DSS Administration

Date of Publish: 2020-03-28

https://docs.hortonworks.com

Contents

Managing required permissions to access Data Steward Studio

Make sure your provide the DSS user the required permissions for Ambari, Knox, Atlas, and Ranger.

In addition to the DSS user with the Data Steward role, a DP Profiler user (dpprofiler) is created internally during the installation of Data Steward Studio for managing and running profiler jobs.

Permissions for the DP Profiler User

Make sure all the required permissions are granted within all the components.

The dpprofiler user should have the following permissions.

- Hive - The dpprofiler user must have access to read and list tables from the Hive metastore.
- Atlas

    - HDP 2.6.5 - The dpprofiler user must have read and write access to types, entities, and terms in Atlas.
    - HDP 3.x version - The dpprofiler user must have read and write access to types, entities, categories, and classifications in Atlas.
- YARN - The dpprofiler user must have access to run jobs in YARN against a configured queue. This includes queues used by each profiler (that can be configured through the Profiler configuration page) and queues used by Livy read and write sessions (that can be configured through Ambari). If you are using the default installation, make sure that these permissions are granted to the default queue.
- Ranger - The user needs to be a Ranger admin user to be able to access all the Ranger policies.

Permissions for the DSS User

The DSS user should have the following permissions.

- Ambari - The DSS user should be an Ambari user assigned a Cluster Operator role.
- Atlas - The DSS user must have read access to types, entities, and classifications.
- Ranger - To view a policy using DSS, the user must be a Ranger admin or a delegated admin with access to the Ranger policy.

Roles required for working with Data Steward Studio

Access to Data Steward Studio requires either a DSS Steward role or DSS Analyst role. To assign these roles to a user, you must log in to DataPlane Platform as a Data Plane Admin and assign the corresponding roles.

DSS Steward role

The DSS Steward has access to all the features within Data Steward Studio and can perform all read and write operations within DSS.

The DSS Steward has the following capabilities and restrictions.

Managing asset collections

- Can create new asset collections and add assets to asset collections.
- Can add comments on asset collections.
- Can rate asset collections.
- Can bookmark asset collections.
- Can view data asset details such as schema, tags, authorization policies, and audit logs.

Managing tags

- Can view, add, and delete managed tags.
- Can view, add, and delete system tags.
- Can view propagated tags.

Managing Dashboard

- Can view Data lake dashboard and components within it.

Managing Profilers

- Can view the status and configurations of all Profilers.
- Can activate or deactivate individual Profilers.
- Can edit and change configurations of Profilers on the Profiler configs page.

DSS Analyst role

The DSS Analyst has read access and can view all the features within Data Steward Studio.

The DSS Analyst has the following capabilities and restrictions.

Managing asset collections

- Can view asset collections.
- Cannot create new asset collections and add assets to asset collections.
- Can add comments on asset collections.
- Can rate asset collections.
- Can bookmark asset collections.
- Can view data asset details such as schema, tags, authorization policies, and audit logs.

Managing tags

- Can view managed tags, but cannot add and delete managed tags.
- Can view system taggs, but cannot add and delete system tags.
- Can view propagated tags.

Managing Dashboard

- Can view Data lake dashboard and components within it.

Managing Profilers

- Cannot activate or deactivate individual Profilers.
- Cannot edit and change configurations of Profilers on the Profiler Configs page.

Managing Asset Collections

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

On the Data Steward Studio menu, click **Asset Collection** to view all the asset collections.

Search for Asset Collection

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

Filter Asset Collections by Tags

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

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.
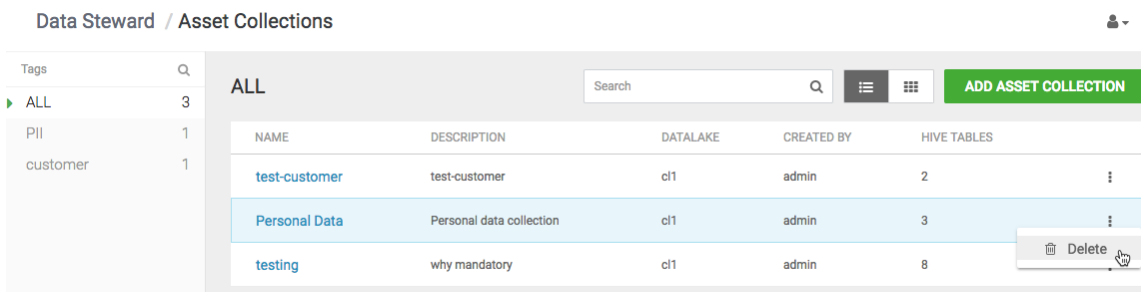
Delete Asset Collections

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.
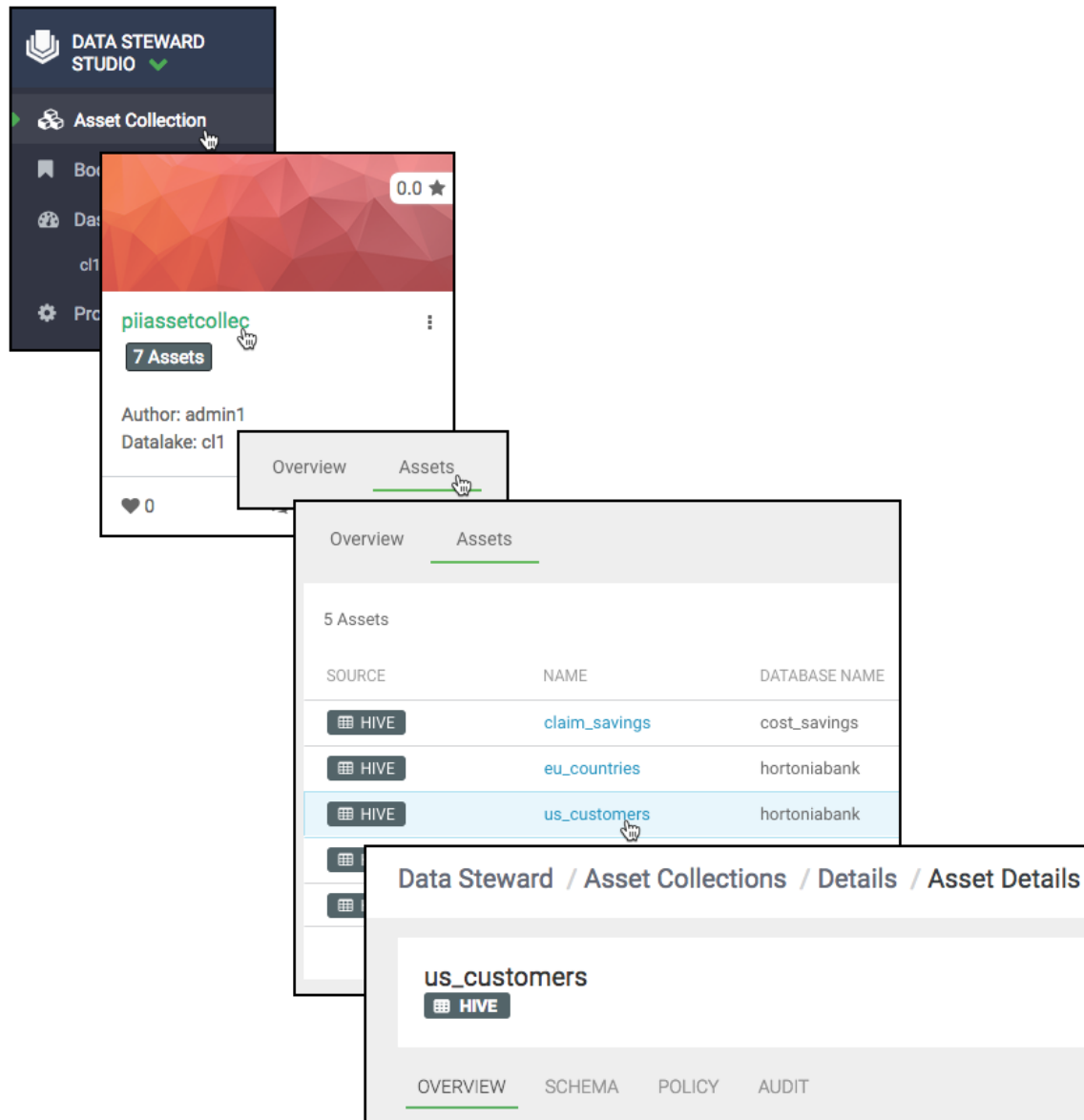
- Overview: Displays an overview for the data asset.

    - 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 the top five 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**:
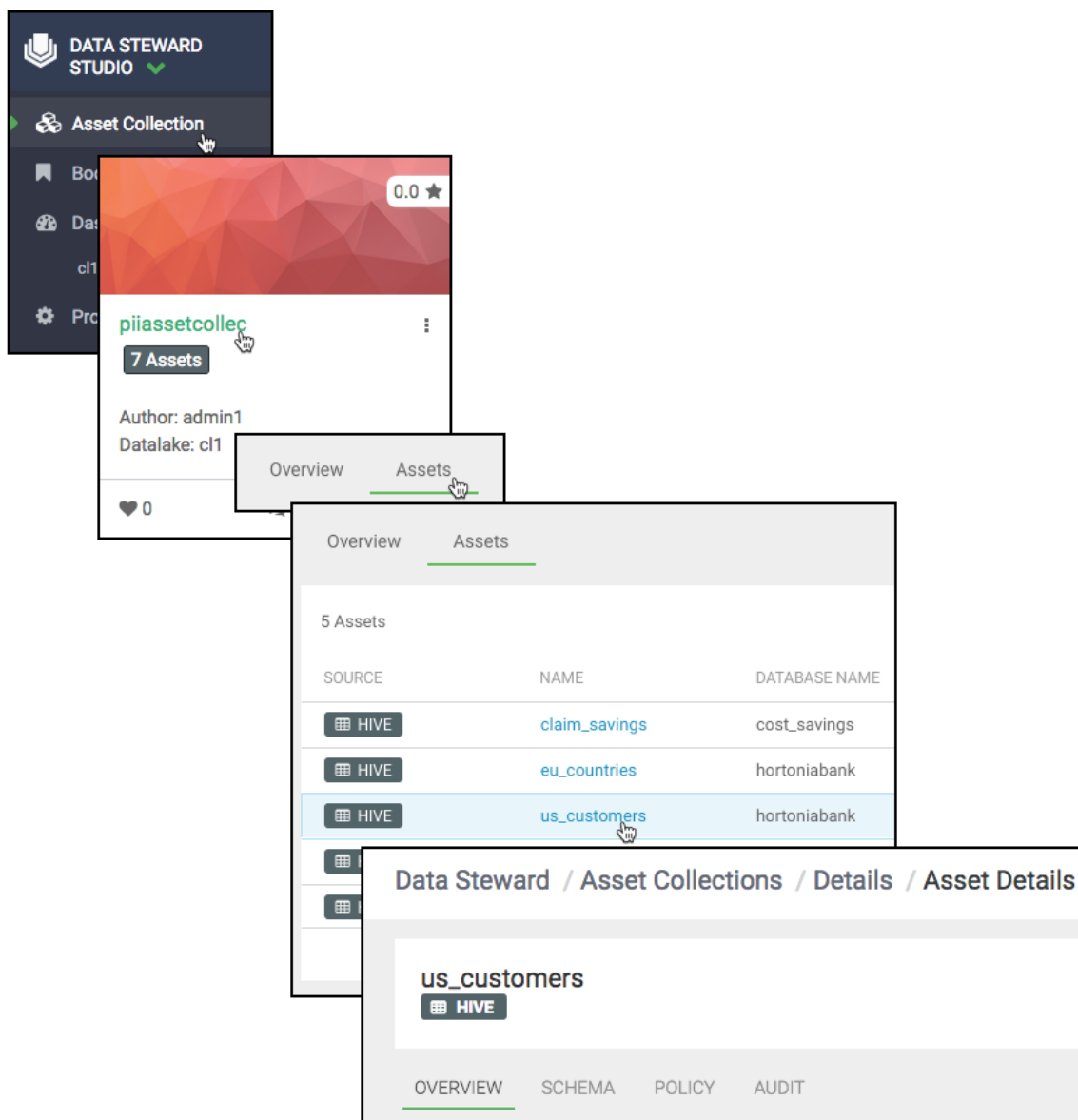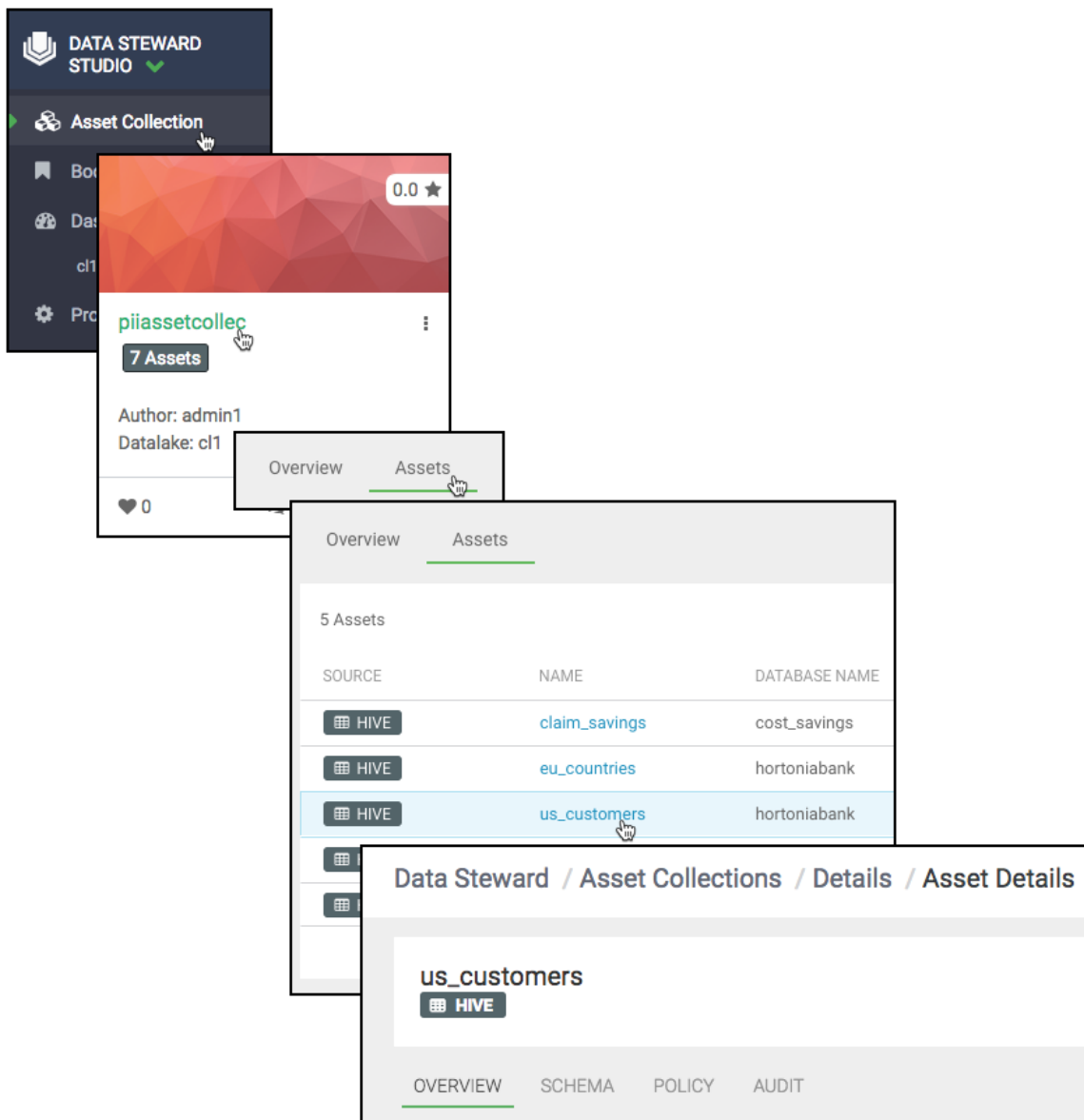
The Asset **Overview** window opens.

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

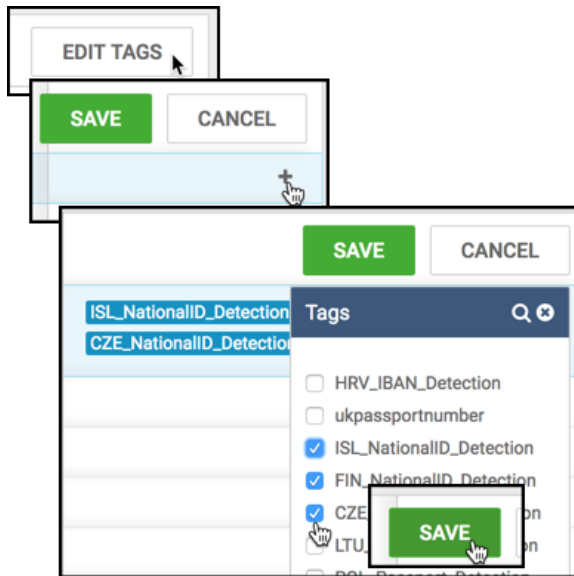1. From Data Steward, click: **Asset Collection** > **Select one asset collection** > **Assets** > **Select one data asset**.

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**.

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

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**:
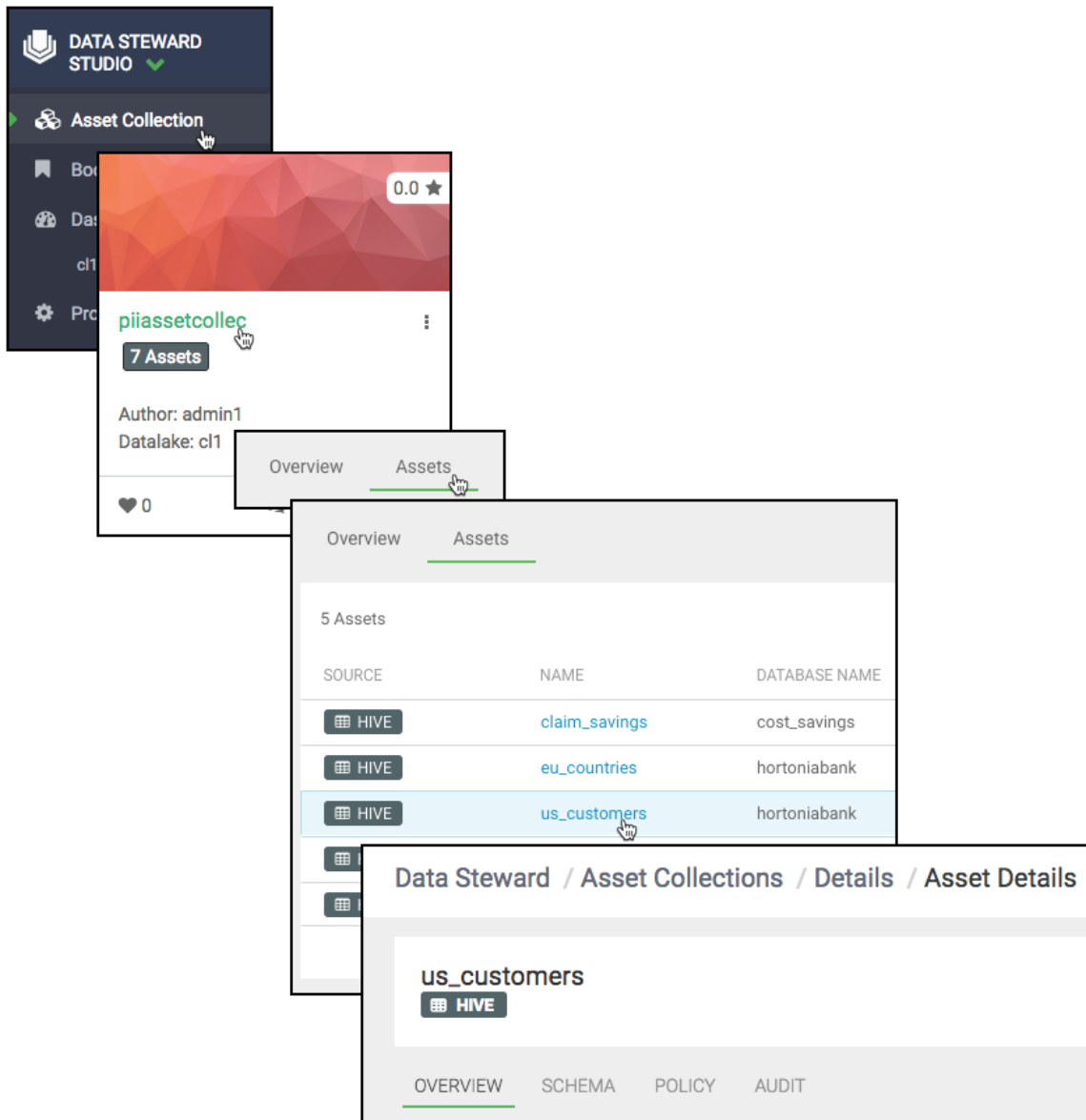
The Asset **Overview** window opens.

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**:
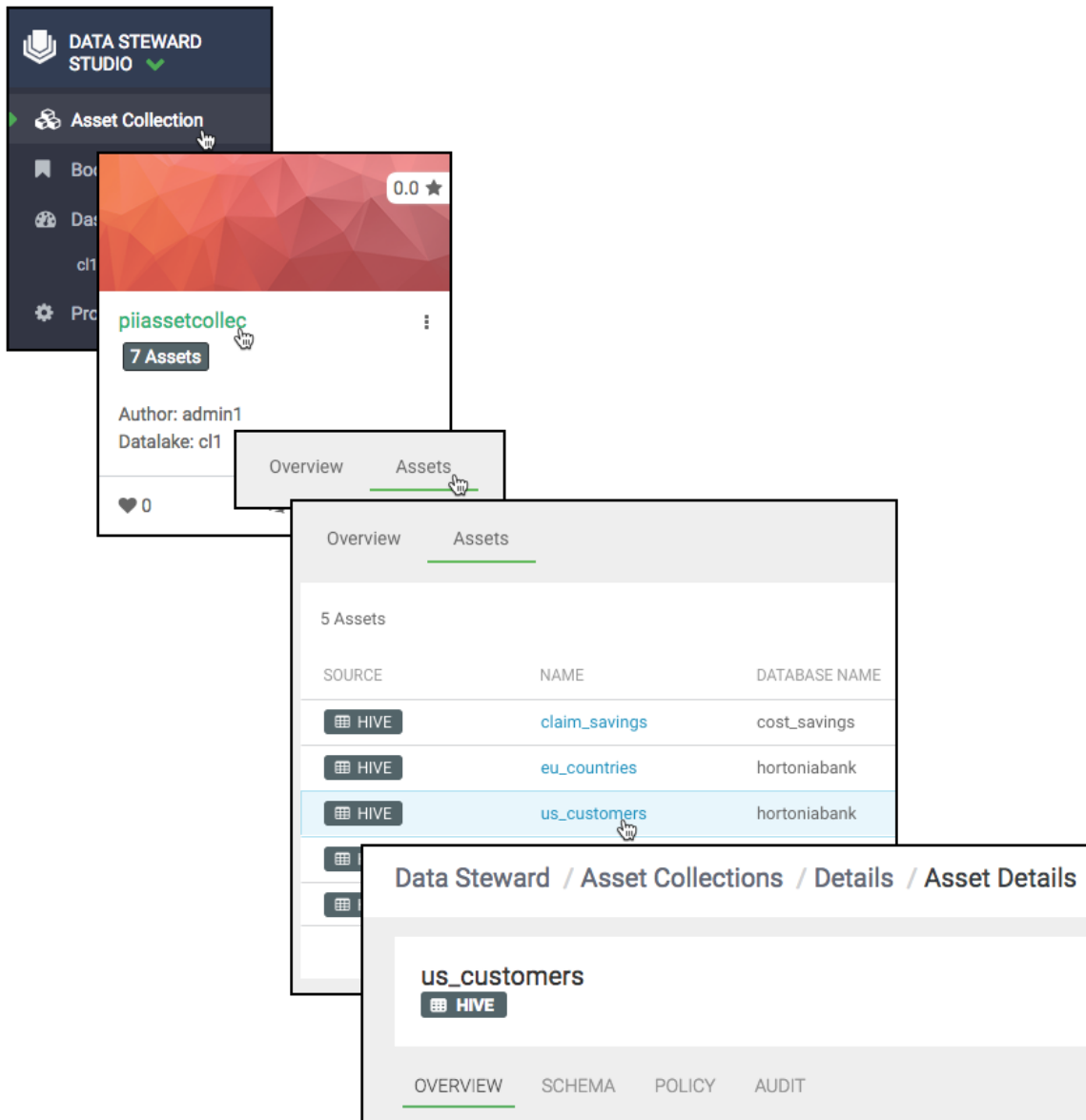
The Asset **Overview** window opens.

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 (for example PII, PCI, and HIPAA), at a glance.
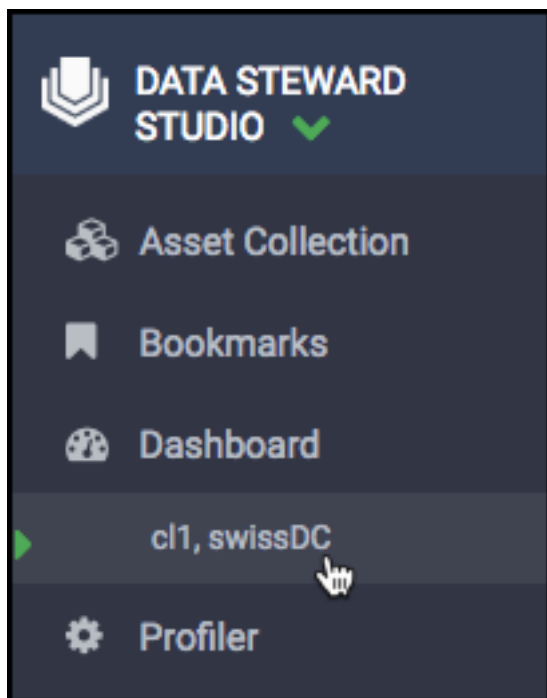
Table 1: Dashboard graphs

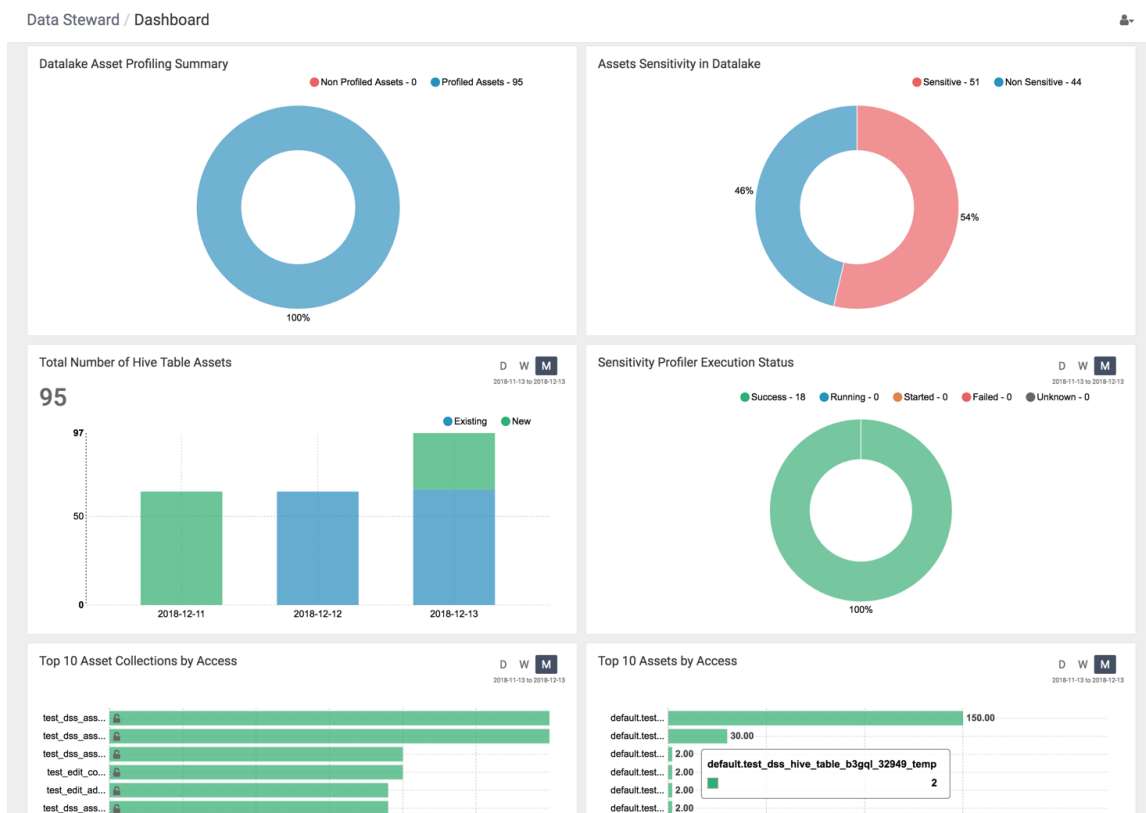| Graph Title	Description
Datalake Asset Profiling Summary	The number and percentage of assets covered by data profiling operations.
Asset Sensitivity in Datalake	The number and percentage of assets that are considered sensitive (for example, PII, PCI, and HIPAA). Based on a defined set of regular expressions, 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.

Managing Profilers

The DSS profiler engine runs data profiling operations as a pipeline on data located in multiple data lakes. These profilers create metadata annotations that summarize the content and shape characteristics of the data assets.
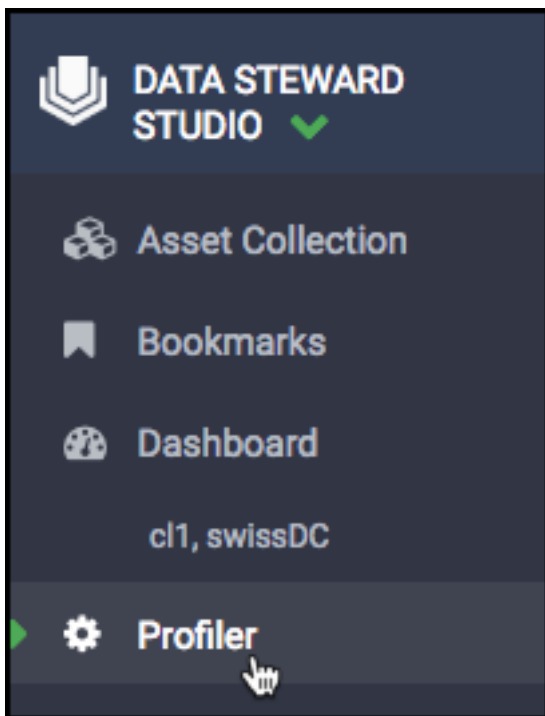
Table 2: List of built-in profilers

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 installation.

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 Configuration

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

Procedure

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

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

   The Profiler Configuration tab is displayed in the right panel.
5. Select the queue from the drop down list.
6. Select a schedule to run the profiler. This is implemented as a quartz cron expression.
7. Update the advanced options.

   - **Number of Executors** - Enter the number of executors to launch for running this profiler.
   - **Executor Cores** - Enter the number of cores to be used for each executor.
   - **Executor Memory** - Enter the amount of memory in GB to be used per executor process.
   - **Driver Cores** - Enter the number of cores to be used for the driver process.
   - **Driver Memory** - Enter the memory to be used for the driver processes.
8. Toggle the state of the profiler from **Active** to **Inactive** as needed.
9. Click **Save** to apply the configuration changes to the selected profiler. The changes should appear in the profiler description.

Additional Configuration for Cluster Sensitivity Profiler

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

Procedure

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

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

   The Profiler Configuration tab is displayed in the right panel.
5. Toggle the **Enable** button to enable Cluster Sensitivity Profiler. Select **Disable** if you do not want to run the Cluster Sensitivity Profiler.
6. Select the Sample Data Size.
   a) From the drop down, select the type of sample data size.
   b) Enter the value based on the previously selected type.
7. Select the queue, schedule, and advanced configuration details as specified in Edit Profiler Configuration.
8. Add Asset Filter Rules as needed to customize the selection and deselection of assets which the profiler profiles.

   - Black-list - The profiler will skip profiling assets that meet any defined blacklist criteria.

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

     a. Click the White-list tab.
     b. Click **Add** to include rules for whitelisting.
     c. Select the key from the drop down list. You can select a database name, name of the asset, name of the owner of the asset, path to the assets, or created date.
     d. 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.
     e. Enter the value corresponding to the key. For example, you can enter a string as mentioned in the previous example.
     f. Click **Done**. Once it is added, you can toggle the state of the new rule to enable it or disable it as needed.

     Note: If an asset meets both whitelist and blacklist rules, the blacklist rule will be given preference.

9. Toggle the state of the profiler from Active to Inactive as needed.
10. Click **Save** to apply the configuration changes to the selected profiler. The changes should appear in the profiler description.

Additional Configuration for Hive Column Profiler

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

Procedure

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

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

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

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

   a. From the drop down, select the type of sample data size.
   b. Enter the value based on the previously selected type.
7. Choose the **Selection Criteria**.

   • **All tables** - The Hive Column Profiler will run on all tables in the asset collection during its next scheduled run.
   • **Only changed tables** - The Hive Column Profiler will run only on tables which are in the asset collection and which have been changed.

   > **Note:**
   >
   > While configuring Hive-site settings in Ambari, make sure the hive.metastore.transactional.event.listeners is set to org.apache.hive.hcatalog.listener.DbNotificationListener parameter. If this parameter is not set as specified, you will not be able to select the option to choose **only changed tables**. An error will appear suggesting you to update the parameter.
8. Click **Save** to apply the configuration changes to the selected profiler. The changes should appear in the profiler description.

Replicate Profiler Configuration

If you have configured your profilers on a cluster and want to use the same configurations on other clusters, you can replicate the profiler configuration to the second cluster using the magic wand feature.

About this task

To replicate profiler configurations:

Procedure

1. Select the cluster which has all the profiler configurations defined.
2. Click the magic wand icon on this cluster.
3. In the tab that appears, select the clusters to which you want to replicate the profiler configurations.
4. Select the parameters that you want to replicate on each cluster. You can choose to copy queue, schedule, or consider advanced Spark options.

5. Click **Apply**.

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**.

Profiler Tag Rules

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

Rules are categorized into three groups:

- System Deployed : These are in-built rules that are available with default mpack installation. You cannot edit these rules.
- Custom Deployed: Tag rules that you create and deploy on clusters after validation will appear under this category. Hover your mouse over the tag rules to deploy or suspend them as needed. You can also edit these tag rules.
- Custom Draft : You can create new tag rules and save them for later validation and deployment on clusters. Such rules appear under this category.

Data Steward / Profilers

JOBS   CONFIGS   TAG RULES

Rule Groups	
System Deployed	78
Custom Deployed	2
Custom Draft	0

Type to search

+ NEW

	Name	Version	Description	Associated Tags	Created By	# of Cluster	Status
☐	SVK_NationalID_Detection	1.1	SVK_NationalID_Detection		Hortonworks	-	Active
☐	LVA_IBAN_Detection	1.1	LVA_IBAN_Detection		Hortonworks	-	Active
☐	ROU_IBAN_Detection	1.1	ROU_IBAN_Detection		Hortonworks	-	Active
☐	NOR_NationalID_Detection	1.1	NOR_NationalID_Detection		Hortonworks	-	Active
☐	FRA_IBAN_Detection	1.1	FRA_IBAN_Detection		Hortonworks	-	Active
☐	name	1.1	name		Hortonworks	-	Active
☐	DEU_IBAN_Detection	1.1	DEU_IBAN_Detection		Hortonworks	-	Active
☐	FIN_NationalID_Detection	1.1	FIN_NationalID_Detection		Hortonworks	-	Active
☐	ESP_Passport_Detection	1.1	ESP_Passport_Detection		Hortonworks	-	Active
☐	DEU_Passport_Detection	1.1	DEU_Passport_Detection		Hortonworks	-	Active
☐	CYP_IBAN_Detection	1.1	CYP_IBAN_Detection		Hortonworks	-	Active
☐	FIN_Passport_Detection	1.1	FIN_Passport_Detection		Hortonworks	-	Active
☐	email	1.1	email		Hortonworks	-	Active
☐	AUT_IBAN_Detection	1.1	AUT_IBAN_Detection		Hortonworks	-	Active
☐	GRC_NationalID_Detection	1.1	GRC_NationalID_Detection		Hortonworks	-	Active
☐	BEL_IBAN_Detection	1.1	BEL_IBAN_Detection		Hortonworks	-	Active
☐	EST_IBAN_Detection	1.1	EST_IBAN_Detection		Hortonworks	-	Active
☐	CHE_NationalID_Detection	1.1	CHE_NationalID_Detection		Hortonworks	-	Active
☐	POL_Passport_Detection	1.1	POL_Passport_Detection		Hortonworks	-	Active
☐	PRT_NationalID_Detection	1.1	PRT_NationalID_Detection		Hortonworks	-	Active
☐	CHE_IBAN_Detection	1.1	CHE_IBAN_Detection		Hortonworks	-	Active
☐	ESP_IBAN_Detection	1.1	ESP_IBAN_Detection		Hortonworks	-	Active
☐	ITA_IBAN_Detection	1.1	ITA_IBAN_Detection		Hortonworks	-	Active
☐	swiftcode	1.1	swiftcode		Hortonworks	-	Active
☐	ukpassportnumber	1.1	ukpassportnumber		Hortonworks	-	Active
☐	SVK_IBAN_Detection	1.1	SVK_IBAN_Detection		Hortonworks	-	Active
☐	ipaddress	1.1	ipaddress		Hortonworks	-	Active
☐	SWE_IBAN_Detection	1.1	SWE_IBAN_Detection		Hortonworks	-	Active
☐	CHE_Passport_Detection	1.1	CHE_Passport_Detection		Hortonworks	-	Active
☐	uknationalinsurancenumber	1.1	uknationalinsurancenumber		Hortonworks	-	Active
☐	LTU_NationalID_Detection	1.1	LTU_NationalID_Detection		Hortonworks	-	Active
☐	NOR_IBAN_Detection	1.1	NOR_IBAN_Detection		Hortonworks	-	Active
☐	Passport_Detection	1.1	Passport_Detection		Hortonworks	-	Active
☐	IBAN_Detection	1.1	IBAN_Detection		Hortonworks	-	Active
☐	GBR_IBAN_Detection	1.1	GBR_IBAN_Detection		Hortonworks	-	Active
☐	IRL_IBAN_Detection	1.1	IRL_IBAN_Detection		Hortonworks	-	Active
☐	FIN_IBAN_Detection	1.1	FIN_IBAN_Detection		Hortonworks	-	Active
☐	NLD_IBAN_Detection	1.1	NLD_IBAN_Detection		Hortonworks	-	Active
☐	BGR_IBAN_Detection	1.1	BGR_IBAN_Detection		Hortonworks	-	Active
☐	SVN_IBAN_Detection	1.1	SVN_IBAN_Detection		Hortonworks	-	Active
☐	LIE_IBAN_Detection	1.1	LIE_IBAN_Detection		Hortonworks	-	Active
☐	AUT_Passport_Detection	1.1	AUT_Passport_Detection		Hortonworks	-	Active
☐	POL_NationalID_Detection	1.1	POL_NationalID_Detection		Hortonworks	-	Active
☐	POL_IBAN_Detection	1.1	POL_IBAN_Detection		Hortonworks	-	Active
☐	PRT_IBAN_Detection	1.1	PRT_IBAN_Detection		Hortonworks	-	Active
☐	LTU_IBAN_Detection	1.1	LTU_IBAN_Detection		Hortonworks	-	Active
☐	LVA_NationalID_Detection	1.1	LVA_NationalID_Detection		Hortonworks	-	Active
☐	HRV_IBAN_Detection	1.1	HRV_IBAN_Detection		Hortonworks	-	Active
☐	DNK_NationalID_Detection	1.1	DNK_NationalID_Detection		Hortonworks	-	Active
☐	GRC_IBAN_Detection	1.1	GRC_IBAN_Detection		Hortonworks	-	Active
☐	CZE_NationalID_Detection	1.1	CZE_NationalID_Detection		Hortonworks	-	Active
☐	LUX_IBAN_Detection	1.1	LUX_IBAN_Detection		Hortonworks	-	Active
☐	telephone	1.1	telephone		Hortonworks	-	Active
☐	bankaccount	1.1	bankaccount		Hortonworks	-	Active
☐	SWE_NationalID_Detection	1.1	SWE_NationalID_Detection		Hortonworks	-	Active
☐	ISL_NationalID_Detection	1.1	ISL_NationalID_Detection		Hortonworks	-	Active
☐	ESP_NationalID_Detection	1.1	ESP_NationalID_Detection		Hortonworks	-	Active
☐	ukdrivernumber	1.1	ukdrivernumber		Hortonworks	-	Active
☐	DNK_IBAN_Detection	1.1	DNK_IBAN_Detection		Hortonworks	-	Active
☐	EST_NationalID_Detection	1.1	EST_NationalID_Detection		Hortonworks	-	Active
☐	ITA_NationalID_Detection	1.1	ITA_NationalID_Detection		Hortonworks	-	Active
☐	GRC_Passport_Detection	1.1	GRC_Passport_Detection		Hortonworks	-	Active
☐	HUN_IBAN_Detection	1.1	HUN_IBAN_Detection		Hortonworks	-	Active
☐	NationalID_Detection	1.1	NationalID_Detection		Hortonworks	-	Active
☐	CZE_IBAN_Detection	1.1	CZE_IBAN_Detection		Hortonworks	-	Active

Additional Operations for Profiler Tags

You can now perform multiple operations on the tag rules.

You can select one or more tags to deploy or suspend the system deployed tag rule status. You can also select multiple tags with different status and deploy/suspend them as applicable.

Data Steward / Profilers

JOBS	CONFIGS	TAG RULES

Rule Groups

System Deployed	78
Custom Deployed	1
Custom Draft	4

Type to search 🔍   **SUSPEND SELECTED**   **+ NEW**

	Name	Description	Associated Tags	Created By	# of Cluster	Status
☐	SVK_NationalID_Detection	SVK_NationalID_Detection	SVK_NationalID_Detection	Hortonworks	0	Suspended
☐	LVA_IBAN_Detection	LVA_IBAN_Detection	LVA_IBAN_Detection	Hortonworks	0	Suspended
☑	ROU_IBAN_Detection	ROU_IBAN_Detection	ROU_IBAN_Detection	Hortonworks	1	Active
☑	NOR_NationalID_Detection	NOR_NationalID_Detection	NOR_NationalID_Detection	Hortonworks	1	Active
☑	FRA_IBAN_Detection	FRA_IBAN_Detection	FRA_IBAN_Detection	Hortonworks	1	Active
☑	name	name	name	Hortonworks	1	Active
☐	DEU_IBAN_Detection	DEU_IBAN_Detection	DEU_IBAN_Detection	Hortonworks	1	Active
☐	FIN_NationalID_Detection	FIN_NationalID_Detection	FIN_NationalID_Detection	Hortonworks	0	Suspended
☐	ESP_Passport_Detection	ESP_Passport_Detection	ESP_Passport_Detection	Hortonworks	1	Active

Select Lakes to suspend for the following 4 rules

ROU_IBAN_Detection
NOR_NationalID_Detection
FRA_IBAN_Detection
name

☑ **cl1, san_US_7019**

CANCEL   **SUSPEND**

Data Steward / Profilers

JOBS	CONFIGS	TAG RULES

Rule Groups

System Deployed	78
Custom Deployed	1
Custom Draft	4

Type to search 🔍   **SUSPEND SELECTED**   **DEPLOY SELECTED**   **+ NEW**

	Name	Description	Associated Tags	Created By	# of Cluster	Status
☐	SVK_NationalID_Detection	SVK_NationalID_Detection	SVK_NationalID_Detection	Hortonworks	0	Suspended
☑	LVA_IBAN_Detection	LVA_IBAN_Detection	LVA_IBAN_Detection	Hortonworks	0	Suspended
☑	ROU_IBAN_Detection	ROU_IBAN_Detection	ROU_IBAN_Detection	Hortonworks	1	Active
☑	NOR_NationalID_Detection	NOR_NationalID_Detection	NOR_NationalID_Detection	Hortonworks	1	Active
☐	FRA_IBAN_Detection	FRA_IBAN_Detection	FRA_IBAN_Detection	Hortonworks	1	Active
☑	name	name	name	Hortonworks	1	Active
☐	DEU_IBAN_Detection	DEU_IBAN_Detection	DEU_IBAN_Detection	Hortonworks	1	Active
☑	FIN_NationalID_Detection	FIN_NationalID_Detection	FIN_NationalID_Detection	Hortonworks	0	Suspended
☐	ESP_Passport_Detection	ESP_Passport_Detection	ESP_Passport_Detection	Hortonworks	1	Active
☑	DEU_Passport_Detection	DEU_Passport_Detection	DEU_Passport_Detection	Hortonworks	1	Active
☐	CYP_IBAN_Detection	CYP_IBAN_Detection	CYP_IBAN_Detection	Hortonworks	1	Active
☑	FIN_Passport_Detection	FIN_Passport_Detection	FIN_Passport_Detection	Hortonworks	0	Suspended
☐	email	email	email	Hortonworks	1	Active
☑	AUT_IBAN_Detection	AUT_IBAN_Detection	AUT_IBAN_Detection	Hortonworks	1	Active
☐	GRC_NationalID_Detection	GRC_NationalID_Detection	GRC_NationalID_Detection	Hortonworks	1	Active
☐	BEL_IBAN_Detection	BEL_IBAN_Detection	BEL_IBAN_Detection	Hortonworks	1	Active
☐	EST_IBAN_Detection	EST_IBAN_Detection	EST_IBAN_Detection	Hortonworks	1	Active
☐	CHE_NationalID_Detection	CHE_NationalID_Detection	CHE_NationalID_Detection	Hortonworks	1	Active

Creating Custom Profiler Rules

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

About this task

Procedure

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

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

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

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

Adding Custom Regular Expressions

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

About this task

Procedure

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

   For example:

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

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

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

Adding Whitelist or Blacklist Files

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

Procedure

1. Create a new whitelist or blacklist file.
2. Click **Resources** in the right panel on the New Custom Profiler Rules page.
3. Click + icon on the Lookups tab. The New Whitelist/Blacklist File page appears.
4. Enter the name of the new whitelist or blacklist file.
5. Click **Choose File** to upload the file.
6. Click **Save**.

Using Behaviors

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

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

1. Regular expressions
2. File based whitelist and blacklist checks
3. Luhn algorithm

Regular expressions

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

Keyword: regex

A regex that matches everything can be defined as follows:

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

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

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

File based blacklist and whitelist checks

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

Keywords: whitelist, blacklist

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

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

For example:

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

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

Luhn algorithm

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

Keyword: luhn_check

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

Using DSL Grammar

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

The two dummy behaviours available in this framework are as follows:

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

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

Binary AND operator

Keyword: and

And works the same way it does it 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 it other languages.

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

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

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

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

Let's expand our DSL to use or 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.

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**.
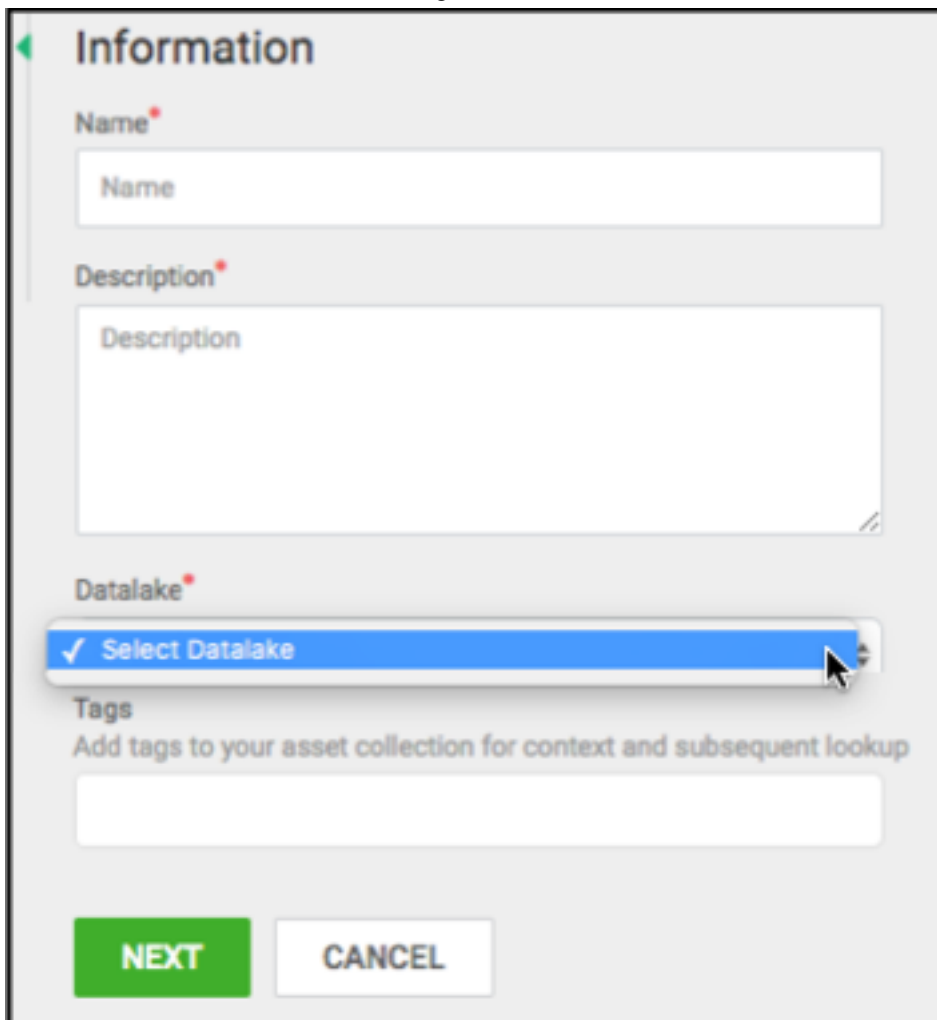
DSS Troubleshooting

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

No data lake 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.
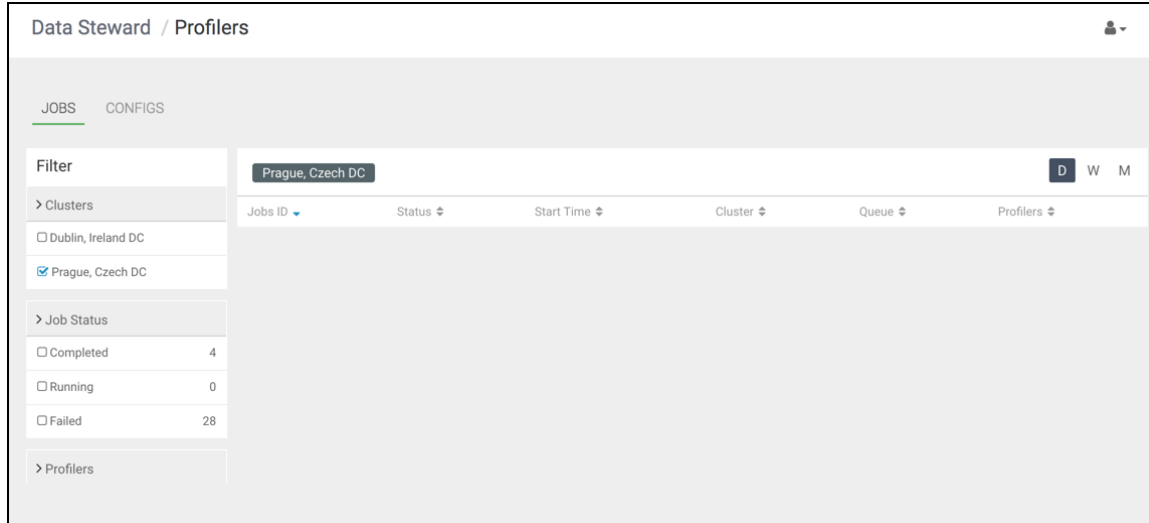

Procedure

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

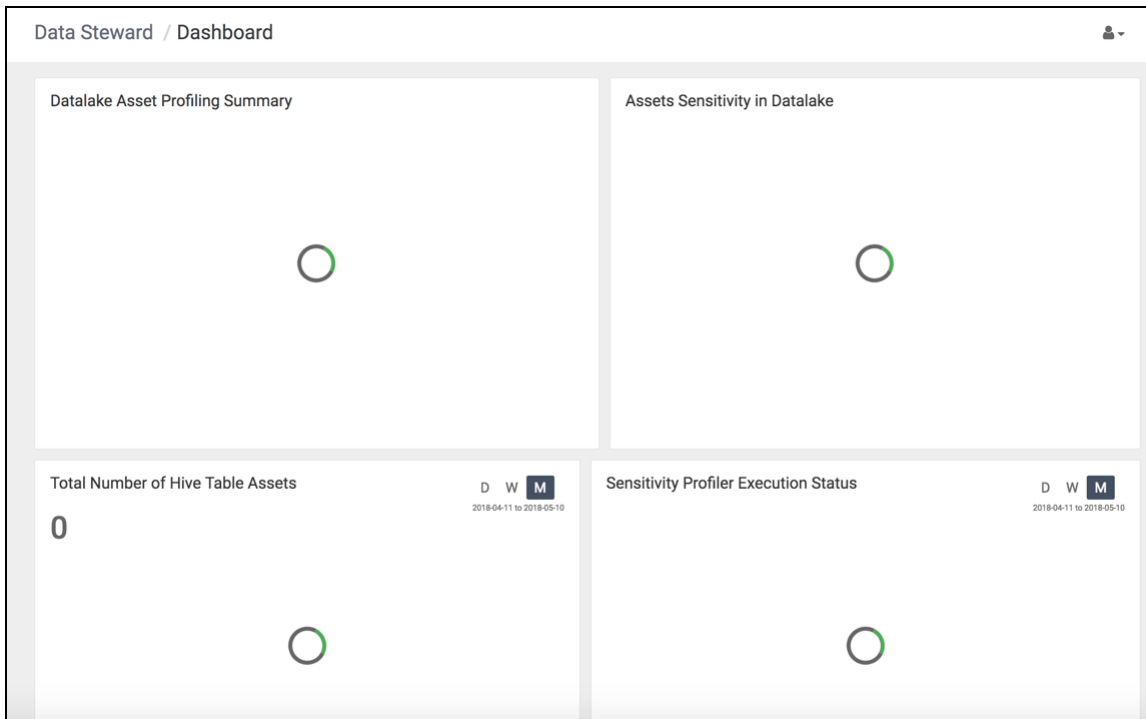In Ambari, the Dataplane Profiler service is down.

Remedy

Procedure

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

Widgets do not load on dashboard

Condition

When loading the **Dashboard**, the widgets do not load 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.