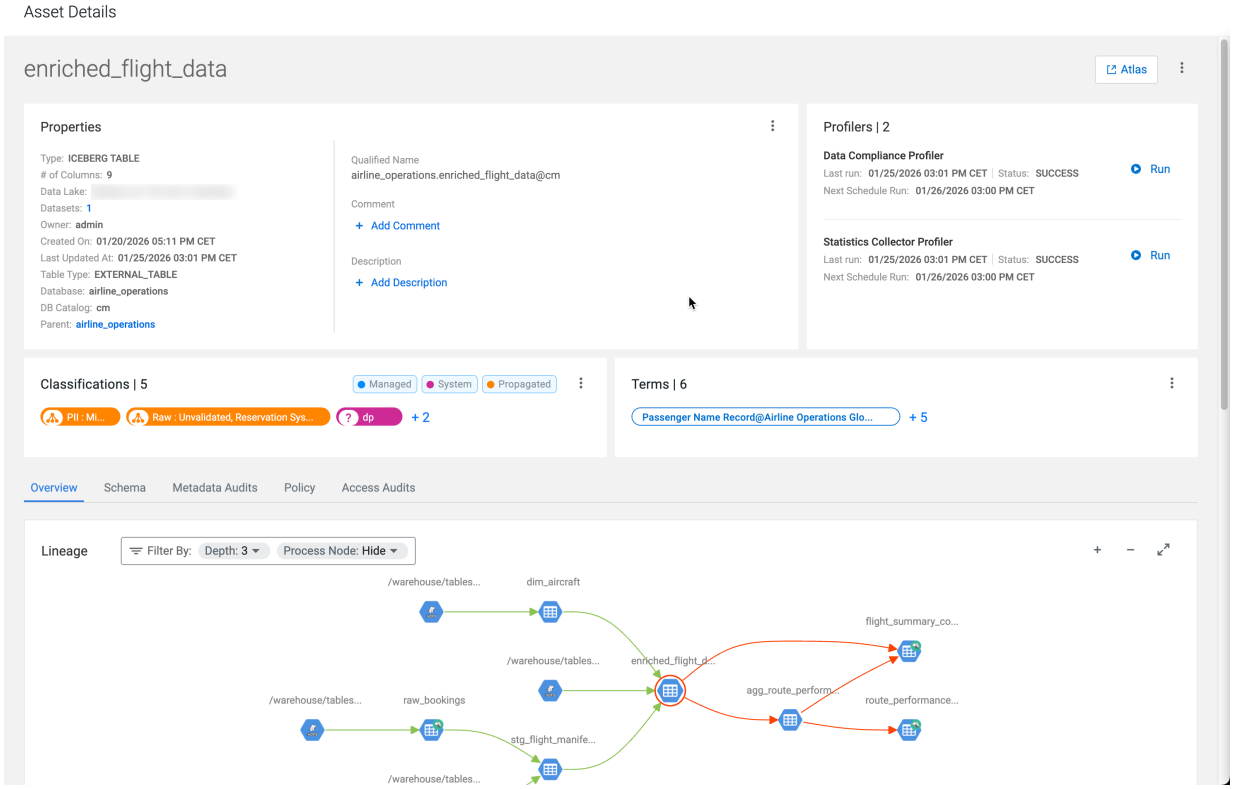
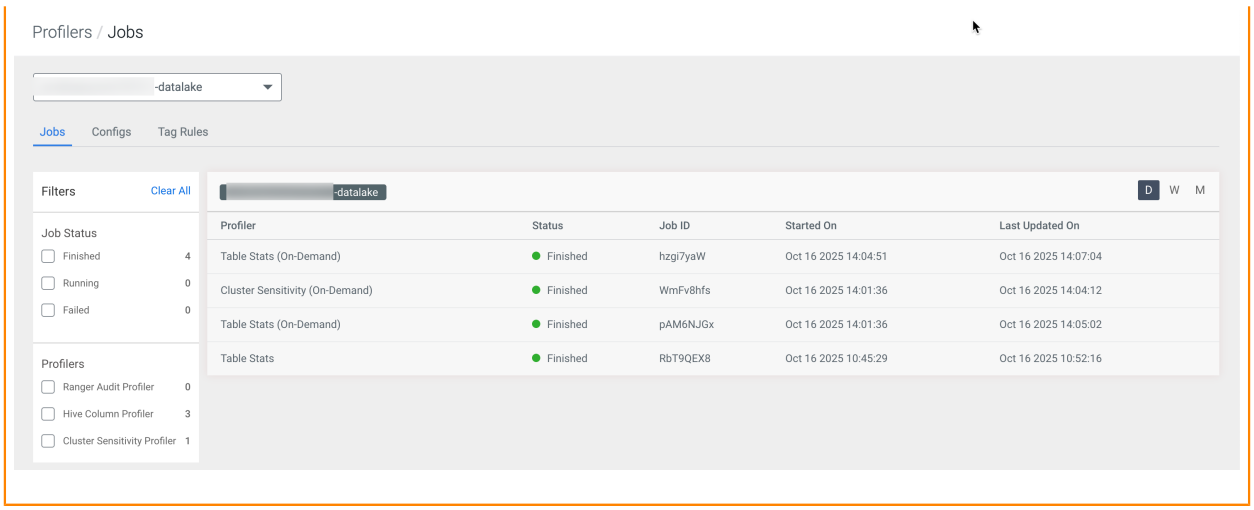


Figure 5: Tracking on-demand profilers



Note:

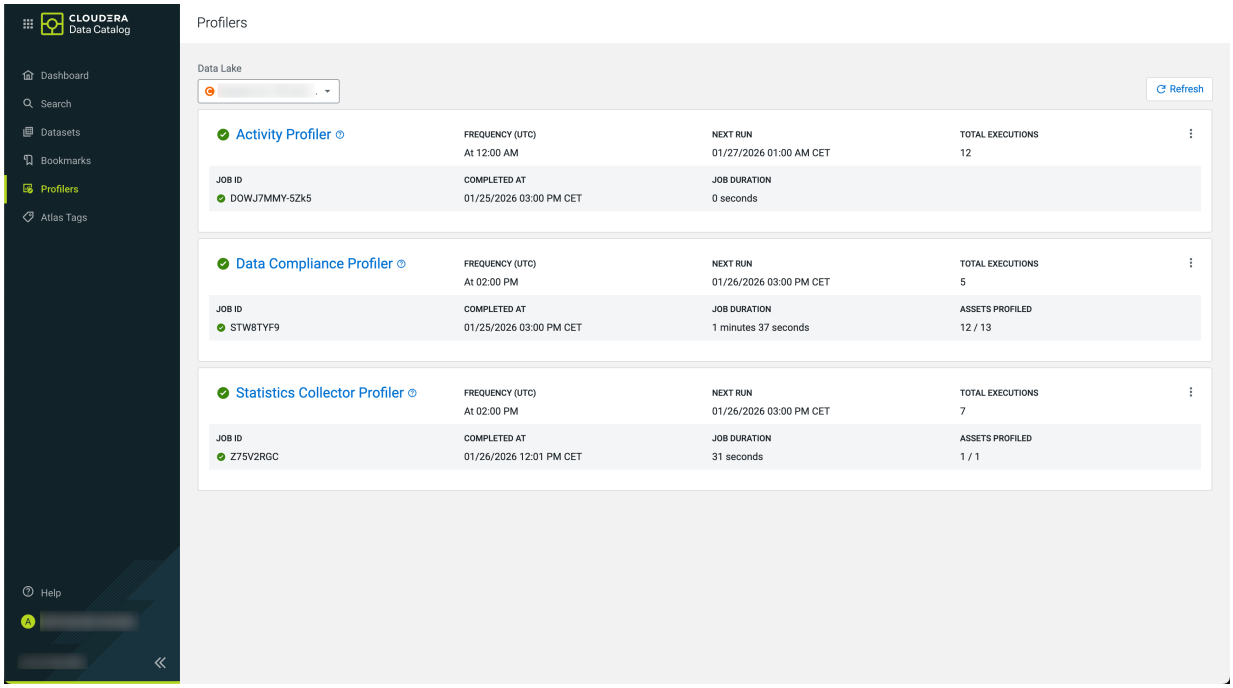
- In Cloudera Data Catalog on premises 1.5.5 SP2 and higher, Iceberg tables can be also profiled with the On-Demand Profiler.
- On-demand profilers do not support incremental profiling. For more information, see Configuring the Data Compliance Profiler and Configuring the Statistics Collector profiler.

Tracking Profiler Jobs

Track and monitor profiler jobs using the Profilers > Jobs page to view job statuses, execution details, and troubleshoot issues. Access key information such as profiler type, job status, and profiling history to manage and analyze your assets effectively.

For After Cloudera Data Catalog 1.5.5 SP2

In **Profilers**, you can see your last run profiling jobs. Use this page to quickly check if your profiler jobs are failing.



For each profiler, you can view the details about:

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



Note: Not available for the Activity Profiler.

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

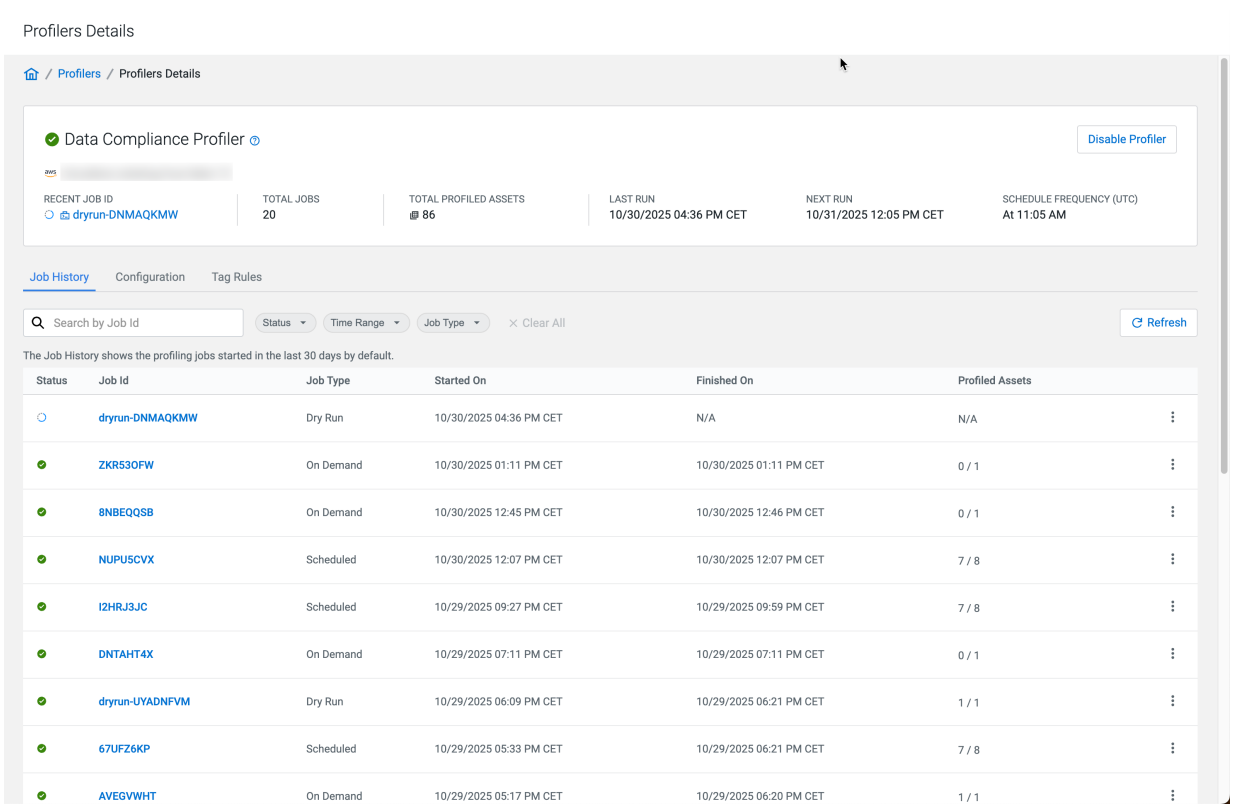
Click >Details to gain more information about your profiler jobs in Profilers Profilers Details .

- Status:
  - Failed
  - Running
  - Finished
- Time Range
- Modes[1]
  - **Dry Run**[2]
  - **On Demand**
  - **Scheduled**

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

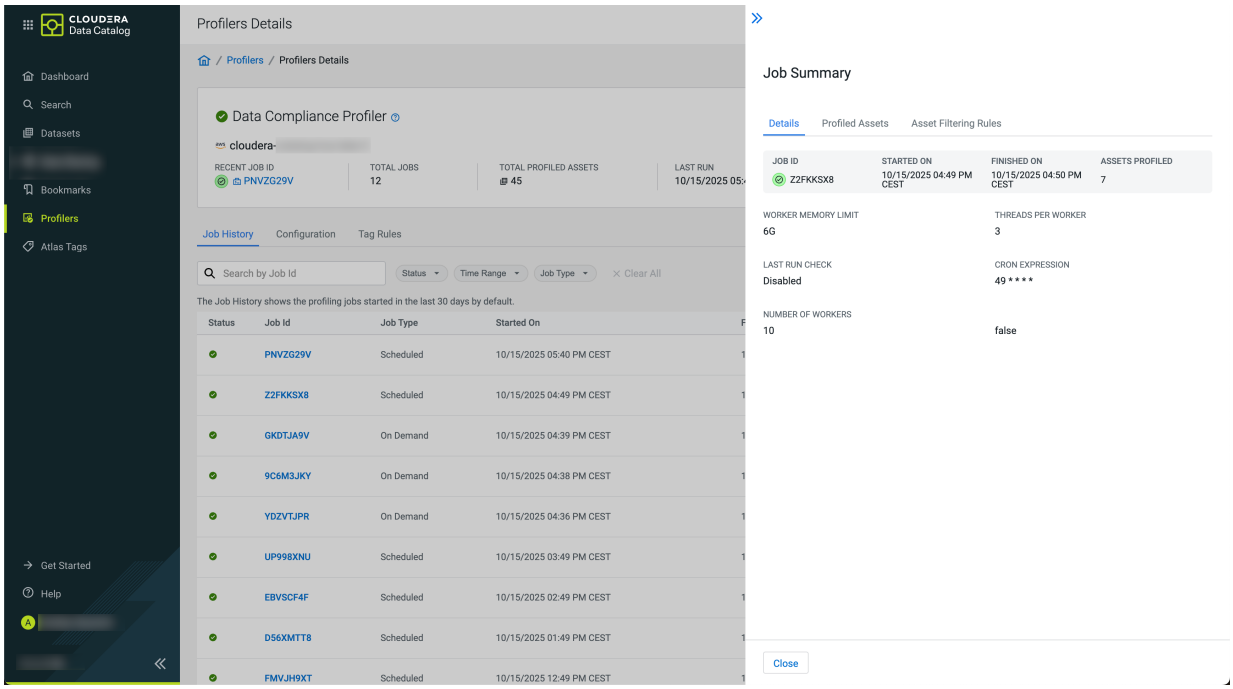
[2] Only available for the Data Compliance profiler.

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

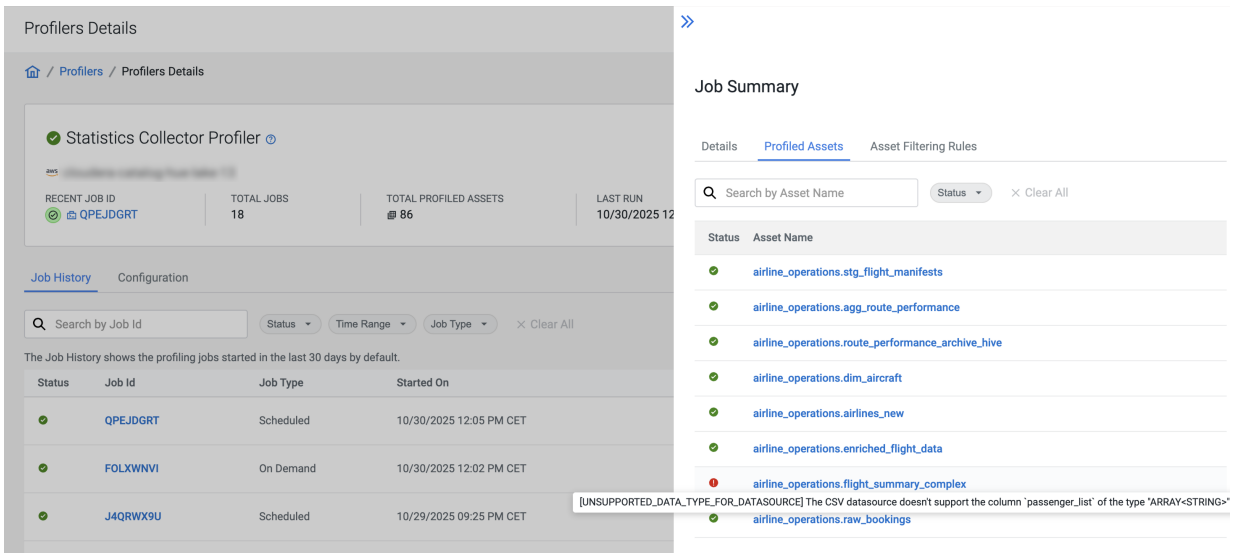


By clicking ⋮ by the individual jobs in **Profilers Details**, you can drill further down to **Job Summary** and **Profiled Assets**.

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

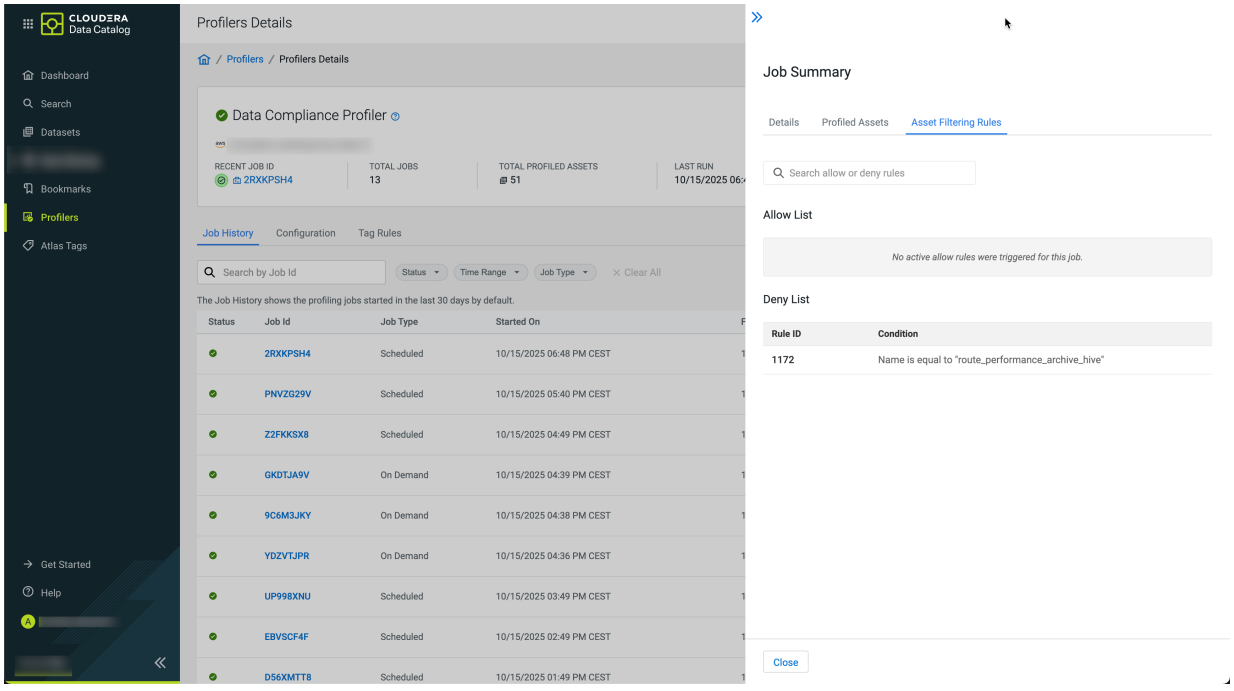


The **Profiled Assets** not only gives you a list of entities that were selected by your profiler to be profiled, but it let's you filter them.



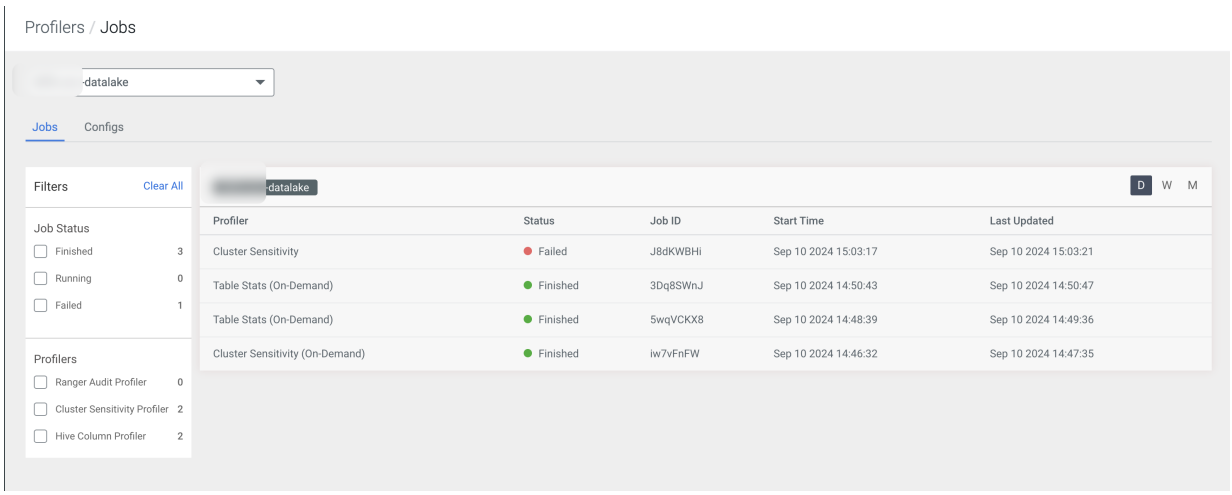
Note: Hovering over the skipped assets shows the reason for not including the particular asset. When trying to profile any unsupported assets, the status SKIPPED is shown in Profiler Details Job History Job Summary Profiled Assets .

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



For Before Cloudera Data Catalog 1.5.5 SP2

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



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

- **Profiler** type
- Job **Status**
- **Job ID**
- **Start Time**
- **Last Updated** Time

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

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

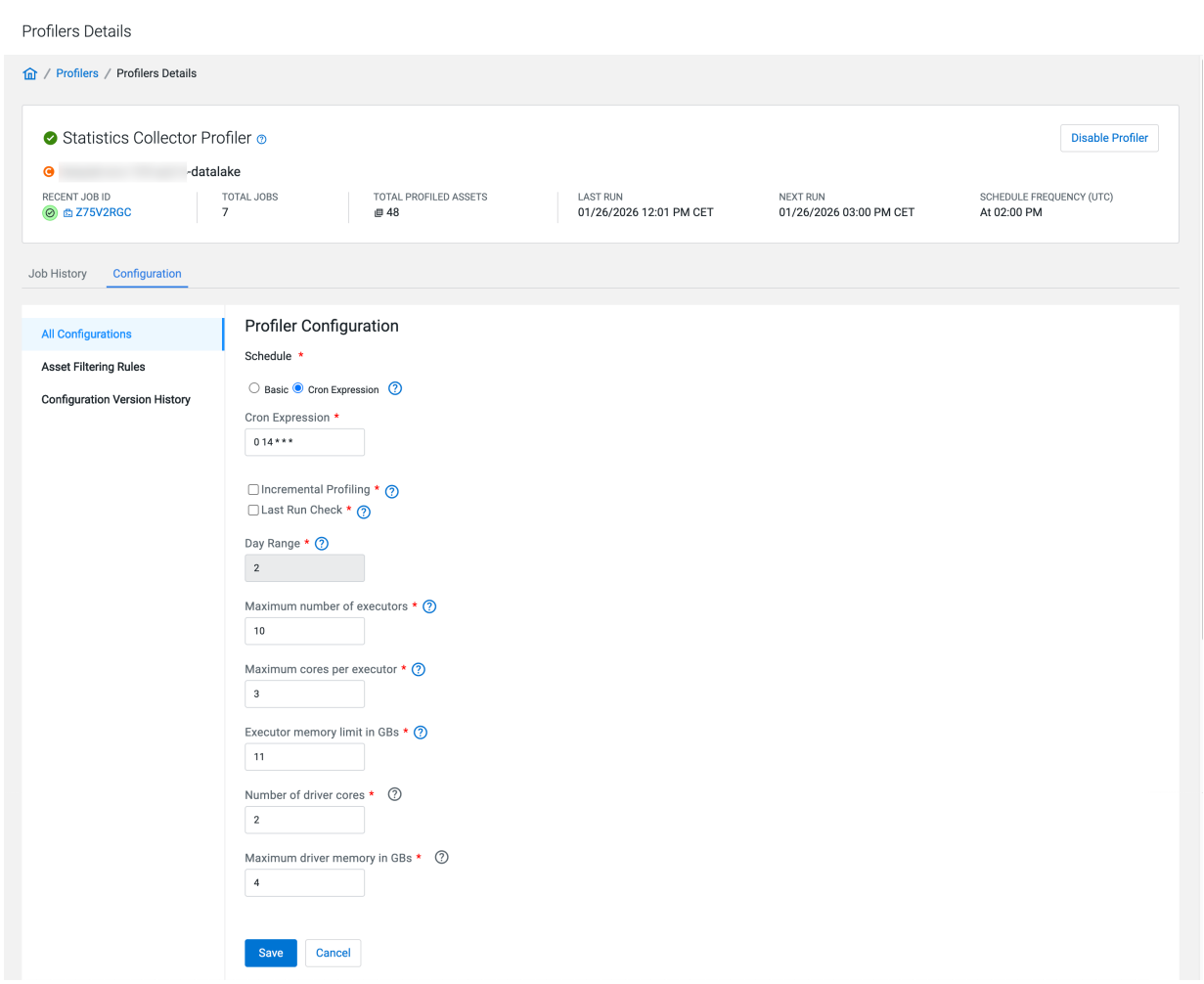
In case of Ranger Audit profiling, there could be a “NA” status for the total number of assets profiled. It indicates that the auditing that happens is dependent on the Ranger policies. In other words, the Ranger policies are actually profiled and not the assets.

Viewing profiler configurations

You can monitor the last status of individual profilers. Also, you can change their resources, sensitivity and scheduling.

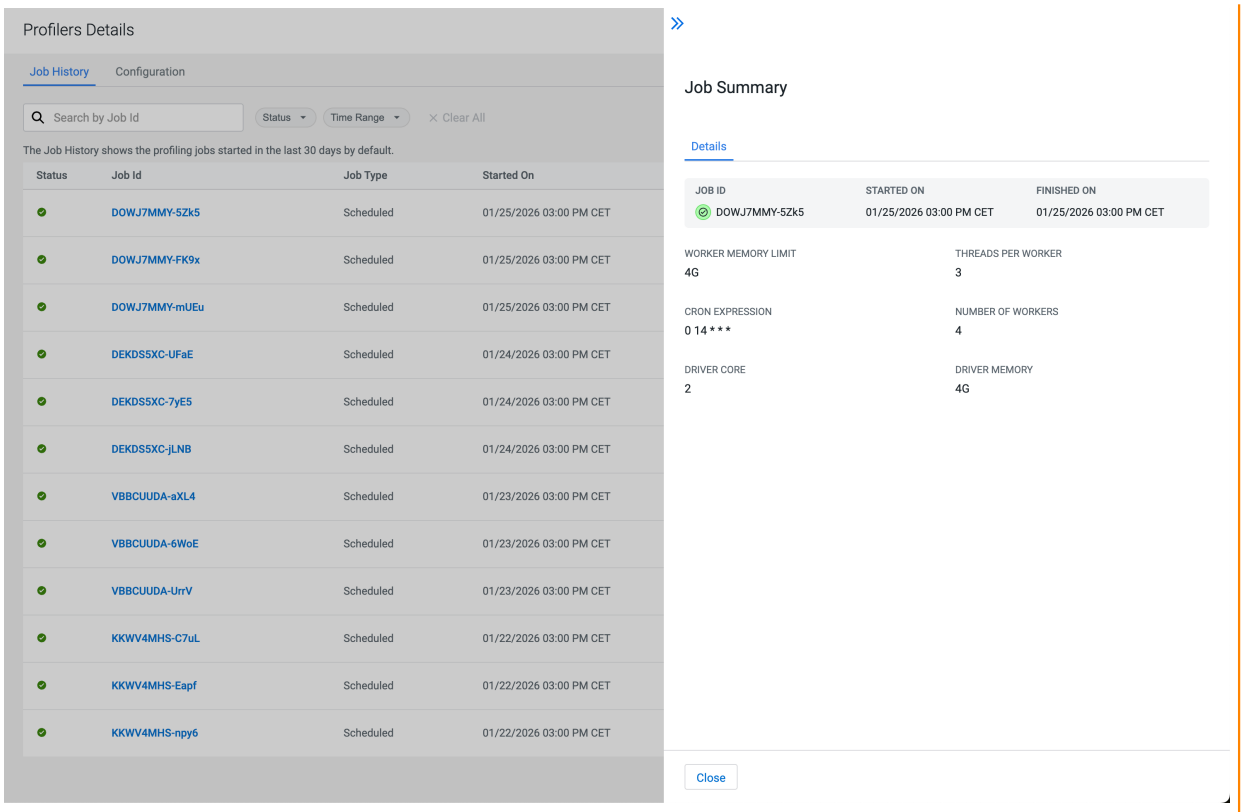
For After Cloudera Data Catalog 1.5.5 SP2

You can check the configuration and its changes for your profiles in Profilers Profiler Details Configuration .



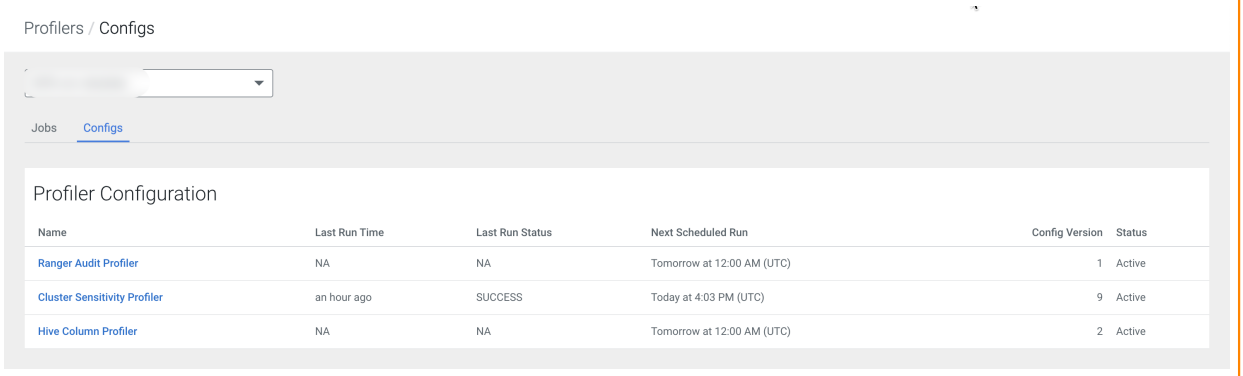
By clicking ⋮ by the individual jobs in **Profilers Details**, you can drill further down to **Job Summary** and **Profiled Assets**.

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



For Before Cloudera Data Catalog 1.5.5 SP2

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



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

Monitoring the profiler configurations has the following uses:

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

Configuring the Cluster Sensitivity Profiler

In addition to the generic configuration, you can configure scheduling and available resources for the Cluster Sensitivity Profiler.

Before you begin

Note: This profiler is only available for Cloudera Data Catalog on premises 1.5.5 SP1 or lower versions.

Procedure

1. Go to **Profilers** and select your data lake.
2. Go to Profilers Configs .
3. Select Cluster Sensitivity Profiler.
   The **Detail** page is displayed.
4. Click the 

toggle to enable or disable the profiler.
5. Select a schedule to run the profiler. This is implemented as a Quartz cron expression.
   For more information, see Understanding the Cron Expression generator on page 57.
6. Select Last Run Check and set a period if needed.

   Note:

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

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

   If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.
7. Set the sampling configurations. When an asset or table is profiled, instead of scanning the whole table, the profiler sample selects only a subset of records.
   a) Set the Sample Count – The number of times the sampling query runs against the table. Cloudera recommends a value between 3 and 30.
   b) Set the Sample Factor – A decimal value used as the row-selection probability when the profiler queries Hive. For each row, a random number is generated; if it is less than or equal to the Sample Factor, the row is included in the sample. Lower values scan less data (faster); higher values produce a more representative sample. The accepted value range is from 0.001 through 0.5.
   c) Set the Sample Records – The maximum number of rows returned per sampling run. This value maps to the LIMIT clause in the SQL query. The accepted value range is from 100 through 100,000.
8. Continue with the **Pod Configurations** and set the Kubernetes job resources.

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

   - Pod CPU limit – Specifies the maximum number of cores that can be allocated to a pod. The value range is from 1-8

- Pod CPU Requirements – Specifies the minimum number of CPUs that will be allocated to a pod when it is provisioned. If the node where a pod is running has enough resources available, a container can use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The value range is from 1-8.
- Pod Memory limit – Specifies the maximum amount of memory that can be allocated to a pod. The value range is from 1 through 256.
- Pod Memory Requirements –Specifies the minimum amount of RAM that will be allocated to a pod when it is provisioned. If the node where a pod is running has enough resources available, a container can use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The value range is from 1-256.

9. Update the **Executor Configurations**.

   Executor configurations specify the runtime configuration. These configurations must be changed if you are changing the pod configurations and when you require additional compute power.

   - **Number of workers** – Specifies the number of processes that are used by the distributed computing framework. The value range is from 1 through 8.
   - **Number of threads per worker** – Specifies the number of threads used by each worker to complete the job. The value range is from 1 through 8.
   - **Worker Memory limit in GB** – Enforces a memory usage threshold for a given worker to prevent memory overutilization. For example, if you have an 8 GB pod and 4 threads, the value of this parameter must be 2 GB. The value range is from 1 through 4.

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

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

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

       1. Select the **Deny List** or **Allow List** tab.
       2. Click Add New to define new rules.
       3. Select one of the following keys from the drop-down list.:
          - Database name
          - Asset name
          - Asset owner
          - Path to the asset
          - Created date
       4. Select the operator from the drop-down list. Depending on the keys selected, you can select an operator such as equals, contains. For example, you can select the name of assets that contain a particular string.
       5. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
       6. Click Done. Once the rule is added, you can enable or disable it as needed by clicking the state toggle.

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

Related Information

The Cluster Sensitivity Profiler

Understanding the Cron Expression generator

Profiler tag rules

You can use preconfigured tag rules or create new rules based on regular expressions and values in your data to limit the number of assets to be profiled by the Cluster Sensitivity Profiler. When a tag rule is matching your data, the selected Apache Atlas classification, also known as a Cloudera Data Catalog tag, is applied. This way you can save compute resources instead of running the profiler on the full dataset.

Note:

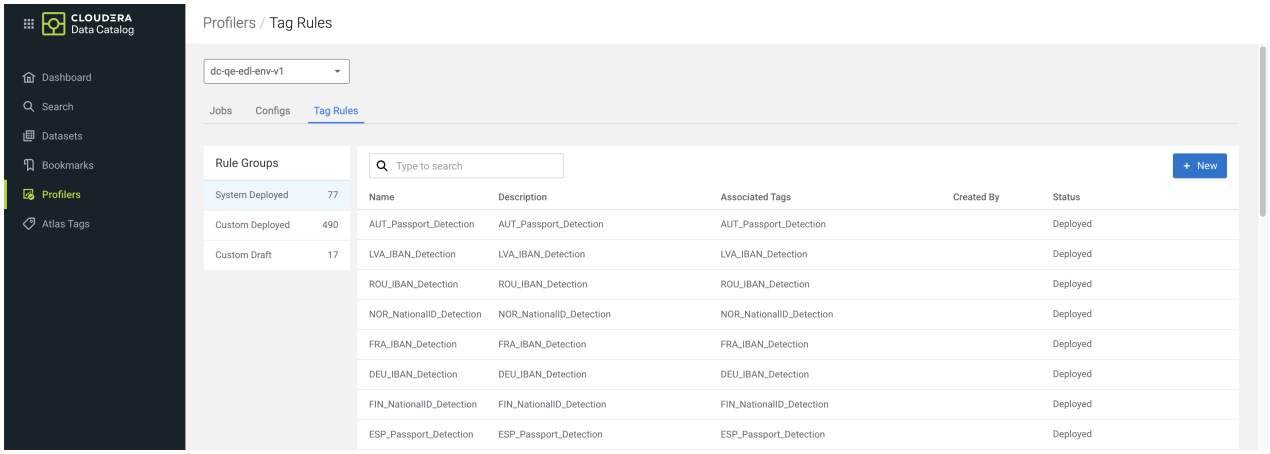
This implementation of the feature is only available for Cloudera Data Catalog on premises 1.5.5 SP1 or lower versions.

Tag rule types

Tag Rules are categorized by type into the following groups:

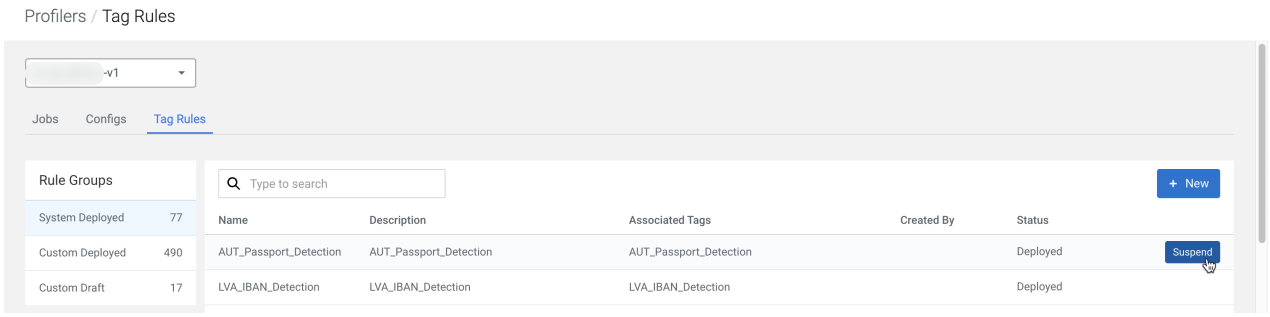
- **System Deployed** – These are built-in rules that cannot be edited. You can only enable or disable them for your data.
- **Custom Deployed** – These are tag rules that you create, edit, and deploy on clusters after validation.

  Hover your mouse over the tag rules to deploy or suspend them as needed.
- **Custom Draft** – These are new tag rules you can create and save for later validation and deployment on clusters.



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

Note: You can temporarily suspend Tag Rules.



Match thresholds and weights

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

Creating custom profiler rules

You can create a custom profiler by adding the required tags, regex entries to specific columns within your tables. You can evaluate the tags and regex entries to Boolean values.

Before you begin

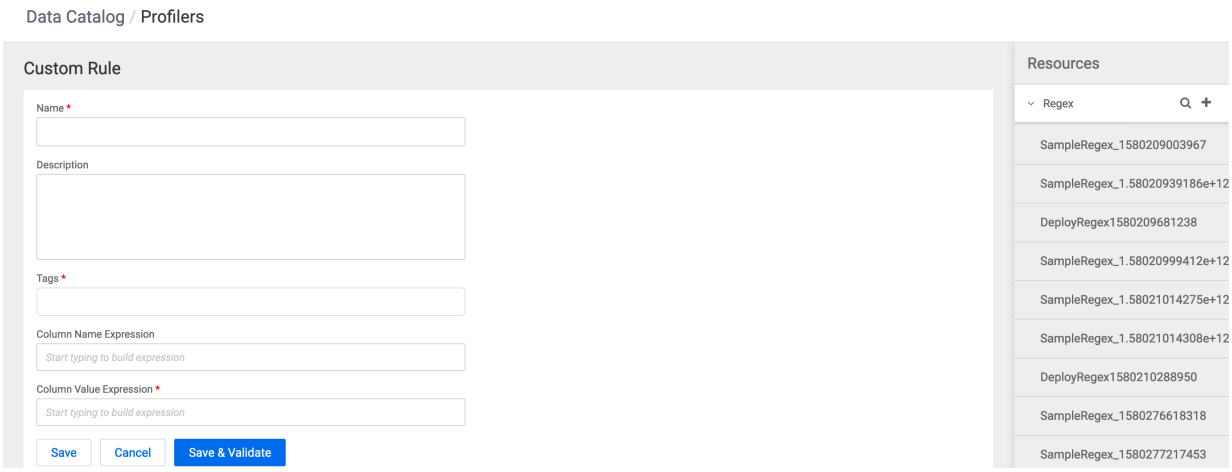
Note: This feature is only available in Cloudera Data Catalog on premises 1.5.5 SP1 or lower versions.

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 to apply to assets that match your regex expressions. You can select tags from the drop-down list 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 Apache 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, Cloudera Data Catalog auto-populates values from the entries already available in the **Resources** tab. For more information about behaviors, see Using DSL Grammar.
8. Click Save and Validate.



   In the validation modal window, enter data to validate your custom profiler tag rule. Ensure that you separate each data entry with a new line.
9. Click Save to create a tag rule, then 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. If a match is found, the expression is evaluated to True.

Before you begin

Note: This feature is only available for Cloudera Data Catalog on premises 1.5.5 SP1 or earlier.

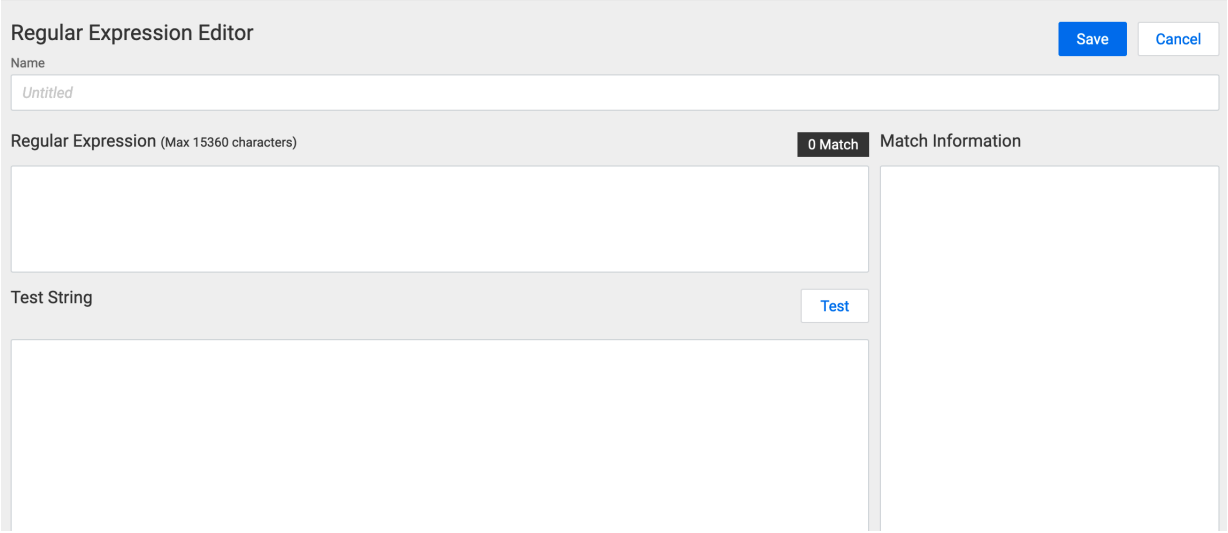
Procedure

1. Go to Profilers Tag Rules .
2. Click New.

3. In the **Resources** in the panel on the **Custom Profiler** page, click +. The **Regular Expression Editor** page appears.
4. Enter the name of the new regular expression.
5. 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
```



6. 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.
7. Click Save to add the new regular expression to the list of regex Resources.

Using DSL grammar

Using Domain-specific Language (DSL) grammar, you can combine different regex expressions in intuitive ways to bring out new functionality while creating custom profiler rules.

Note: This feature is only available for Cloudera Data Catalog on premises 1.5.5 SP1 or earlier.

The two behaviors available in this framework are the following:

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.

Unary NOT operator

The NOT operator negates the value of a regex expression (from TRUE to FALSE and vice versa).

```
not(falseIdentity) == trueIdentity
```

```
not(trueIdentity) == falseIdentity
```

Use the not operator as follows.

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

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

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

```
val rule4=rule3 or trueIdentity
```

```
rule4 and not(falseIdentity)
```

The above expression evaluates to true.

Creating tag rules

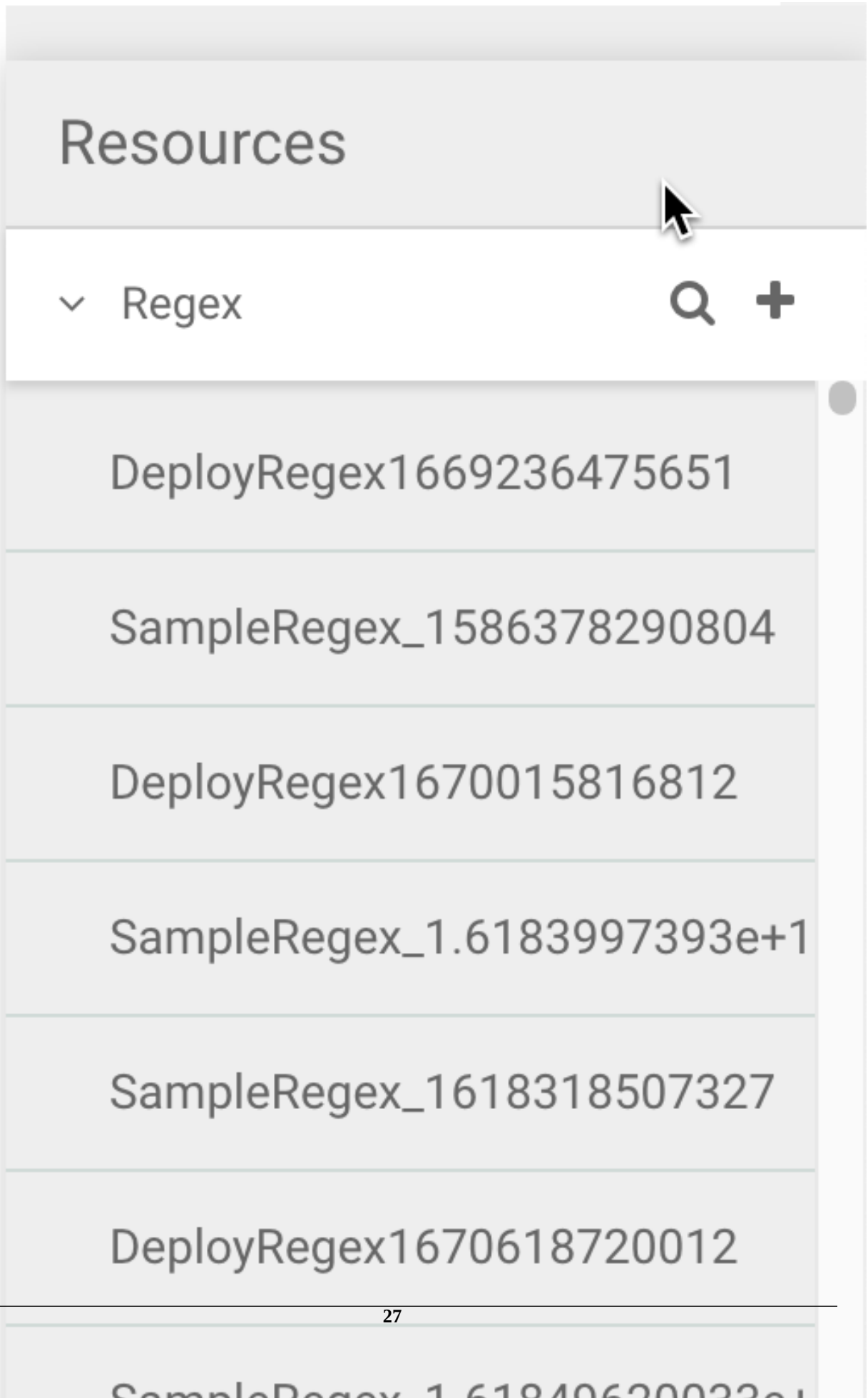
With tag rules, you can apply Apache Atlas classifications to your assets based on regex expressions.

Before you begin

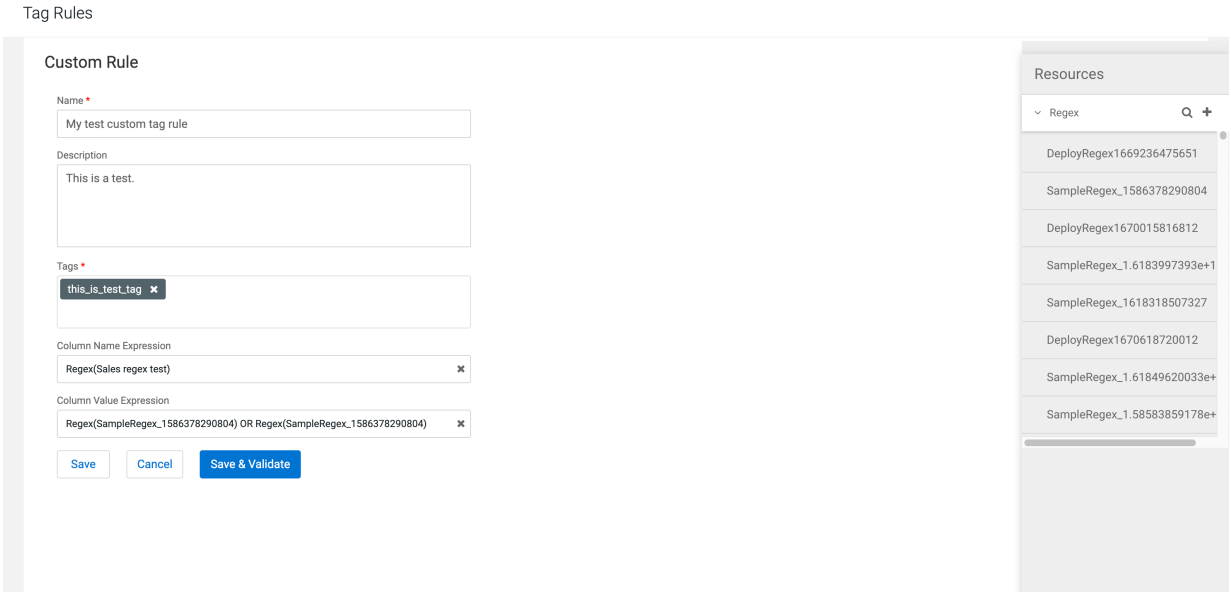
This implementation of the feature is only available for Cloudera Data Catalog on premises 1.5.5 SP1 or earlier.

Procedure

1. To start applying tags, go to **Profilers** and select your Data Lake.
2. Go to Profilers Tag Rules .
3. Click + New.
4. Name your tag rule and add a description to it.
5. Select the tags to be applied from the list of available tags synchronized from the list of Atlas classifications. Multiple tags can be selected.
6. In **Column Name Expression**, select at least one regular expression to use a match it against for column names. Select from the same regular expression you had created under the **Resources** pane.

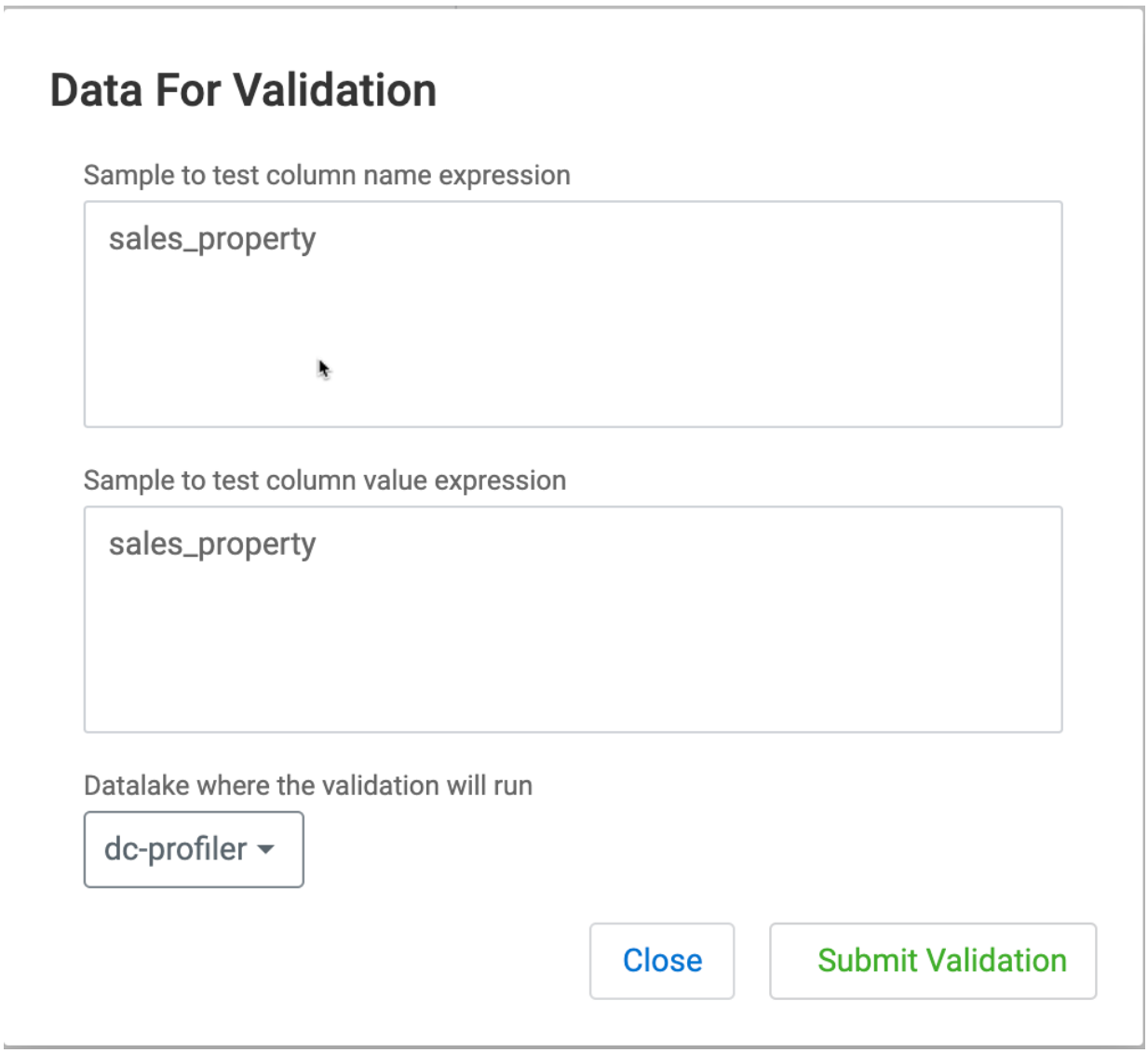

Resources
Regex
DeployRegex1669236475651
SampleRegex_1586378290804
DeployRegex1670015816812
SampleRegex_1.6183997393e+1
SampleRegex_1618318507327
DeployRegex1670618720012

Note: You can select multiple expressions connected by AND, OR, NOT logical operators.



7. In **Column Value Expression**, select at least one regular expression to use a match it against for column names.

   The **Column Name Expression** matches are considered with a 15% weightage in the final score when calculating if the tag needs to be applied. The **Column Value Expression** matches receive the remaining 85% weightage. The column name expression results are binary (TRUE, FALSE), while by column value a certain ratio of all values can be matched.
8. Click Save & Validate.
9. Enter some sample data manually to check the validity of your regular expression, then click Submit Validation.



The status for the newly created regular expression validation is displayed on the **Tags Rules** tab. Once the validation is successful, you can deploy the rule.

Configuring the Data Compliance Profiler

In addition to the generic configuration, you can configure scheduling and available resources for the Data Compliance.

Before you begin

Note: This profiler is only available for Cloudera Data Catalog on premises 1.5.5 SP2 or higher versions.

Procedure

1. Go to **Profilers** and select your data lake.
2. Go to Profilers Data Compliance Profiler Details Configuration All Configurations .
3. Select a schedule to run profiler using either UNIX Cron Expression or the Basic scheduler.

Note: Both the Basic and Cron Expression scheduler use the UTC timezone instead of the local timezone of the user.

Figure 6: Profiler schedule with cron expression

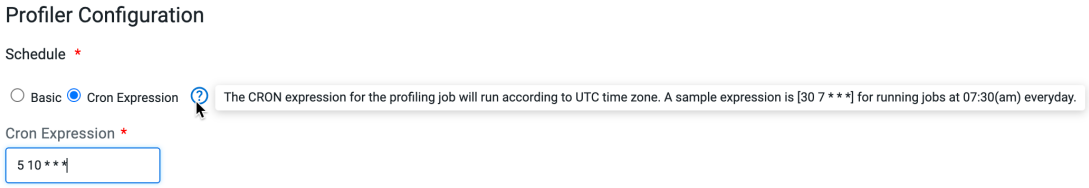
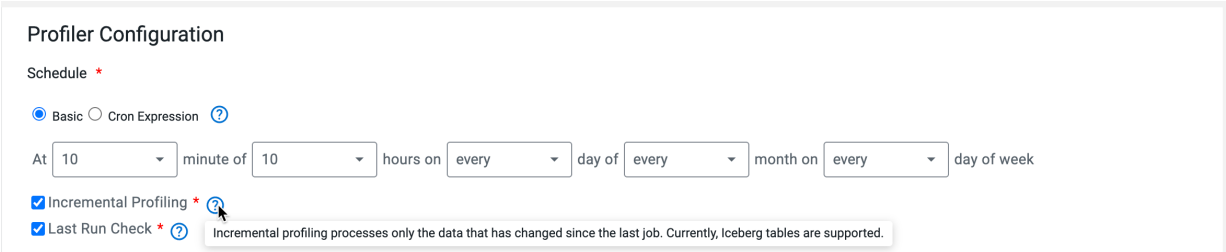


Figure 7: Profiler schedule with natural language



4. Select Incremental Profiling when needed.

   Using Incremental Profiling can decrease the compute resources and the time needed for the profiling job by processing only the Iceberg tables updated or added since the previous job.

   With Incremental Profiling, you can refine the results from the Last Run Check. Incremental Profiling checks the data (rows) in assets, while Last Run Check filters complete assets.
5. Select Last Run Check and set a period in Day Range if needed.

   Note:

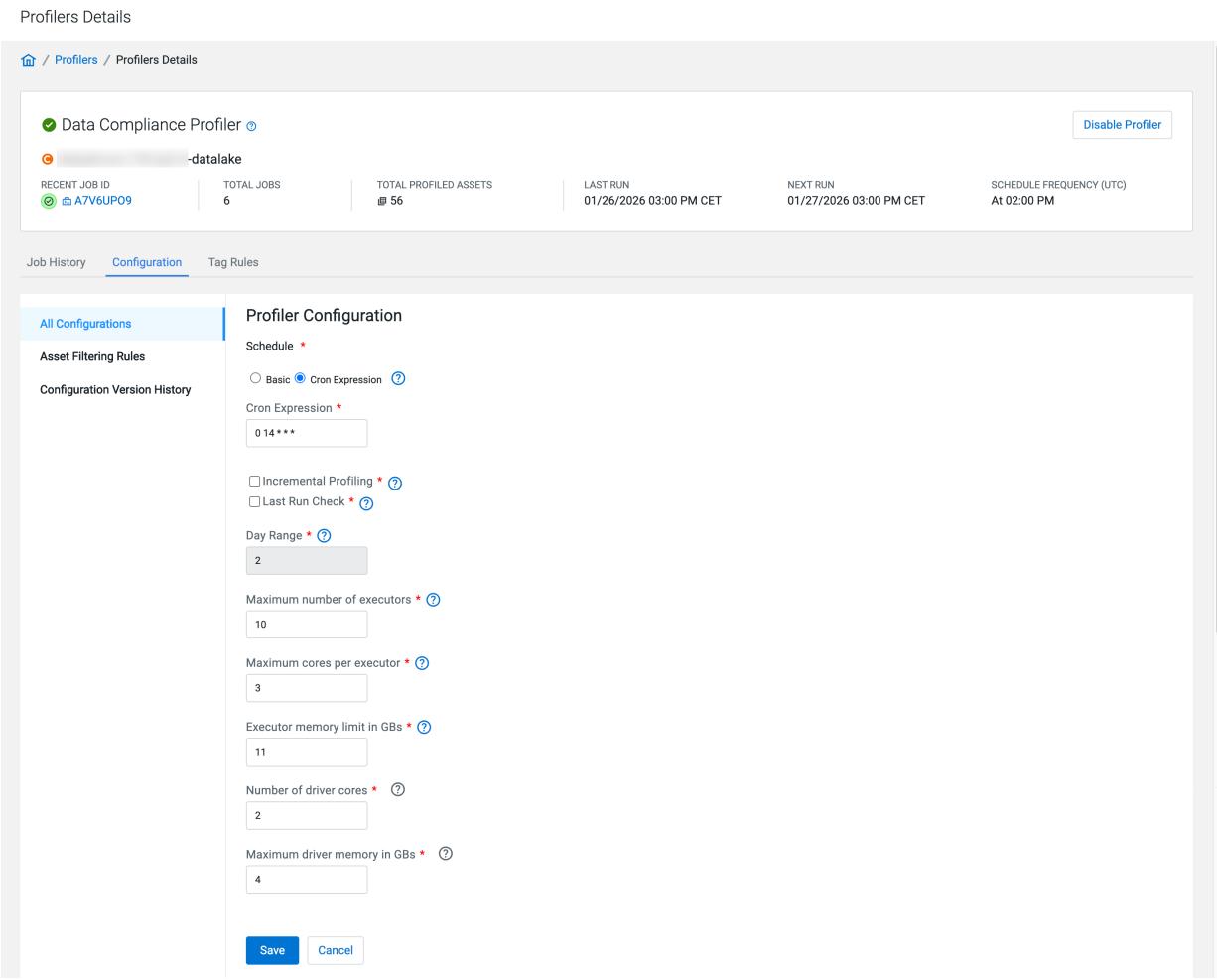
   The Last Run Check configuration enables profilers to avoid profiling the same asset on each scheduled run.

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

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

   The Last Run Check precedes Incremental Profiling.
6. Review sampling behavior.

   When an asset or table is profiled, instead of scanning the whole table, the profiler selects a subset of records. Sampling is automatic: the profiler loads data from HDFS and determines the sample size based on table size.
7. Configure the resources.



Note: Is your infrastructure is deployed completely in AWS, profilers might hang if the underlying AWS cloud provider environment cannot provide the necessary memory for the executor instances. In this case, reconfigure your executors with 4-5 GB memory in Profiler Details Configuration .

a) Set the Maximum number of executors.

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

b) Set the Maximum cores per executor.

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

c) Set the Executor memory limit in GBs.

d) Set the Number of driver cores.

   Specifies the maximum number of driver cores. Increase the number of cores to improve the speed of profiler job scheduling.

e) Set the Maximum driver memory in GBs.

   Specifies the maximum amount of memory that can be allocated to a driver core. Increasing the available memory accelerates the profiling of larger and more complex tables and prevents out-of-memory errors.

8. Click Save to apply the configuration changes to the selected profiler.
9. Add **Asset Filtering Rules** as needed to customize the selection and deselection of assets which the profiler profiles.

   - Profiler configurations apply to both scheduled and on-demand profiler jobs.

- Asset filtering rules apply to assets, such as tables, and not to complete databases. This means that filtering rules work at table level and not at a database level. You are able to filter out all tables of a specific database, but each table will be checked whether its parent is the selected database.
- Multiple asset filtering rules are evaluated together as if connected by the OR operator.
- You cannot enable conflicting Allow and Deny list rules at the same time. Enabling conflicting rules results in the following error message:



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

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

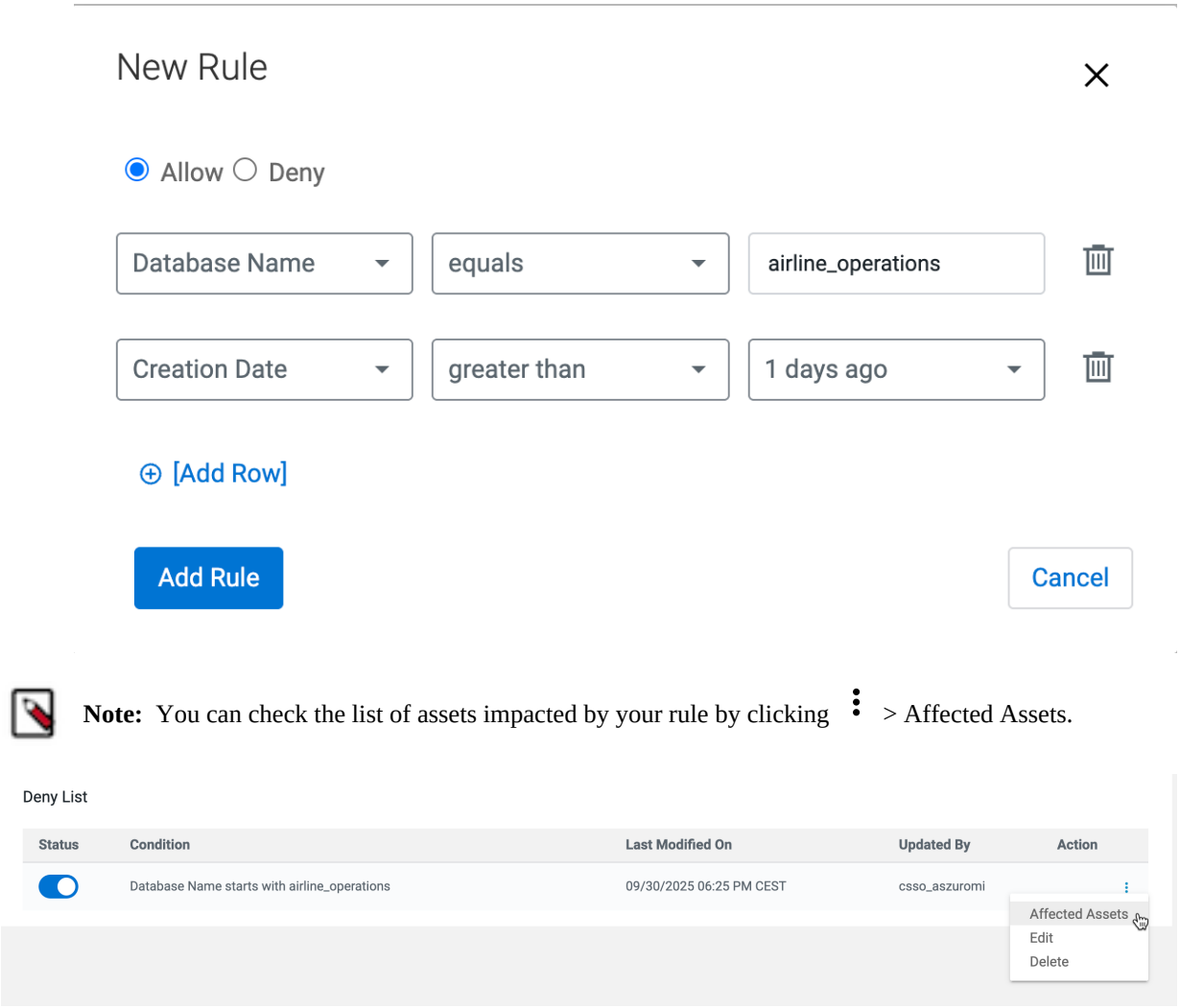
1. Click Add New Rule to define new rules.
2. Use the radio buttons to define your new rule for the Allow List or Deny List.
3. Select one of the following keys from the drop-down list and select the relevant operator:

Key	Operator
Database name	• equals • starts with • ends with
Name (of asset)	• equals • contains • starts with • ends with
Owner (of asset)	
Creation date[3]	• greater than • less than

Note: **Name** refers here to the actual name of the asset and not to its **Qualified Name**.

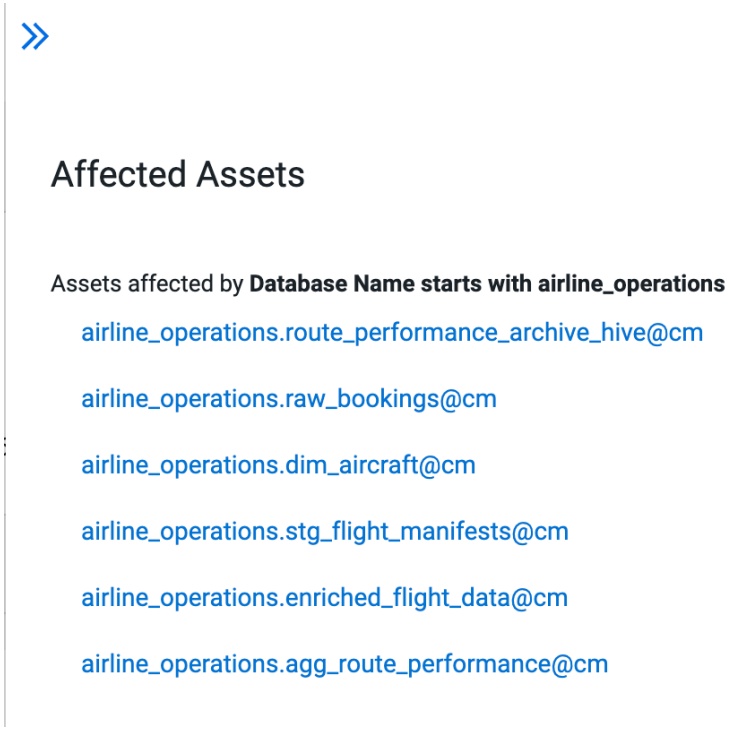
4. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
5. Click Add Rule. Once a rule is added (enabled by default), you can enable or disable it as needed by clicking the state toggle.

[3] By Creation Date, Greater than 7 days means an asset older than seven days. Less than 7 days means an asset younger than seven days.

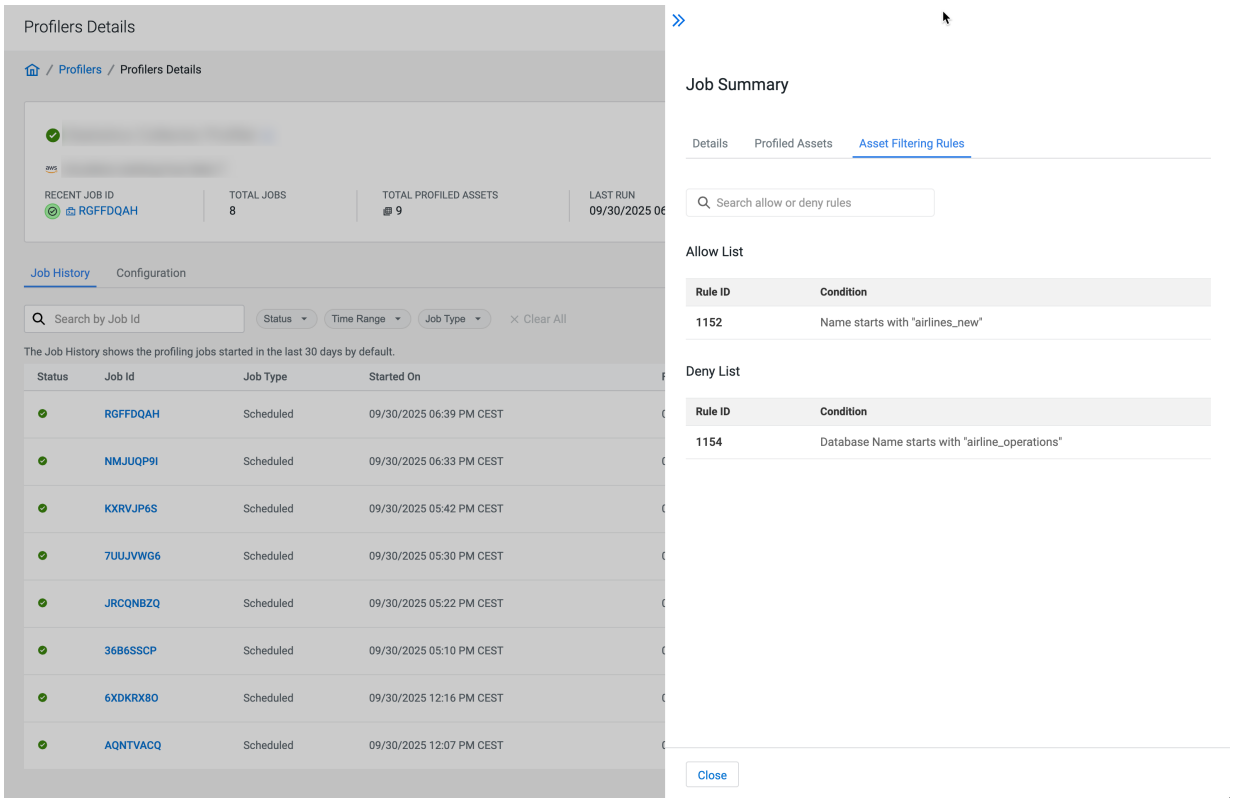


Note: You can check the list of assets impacted by your rule by clicking ⋮ > Affected Assets.

Figure 8: Affected Assets in Asset Filtering Rules configuration



Job Summary shows the asset filtering rules applied for the particular profiling job.



Related Information

Data Compliance Profiler

Profiler tag rules

You can use preconfigured tag rules or create new rules based on regular expressions and values in your data to limit the number of assets to be profiled by the Data Compliance. When a tag rule is matching your data, the selected Apache Atlas classification, also known as a Cloudera Data Catalog tag, is applied. This way you can save compute resources instead of running the profiler on the full dataset.

Note:

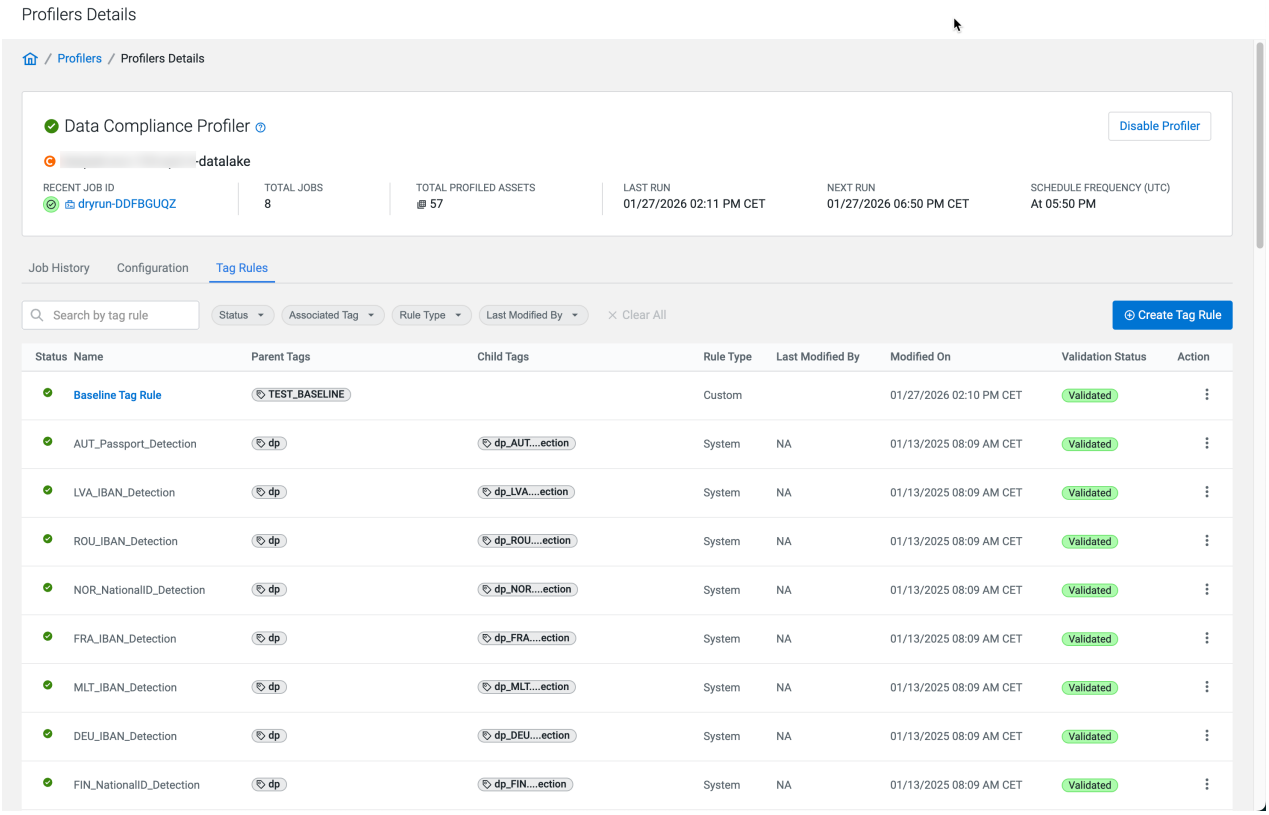
This implementation of the feature is only available for Cloudera Data Catalog on premises 1.5.5 SP2 or later.

Tag rule types

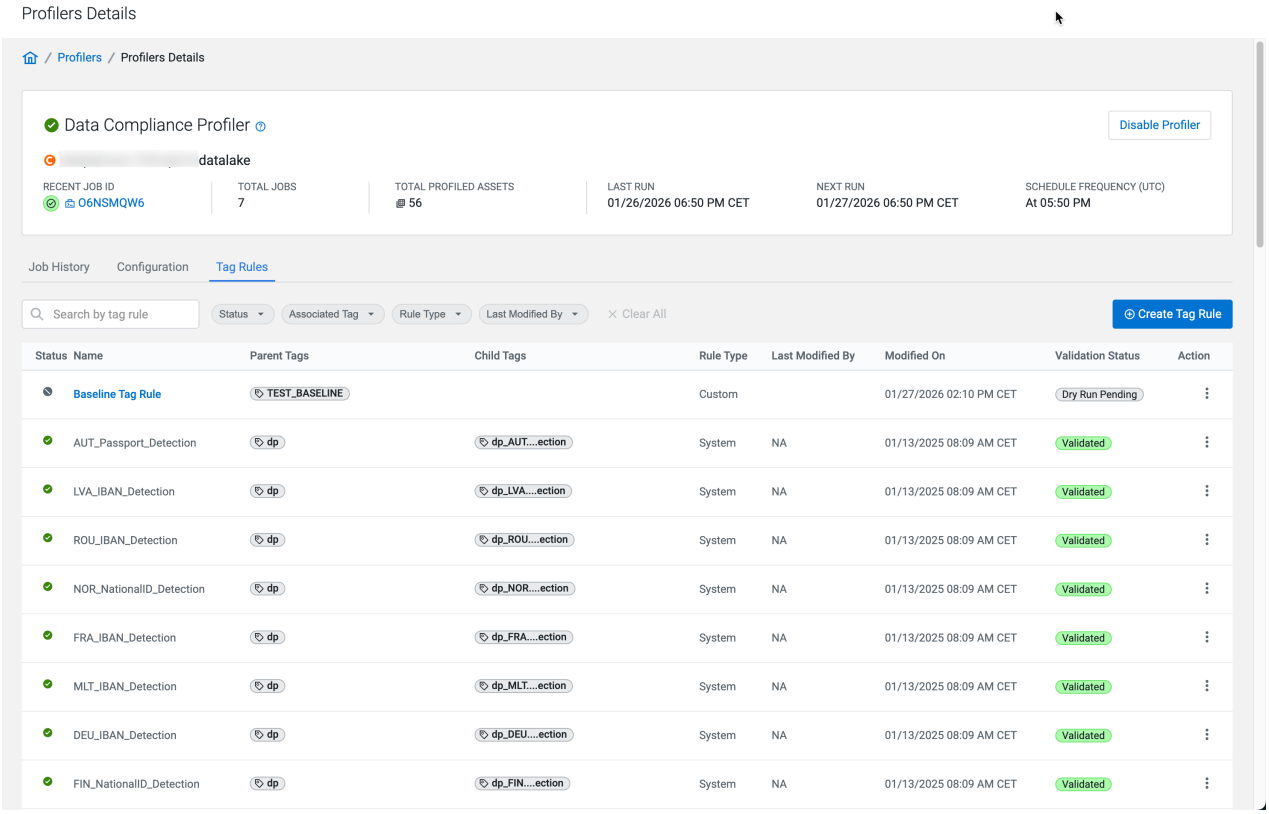
Tag Rules are categorized by type into the following groups:

- **System Defined** – These are built-in rules that cannot be edited. You can only enable or disable them for your data.
  - These rules use a 70% match threshold for column values against the regex. The final score is calculated with a weight of 85% for column values and 15% for the column name.
- **Custom** – These are tag rules that you create, edit, and deploy on clusters after validation.

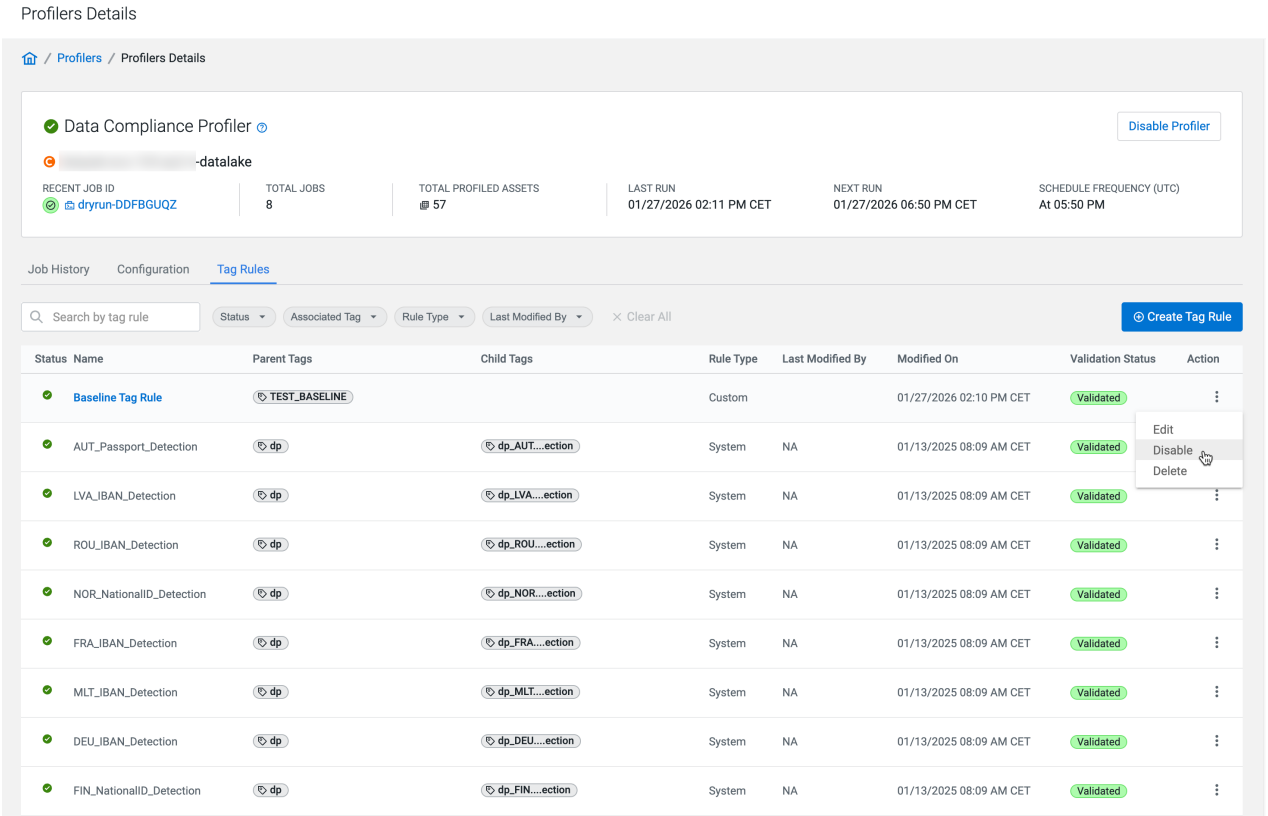
  Click the ⋮ icon in the **Action** column to enable your custom tag rules. You can also edit these tag rules.



After creating your rule, you have to validate them with test data, then Deploy them from **Dry Run Pending** status.



Note: You can temporarily disable Tag Rules.



Match thresholds and weights

In Compute Cluster-enabled environments, you can adjust the **Column Value Weightage** for tag rules defined with regex patterns. The column value weight percentage complements the column name weight to 100%. For example, if you set the column value weight to 80%, the column name adds either 20 or zero to the match score. This is because column name matching is binary (it either matches or does not match), while column value matching is dynamic, based on the percentage of matching rows.

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

Tag rule testing

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

In Compute Cluster-enabled environments, validate tag rules using data uploaded in a file, then save the tag rules to reach the Dry Run Pending status. Before deploying, you must also test them with a Dry Run on a subset of your data (up to 10 tables) in the Data Lake. A Dry Run is a special on-demand profiling job.

Tag handling by tag rules

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

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

Note:

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

Approving Data Compliance Profiler tags

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

About this task

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

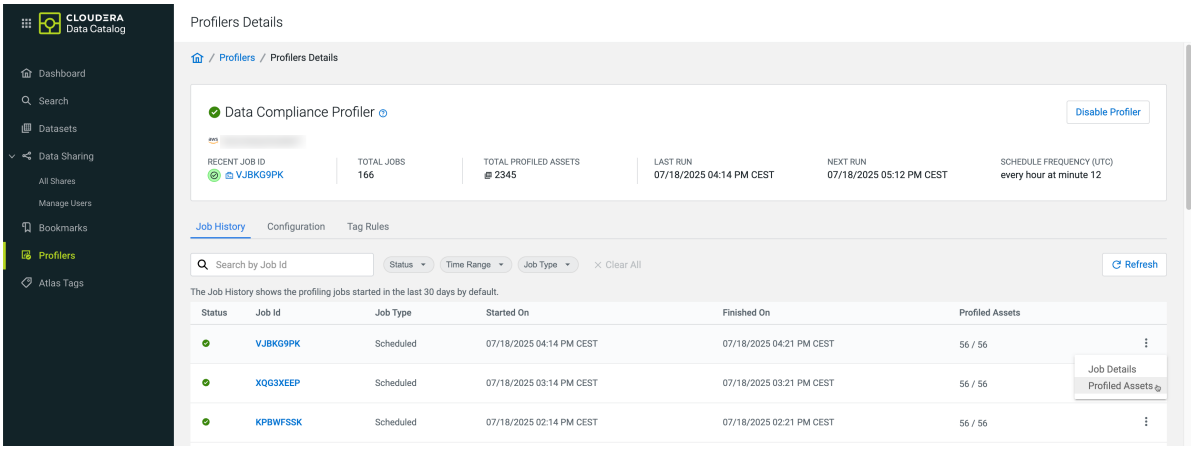
This gives you the ability, for example, to correct the tagging of assets mistakenly marked as PII. Incorrectly applied tags can affect your tag-based policies, unexpectedly changing the access of your users.

Before you begin

Note: This feature is only available for the Data Compliance Profiler in Cloudera Data Catalog on premises 1.5.5 SP2 or later.

Procedure

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



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

Option	
Asset Name	Opens the **Asset Details**.
Suggested Tags Count	Opens the list of identified tags for the asset. If the asset has columns, you can see the tags assigned to each of them.

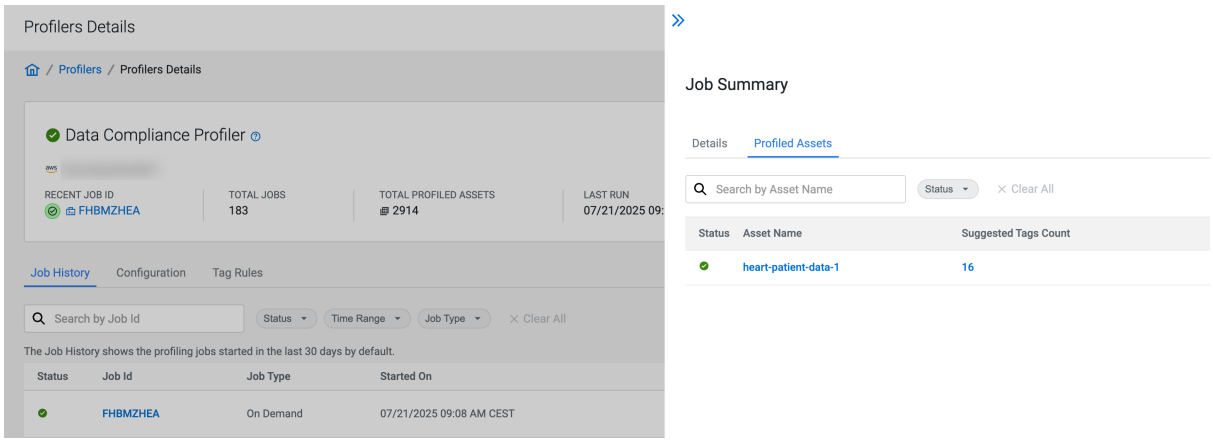
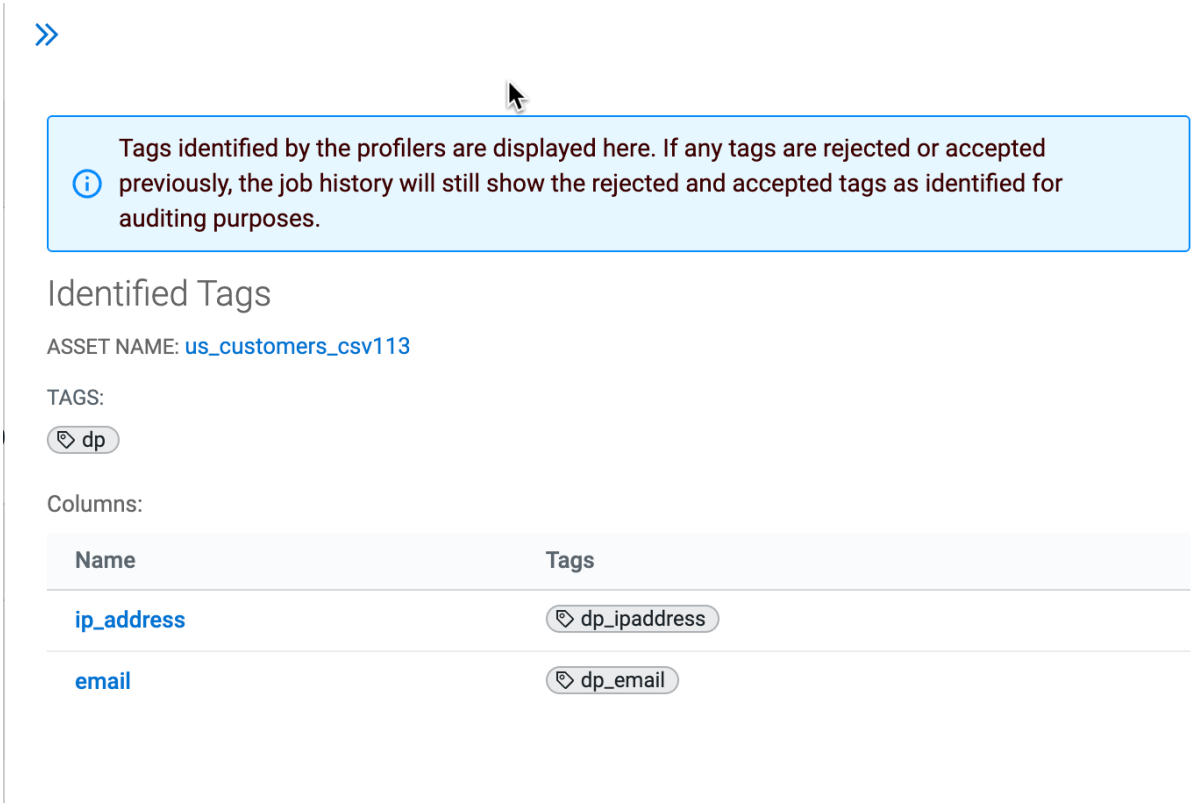


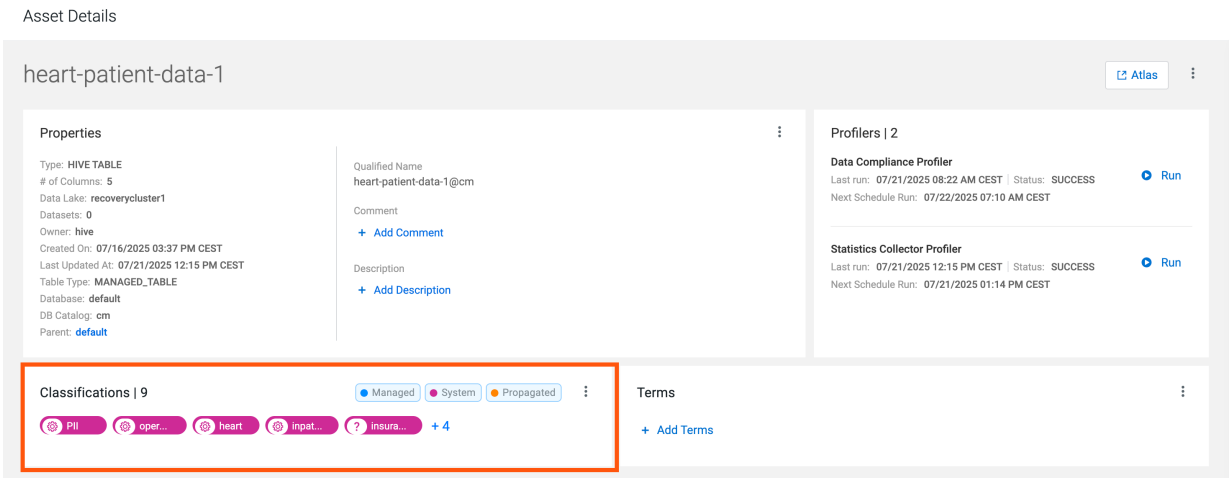
Figure 9: List of tags in Suggested Tags Count



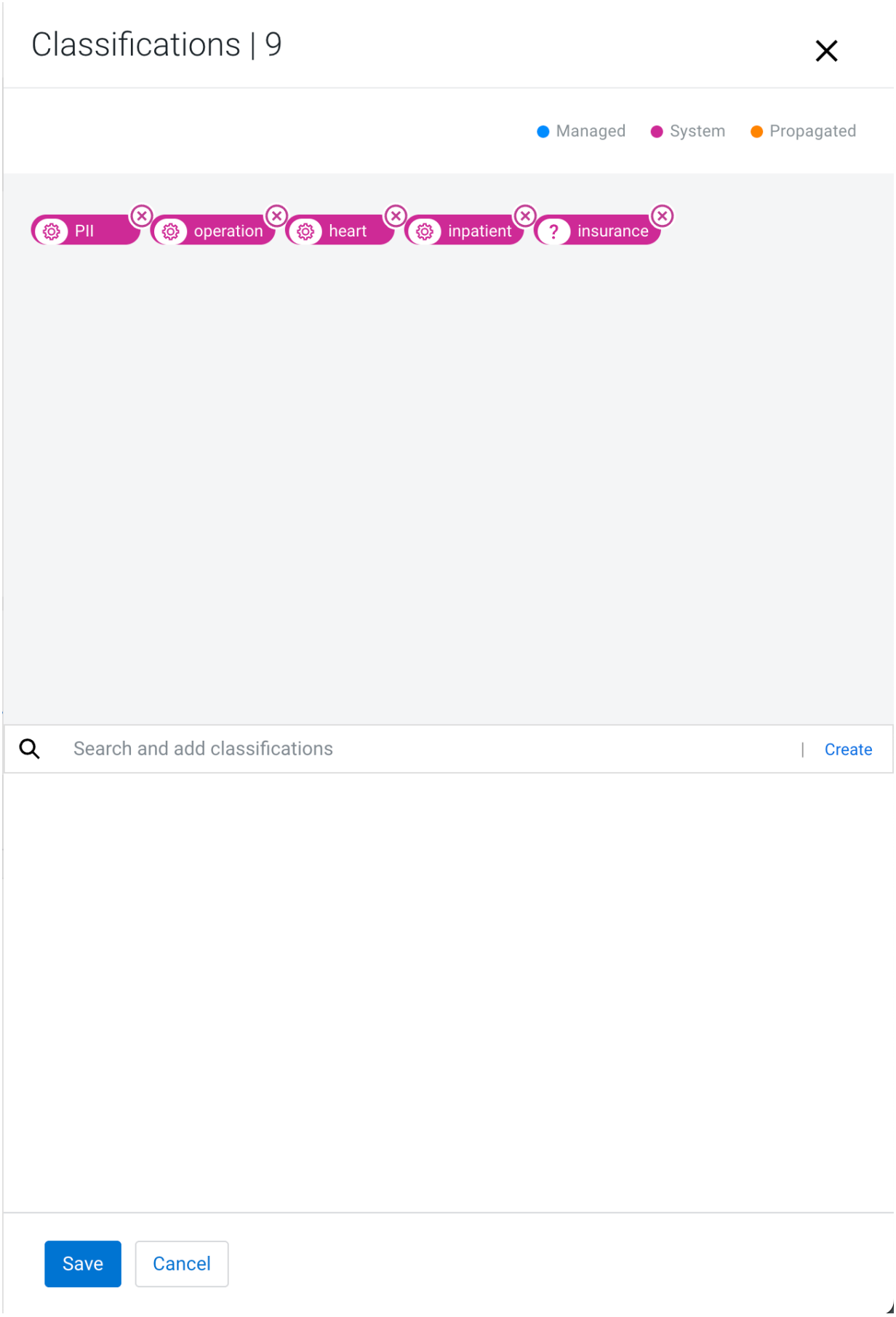
3. Click the asset whose suggested tags you want to review.
4. Click Asset Details Classifications ⋮ Edit .

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

   Approved tags are marked with the ⚙ icon.



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


Classifications | 9
Managed
System
Propagated
PII
operation
heart
inpatient
insurance
Search and add classifications
Create
Save
Cancel

5. Click Save to apply your changes.

Creating tag rules

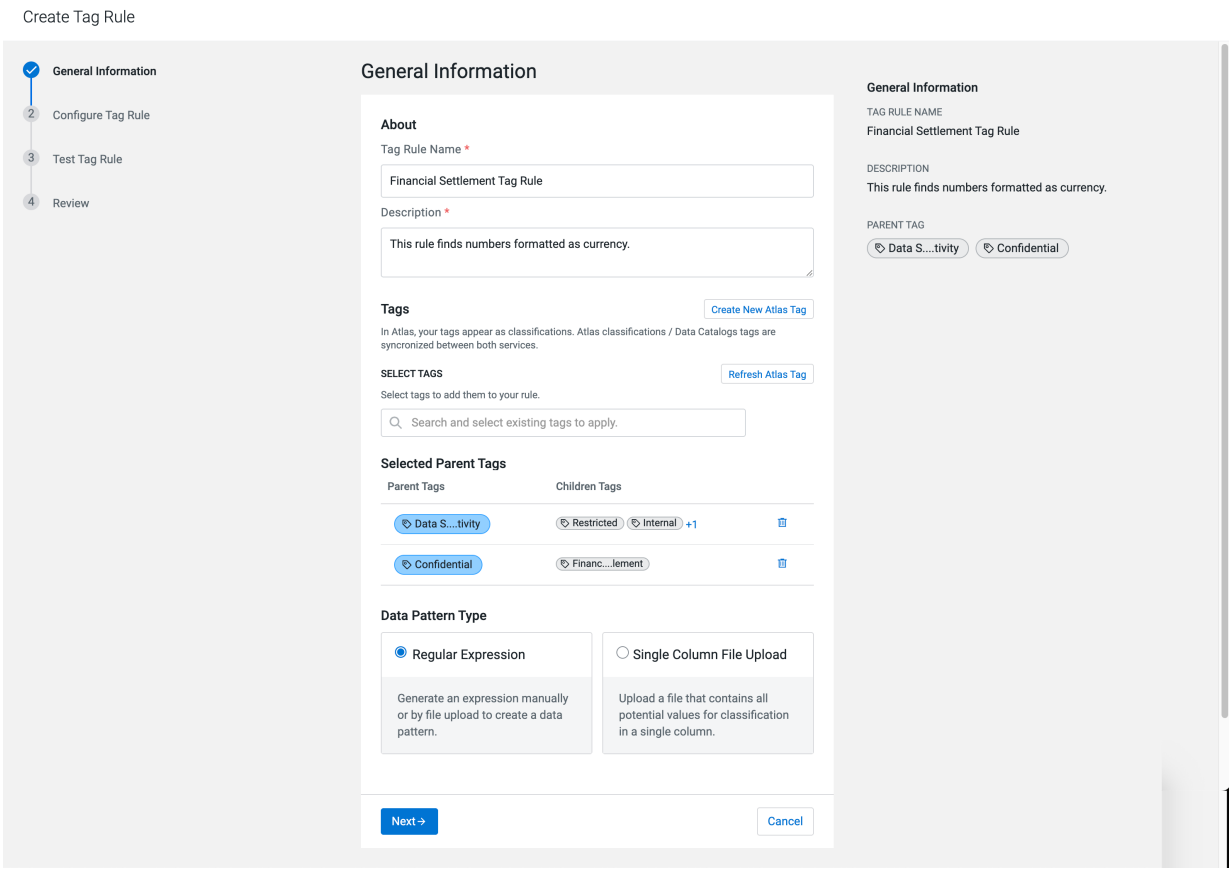
With tag rules, you can apply Apache Atlas classifications to your assets based on regex expressions or similarity to a set of values in a table.

Before you begin

This implementation of the feature is only available for Cloudera Data Catalog on premises 1.5.5 SP2 or later.

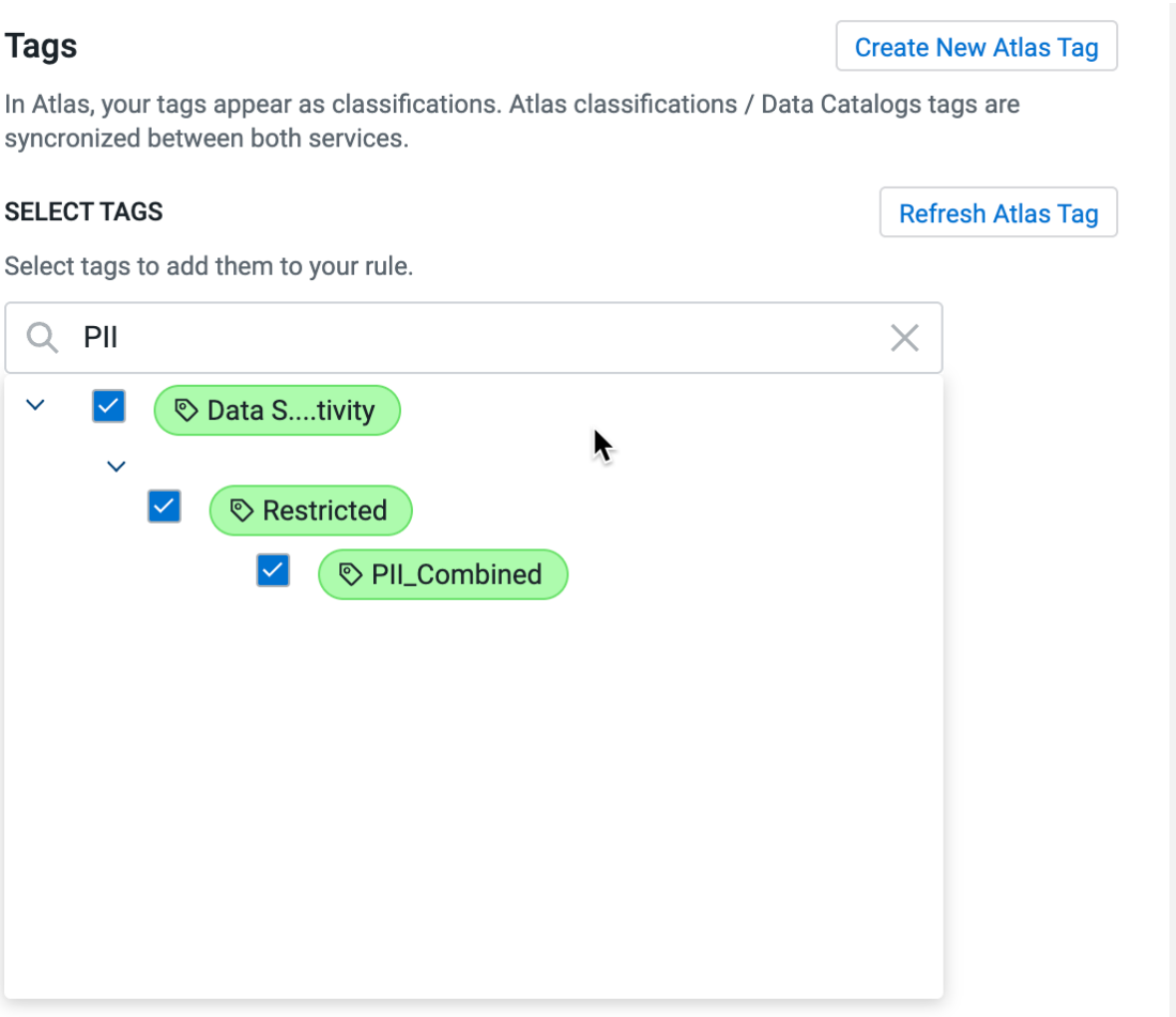
Procedure

1. To start applying tags, go to **Profilers** and select your Data Lake.
2. Go to Profilers Data Compliance Tag Rules .
3. Click + Create Tag Rule.
4. Name your tag rule and add a description to it in **General Information**.



5. Select the tags to be applied from the list of available tags synchronized from the list of Atlas classifications.

   If you select a child tag, its parent tag is also automatically selected. By default, if the child tag is applied to a column, the table receives the parent tag.



6. Select your **Data Pattern Type**:

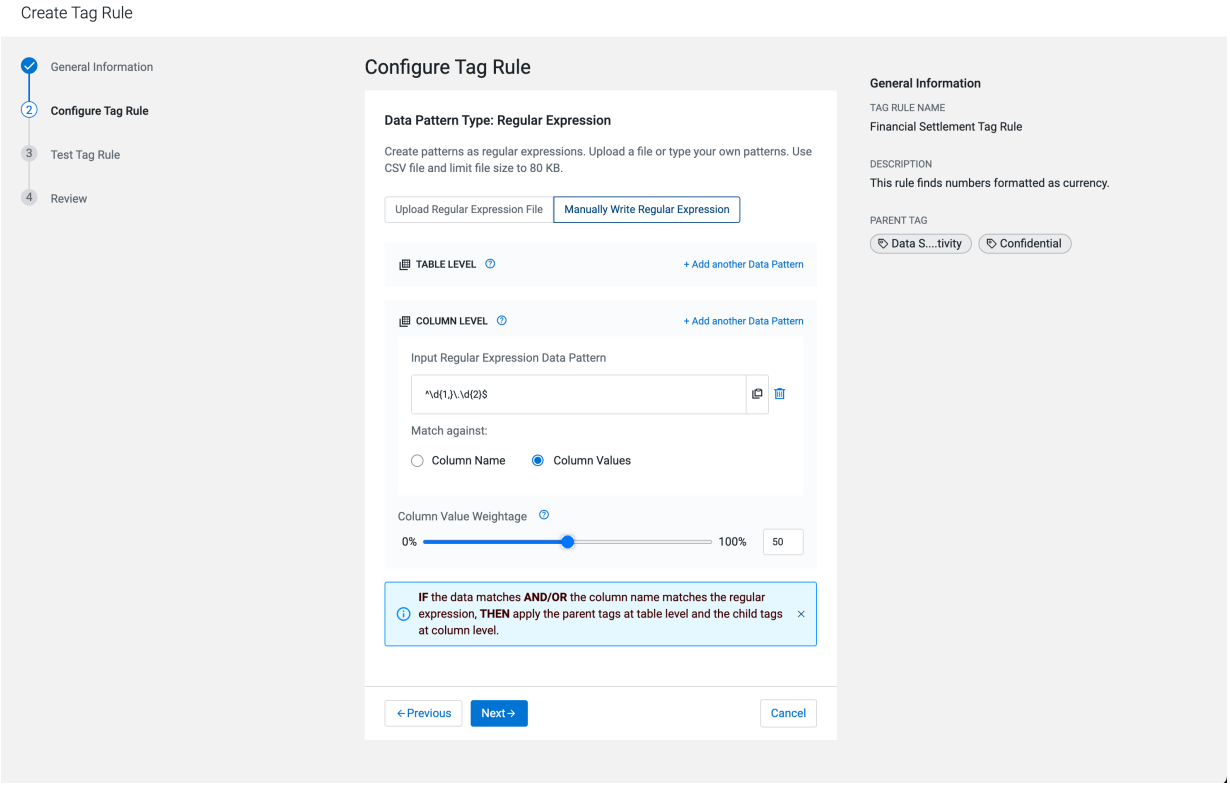
Option	
Regular Expression	You can upload a text file containing your regex expression or directly type it in the **Configure Tag Rule** page. The required format of the CSV file can be seen by clicking Download Sample Tag Rule. Continue in step 7 on page 43.

Single Column File Upload	Upload a CSV file with values to be matched against the actual values in your tables. After uploading your file, continue with step 11 on page 45.
	Use the available template to come up with your own regular expressions, for example:
	`regex matchType threshold`
	`[a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.`
	`[a-zA-Z]{2,} columnValue 50`
	`(?i)email columnName 50`
	`\d{3}-\d{4} columnValue 50`
	`(?i)phone columnName 50`

Creating regular expression based tag rule:

7. Define your regular expression for the table name.

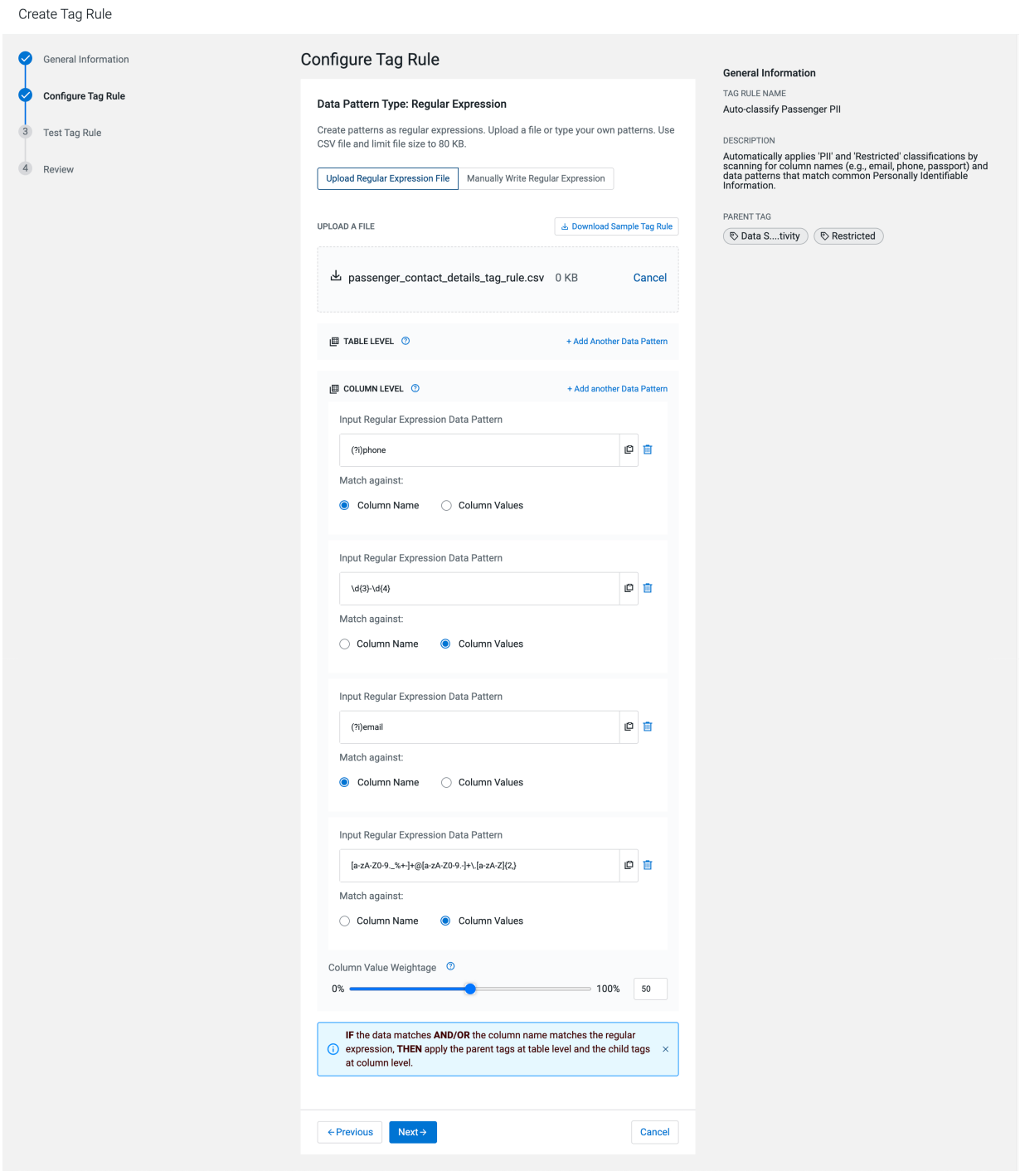
Figure 10: Column level regular expression for matching values



Note: Cloudera recommends using PCRE2 compatible regular expressions. Non-compliant regular expressions may show reduced performance.

For more information, see PCRE - Perl Compatible Regular Expressions.

8. When using **Column Level** regex expressions, you can define multiple expression for both of the following:
   - Column Name
   - Column Values



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

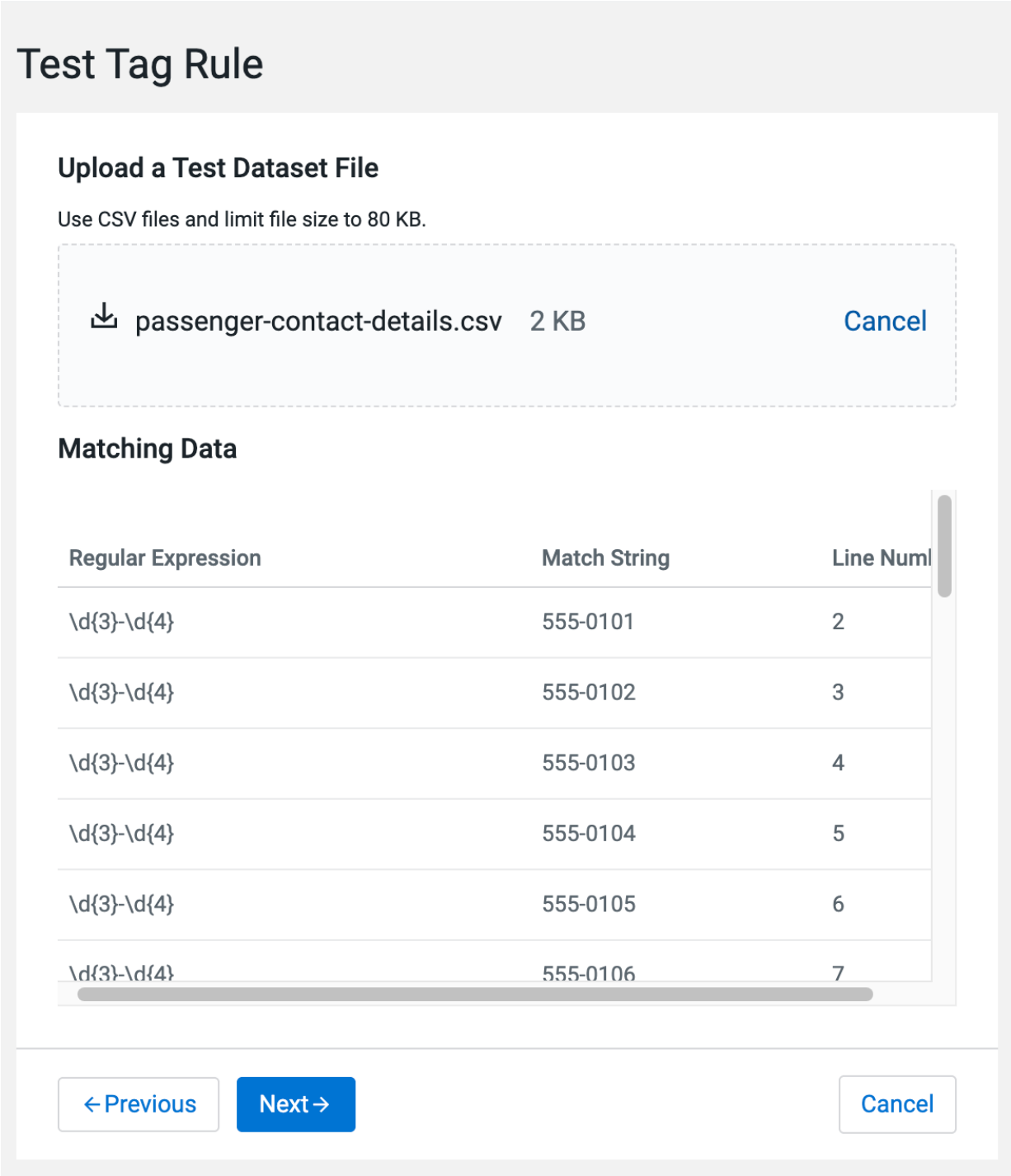
9. Define the Column Value Weightage in percentage with the slider.

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

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

Tag rule testing:

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



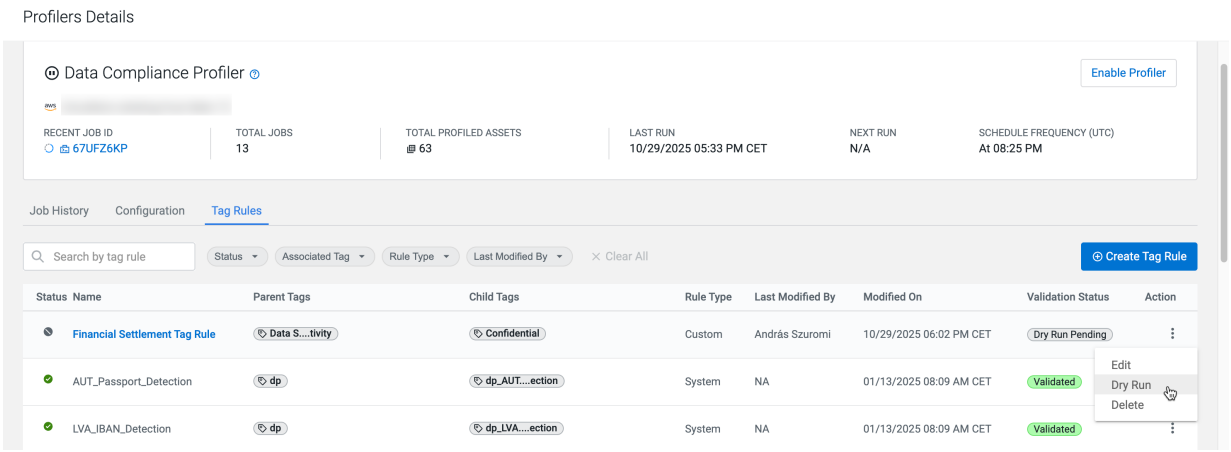
Note: A final test called "Dry Run" is still needed to be passed to enable your tag rule.

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

a) Click Confirm to finalize your tag rule.

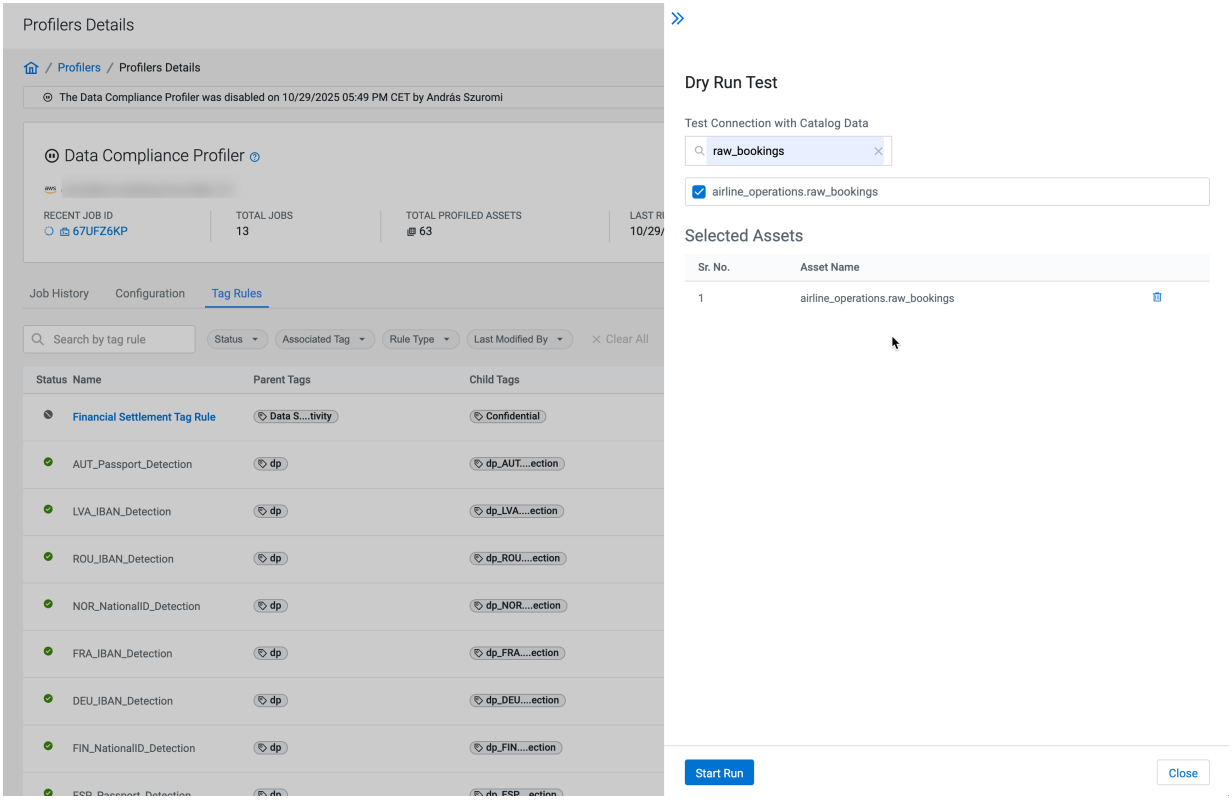
Your tag rule is created with **Status** Disabled( ) and the **Test Status** will be Test Pending.

12. Click ⋮ > Dry Run.



The **Dry Run Test** pane opens.

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



Your tag rule becomes VALIDATED after a successful dry run.

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

Configuring the Hive Column Profiler

In addition to the generic configuration, you can configure additional parameters for the Hive Column Profiler.

Before you begin

Note: This profiler is only available for Cloudera Data Catalog on premises 1.5.5 SP1 or lower versions.

Procedure

1. Go to Profilers Configs .
2. Select Hive Column Profiler.
   The **Detail** page is displayed.
3. Click the 

toggle to enable or disable the profiler.
4. Select a schedule to run the profiler. This is implemented as a Quartz cron expression.
   For more information, see Understanding the Cron Expression generator on page 57.
5. Select Last Run Check and set a period

   Note:

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

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

   If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.
6. Set the sampling configurations. When an asset or table is profiled, instead of scanning the whole table, the profiler sample selects only a subset of records.
   a) Set the Sample Count – The number of times the sampling query runs against the table. Cloudera recommends a value between 3 and 30.
   b) Set the Sample Factor – A decimal value used as the row-selection probability when the profiler queries Hive. For each row, a random number is generated; if it is less than or equal to the Sample Factor, the row is included in the sample. Lower values scan less data (faster); higher values produce a more representative sample. The accepted value range is from 0.001 through 0.5.
   c) Set the Sample Records – The maximum number of rows returned per sampling run. This value maps to the LIMIT clause in the SQL query. The accepted value range is from 100 through 100,000.
7. Continue with the **Pod Configurations** and set the Kubernetes job resources.

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

   - Pod CPU limit – Specifies the maximum number of cores that can be allocated to a pod. The value range is from 1 through 8.
   - Pod CPU Requirements – Specifies the minimum number of CPUs that will be allocated to a pod when it is provisioned. If the node where a pod is running has enough resources available, a container can use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The value range is from 1 through 8.
   - Pod Memory limit – Specifies the maximum amount of memory that can be allocated to a pod. The value range is from 1 through 256.
   - Pod Memory Requirements – Specifies the minimum amount of RAM that will be allocated to a pod when it is provisioned. If the node where a pod is running has enough resources available, a container can use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The value range is from 1 through 256.

8. Update the **Executor Configurations**.

   Executor configurations specify the runtime configuration. These configurations must be changed if you are changing the pod configurations and when you require additional compute power.

   - **Number of workers** – Specifies the number of processes that are used by the distributed computing framework. The value range is from 1 through 8.
   - **Number of threads per worker** – Specifies the number of threads used by each worker to complete the job. The value range is from 1 through 8.
   - **Worker Memory limit in GB** – Enforces a memory usage threshold for a given worker to prevent memory overutilization. For example, if you have an 8 GB pod and 4 threads, the value of this parameter must be 2 GB. The value range is from 1 through 4.

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

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

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

      1. Select the **Deny List** or **Allow List** tab.
      2. Click Add New to define new rules.
      3. Select one of the following keys from the drop-down list:

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

      4. Select the operator from the drop-down list. Depending on the keys selected, you can select an operator such as equals, contains. For example, you can select the name of assets that contain a particular string.
      5. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
      6. Click Done. Once the rule is added, you can enable it or disable it as needed by clicking the state toggle.

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

Related Information

The Hive Column Profiler

Understanding the Cron Expression generator

Configuring the Statistics Collector profiler

Configure additional parameters for the Statistics Collector to optimize profiling tasks. Adjust settings such as scheduling, incremental profiling, and resource allocation to enhance performance and accuracy.

Before you begin



Note: This profiler is only available for Cloudera Data Catalog on premises 1.5.5 SP2 or higher versions.

Procedure

1. Go to **Profilers** and select your Data Lake.
2. Go to Profilers Statistics Collector Profiler Details Configuration All Configurations .
3. Select a schedule to run profiler using either UNIX Cron Expression or the Basic scheduler.

Note: Both the Basic and Cron Expression scheduler use the UTC timezone instead of the local timezone of the user.

Figure 11: Profiler schedule with cron expression

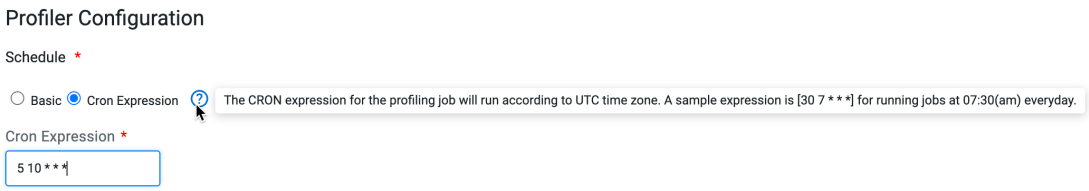
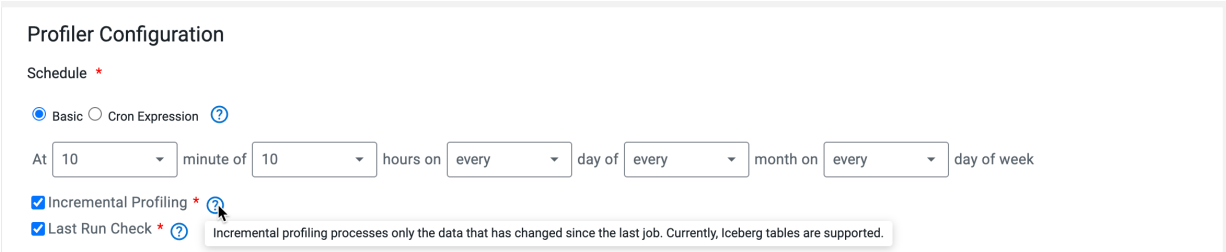


Figure 12: Profiler schedule with natural language



4. Select Incremental Profiling when needed.

   Using Incremental Profiling can decrease the compute resources and the time needed for the profiling job by processing only the Iceberg tables updated or added since the previous job.

   With Incremental Profiling, you can refine the results from the Last Run Check. Incremental Profiling checks the data (rows) in assets, while Last Run Check filters complete assets.

   Note: By Statistics Collector Profilers, the profiler compares the aggregated metrics between old and newly added data. Depending on the differences, this can slightly skew results. Cloudera recommends processing the entire dataset from time to time for the most accurate results.

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

   Note:

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

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

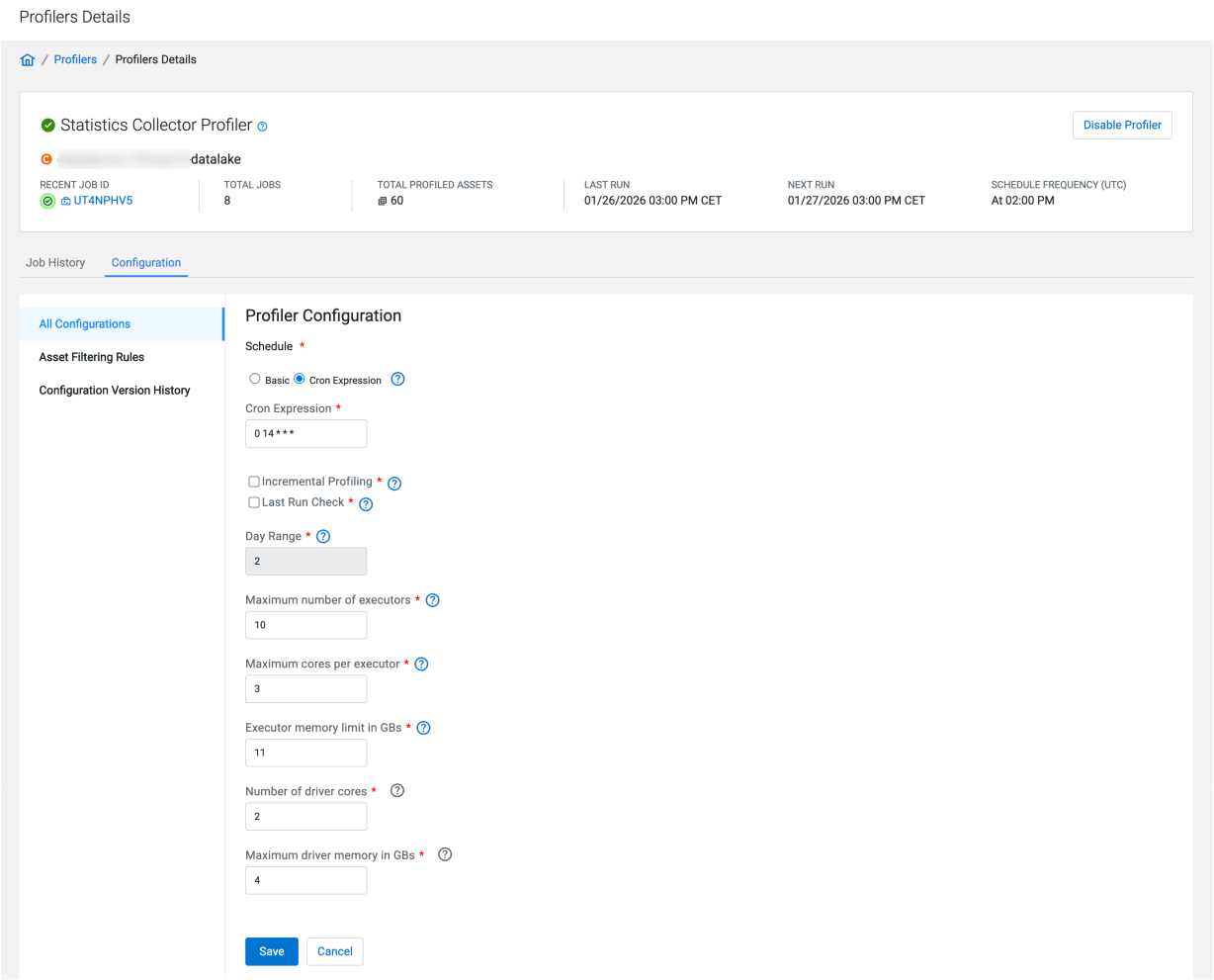
   If the Last Run Check configuration is disabled, assets will be picked up for profiling as per the job cron schedule, honoring the asset filter rules.

   The Last Run Check precedes Incremental Profiling.

6. Review sampling behavior.

   When an asset or table is profiled, instead of scanning the whole table, the profiler selects a subset of records. Sampling is automatic: the profiler loads data from HDFS and determines the sample size based on table size.

7. Configure the resources.



Note: Is your infrastructure is deployed completely in AWS, profilers might hang if the underlying AWS cloud provider environment cannot provide the necessary memory for the executor instances. In this case, reconfigure your executors with 4-5 GB memory in Profiler Details Configuration .

a) Set the Maximum number of executors.

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

b) Set the Maximum cores per executor.

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

c) Set the Executor memory limit in GBs.

d) Set the Number of driver cores.

Specifies the maximum number of driver cores. Increase the number of cores to improve the speed of profiler job scheduling.

e) Set the Maximum driver memory in GBs.

Specifies the maximum amount of memory that can be allocated to an driver core. Increasing the available memory accelerates the profiling of larger and more complex tables and prevents out-of-memory errors.

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

- Profiler configurations apply to both scheduled and on-demand profiler jobs.

- Asset filtering rules apply to assets, such as tables, and not to complete databases. This means that filtering rules work at table level and not at a database level. You are able to filter out all tables of a specific database, but each table will be check whether its parent is the selected database.
- Multiple asset filtering rules are evaluated together as if connected by the OR operator.
- In Compute Cluster environments, you cannot enable conflicting Allow and Deny list rules at the same time. Enabling conflicting rules results in the following error message:



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

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

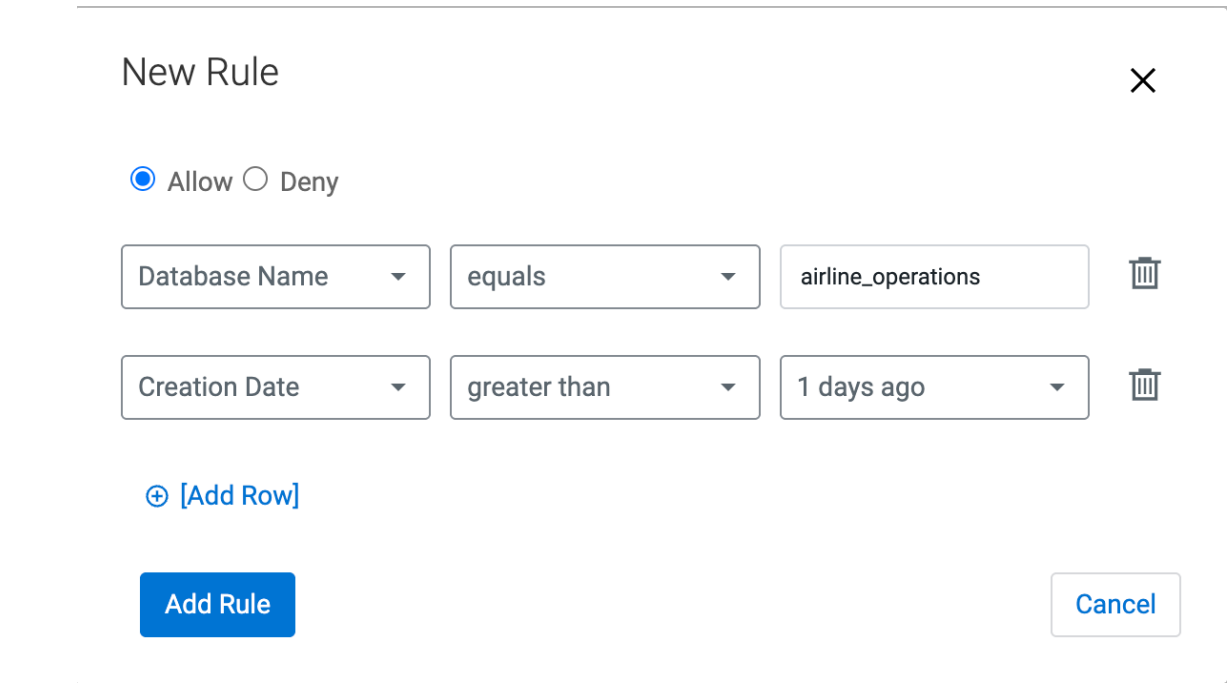
1. Click Add New Rule to define new rules.
2. Use the radio buttons to define your new rule for the Allow List or Deny List.
3. Select one of the following keys from the drop-down list and select the relevant operator:

Key	Operator
Database name	• equals • starts with • ends with
Name (of asset)	• equals • contains • starts with • ends with
Owner (of asset)	
Creation date[4]	• greater than • less than

Note: **Name** refers here to the actual name of the asset and not to its **Qualified Name**.

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

[4] By Creation Date, Greater than 7 days means an asset older than seven days. Less than 7 days means an asset younger than seven days.



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

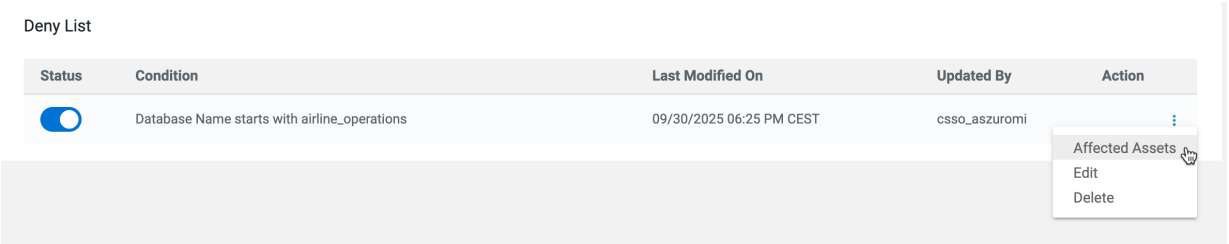
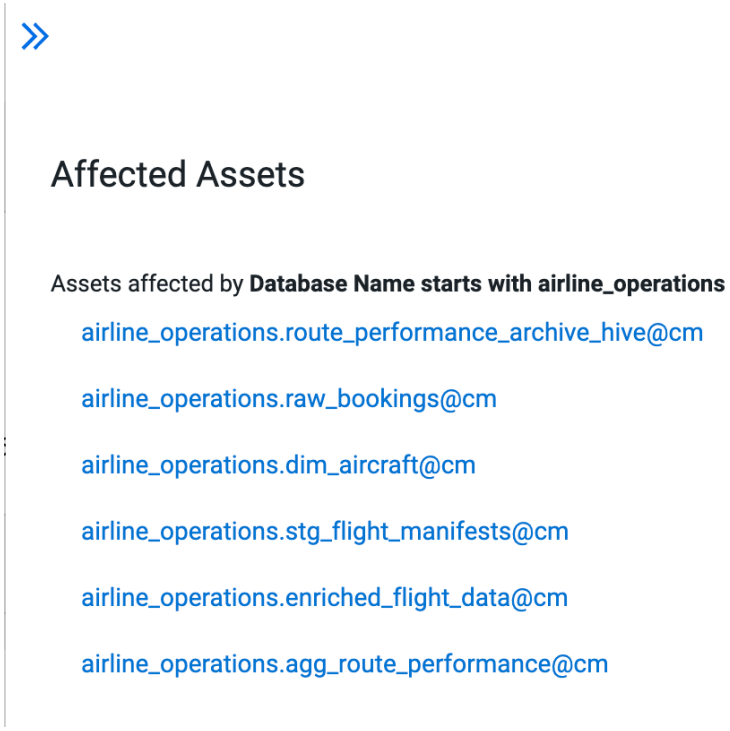
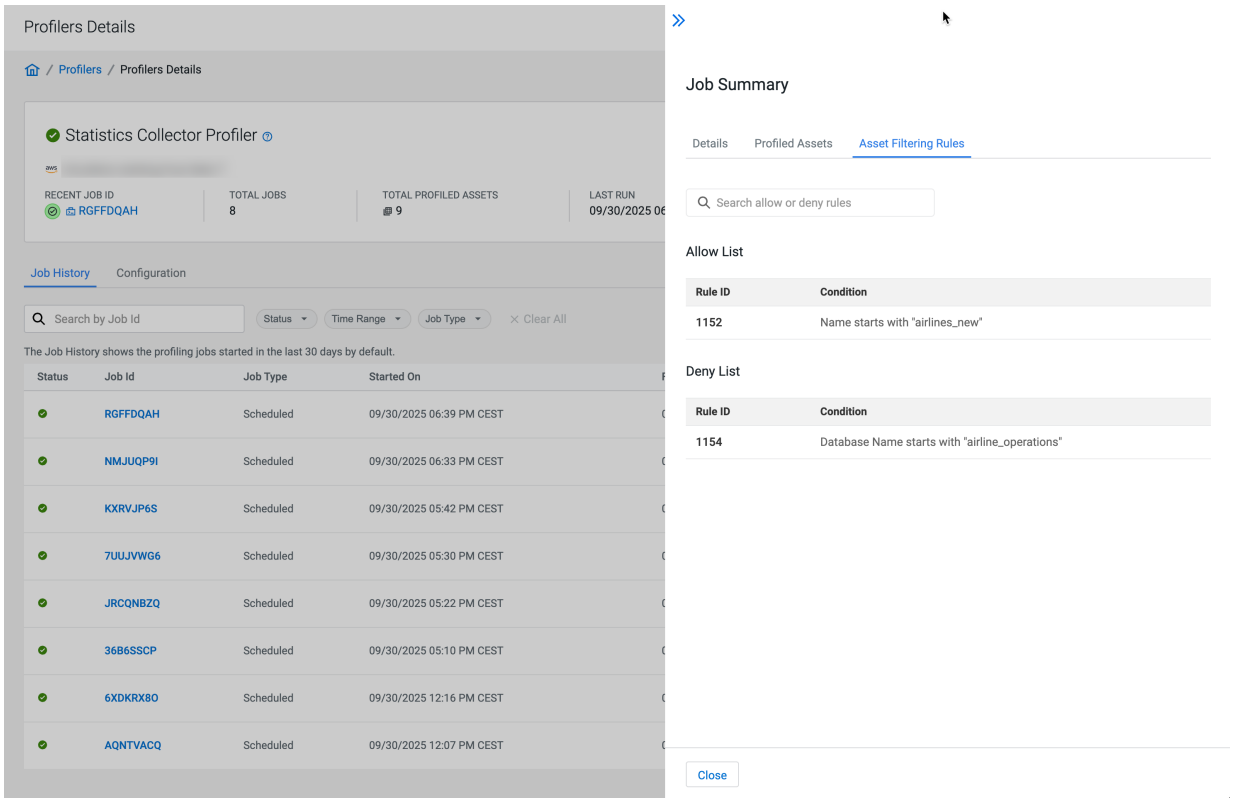


Figure 13: Affected Assets in Asset Filtering Rules configuration



Job Summary shows the asset filtering rules applied for the particular profiling job.



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

Related Information

The Statistics Collector Profiler

Configuring the Ranger Audit Profiler

In addition to the generic configuration, you can configure scheduling and available resources for the Ranger Audit Profiler.

Before you begin

Note: This profiler is only available for Cloudera Data Catalog on premises 1.5.5 SP1 or lower versions.

Procedure

1. Go to **Profilers** and select your data lake.
2. Go to Profilers Configs .
3. Select Ranger Audit Profiler.
   The **Detail** page is displayed.
4. Click the 

toggle to enable or disable the profiler.
5. Select a schedule to run the profiler using a Quartz cron expression.
6. Set the **Input block size**.

   Note: The value range is 100,000 to 10,000,000.

7. Continue with the **Pod Configurations** and set the Kubernetes job resources.

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

   - **Pod CPU Limit** – Specifies the maximum number of cores that can be allocated to a Pod. The accepted values range from one through eight.
   - **Pod CPU Requirement** – Specifies the minimum number of CPUs that will be allocated to a pod when it is provisioned. If the node where a pod is running has enough resources available, r a container can use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The value range is from 1 through 8.
   - **Pod Memory Limit** – Specifies the maximum amount of memory that can be allocated to a pod. The value range is from 1 through 256.
   - **Pod Memory Requirement** – Specifies the minimum amount of RAM that will be allocated to a pod when it is provisioned. If the node where a pod is running has enough resources available, a container can use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit. The value range is from 1 through 256.

8. Update the **Executor Configurations**.

   Executor configurations specify the runtime configurations. These configurations must be changed if you are changing the pod configurations and when you require additional compute power.

   a.
   - **Number of workers** – Specifies the number of processes that are used by the distributed computing framework. The value range is from 1 through 8.
   - **Number of threads per worker** – Specifies the number of threads used by each worker to complete the job. The value range is from 1 through 8.

- **Worker Memory limit in GB** – Enforces a memory usage threshold for a given worker. For example, if you have an 8 GB pod and 4 threads, the value of this parameter must be 2 GB. The value range is from 1 through 4.

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

Related Information

The Ranger Audit Profiler

Understanding the Cron Expression generator

Configuring the Activity Profiler

In addition to the generic configuration, you can configure scheduling and available resources for the Activity Profiler.

Before you begin

Note: This profiler is only available for Cloudera Data Catalog on premises 1.5.5 SP2 or higher versions.

Procedure

1. Go to **Profilers** and select your data lake.
2. Go to Profilers Activity Profiler Profiler Details Configuration All Configurations
3. Select a schedule to run profiler using either UNIX Cron Expression or the Basic scheduler.

Note: Both the Basic and Cron Expression scheduler (Unix in Compute Cluster enabled environments cron jobs) use the UTC timezone instead of the local timezone of the user.

Figure 14: Profiler schedule with cron expression

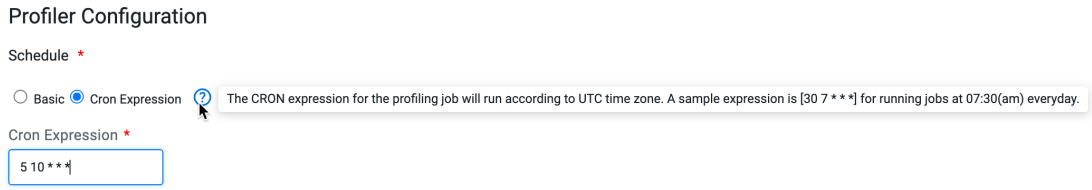
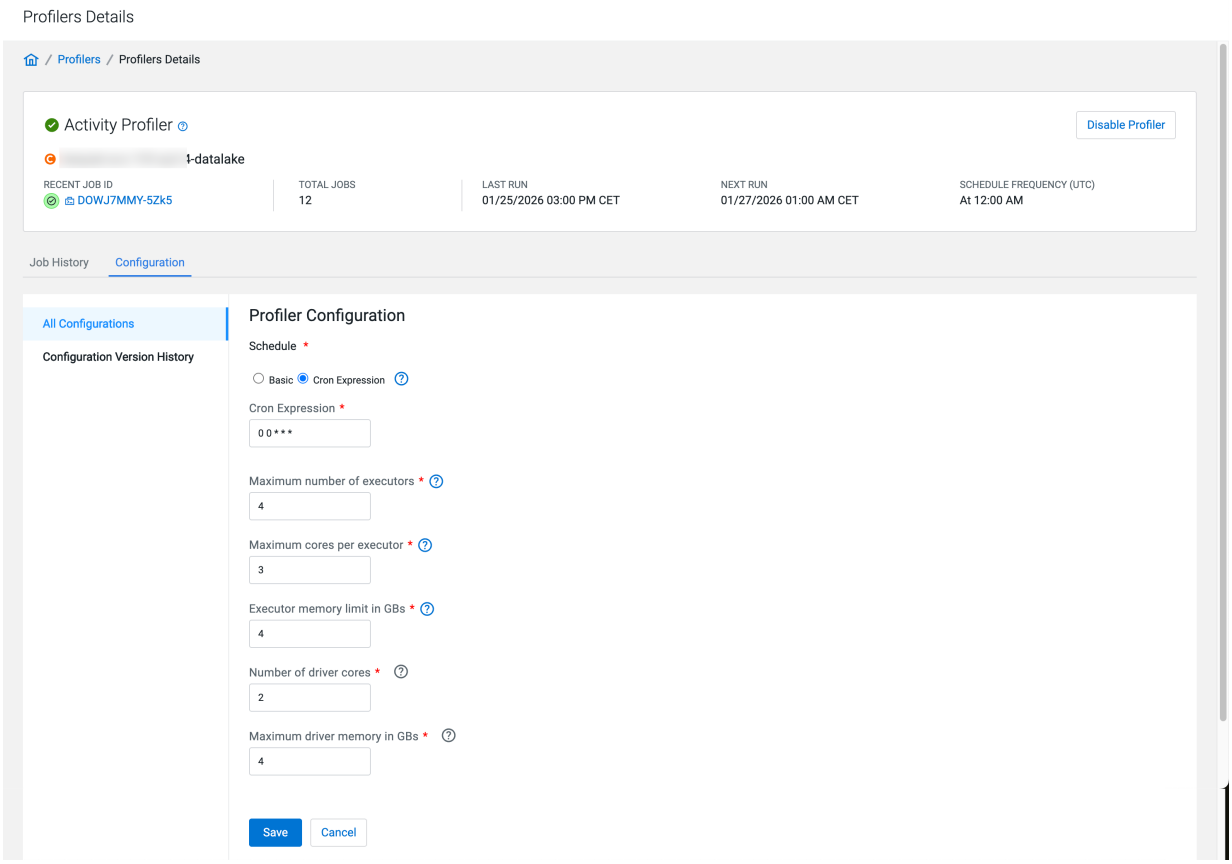


Figure 15: Profiler schedule with natural language



Note: Only the metadata events of the last four hours are processed for the first run. After the first run, all events between runs are processed.

4. Configure the resources.

   Note: Is your infrastructure is deployed completely in AWS, profilers might hang if the underlying AWS cloud provider environment cannot provide the necessary memory for the executor instances. In this case, reconfigure your executors with 4-5 GB memory in Profiler Details Configuration .

   a) Set the Maximum number of executors.

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

   b) Set the Maximum cores per executor.

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

   c) Set the Executor memory limit in GBs.

   d) Set the Number of driver cores.

      Specifies the maximum number of driver cores. Increase the number of cores to improve the speed of profiler job scheduling.

   e) Set the Maximum driver memory in GBs.

      Specifies the maximum amount of memory that can be allocated to an driver core. Increasing the available memory accelerates the profiling of larger and more complex tables and prevents out-of-memory errors.

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

Related Information

The Activity Profiler

Understanding the Cron Expression generator

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

The Unix cron expression uses the following typical format:

Each * in the cron represents a unique value.

Cron Expression: 0 18 * * *

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

For example, consider a cron with the following values:

CRON Expression: 30 10 15 5 *

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

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

Consider another example:

30 10 * 5 7

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

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

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

Related Information

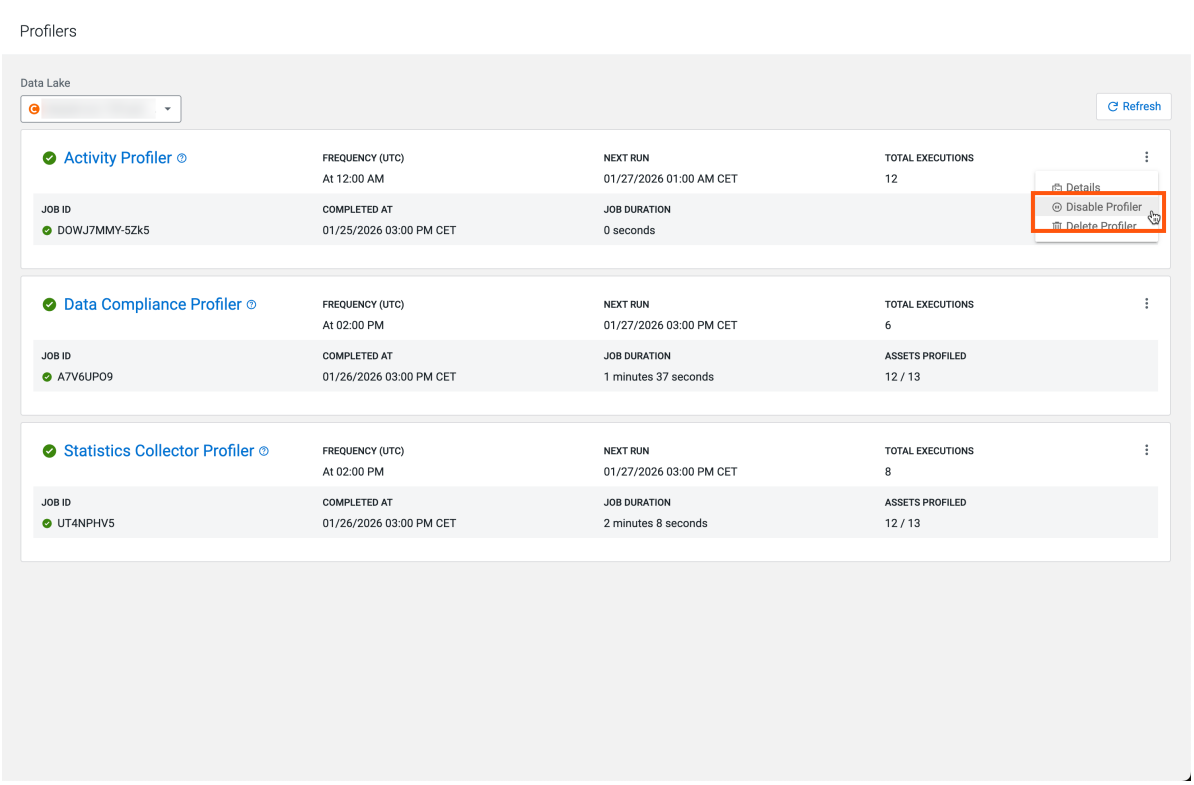
The Open Group Base Specifications Issue 8: crontab

Enabling or disabling profilers

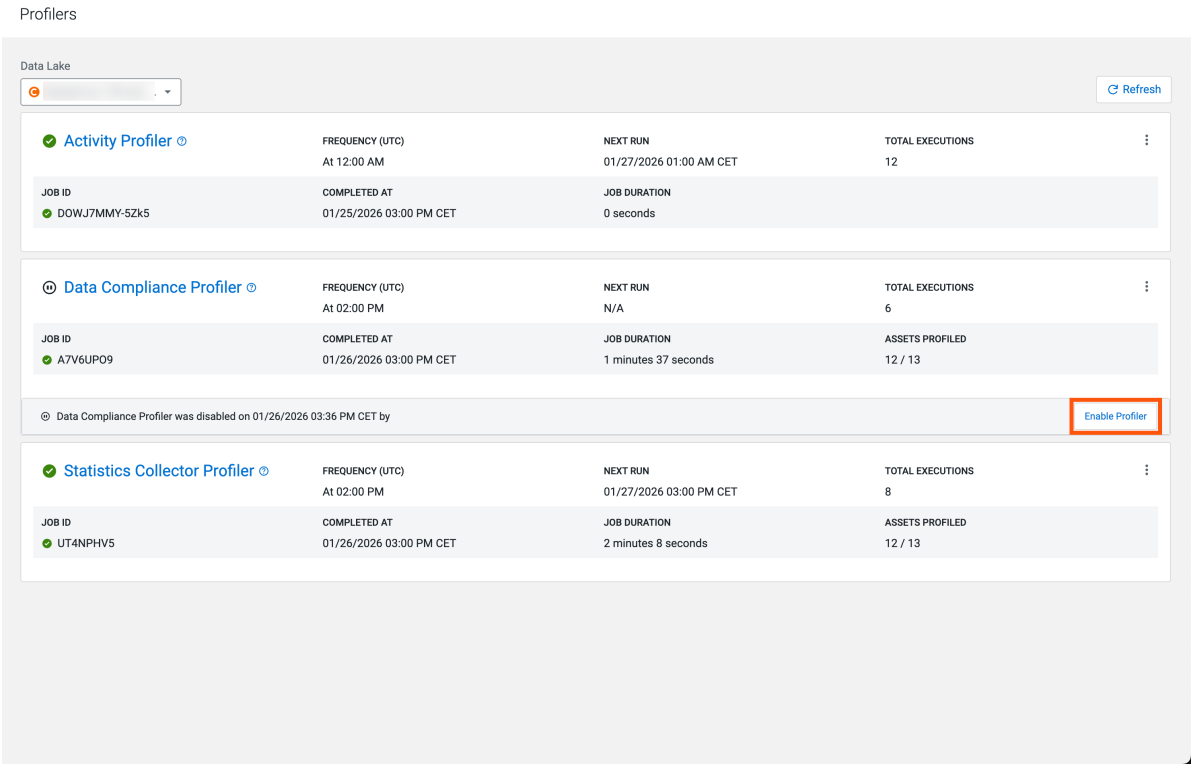
Learn how to enable or disable profilers in Cloudera Data Catalog to manage resource usage effectively.

For After Cloudera Data Catalog 1.5.5 SP2

1. Go to **Profilers**.
2. Click ⋮ > Disable Profiler.



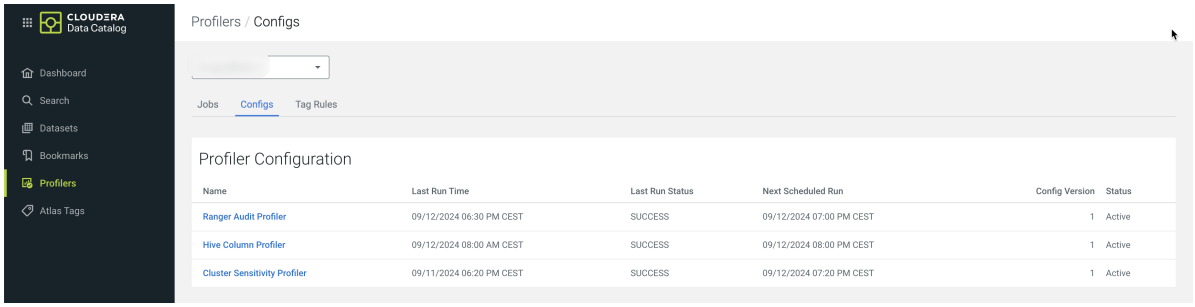
3. Click Confirm.
4. You can click Enable Profiler to continue using it.



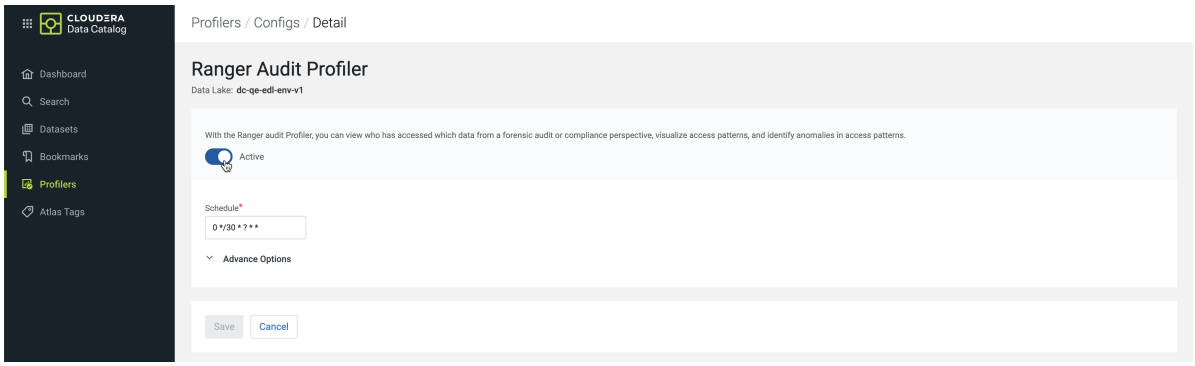
Note: Disabling a profiler suspends its Kubernetes CRON job, so no new profiling runs are scheduled until you enable it again. The profiler keeps its schedule and configuration while disabled, and no profiler data or previously collected profiling results are deleted. Disabling a profiler is not the same as deleting it; enabling restores scheduled profiling with all settings and history intact.

For Before Cloudera Data Catalog 1.5.5 SP2

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



3. Switch the toggle to the desired state.



Note: Disabling a profiler only pauses its scheduling; the profiler remains visible, and its configuration, Cluster Sensitivity Profiler rules, asset selectors, schedule, and profiling history are retained. Enabling the profiler resumes profiling with all settings intact. Disabling a profiler is not the same as deleting it.