Hortonworks DataPlane Service

Data Steward Studio Administration

(October 31, 2017)

docs.hortonworks.com

Hortonworks DataPlane Service: Data Steward Studio Administration

Please visit the Hortonworks Data Platform page for more information on Hortonworks technology. For more information on Hortonworks services, please visit either the Support or Training page. Feel free to contact us directly to discuss your specific needs.

Hortonworks reserves the right to change any products described herein at any time, and without notice. Hortonworks assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by Hortonworks.

Trademark: Hortonworks DataPlane Service and DPS are trademarks of Hortonworks, Inc. in the United States and other countries. All other trademarks are the property of their respective owners.

Table of Contents

1. About This Guide

The goal of this guide is to provide information and steps required for administering, configuring, using, and troubleshooting Hortonworks DataPlane Service Data Steward Studio, a service that enables users to understand and govern data across enterprise data lakes.

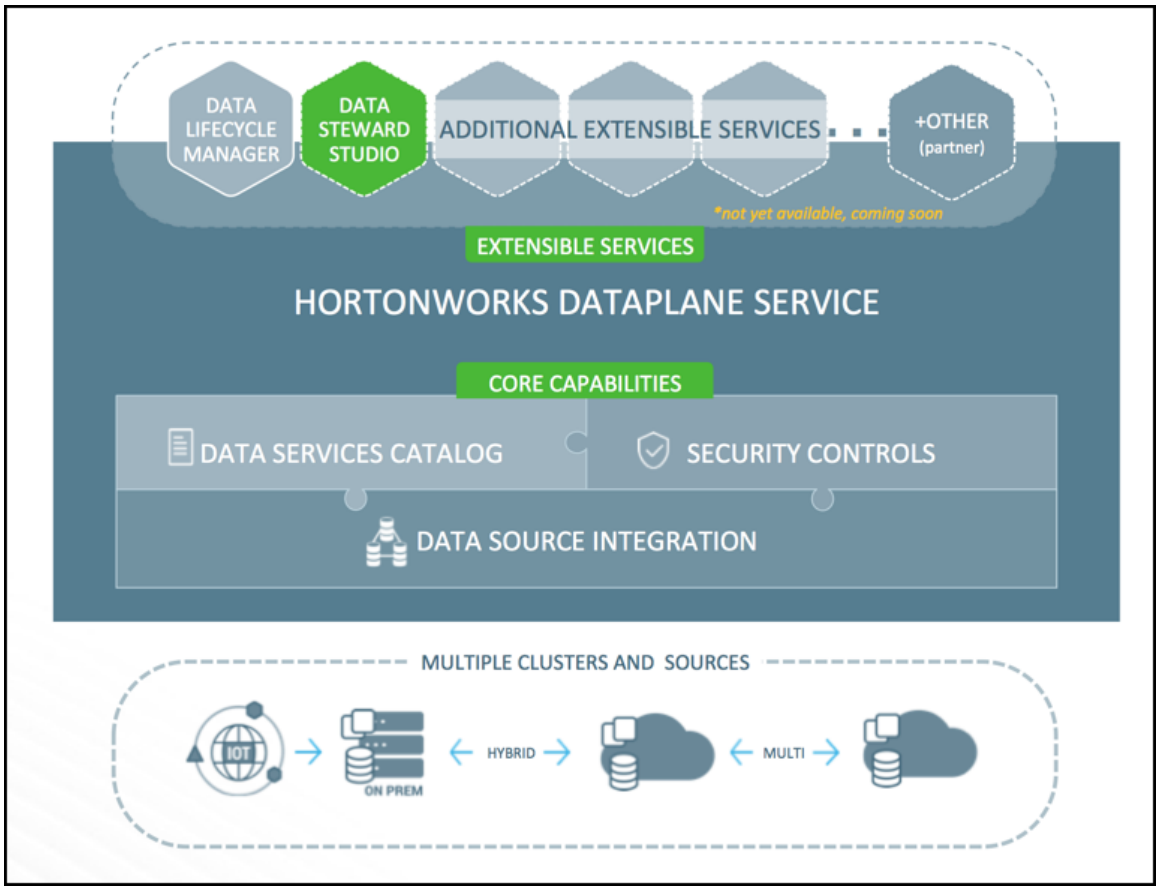
The primary audience of this guide are the administrators and users of Data Steward Studio (DSS).

What's in this guide:

- Overview of DSS features, functionality, and concepts
- Data Management: Data Assets and Asset Collections
- Troubleshooting help

2. Overview

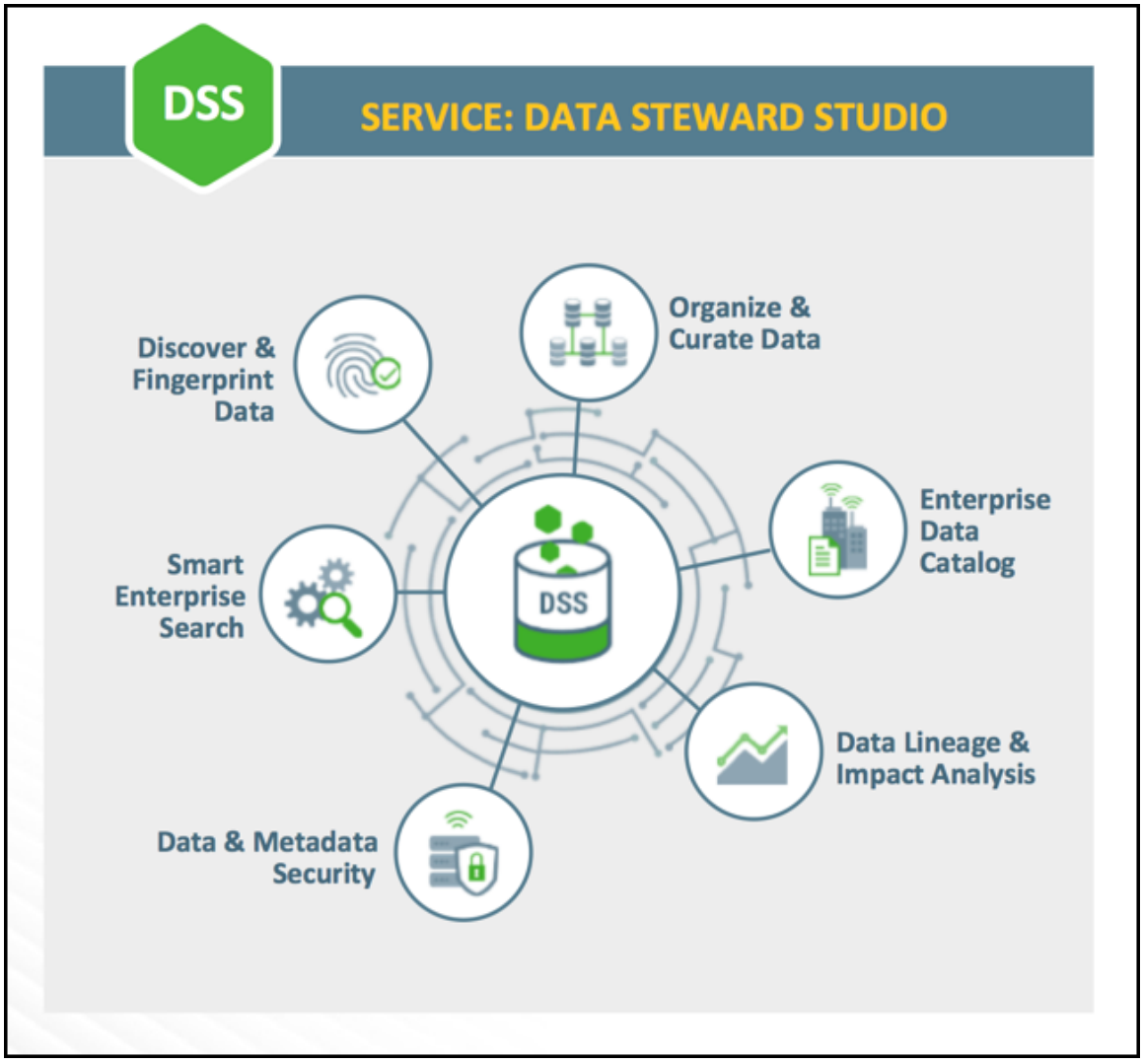
Hortonworks DataPlane Service (DPS™) is a common set of services to manage, secure, and govern data assets across multiple tiers and types. It does this for data at rest and in multiple clusters and tiers (on-premises, cloud, and to the point of origin (IOT)).



Data Steward Studio (DSS) is one of several services available for Hortonworks DataPlane Service; it provides a suite of capabilities that allows users to understand and govern data across enterprise data lakes.

2.1. Data Steward Studio Overview

Data Steward Studio (DSS) is one of several services available for Hortonworks DataPlane Service; it provides a suite of capabilities that allows users to understand and govern data across enterprise data lakes.



The goal of the Data Steward Studio is the help data stewards across the enterprise to:

- Organize and curate data globally
  - Organize data based on business classifications, purpose, protections needed, etc.
  - Promote responsible collaboration across enterprise data workers
- Understand where relevant data is located
  - Cataloging and searching to locate relevant data of interest (sensitive data, commonly used, high risk data, etc.)
- Understand how data is interpreted for use
  - Basic descriptions: Schema, classifications (business cataloging), encodings
  - Statistical models and parameters
  - User annotations, wrangling scripts, view definitions etc.

- Understand how data is created and modified
  - Visualize upstream lineage and downstream impact
  - Understand how schema or data evolve
  - View and understand data supply chain (pipelines, versioning, evolution)
- Understand how data access is secured/protected and audit usage
  - Understand who can see which data and metadata (e.g. based on business classifications) under what conditions (security policies, data protection, anonymization)
  - View who has accessed what data from a forensic audit/compliance perspective
  - Visualize access patterns and identify anomalies

For a list of features that are available in this release, see Evaluation Software Features [4].

2.2. Evaluation Software Features

The following capabilities are available in this Tech Preview:

The technical preview of Data Steward Studio will provide organization and curation capabilities for creating Data Asset Collections based on Hive Tables. Additionally, Data Steward Studio provides an Asset 360 View which has a comprehensive picture of various types of metadata for Hive tables (E.G., technical properties, lineage, security policies, classifications or tags, schema or structure.) Additionally, users can drill down into a view that shows statistical summaries of data in the columns contained within the Hive tables as well as summarized views of the usage of a Hive table from access and audit logs.

Data Steward Studio 1.0 is a Hortonworks Evaluation Software. Data Steward Studio is available within DataPlane, but is not ready for production deployment. Hortonworks encourages you to explore this technical preview release in non-production environments and provide feedback on the technical preview via your Hortonworks support channels or Hortonworks Community Forums.

2.3. Understanding Asset Collections

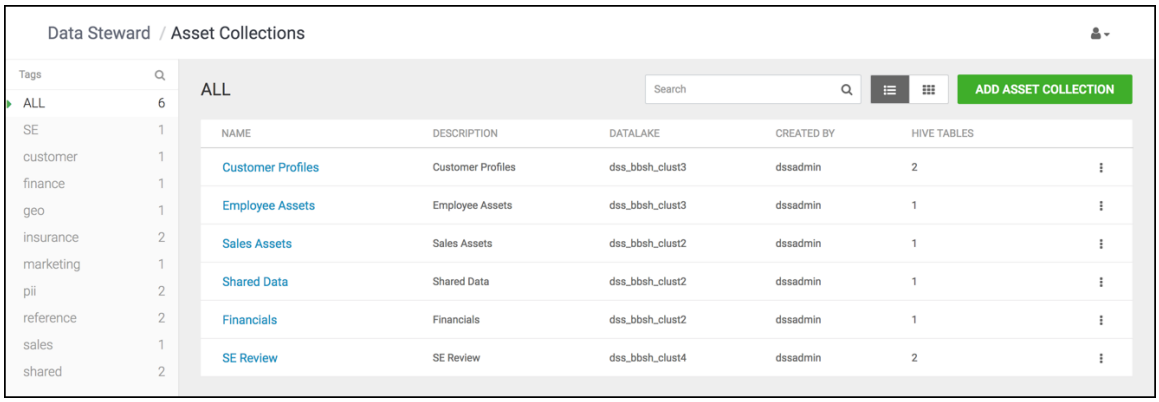
An asset collection is a list of assets that have been grouped and assigned unique search criteria by a data steward for purposes of management and administration.

With Asset Collections, you can:

- **Organize**: Group data assets into asset collections based on business classifications, purpose, protections, relevance, etc.
- **Search**: Find tags or assets in your data lake using Hive assets, attribute facets, or free text.

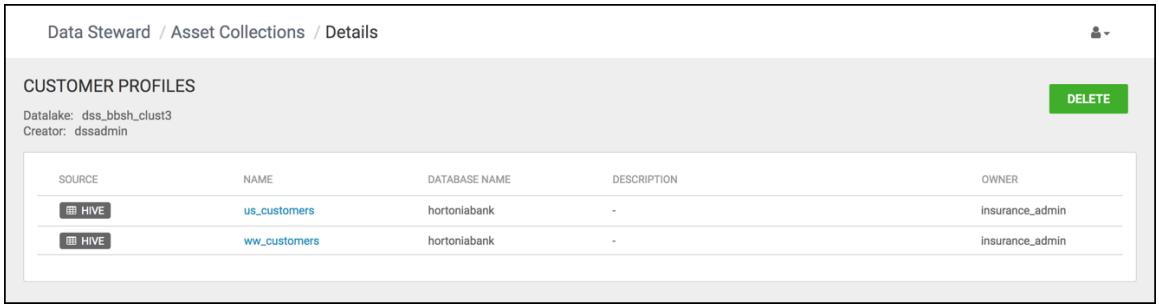
- **Summarize**: View personalized dashboards with an overview of data assets within an asset collection.
- **Understand**: Audit data asset security and use for anomaly detection, forensic audit/ compliance & proper control mechanisms

Asset Collections are immutable once created. Once an Asset Collection has been created by running a query filter, the assets contained within it are not updated when subsequent changes (such as additional assets being added or deleted) that pass the filtering criteria will not be updated in the Asset Collection.



2.4. Asset 360 View: Understanding Data Assets

A data asset is a specific instance of a data type, including the related attributes and metadata. A data asset is a physical asset located in the Hadoop ecosystem such as a Hive tablewhich contains business or technical data. A data asset is a specific instance of a given type and includes relevant attributes and metadata. An asset can belong to only one asset collection. Data assets are also known as “entities” in Apache Atlas.



2.5. Understanding the DSS Profiler

DSS includes a profiler engine that can run various data profiling operations as a pipeline on data located in multiple data lakes. The profiler agent is installed in a data lake and can be set up on a specific schedule to generate various types of data profiles that create metadata annotations that summarizes the content and shape characteristics for data assets.

The technical preview version of DSS will include two built-in profilers: a Hive column univariate statistical profiler and a Ranger audit log summarizer. Both of these profiler agents are included as part of DSS and must be installed on the cluster.

When an Asset Collection is created, all data assets in that collection are added to a scheduler in the profiler backend. You cannot manually trigger the Profiler to run; the schedule is hardcoded to run once every hour.

2.6. DSS Terminology

For a complete list of DataPlane terminology, see: Hortonworks DataPlane Service Terminology.

Hortonworks DataPlane Service (HDS)	The family of components that include the Core service platform and all services that plug into it.
Data Center	The facility that contains the computer, server, and storage systems and associated infrastructure, such as routers, switches, and so forth. Corporate data is stored, managed, and distributed from the data center. In an on-premise environment, a data center hosts one or more Hadoop clusters.
Profiler	Enables the data steward to gather and view information about different relevant characteristics of data such as shape, distribution, quality, and sensitivity which is important to understand and use the data effectively. For example, view the distribution between males and females in column “Gender”, or min/max/mean/null values in a column named “avg_income”. Profiled data is generated on a periodic basis from the profilers, which run at regularly scheduled intervals. Works with data sourced from Apache Ranger Audit Logs, Apache Atlas Metadata Store, and Hive.
Data Lake	A trusted and governed data repository that stores, processes, and access to many kinds of enterprise data to support data discovery, data preparation, analytics, insights, and predictive analytics. In the context of Hortonworks DPS, a data lake can be realized in practice with an Apache Ambari managed Hadoop cluster that runs Apache Atlas for metadata and governance services, and Apache Knox and Apache Ranger for security services.
Data Asset	A data asset is a physical asset located in the Hadoop ecosystem such as a Hive table which contains business or technical data. A data asset could include a specific instance of an Apache Hive database, table, or column. An asset can belong to only one asset collection. Data assets are equivalent to “entities” in Apache Atlas.
Asset Collection	Asset collections allow users of DSS to manage and govern various kinds of data objects as a single unit through a unified interface. Asset collections help organize and curate information about many assets

based on many facets including data content and metadata, such as size/schema/tags/alterations, lineage, and impact on processes and downstream objects in addition to the display of security and governance policies.

The content of an asset collection is a static list that can only be modified by a user. So, adding new assets to a collection must be done manually.

2.7. Supported Configurations

See the DataPlane Service Support Matrices.

3. DSS Data Management

You can create and delete Asset Collections (Managing Asset Collections [8]) and view information about Data Assets contained in an Asset Collection (Viewing Data Assets (Asset 360) [11]).

3.1. Managing Asset Collections

You can Create Asset Collections [8] and Delete Asset Collections [10].

You cannot edit Asset Collections after you have created them. You would delete the Asset Collection, and create it again.

3.1.1. Create Asset Collections

About this task

You can group data assets into Asset Collections. This enables you to organize data based on business classifications, purpose, protections needed, etc.

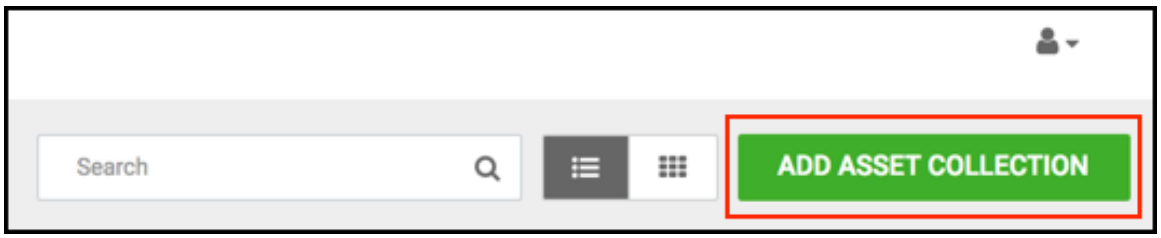
Note

You cannot edit Asset Collections after you have created them. You would delete the Asset Collection, and create it again.

Steps

From the Asset Collection page:

1. Click the **Add Asset Collection** button.



The **Add** page opens.

2. Fill in the following information:

Field Name	Description	Example Value
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

Field Name	Description	Example Value
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

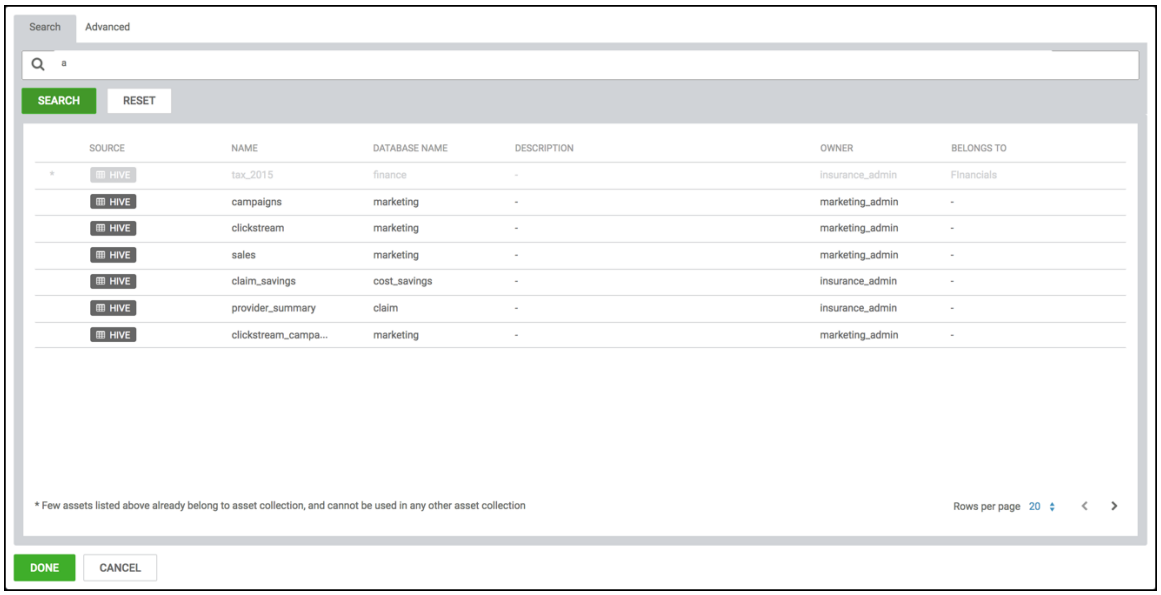
3. Add assets to the Asset Collection:

   a. Click **Add Assets**.

      The **Asset Search** window opens.

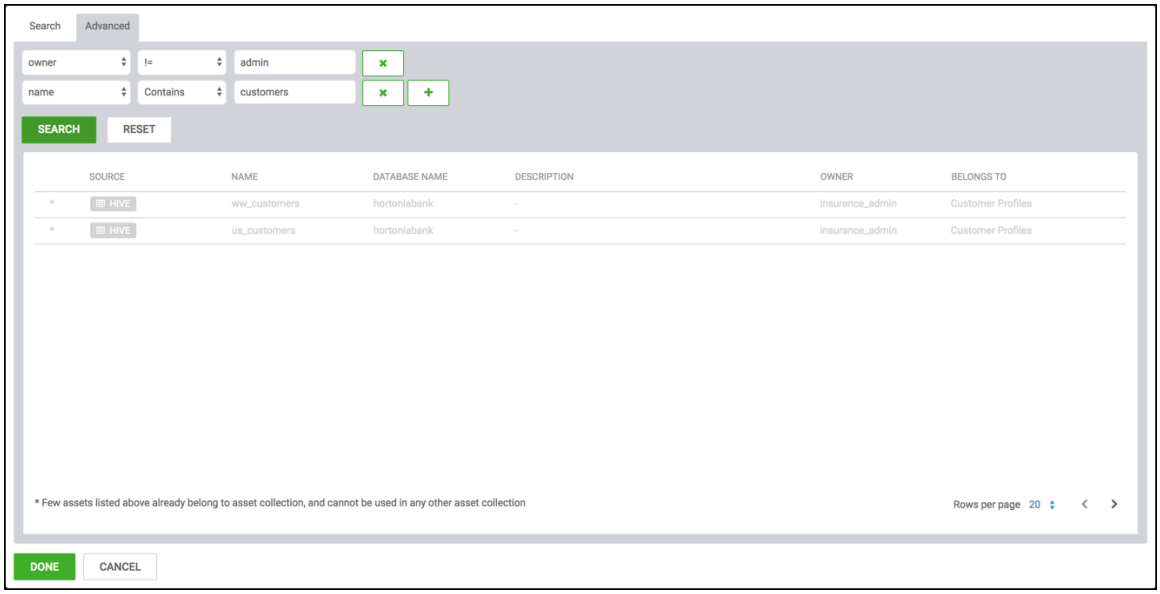
   b. Search for assets using Basic or Advanced search.

      Basic Search



      Advanced Search

      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.



c. Click **Done**.

   All assets that display from the search result will be selected, with the exception of greyed out assets; assets can only belong to one Asset Collection at a time.

d. Click **Next**.

   The **Summary** page opens.

4. Click **Save**.

   You are returned to the **Asset Collections** home page.

3.1.2. Delete Asset Collections

About this task

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.

You might want to delete an Asset Collection if you no longer need to track that Asset Collection, or if you wish to reassign assets to another Asset Collection.

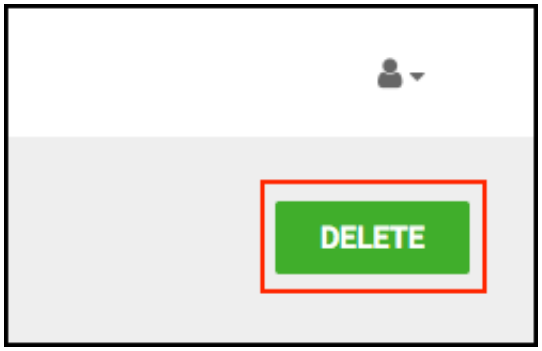
Steps

From the `Data Steward>Asset Collections` page:

1. Click on the name of the Asset Collection you wish to delete.

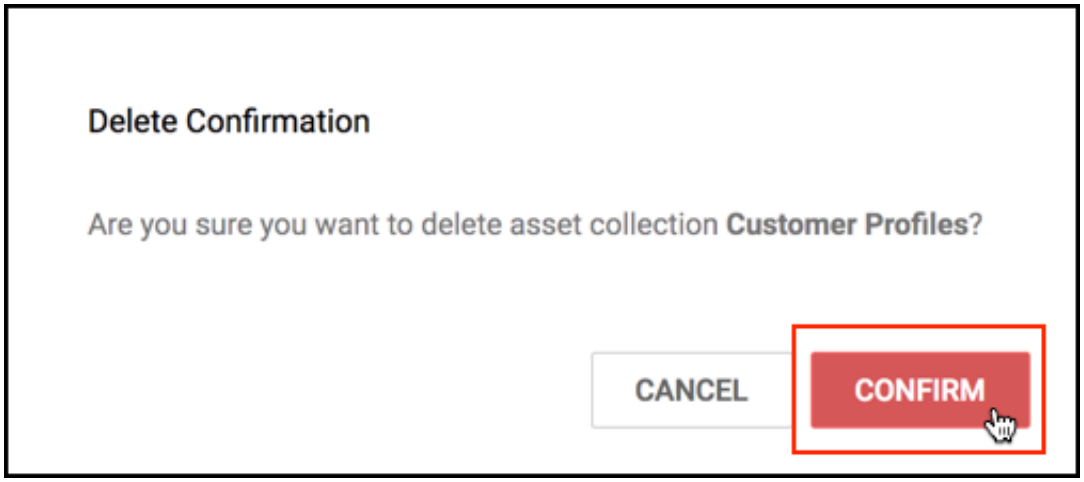
   The **Asset Collection Details** page opens.

2. Click the **Delete** button.



A confirmation window opens.

3. Click **Confirm**.



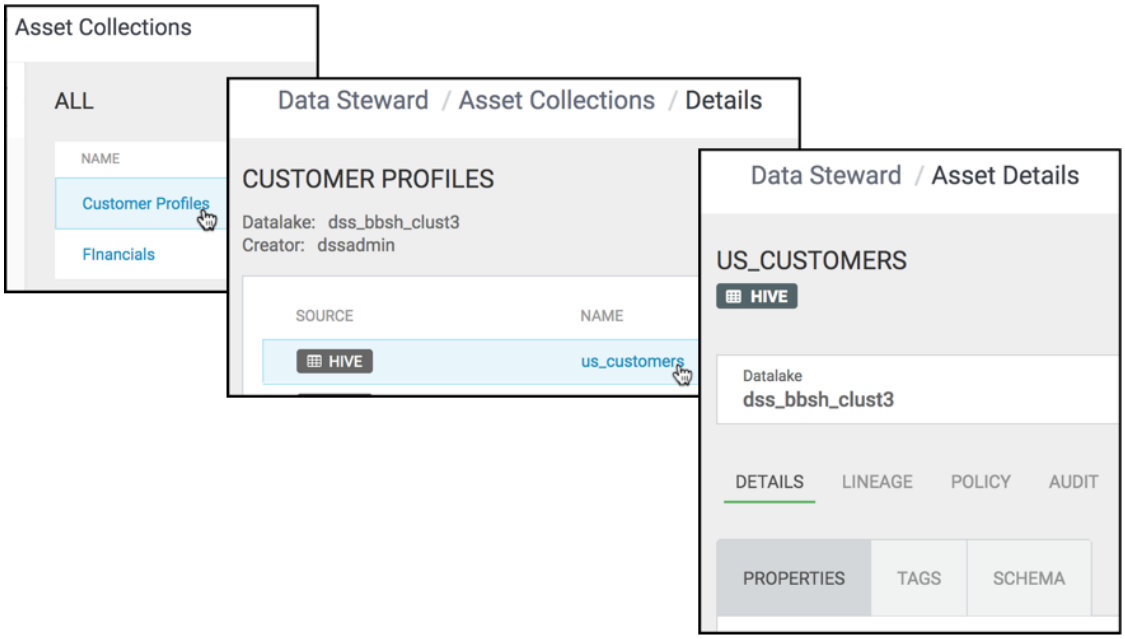
You are returned to the **Asset Collections** home page.

3.2. Viewing Data Assets (Asset 360)

The Asset 360 page is a centralized hub providing the following information related to an asset:

- Data Asset Details: table properties (View Data Asset Properties [12]), tags from Apache Atlas (View Data Asset Tags [14]), and schema details (View Data Asset Schema [15]).
- View Data Asset Lineage [17]: The lineage view 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.
- View Authorization Policies on a Data Asset [19]: 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
- View Data Asset Audit Logs [21]: Audit View 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.

The Asset 360 page can be accessed from Data Steward>Asset Collections>(click the name of one Asset Collection)>(click the name of one asset).



3.2.1. View Data Asset Properties

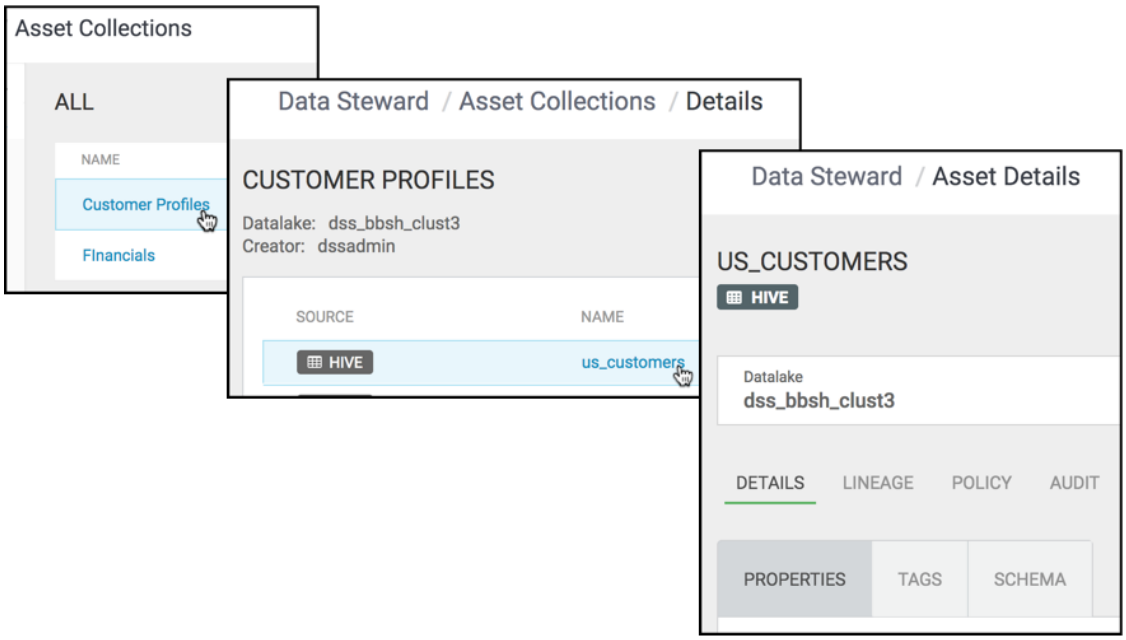
About this task

The Asset 360 page contains a personalized dashboard with an overview of data assets within an asset collection. You can view all the Apache Atlas metadata associated with a particular data asset on the Asset 360 page>Properties page.

Steps

To view data asset properties:

1. From Data Steward>Asset Collections, select one Asset Collection, then select one data asset.



The **Asset 360** window opens.

2. Under the **Details** tab, click **Properties**.

The **Properties** table shows the data asset metadata properties as retrieved from Apache Atlas.

3.2.2. View Data Asset Tags

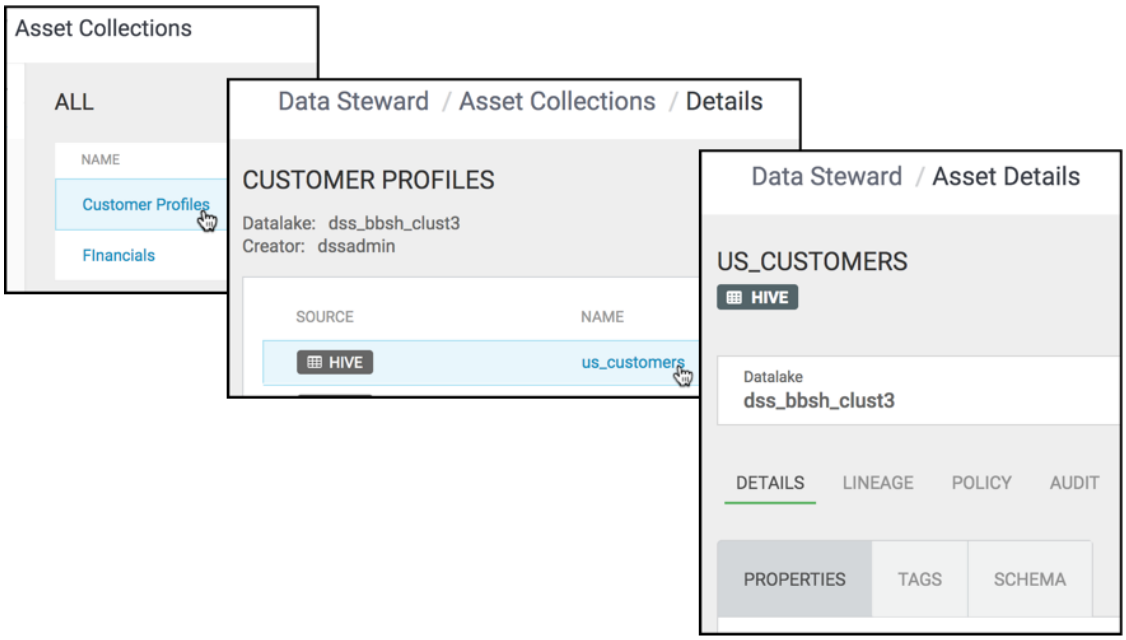
About this task

The Asset 360 page contains a personalized dashboard with an overview of data assets within an asset collection. You can view all the classifications associated with the asset via Apache Atlas on the Asset 360 page>Tags page.

Steps

To view data asset tags details:

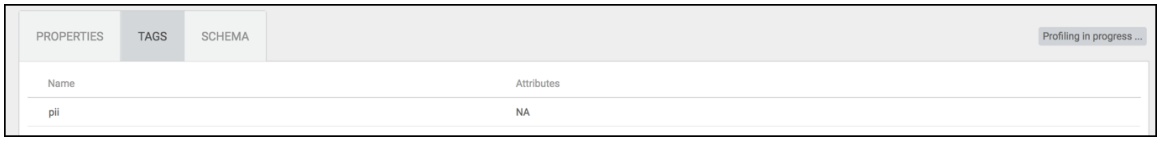
1. From Data Steward>Asset Collections, select one Asset Collection, then select one data asset.



The **Asset 360** window opens.

2. Under the **Details** tab, click **Tags**.

The **Properties** table shows the data asset tags as retrieved from Apache Atlas.



3.2.3. View Data Asset Schema

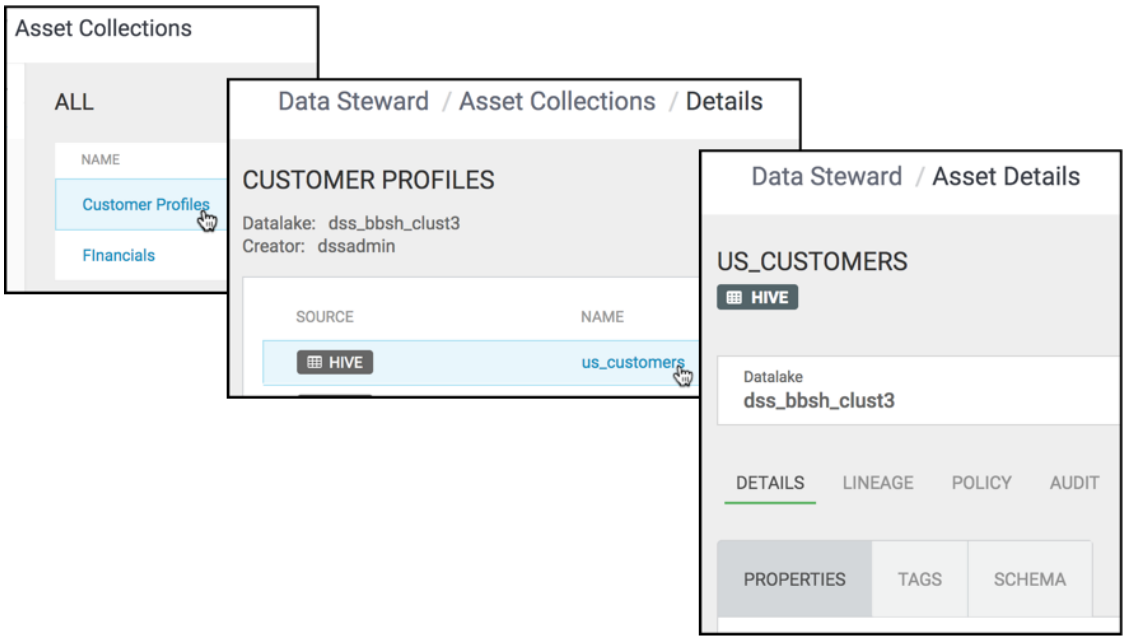
About this task

The Asset 360 page contains a personalized dashboard with an overview of data assets within an asset collection. From the Asset 360 page>Schema page, you can view the schema of the data asset for structured data (such as Hive tables) from the relevant metadata repositories (such as Atlas). You can also view the shape or distribution characteristics of the columnar data within a schema based on the Hive column profiler. By providing statistical models and parameters, this helps you understand how data is interpreted for use.

Steps

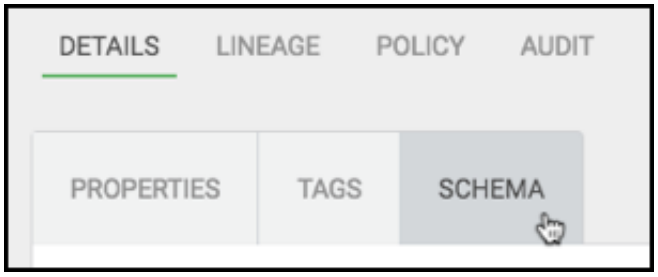
To view data asset schema details:

1. From Data Steward>Asset Collections, select one Asset Collection, then select one data asset.

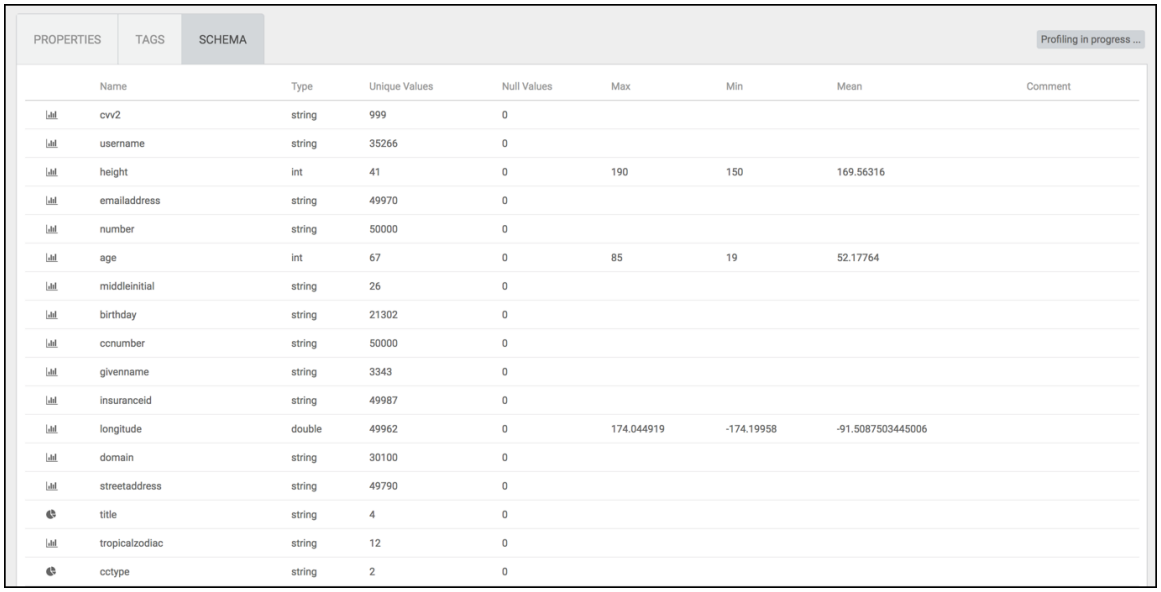


The **Asset 360** window opens.

2. Under the **Details** tab, click **Schema**.



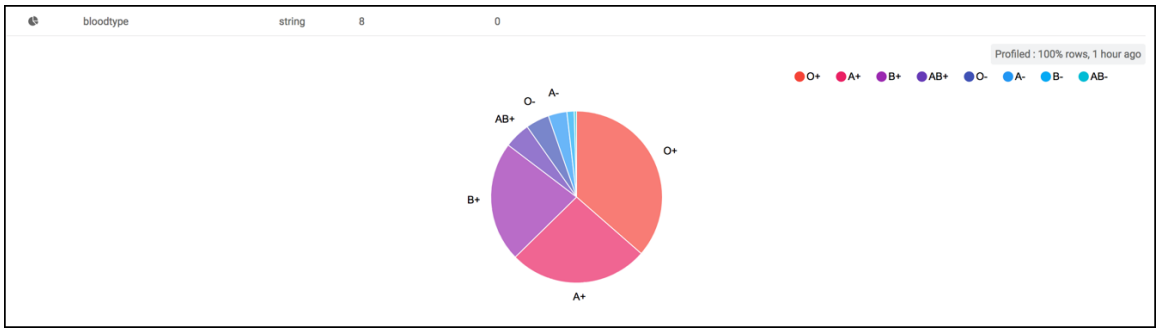
The **Schema** table shows the data asset schema as retrieved from Apache Atlas.



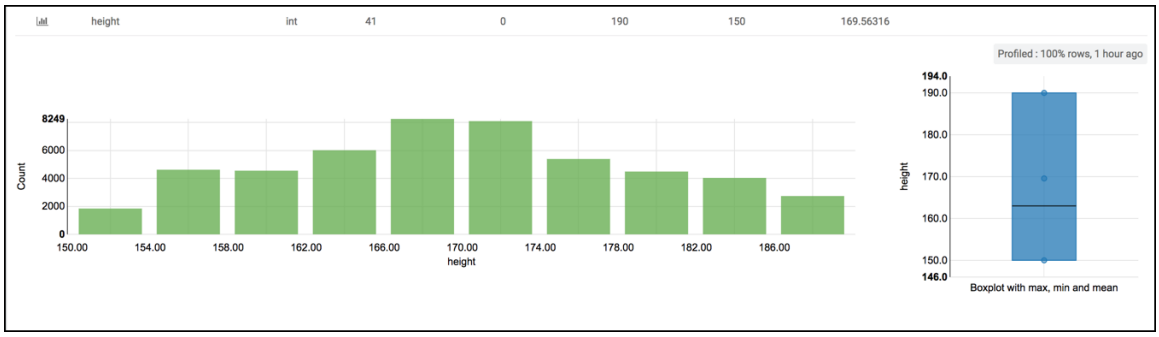
Name	Type	Unique Values	Null Values	Max	Min	Mean	Comment
cvv2	string	999	0				
username	string	35266	0				
height	int	41	0	190	150	169.56316	
emailaddress	string	49970	0				
number	string	50000	0				
age	int	67	0	85	19	52.17764	
middleinitial	string	26	0				
birthday	string	21302	0				
ccnumber	string	50000	0				
givenname	string	3343	0				
insuranceid	string	49987	0				
longitude	double	49962	0	174.044919	-174.19958	-91.5087503445006	
domain	string	30100	0				
streetaddress	string	49790	0				
title	string	4	0				
tropicalzodiac	string	12	0				
cctype	string	2	0				

3.2.3.1. Graph Types, Uses, and Examples

The Asset 360 page contains a personalized dashboard with an overview of data assets within an asset collection. You can view all the Apache Atlas schema details associated with a particular data asset on the Asset 360 page>Schema page. There are different charts available to help visualize the shape and distribution of the data within the column as well as summary statistics (such as means, null count, cardinality of the data) contained within a column.



Pie charts are presented for categorical data with limited number of categories or classes. Examples include data such as blood group, gender, and titles.



When the data within columns is numeric, a histogram of the distribution of values binned into 10 groups (decile frequency histogram) and a box plot with a 5-number summary (mean, median, quartiles, max, and min values).

3.2.4. View Data Asset Lineage

About this task

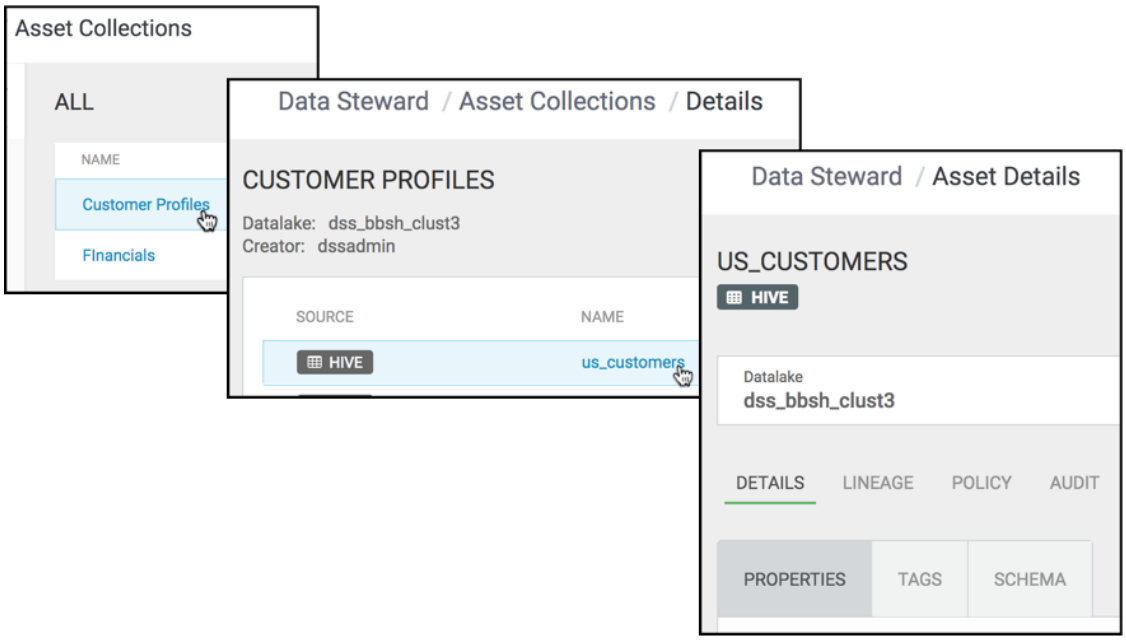
The Asset 360 page contains a personalized dashboard with an overview of data assets within an asset collection. The Asset 360 page>Lineage page shows the chain of custody for the data from relevant metadata repositories such as Apache Atlas. This helps you understand how data is created, modified, and evolves. It also helps you view and understand data supply chain (pipelines, versioning, evolution)

The lineage view 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.

Steps

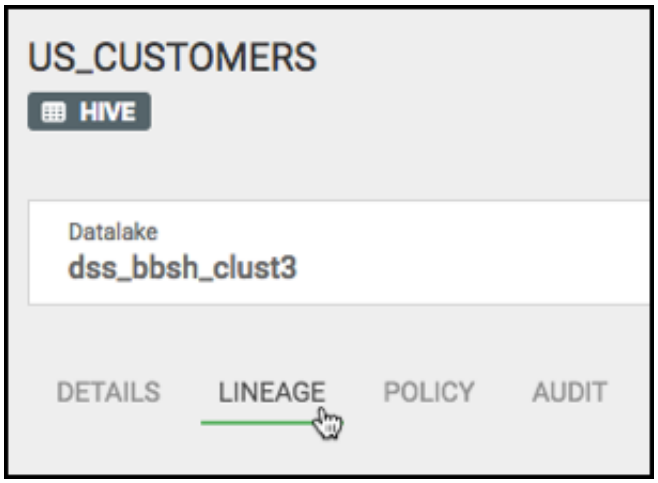
To view data asset lineage:

1. From Data Steward>Asset Collections, select one Asset Collection, then select one data asset.



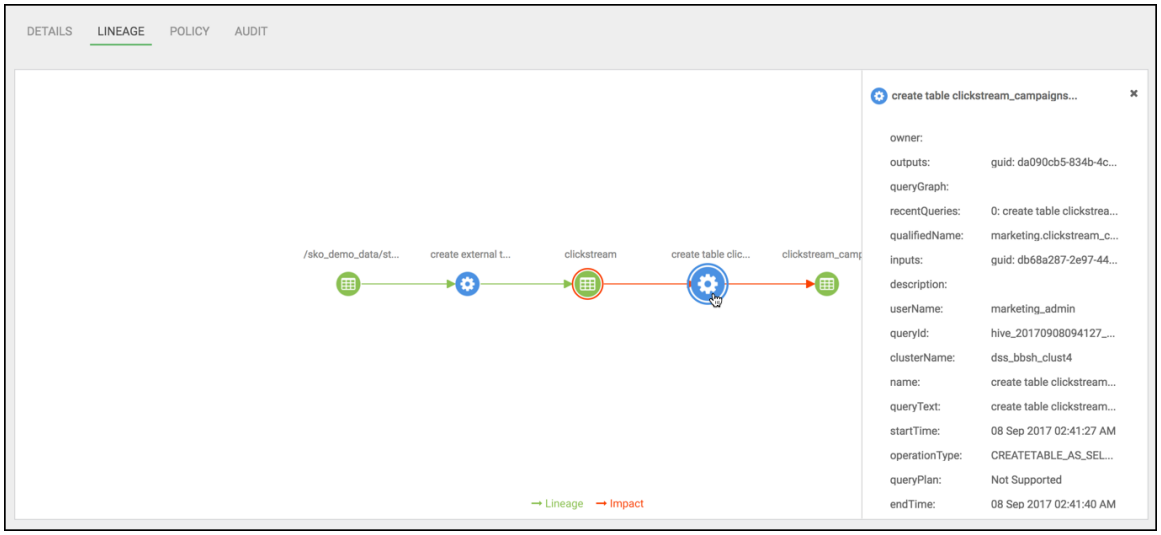
   The **Asset 360** window opens.

2. Click **Lineage**.



   The **Lineage** table shows the data asset lineage as retrieved from Apache Atlas.

3. To view lineage details, click on any process or entity node.



3.2.5. View Authorization Policies on a Data Asset

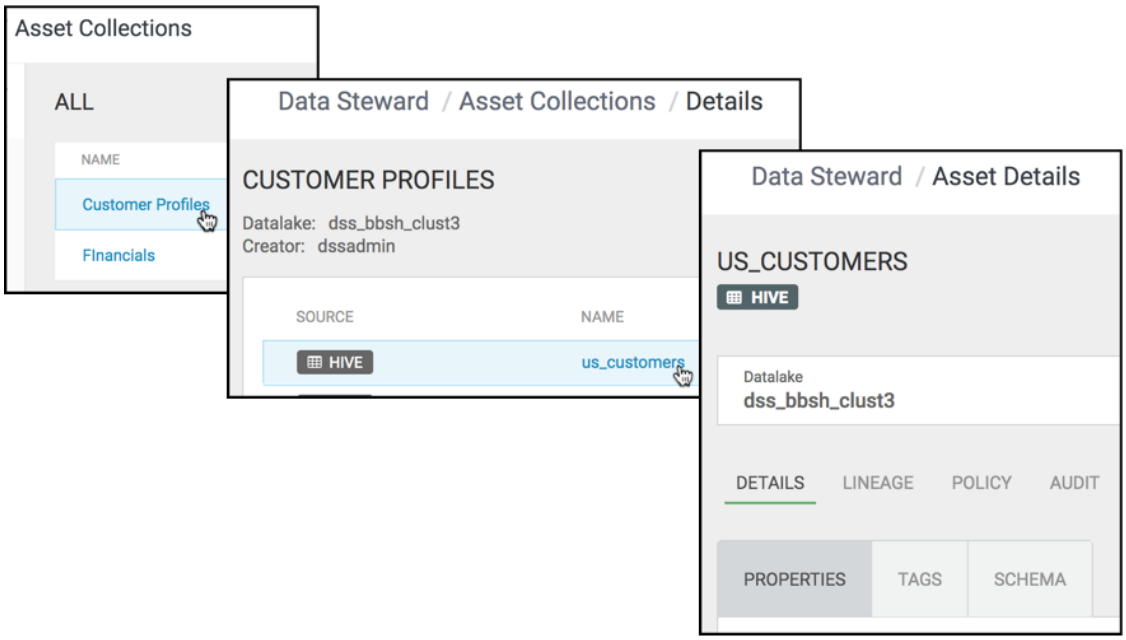
About this task

The Asset 360 page contains a personalized dashboard with an overview of data assets within an asset collection. You can view all the Apache Ranger policy details associated with a particular data asset on the Asset 360 page>Policy page. 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).

Steps

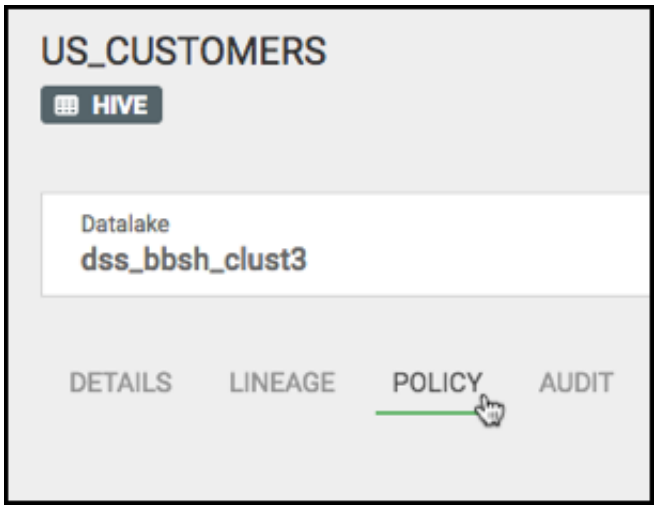
To view data asset lineage:

1. From Data Steward>Asset Collections, select one Asset Collection, then select one data asset.

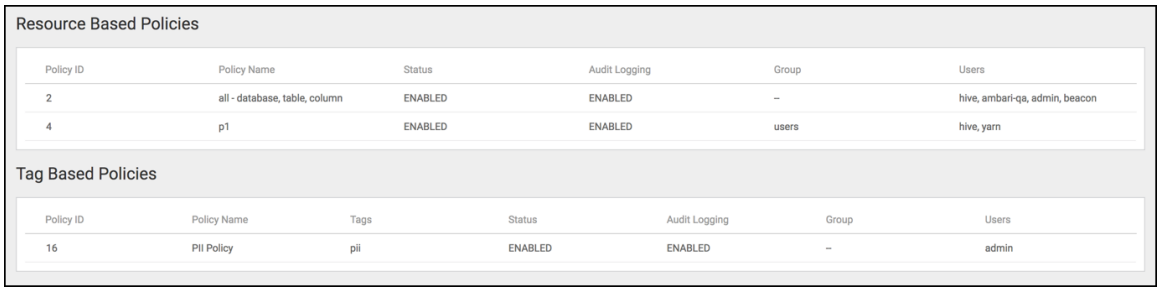


The **Asset 360** window opens.

2. Click **Policy**.



The **Policy** table shows the data asset policies as retrieved from Apache Ranger.



3.2.6. View Data Asset Audit Logs

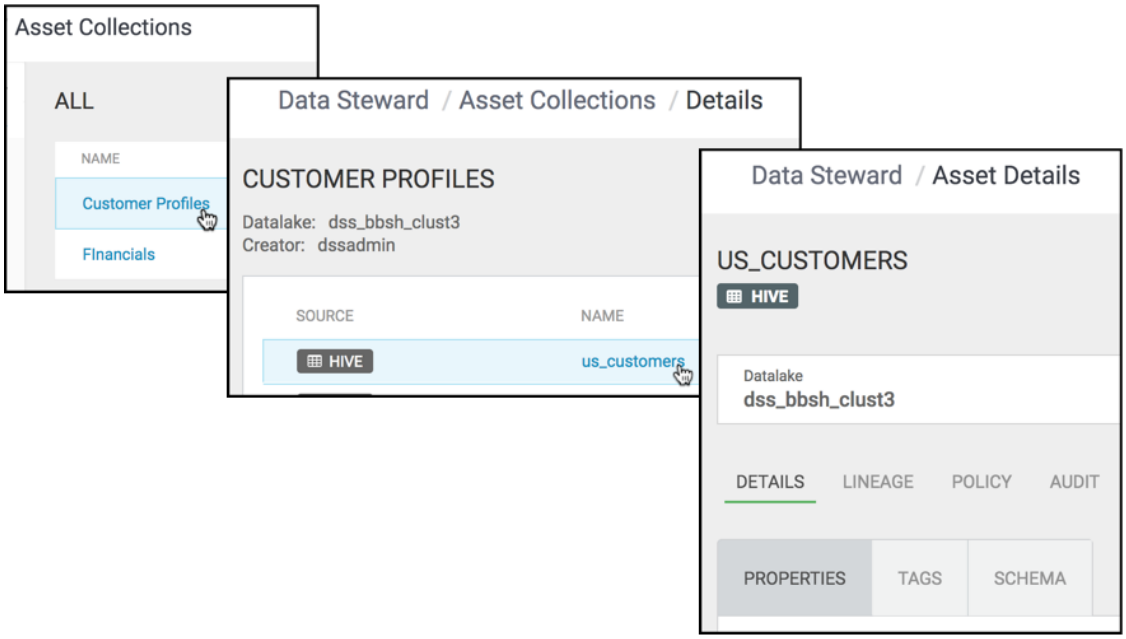
About this task

The Asset 360 page contains a personalized dashboard with an overview of data assets within an asset collection. You can view all the Apache Ranger audit events associated with a particular data asset on the Asset 360 page>Audit page. This helps you view who has accessed what data from a forensic audit/compliance perspective, and visualize access patterns and identify anomalies.

Steps

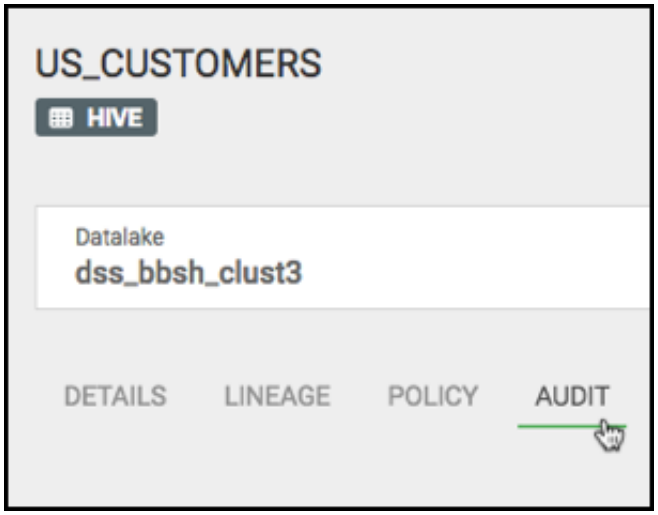
To view data asset audit logs:

1. From Data Steward>Asset Collections, select one Asset Collection, then select one data asset.

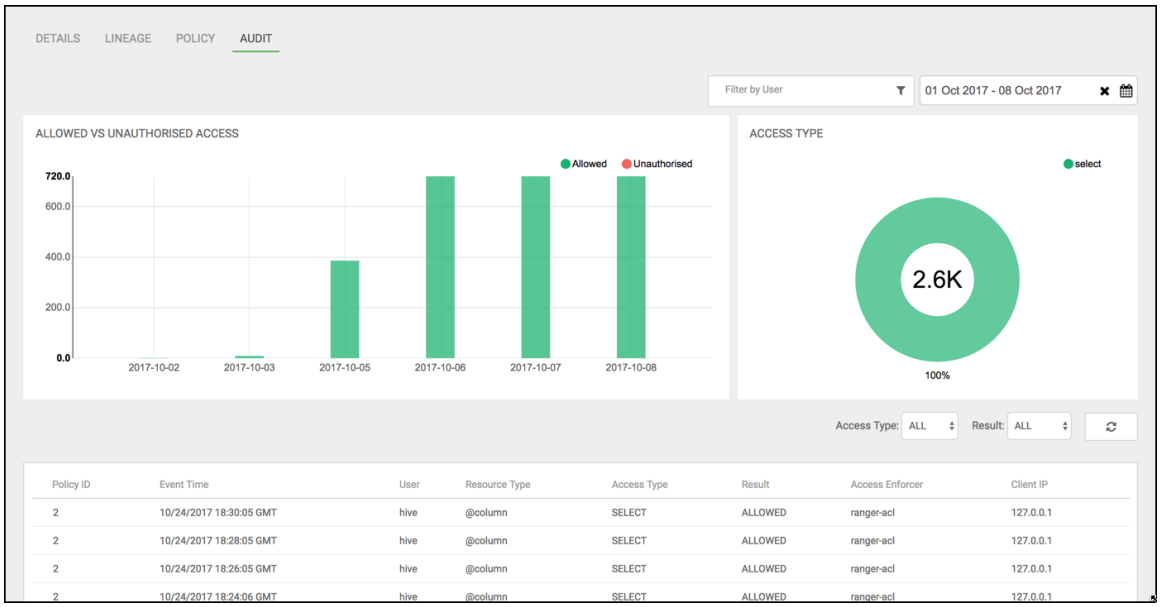


   The **Asset 360** window opens.

2. Click **Audit**.



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.



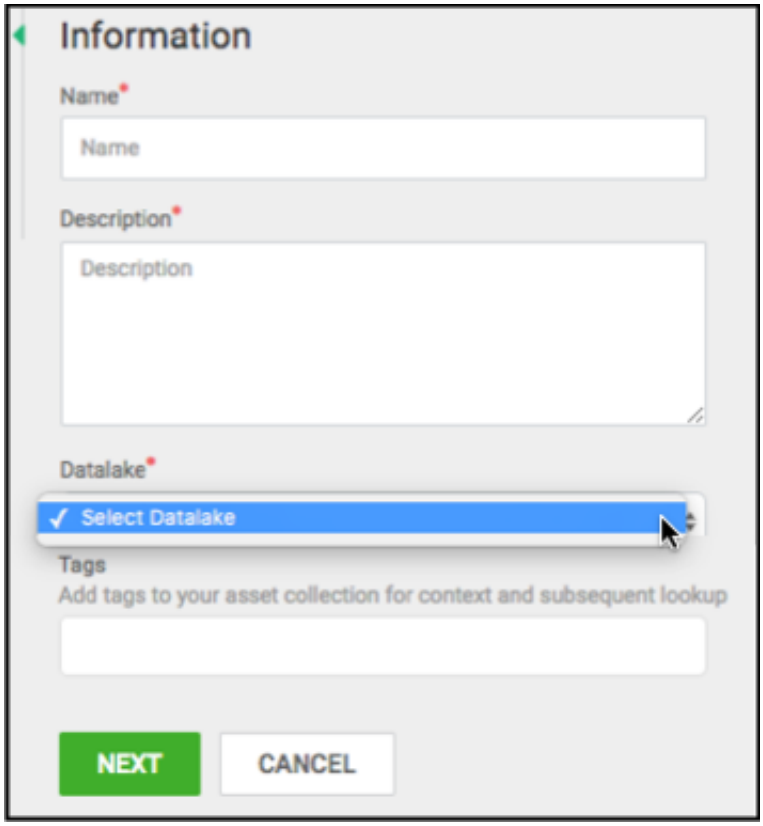
4. Troubleshooting

This chapter contains common issues (with workarounds) and error message help.

4.1. No datalake available when creating an Asset Collection

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

Description: 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 DataPlane, no datalakes will be available.



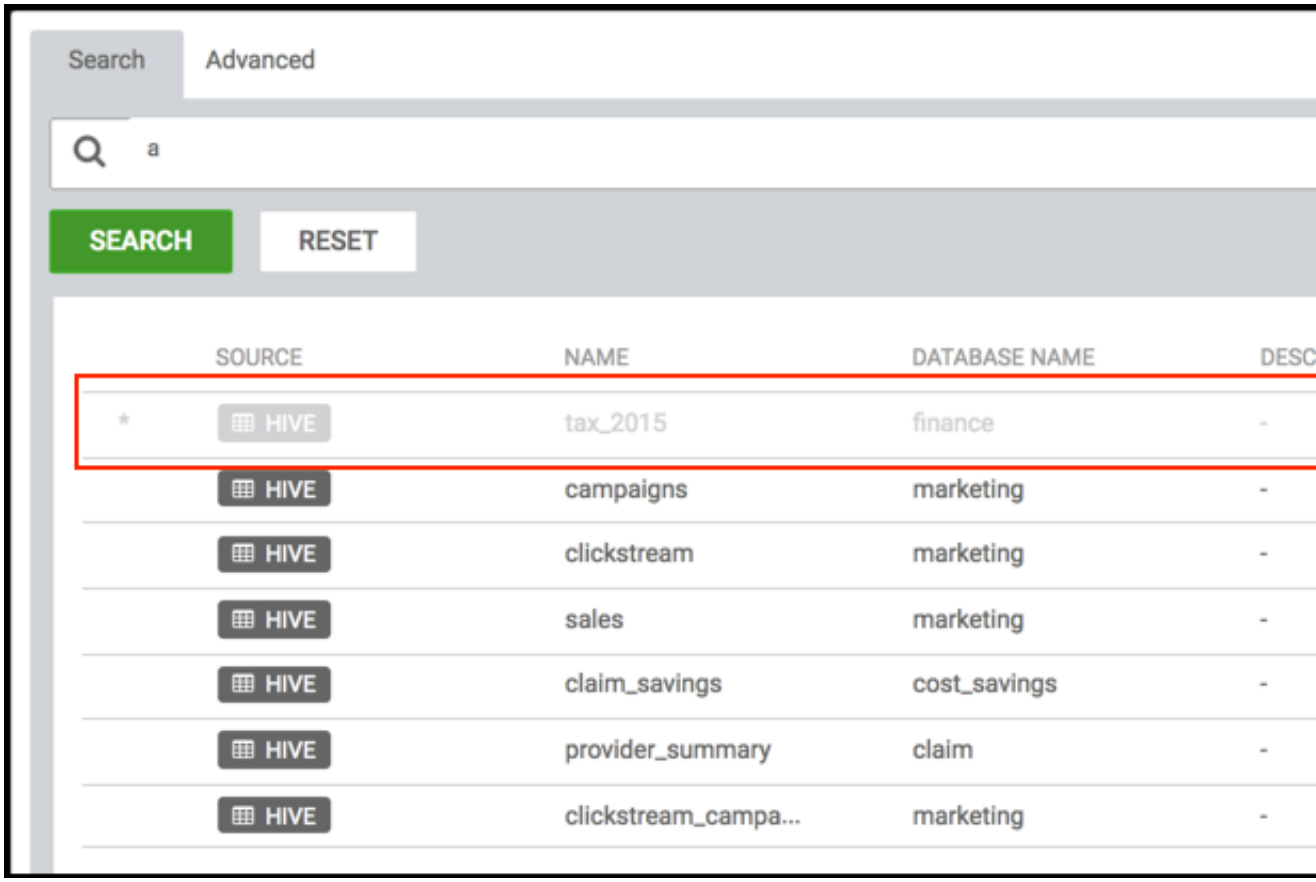
Workaround: You might need to either register the cluster, or install Apache Atlas and Apache Ranger on the cluster:

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

4.2. Data Asset is greyed out

Issue: When assigning Data Assets to an Asset Collection, some data assets are greyed out an unselectable.

Description: Data Assets can only be assigned to one Asset Collection at a time; those that are already assigned are greyed out.



Workaround:

- Delete the Asset Collection the Data Asset belongs to, and re-create the Asset Collections with new assignments.
- Create the Asset Collection without the unavaible Data Assets.

4.3. Edit an Asset Collection

Issue: If you wish to remove or add Data Assets to an Asset Collection, you would like to edit the AC.

Description: You cannot edit Asset Collections after you have created them.

Workaround: Delete the Asset Collection the Data Asset belongs to, and re-create the Asset Collections with new assignments.