DSS Administration

Date of Publish: 2018-05-16

http://docs.hortonworks.com

Contents

Managing Asset Collections

You can create, edit, and delete Asset Collections and delete Asset Collections.

Create Asset Collections

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

Procedure

1. From the **Asset Collection** page, click **Add Asset Collection**.

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

Field Name	Description	Example Values
Name	Enter an appropriate Asset Collection name. This name cannot be duplicated across the system. (Mandatory)	Customer Profiles, Sales Assets, Financials
Description	Describe the purpose or intent of the Asset Collection. (Mandatory)	Contains customer profiles: data assets for US and WW.
Datalake	Assign the Asset Collection to one Datalake. Choose from a list of available Datalakes. (Mandatory)	dss_bbsh_clust3
Tags	Add tags to your asset collection 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 asset collection. Select private if only you want to have access to this asset collection. **Note:** You can later change the status of the asset collection. Click the lock icon on the Asset Collection Details page to change the access state of the asset collection.	

3. Click **Next**.

   The Asset Collection Details page appears for the new asset collection.
4. Click **Add Assets** to add related data assets into your asset collection.

   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 asset collection.

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 asset collection and the Search page is refreshed.
8. Close the Search tab.

   The **Asset Collection Details** page appears.
9. Click **Save**.

Edit Asset Collections

You can edit asset collections by adding or removing assets and changing the access state of the asset collection.

Procedure

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

Delete Asset Collections

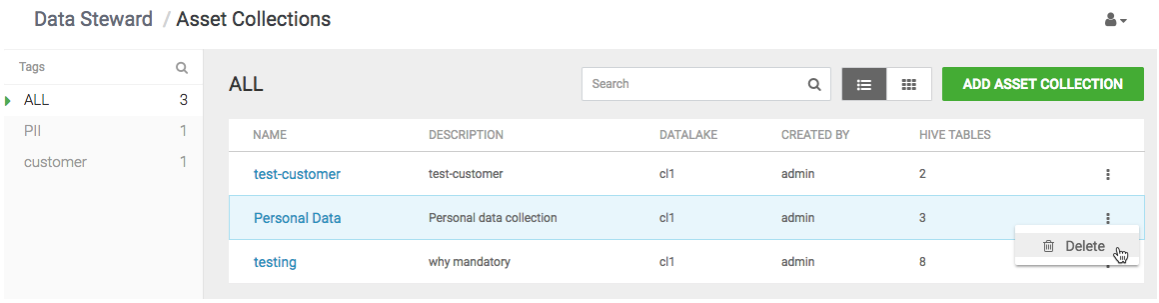
You might want to delete an Asset Collection if you no longer need to track those assets in that collection, or if you want to reassign those assets to another collection. You can delete Asset Collections at any time. Deleting an Asset Collection does not delete the assets contained therein, it only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

Procedure

1. From **Data Steward** > **Asset Collections** page, click the **More Options** icon

   ( )

   beside the name of the Asset Collection you want to delete.
2. Click **Delete**:



3. Click **Confirm**.
   You are returned to the **Asset Collections** home page.

Collaborate with other users

Data Stewards can collaborate and share insights with other users in the enterprise regarding various asset collections.

As a data steward, you can rate asset collections and view the average rating of an asset collection. This can help other users to find asset collections 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 the asset collection page, you can see additional details about the asset collection. 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 asset collection is bookmarked by the current user or not.

You can perform the following collaboration actions for each asset collection.

Like an asset collection

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

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

Comment and discuss about an asset collection

You might want to share your knowledge or insights about this asset collection with other users. Data Steward Studio allows you to collaborate with other users by adding comments.

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

Bookmark the asset collection

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

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

Rate the asset collection

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

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

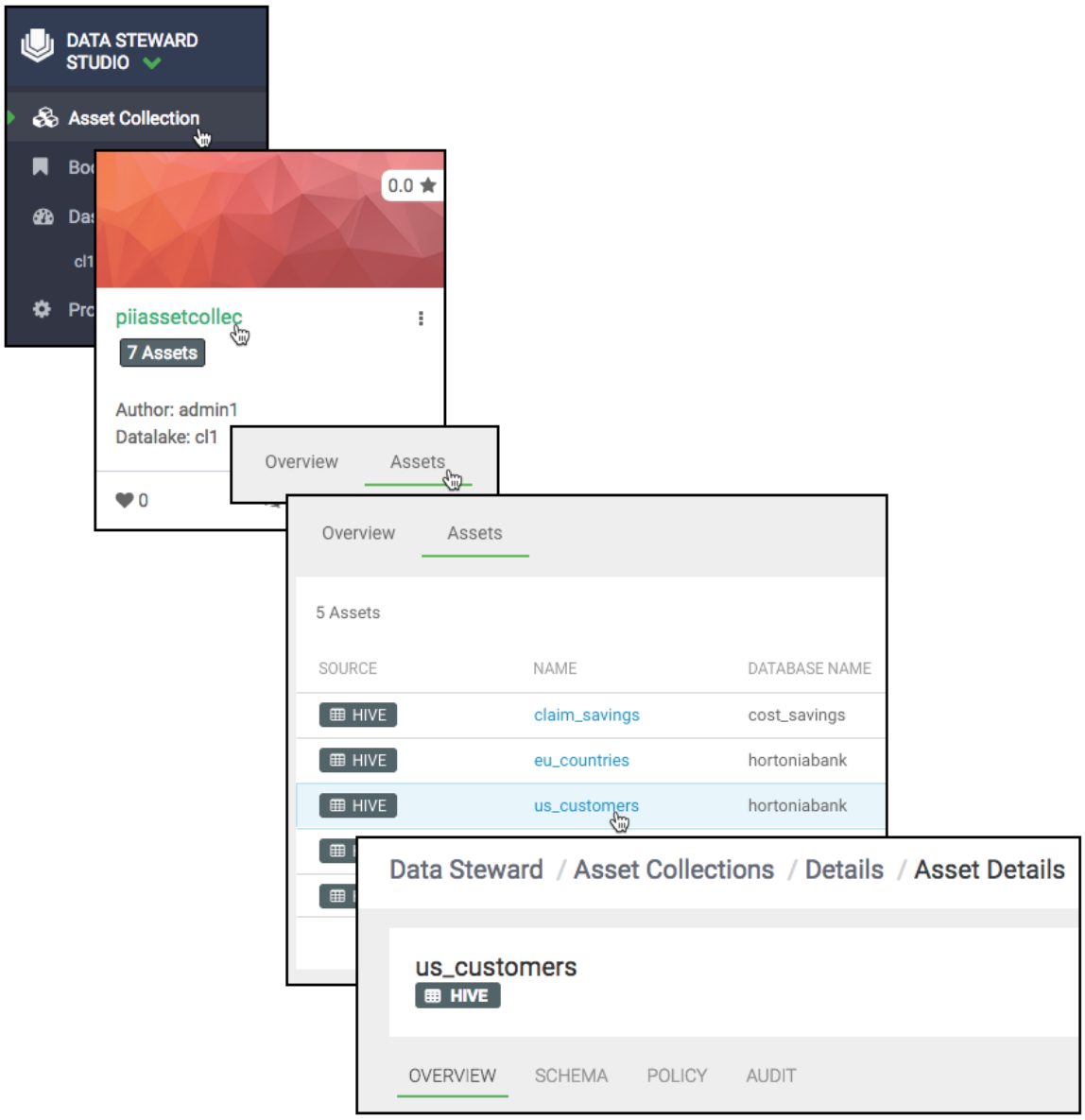
View the tags of an asset collection

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

Viewing Data Asset Details (Asset 360)

The Asset 360 page comprises four tabs (Overview, Schema, Policy, and Audit). These tabs contain dashboards that provide an overview of your asset collection.

The Asset 360 page can be accessed from **Asset Collection** > **Select one asset collection** > **Assets** > **Select one data asset**. This brings you to the Overview tab, the first of the four tabs that form the Asset 360 page.



- Overview: Displays an overview for the data asset.
  - Table properties: Number of rows, number of columns, sensitive columns, number of partitions, owner, tags, profilers

    - 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.
    - Users: Displays top 10 users for the data asset.
    - Access types: By action and operation type.
- Schema: Displays the schema of the data asset for structured data (such as Hive tables) from the relevant metadata repositories (such as Atlas). You can also view the shape or distribution characteristics of the columnar data within a schema based on the Hive column profiler.
- 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 both most recent access audits from Apache Ranger and also summarized views of audits by type, user, and time window based on profiling of audit data.

View Data Asset Overview

Asset 360 > **Overview** displays all the Apache Atlas metadata associated with a particular data asset.

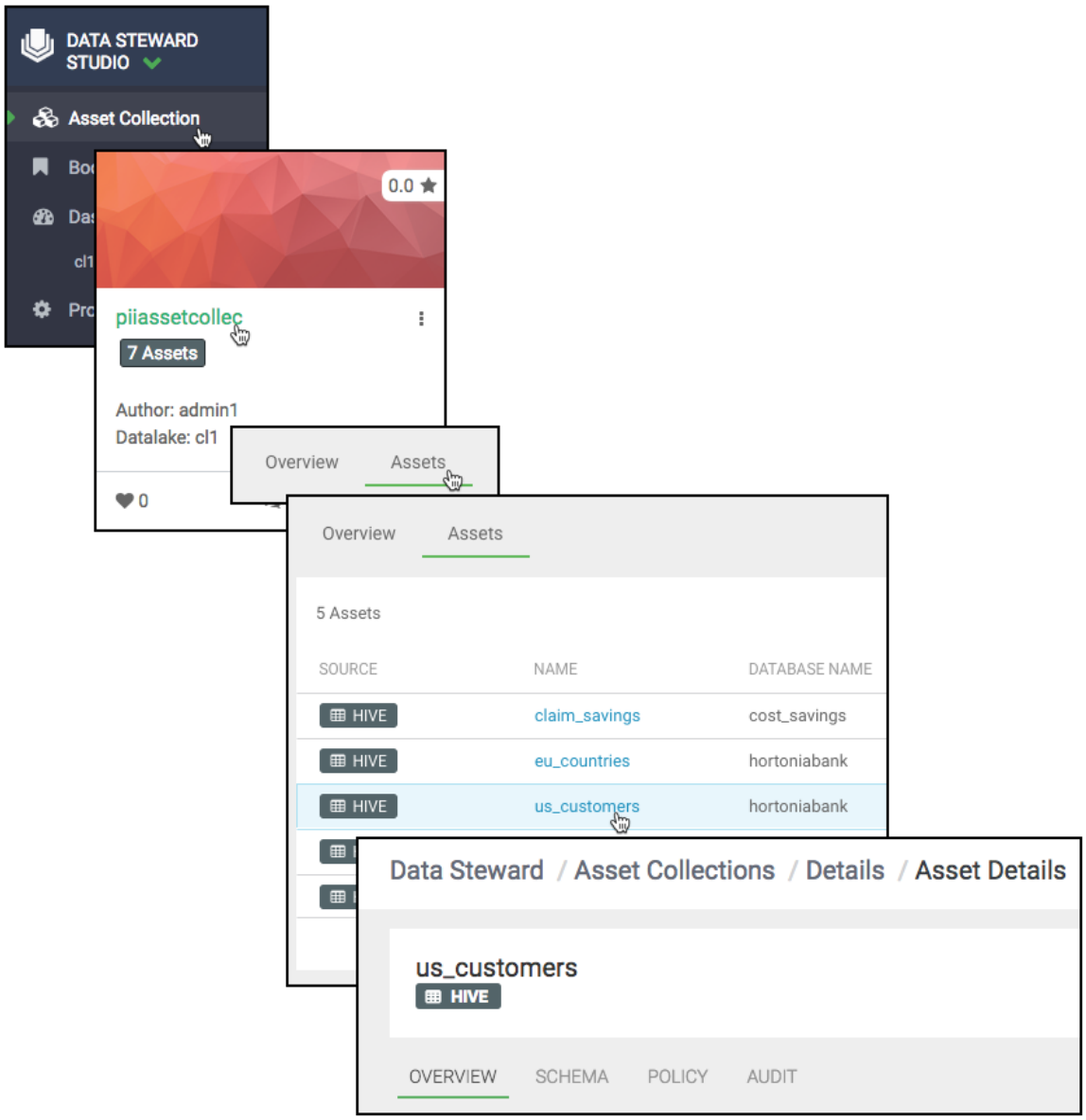
About this task

The Data Asset Overview page shows:

- Table properties: Number of rows, number of columns, sensitive columns, number of partitions, owner.
- System Tags: Displays tags associated with your asset to help with cataloging, searching, and retrieving.
- Profilers: Shows the status of profilers: active/inactive and time last run.
- 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.
- Users: Displays top 10 users for the data asset.
- Access types: By action and operation type.

Procedure

From Data Steward, click: **Asset Collection** > **Select one asset collection** > **Assets** > **Select one data asset**:



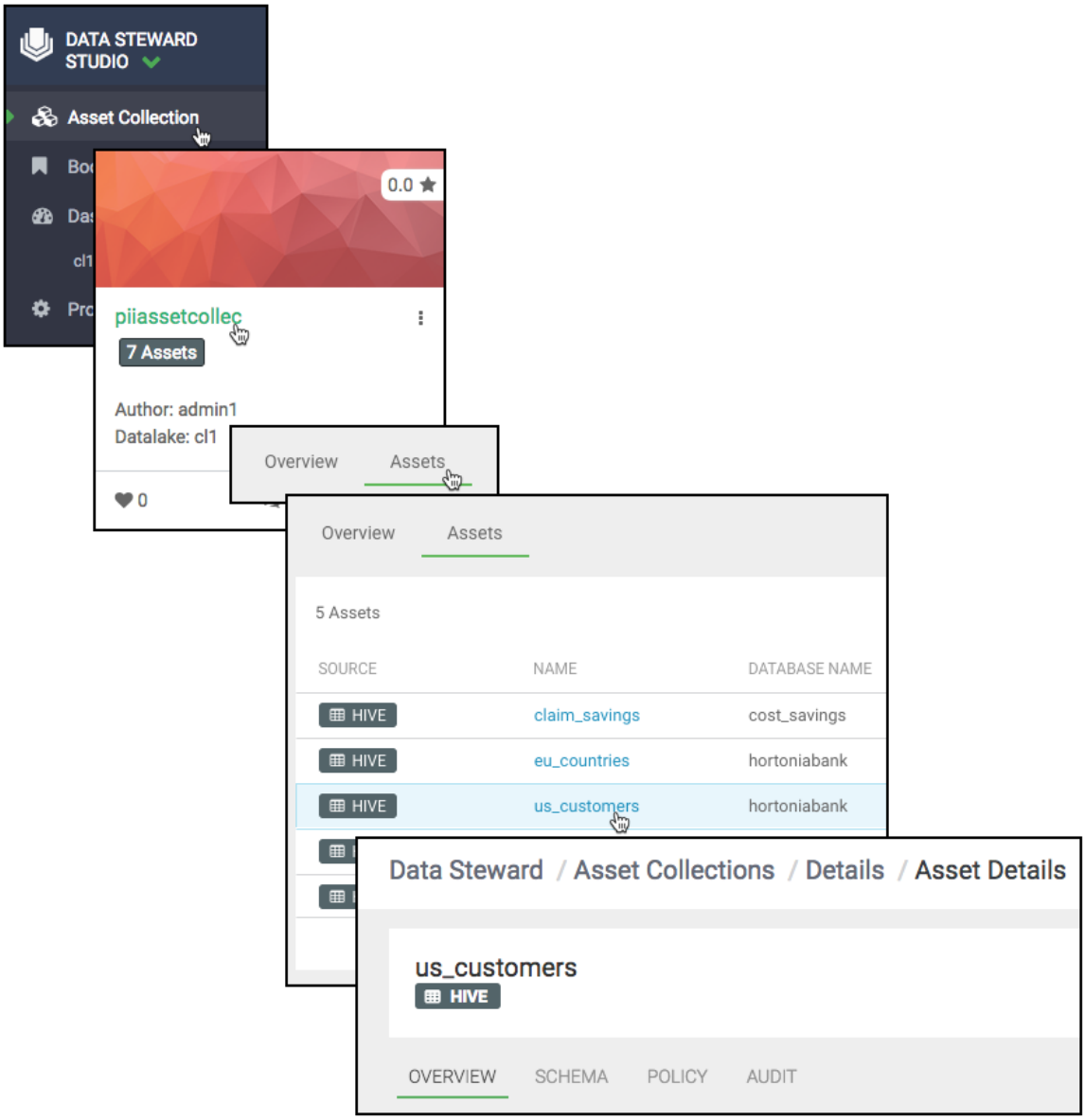
The Asset **Overview** window opens.

View Data Asset Schema

From **Asset 360** > **Schema**, you can view the schema of the data asset for structured data (such as Hive tables) from the relevant metadata repositories (such as Atlas).

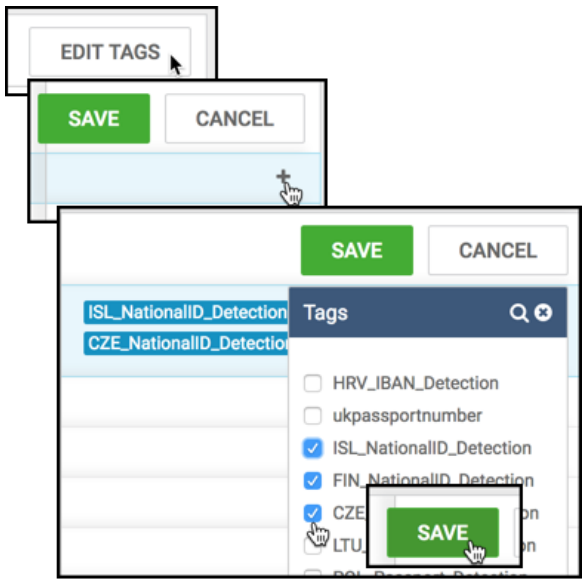
Procedure

1. From Data Steward, click: **Asset Collection** > **Select one asset collection** > **Assets** > **Select one data 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**.

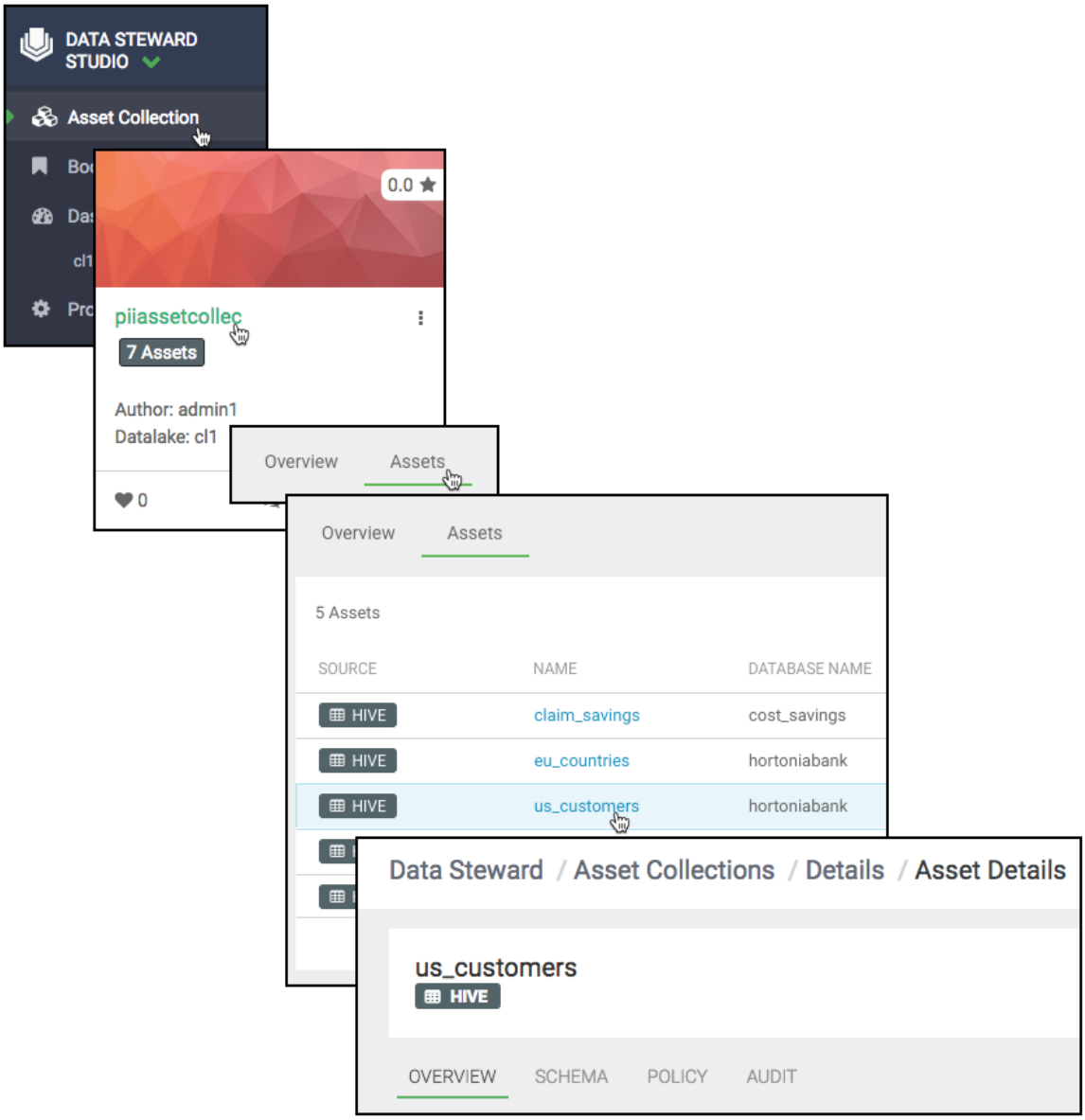


View Authorization Policies on a Data Asset

Asset 360 > **Policy** 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).

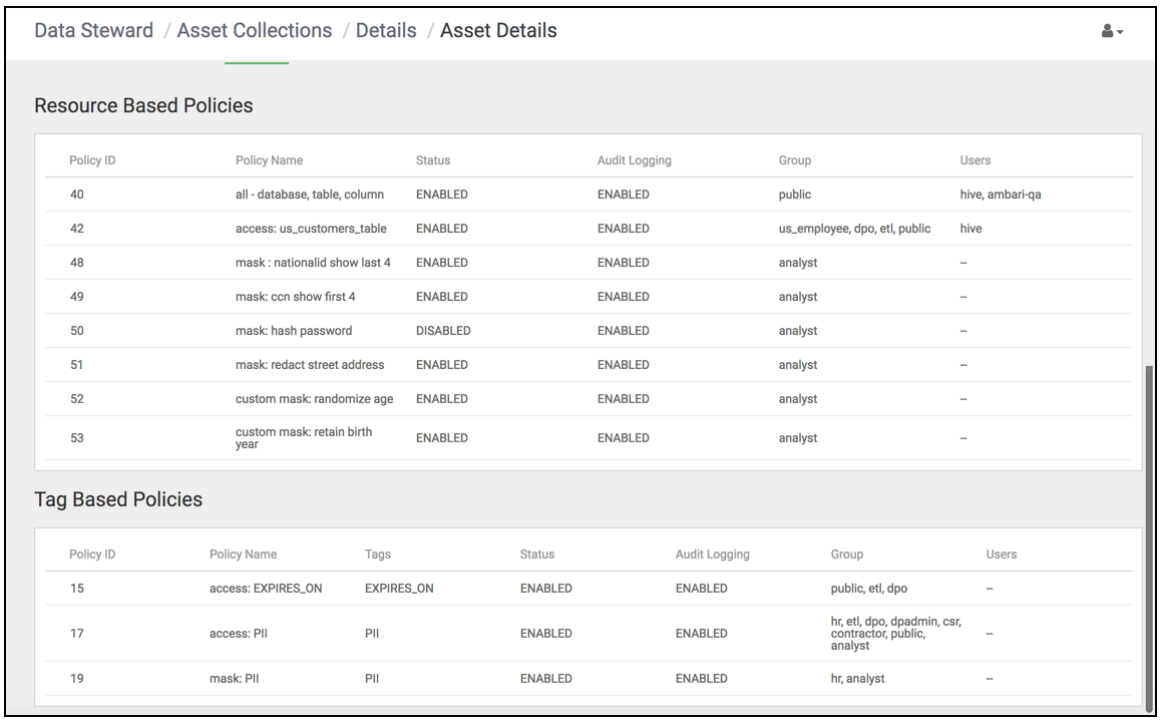
Procedure

1. From Data Steward, click: **Asset Collection** > **Select one asset collection** > **Assets** > **Select one 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.

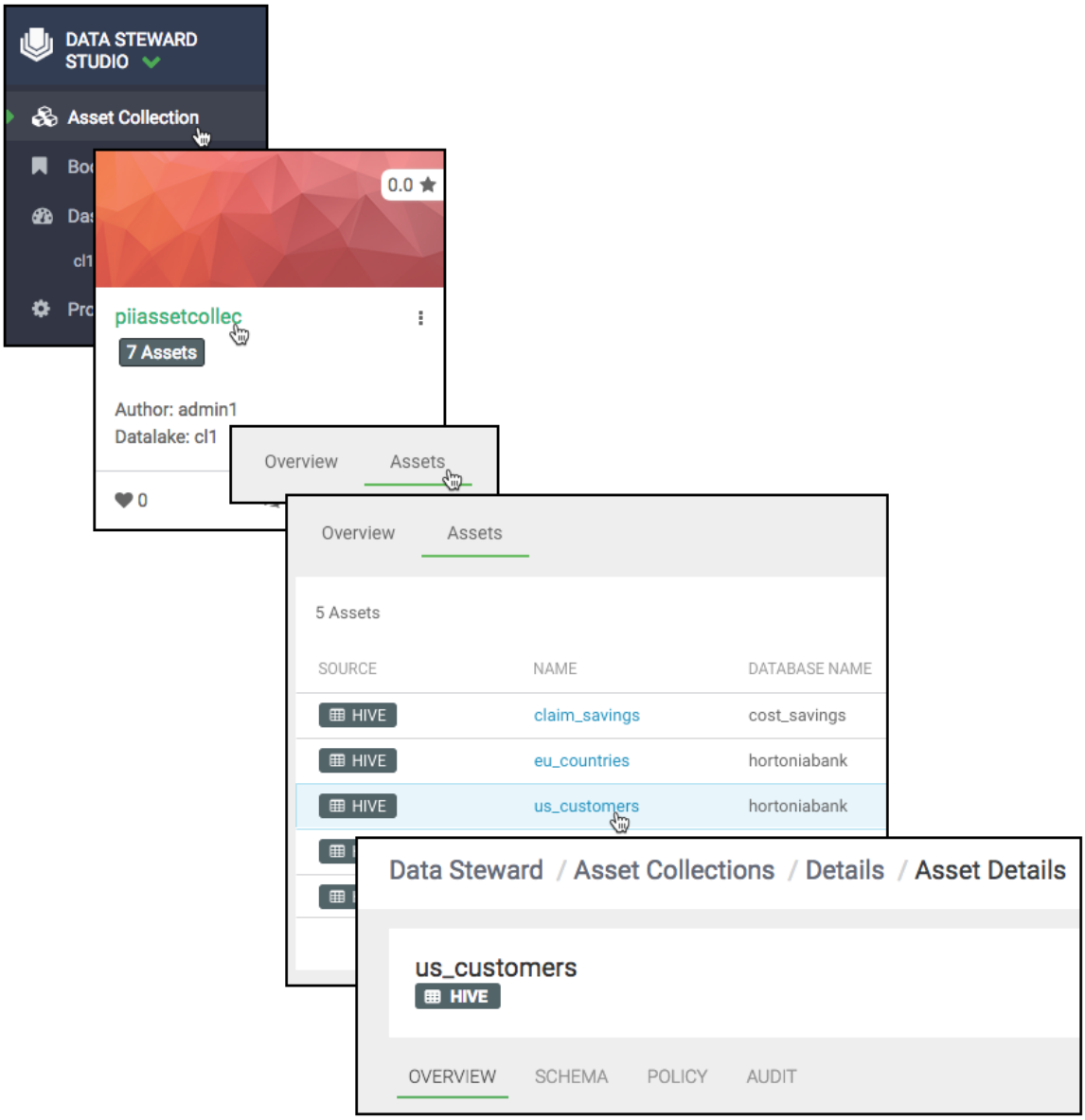


View Data Asset Audit Logs

Asset 360 > **Audit** 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.

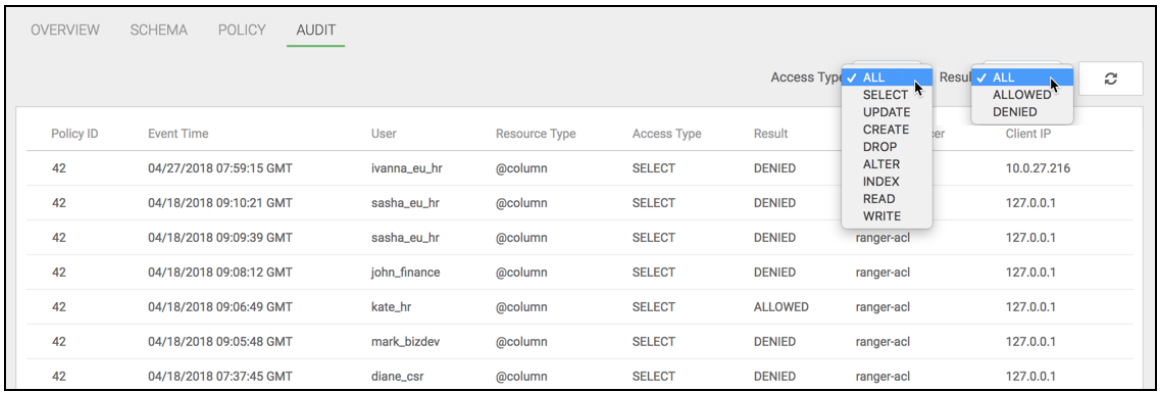
Procedure

1. From Data Steward, click: **Asset Collection** > **Select one asset collection** > **Assets** > **Select one data asset**:



The Asset **Overview** window opens.

2. Click the **Audit** tab.
   The Audit table shows the most recent raw audit event data as well as summarized views of audits by type of access and access outcome (authorized/unauthorized). Such summarized views are obtained by profiling audit records in the data lake with the audit profiler.



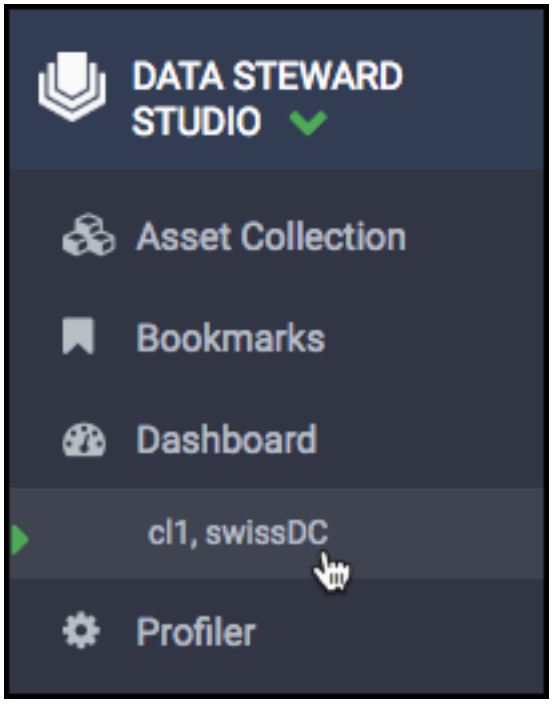
3. (Optional) You can filter the audit results by Access Type or Result.

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

Result: ALLOWED, DENIED.

Viewing a Data Lake Dashboard

The Data Steward Studio Dashboard gives you an overview of your data lake's profiles and assets: Hive tables, execution, sensitivity, and access. This helps you understand asset profile coverage, access data, and asset sensitivity proportion (e.g. PII, PCI, HIPAA), at a glance.



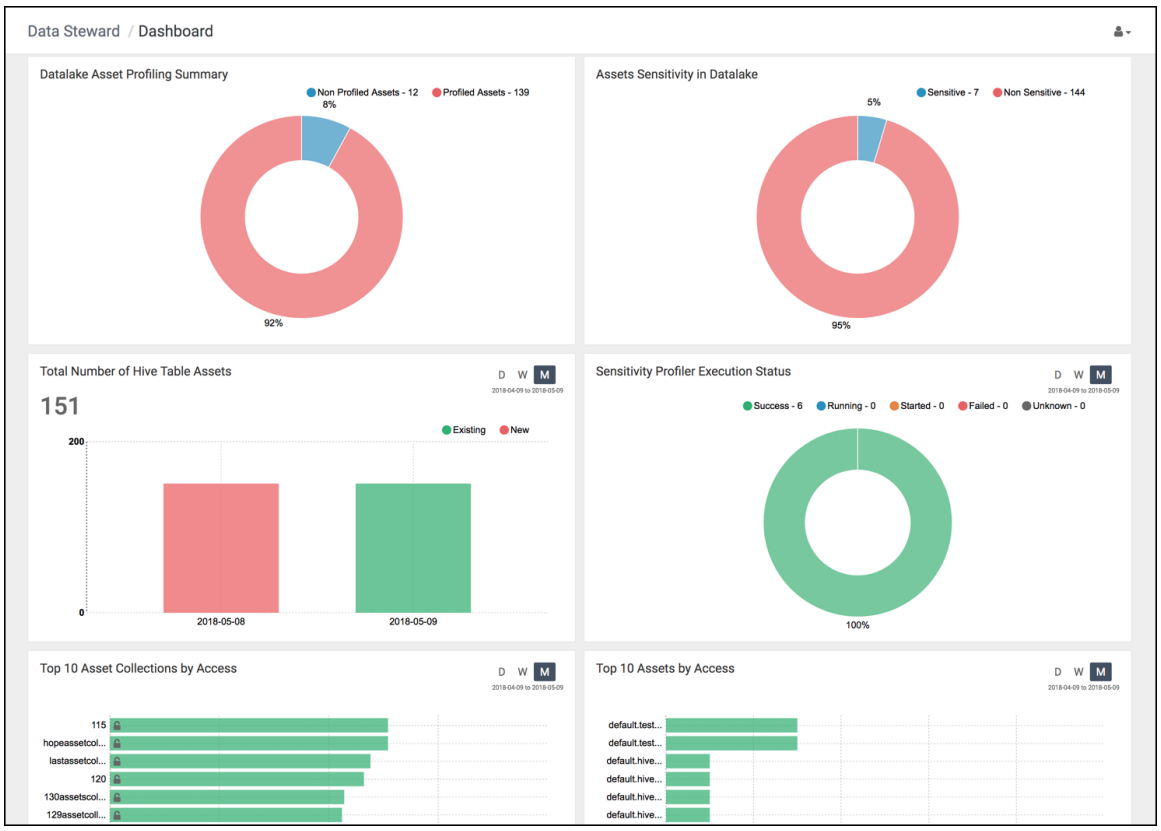


Table 1: Dashboard graphs

Graph Title	Description
Datalake Asset Profiling Summary	The number and percentage of assets covered by data profiling operations.
Asset Sensitivity in Datalake	The number and percentage of assets that are considered sensitive (e.g., PII, PCI, HIPAA). Based on a defined set of regular expressions, DSS runs a profiler job against Hive columns to determine whether values of the column satisfy the criteria for specific types of sensitive data and classify the columns accordingly.
Total Number of Hive Table Assets	Shows how your Hive table assets are growing over time.
Sensitivity Profiler Execution Status	This graph provides information about the monthly status of a particular profiler's execution: How many assets were run on that day, and how many completed successfully.
Top 10 Asset Collections by Access	Most accessed Asset Collections and how many times they were accessed.
Top 10 Assets by Access	Most accessed assets, who is accessing them, and how many times.

Related Information

Sensitive information types in Exchange 2016

Managing Profiles

The DSS profiler engine runs data profiling operations as a pipeline on data located in multiple data lakes. These profiles create metadata annotations that summarizes the content and shape characteristics for data assets.

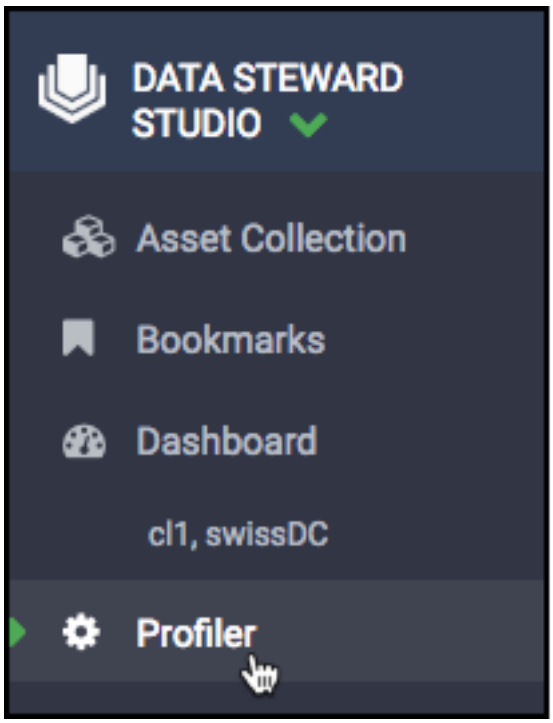


Table 2: List of built-in profilers

Name	Profiler	Description
Hive Column	tablestats hivecolumn	A Hive column univariate statistical profiler.
Hive Metastore	hive_metastore_profiler	Retrieves information about the number of hive tables that have been added every day.
Sensitive	sensitiveinfo	A sensitive data profiler- PII, PCI, HIPAA, etc.
Ranger Audit	audit	A Ranger audit log summarizer.

You can edit some of the profiler configurations in Ambari via the Datalake Profiler component. Currently, you can only use pre-built profilers. You can only schedule profilers during install.

Related reference

Ambari Dataplane Profiler Configs

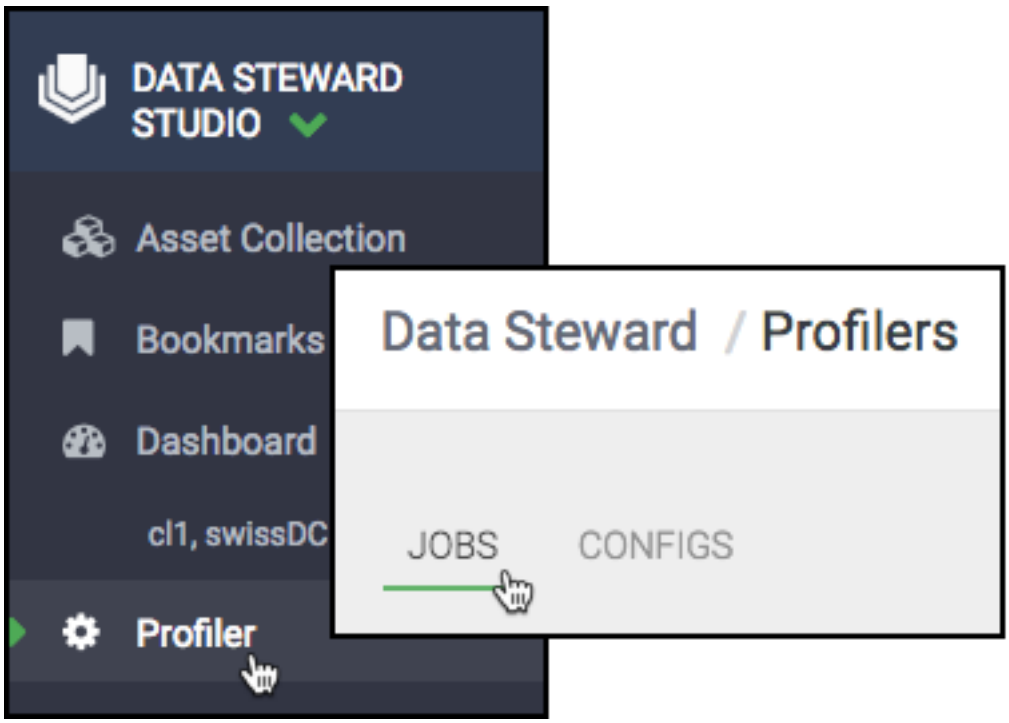
Viewing Profiler Jobs

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

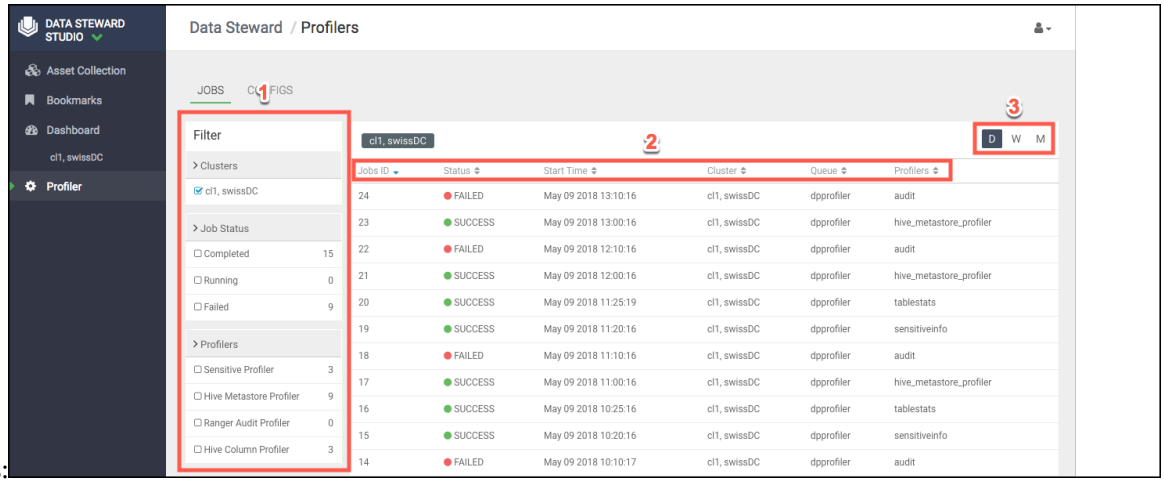
Each profiler, such as the Sensitive Profiler or Hive Column Profiler, runs a Spark job on a user-defined schedule in a user-defined queue. The queue is defined via the profiler configuration. You can view the status of each of those jobs for all your clusters.

Monitoring the profiler jobs has the following uses:

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



You can take the following actions:



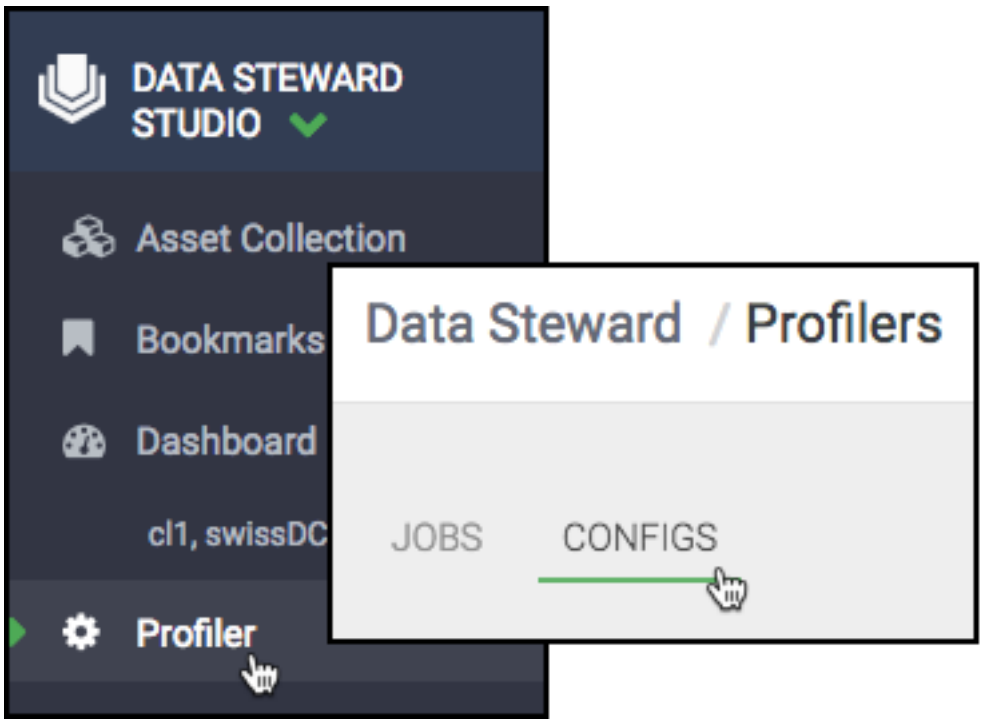
1. Filter by cluster, job status, or profiler.
2. Sort by jobs ID, status, start time, cluster, queue, or profilers.
3. Expand or narrow to show a day, week, or month of jobs.

Viewing Profiler Configurations

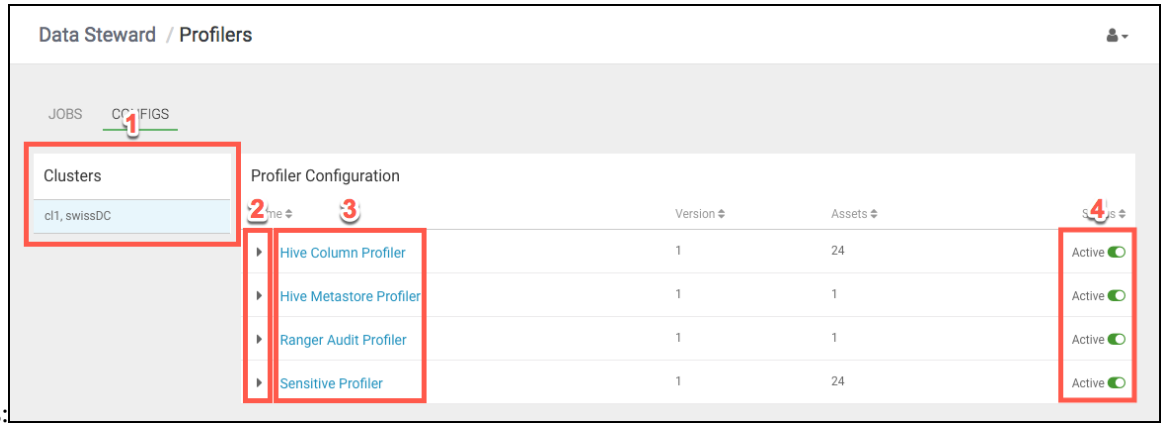
You can monitor the overall health of individual profilers by viewing their status on **Profiler** > **Configs**.

Monitoring the profiler configurations has the following uses:

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



You can take the following actions:



1. Filter by cluster.
2. Expand the execution status of an individual profiler. The percentage specifies how many assets have been profiled by this profiler on that day; the color denotes whether they were all successful, or not.
3. Edit the profiler configuration.
4. Toggle each profiler on/off.

Edit Profiler Configurations

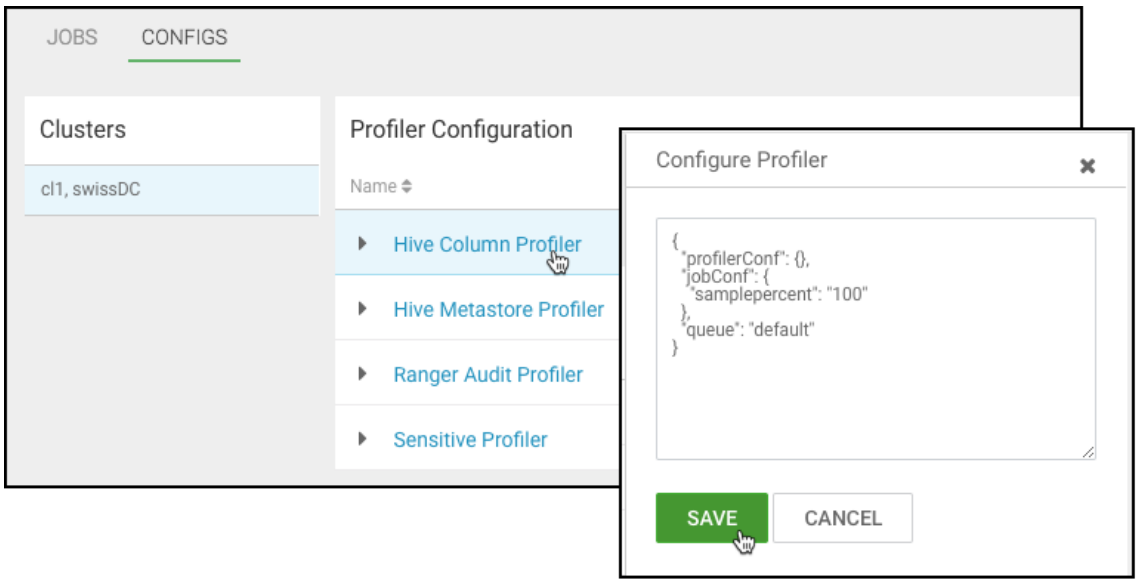
In addition to turning on and off the profiler configurations, the individual profilers can be run with their own execution parameters. These parameters are for submission of the profiler job onto Spark.

About this task

The values entered through the interface must be a valid JSON string.

Procedure

1. From **Profiler** > **Configs**, click the name of the profiler whose configuration you wish to edit.



Property	Description	Type	Sample
driverMemory	Amount of memory to use for the driver process	string	3g
driverCores	Number of cores to use for the driver process	int	2
executorMemory	Amount of memory to use per executor process	string	3g
executorCores	Number of cores to use for each executor	int	8
numExecutors	Number of executors to launch for this session	int	8

2. Edit the profiler and click **Save**.

Example

Profiler	Example
Hive Column Profiler	`{ "profilerConf": {}, "jobConf": { "samplepercent": "100" }, "queue": "default" }`
Hive Metastore Profiler	`{ "profilerConf": {}, "jobConf": {}, "queue": "default" }`
Ranger Audit Profiler	`{ "profilerConf": {}, "jobConf": {}, "queue": "default" }`
Sensitive Profiler	`{ "profilerConf": {}, "jobConf": { "sampleSize": "100", "saveToAtlas": "true" }, "queue": "default" }`

Related Information

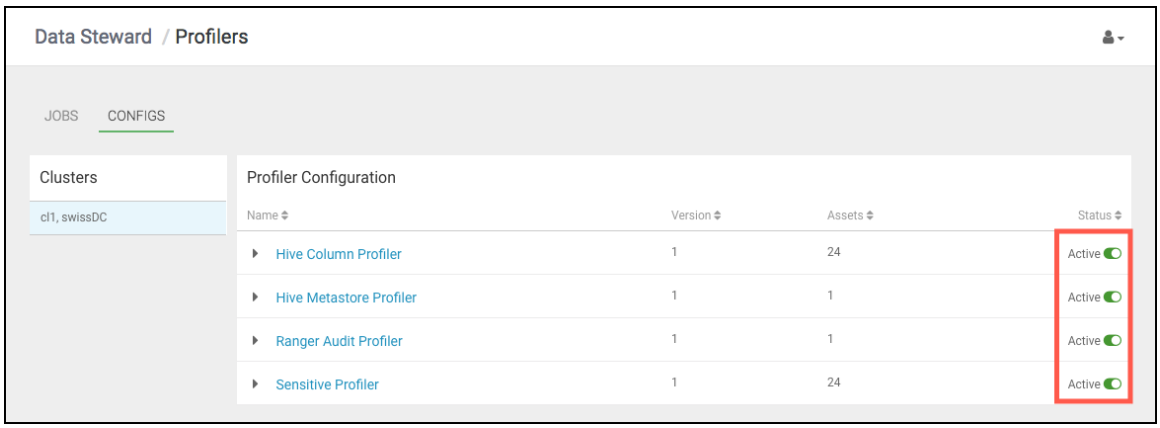
Submitting Spark Applications Through Livy

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 from the Configs tab.

Procedure

From **Profiler** > **Configs**, toggle the profiler **Active** or **Inactive**.



Accept or Reject Auto-Suggested Tags

The Sensitive Data Profiler detects data types and automatically suggests tags. If accepted, the tags are pushed back to Apache Atlas. Automatically suggested tags display as purple on the Schema tab.

Procedure

- To accept tags: **Schema tab** > **Edit Tags** > **Click purple tags** > **Save**.

  Once accepted by the user, the tag is saved back to Atlas in the cluster. In Atlas, the tag will be prefixed with "dp_" to denote that it comes from Dataplane Service.
- To reject tags: **Schema tab** > **Edit Tags** > **hover over purple tags** > **click (x) icon**.

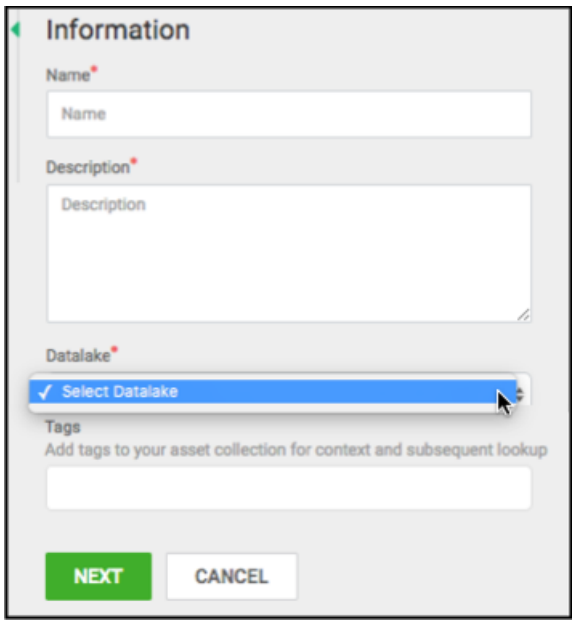
DSS Troubleshooting

This chapter contains common issues (with workarounds) and error message help for Data Steward Studio (DSS).

No datalake available when creating an Asset Collection

When creating an Asset Collection, no datalake displays in the drop-down menu.

A datalake is a cluster that has Apache Atlas and Apache Ranger installed. If registered clusters do not have Apache Atlas installed or there are no clusters registered to Hortonworks DataPlane Service, then no datalakes are available.



Procedure

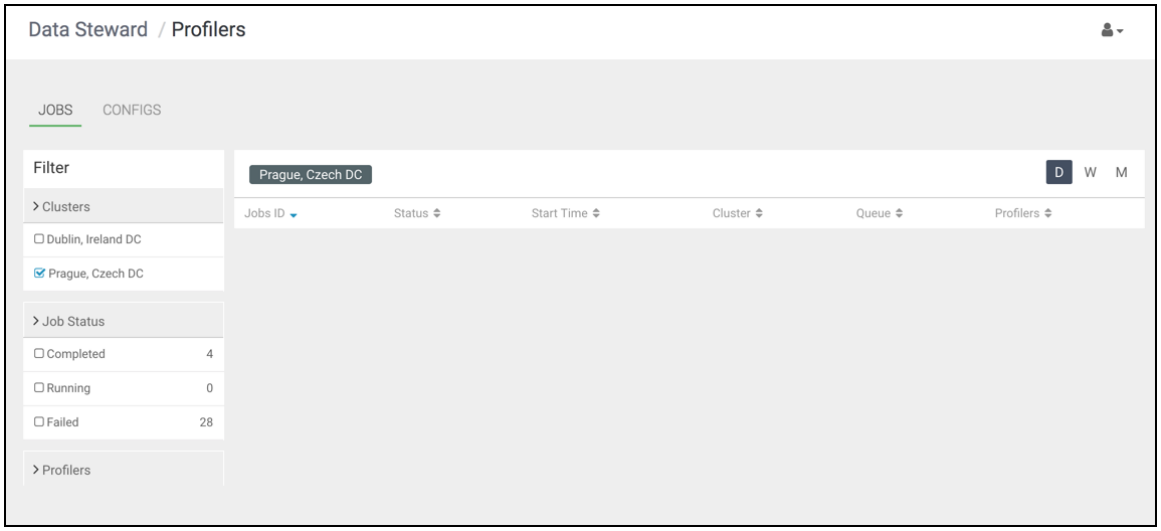
Register the cluster or install Apache Atlas and Apache Ranger on the cluster:

- Register a cluster in DataPlane
- Install Apache Atlas
- Install Apache Ranger

Profiler data does not load

Condition

When loading the **Profilers** tabs, profiler data does not load.



Cause

In Ambari, the Dataplane Profiler service is down.

Remedy

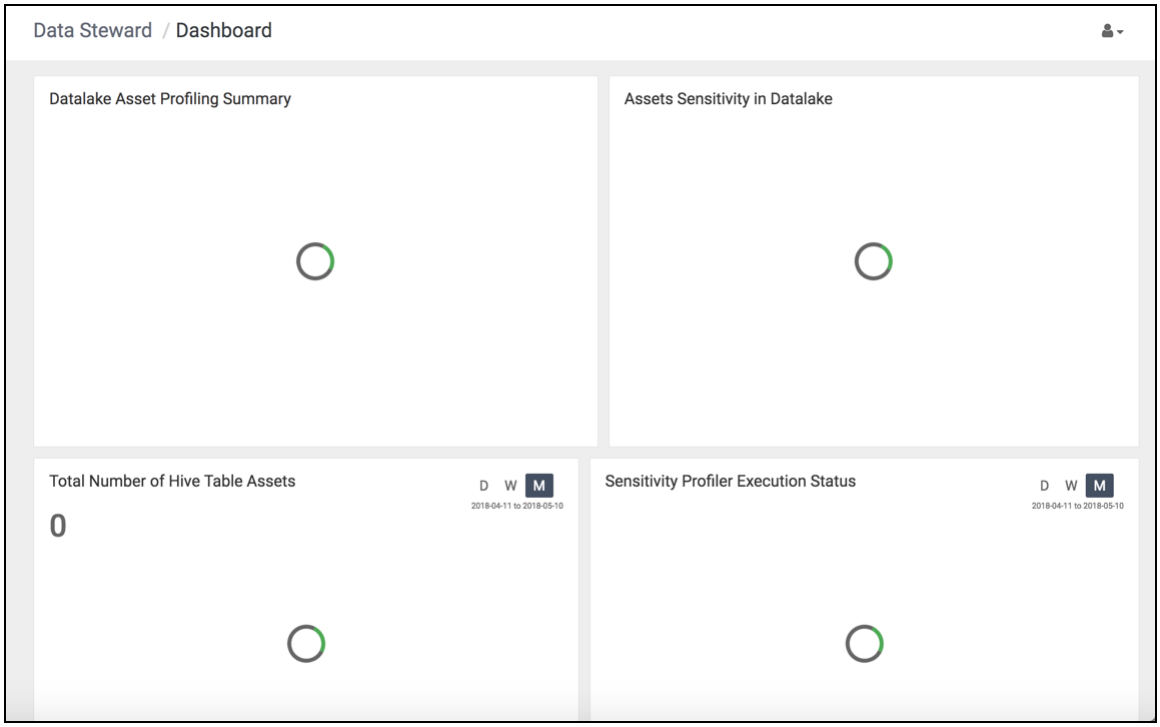
Procedure

Go to **Ambari** > **Dataplane Profiler** and turn the service on.

Widgets don't paint on dashboard

Condition

When loading the **Dashboard**, the widgets don't paint correctly (no data loaded).



Cause

There are errors occurring on the Profiler jobs.

Procedure

- Check for failed profiler jobs:
  a) From **Profiler** > **Jobs**, filter to the cluster whose dashboard is failing.
  b) Filter the job status to **Failed**.
  c) Use these failed profiler jobs to help troubleshoot the root cause.
- Verify that the Dataplane Profiler service is running in Ambari.
  a) Go to **Ambari** > **Dataplane Profiler** and check the status of the service.
  b) If the service is down, turn it on.