

Data Catalog 1.5.2

Data Catalog Overview

Date published: 2023-10-10

Date modified: 2023-11-02

CLOUDERA

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

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

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

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

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

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

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

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

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

Contents

About Data Catalog.....	4
Before you start.....	5
CDP Private Cloud Base cluster requirements.....	5
Prerequisite to access Data Catalog service.....	5
How to access Data Catalog service.....	7
Data Catalog Overview.....	8
Data Catalog Terminology.....	10
Authorization for viewing Assets.....	11
Restricting access for certain users of Data Catalog.....	12
Understanding Datasets.....	14
Understanding Data Assets.....	14
About the Data Catalog Profiler.....	14
Understanding the Cluster Sensitivity Profiler.....	15
Understanding the Hive Column Profiler.....	16
Understanding the Ranger Audit Profiler.....	17

About Data Catalog

Data Catalog is a service within Cloudera Data Platform that enables you to understand, manage, secure, and govern data assets across the enterprise.

Data Catalog helps you understand data lying in your data lake (Private Cloud Base Cluster). You can search to locate relevant data of interest based on various parameters. Using Data Catalog, you can understand how data is interpreted for use, how it is created and modified, and how data access is secured and protected.



Data Catalog enables data stewards across the enterprise to work with data assets in the following ways:

- 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
 - Catalog and search to locate relevant data of interest (sensitive data, commonly used, high risk data, etc.)
 - Understand what types of sensitive personal data exists and where it is located
- Understand how data is interpreted for use
 - View basic descriptions: schema, classifications (business cataloging), and encodings

- View statistical models and parameters
- View 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, and evolution)
- Understand how data access is secured, protected, and audited
 - Understand who can see which data and metadata (for example, based on business classifications) and under what conditions (security policies, data protection, anonymization)
 - View who has accessed what data from a forensic audit or compliance perspective
 - Visualize access patterns and identify anomalies

Related Information

[Data Catalog Terminology](#)

Before you start

Before you access Cloudera Data Catalog on Cloudera Data Platform (CDP) Private Cloud Data Services, you must deploy the Data Catalog service.

Before deploying Data Catalog, make sure you have reviewed and compiled with the requirements in the installation guide for your environment:

Data Catalog is supported on both **OpenShift** and **ECS** clusters. For more information, see:

[Installing on OpenShift](#)

[The Embedded Container Service \(ECS\)](#)

CDP Private Cloud Base cluster requirements

The CDP Private Cloud Base cluster that is used for the Cloudera Data Catalog service.

As Atlas and Ranger services are installed in the Base cluster, Data Catalog communicates with these services to fetch metadata and security policies. Data Catalog Profilers connect with Hive and HDFS services running in the Base cluster for data profiling operations.

For more information, check the following links:

- [Supported Base clusters for Data Services - 7.1.9 / 7.1.9 Cumulative Hotfix 1](#)
- [Supported Cloudera Manager versions - 7.11.3 / 7.11.3 Cumulative Hotfix 1](#)
- [Supported Private Cloud Data Services versions](#)

Prerequisite to access Data Catalog service

To access Data Catalog service, you must have the required credentials.

Follow these instructions to provide the required access to the Data Catalog users.

The PowerUser must provide the requisite access to subscribers who plan to use Data Catalog as per the requirement.



Note: To launch and delete profilers you must be a PowerUser.

There are a couple of roles based on which you can use Data Catalog:

- DataCatalogCspRuleManager
- DataCatalogCspRuleViewer

Using DataCatalogCspRuleManager role, you can create, deploy new Custom Sensitivity Profiler rules, create new Regex, and run validations on newly created rules.

Using DataCatalogCspRuleViewer role, you can list and view the existing Custom Sensitivity Profiler rules.

Additionally, using Cloudera Manager, you must configure Atlas and Ranger services. Use the following instructions to complete the process.

1. Cloudera Manager > Clusters > Atlas > Configuration .
2. Enable the Kerberos Authentication for the Atlas service.

Enable Kerberos Authentication

ATLAS-1 (Service-Wide) 

atlas.authentication.method.kerberos

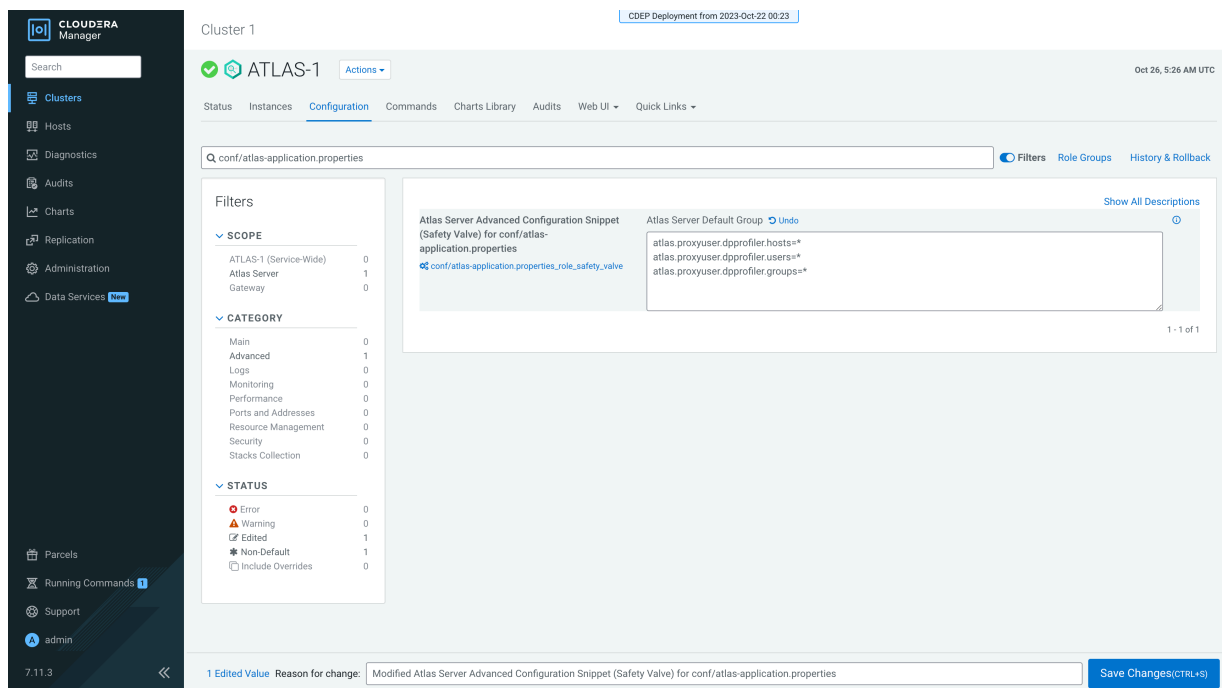
 [kerberos.auth.enable](#)

3. In the search bar, look up the following property:conf/atlas-application.properties_role_safety_valve and enter the values for the Atlas service:

atlas.proxyuser.dpprofiler.hosts=*

atlas.proxyuser.dpprofiler.users=*

atlas.proxyuser.dpprofiler.groups=*



Cluster 1 CDEP Deployment from 2023-Oct-22 00:23

ATLAS-1 Oct 26, 5:26 AM UTC

Status Instances **Configuration** Commands Charts Library Audits Web UI Quick Links

Q conf/atlas-application.properties Filters Role Groups History & Rollback

Filters	Count
SCOPE	
ATLAS-1 (Service-Wide)	0
Atlas Server	1
Gateway	0
CATEGORY	
Main	0
Advanced	1
Logs	0
Monitoring	0
Performance	0
Ports and Addresses	0
Resource Management	0
Security	0
Stacks Collection	0
STATUS	
Error	0
Warning	0
Edited	1
Non-Default	1
Include Overrides	0

Atlas Server Advanced Configuration Snippet (Safety Valve) for conf/atlas-application.properties

Atlas Server Default Group [Undo](#)

```
atlas.proxyuser.dpprofiler.hosts=*
atlas.proxyuser.dpprofiler.users=*
atlas.proxyuser.dpprofiler.groups=*
```

1-1 of 1

1 Edited Value Reason for change: Modified Atlas Server Advanced Configuration Snippet (Safety Valve) for conf/atlas-application.properties Save Changes(CTRL+S)

4. For the Ranger configuration updates, go to Cloudera Manager > Clusters > Ranger > Configuration .

- In the search bar, look up the following property: `conf/ranger-admin-site.xml` and enter the values for the Ranger service:

Name: `ranger.proxyuser.dpprofiler.hosts`

Value: *

Name: `ranger.proxyuser.dpprofiler.users`

Value: *

Name: `ranger.proxyuser.dpprofiler.groups`

Value: *



Note: You can add new rows using the + icon.

The screenshot shows the Cloudera Manager Configuration page for Ranger Admin. The search bar contains the query `conf/ranger-admin-site.xml`. The configuration table is as follows:

Name	Value	Description	Final
<code>ranger.proxyuser.dpprofiler.hosts</code>	*		<input type="checkbox"/>
<code>ranger.proxyuser.dpprofiler.users</code>	*		<input type="checkbox"/>
<code>ranger.proxyuser.dpprofiler.groups</code>	*		<input type="checkbox"/>

Later, restart Atlas and Ranger services respectively.

How to access Data Catalog service

Before you commence using Data Catalog service note that when the Data Services installation is complete, the CDP Private Cloud Base cluster registration is automatically enabled and Data Catalog as a service is installed on the ECS cluster.

You can navigate to the Data Catalog service by clicking on the CDP homepage on your CDP console.



Data Catalog is supported on both **OpenShift** and **ECS** clusters. For more information, see:

[Installing on OpenShift](#)

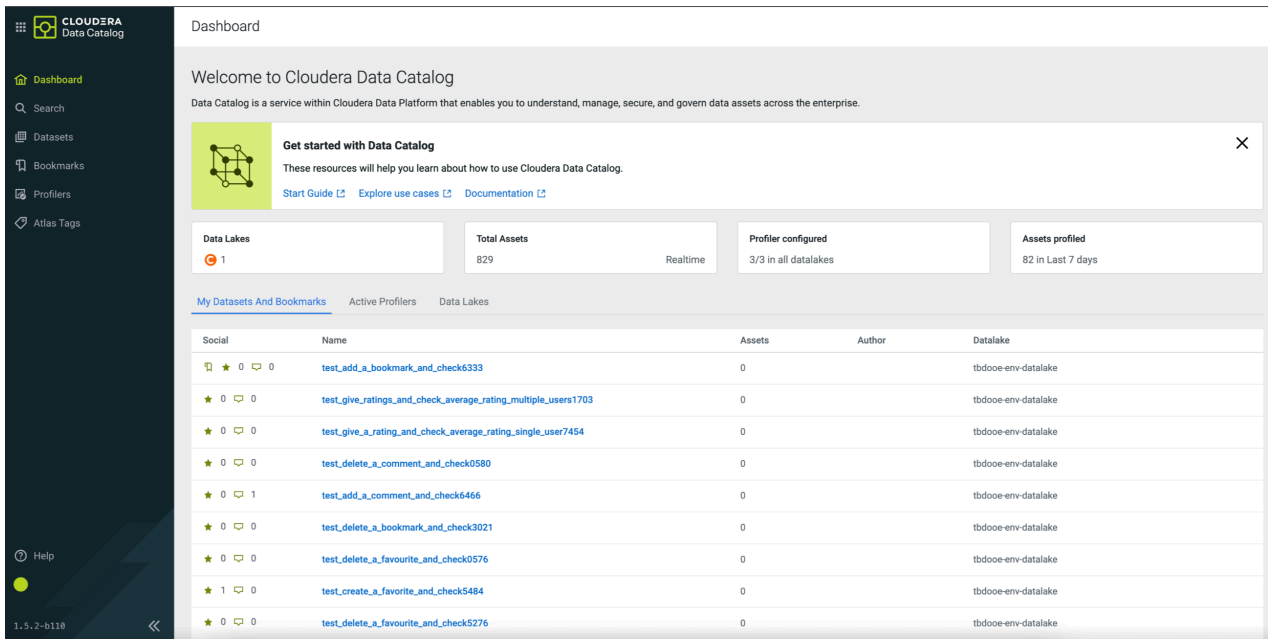
[The Embedded Container Service \(ECS\)](#)

[About CDP Private Cloud Data Services](#)

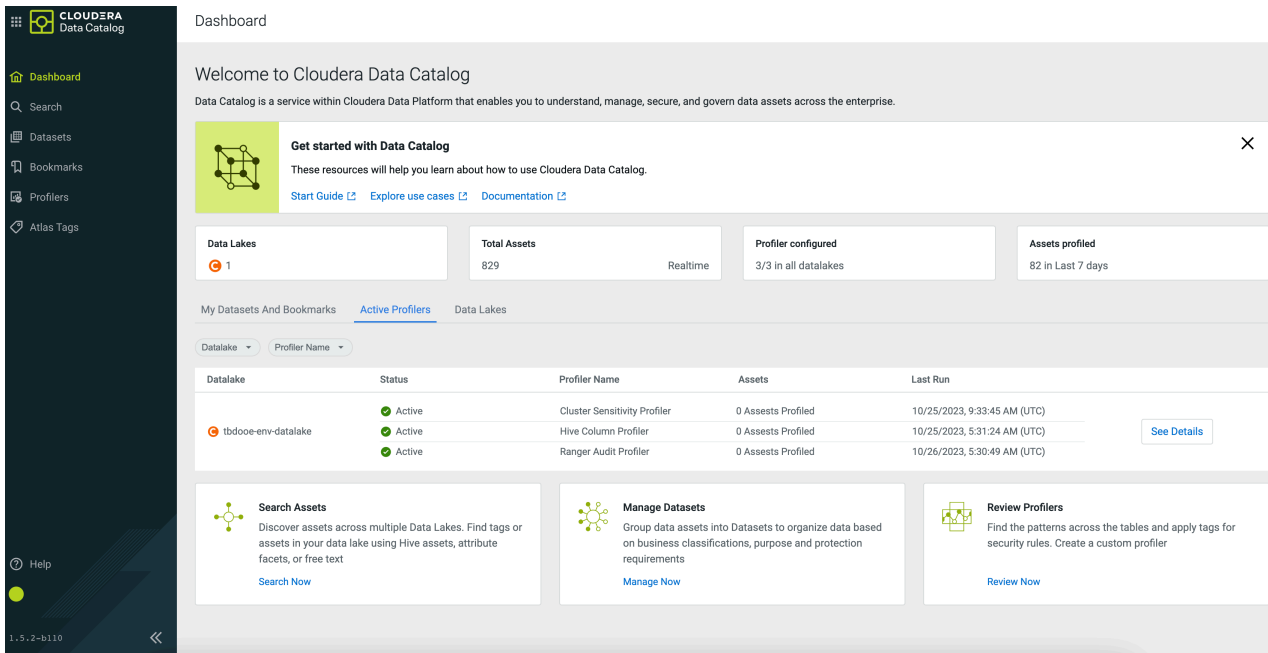
Data Catalog Overview

Overview layout provides quick access to vital service information at a glance, in the form of visual, actionable navigation for multiple operations. The user-friendly navigation enables viewing, filtering, and acting upon data quickly and in a simple manner.

Data Stewards can view the homepage at a glance, and also focus on the most important tasks, enabling faster decision making as well as immediate action. The application lets you perform multiple actions for different types of content that helps in visualizing information with ease.



The displayed sections (panes) are fully interactive, with clickable areas for easy navigation to relevant parts of applications. Users can access individual sections and narrow down the information displayed,



The overview page contains information pertaining to the Data Lakes, the total number of assets that are profiled, along with the assets that are scanned for data.

Dashboard

Welcome to Cloudera Data Catalog

Data Catalog is a service within Cloudera Data Platform that enables you to understand, manage, secure, and govern data assets across the enterprise.

Get started with Data Catalog

These resources will help you learn about how to use Cloudera Data Catalog.

[Start Guide](#) [Explore use cases](#) [Documentation](#)

Data Lakes 1	Total Assets 829 Realtime	Profiler configured 3/3 in all datalakes	Assets profiled 82 in Last 7 days
------------------------	--	--	---

My Datasets And Bookmarks Active Profilers **Data Lakes**

Status	Provider	Name	Version
Running		ftdooe-env-datalake	7.1.9

Search Assets
Discover assets across multiple Data Lakes. Find tags or assets in your data lake using Hive assets, attribute facets, or free text
[Search Now](#)

Manage Datasets
Group data assets into Datasets to organize data based on business classifications, purpose and protection requirements
[Manage Now](#)

Review Profilers
Find the patterns across the tables and apply tags for security rules. Create a custom profiler
[Review Now](#)

Help
1.5.2-b110

Additionally, you can manage the datasets, launch profilers, and search or discover assets.

Data Catalog Terminology

An overview of terminology used in Data Catalog service.

Profiler

Enables the Data Catalog service to gather and view information about different relevant characteristics of data such as shape, distribution, quality, and sensitivity which are 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 provides access to many kinds of enterprise data to support data discovery, data preparation, analytics, insights, and predictive analytics. In the context of Cloudera Data Platform, a Data Lake can be realized in practice with an Cloudera Manager enabled CDP cluster that runs Apache Atlas for metadata and governance services, and Apache Ranger for security services.

ECS

The Embedded Container Service (ECS) service enables you to run CDP Private Cloud Data Services by creating container-based clusters in your data center. In addition to the option to use OpenShift, which requires that you deploy and manage the Kubernetes infrastructure, you can also deploy an Embedded Container Service cluster, which creates and manages an embedded Kubernetes infrastructure for use with CDP Private Cloud Data Services.

OpenShift Container (OCP)

OpenShift is an enterprise platform for container orchestration.

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 multiple asset collections. Data assets are equivalent to “entities” in Apache Atlas.

Datasets

Datasets allow users of Data Catalog 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.

You can launch a Profiler cluster for a Data Lake. Adding new assets to (or removing from) a dataset must be done manually.

Related Information

[About Data Catalog](#)

Authorization for viewing Assets

Data Catalog users must have appropriate authorization set-up in Ranger to view assets.

Hive Ranger Policy - You must set-up Hive Ranger policies as per your requirement to work with Hive assets in Data Catalog.

For example, the following diagram provides a sample Hive Ranger policy.

The screenshot shows the 'Create Policy' interface. The 'Policy Details' section includes:

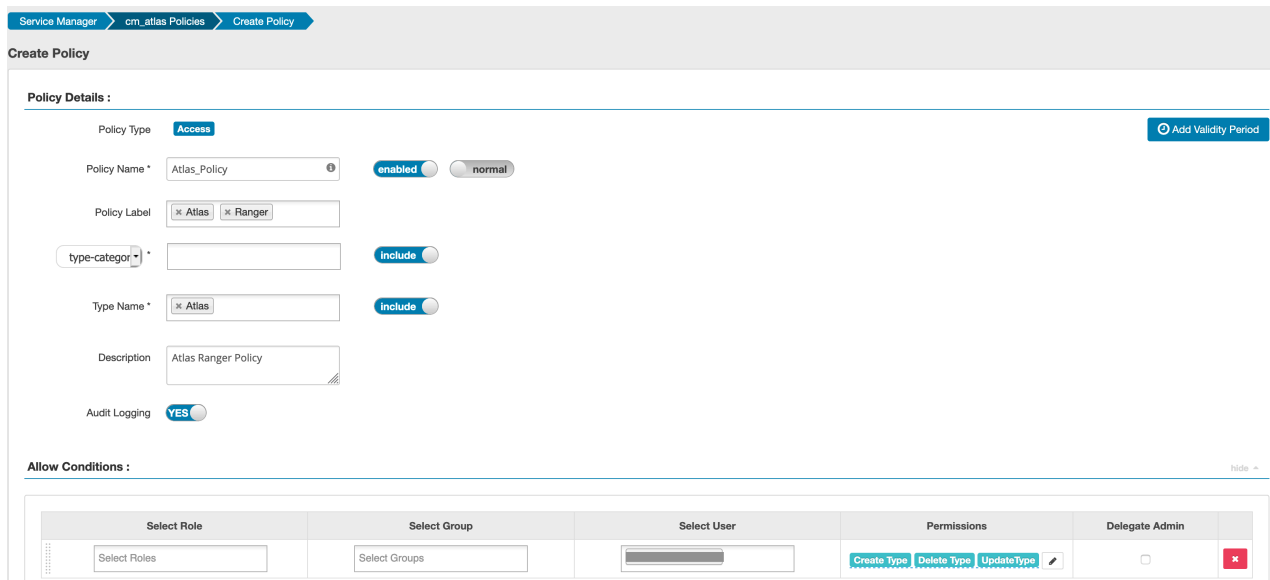
- Policy Type: **Access**
- Policy Name: (enabled)
- Policy Label:
- database: (include)
- Description:
- Audit Logging: YES

The 'Allow Conditions' section includes a table with the following columns:

Select Role	Select Group	Select User	Permissions	Delegate Admin
<input type="text"/>	<input type="text" value="Select Groups"/>	<input type="text"/>	<input type="checkbox"/> All <input type="checkbox"/> Alter <input type="checkbox"/> Create <input type="checkbox"/> Drop <input type="checkbox"/> Index <input type="checkbox"/> Lock <input type="checkbox"/> Read <input type="checkbox"/> Refresh <input type="checkbox"/> ReplAdmin <input type="checkbox"/> select <input type="checkbox"/> Service Admin <input type="checkbox"/> Temporary UDF Admin <input type="checkbox"/> update <input type="checkbox"/> Write	<input type="checkbox"/>

Atlas Ranger Policy- You must set-up Ranger policies for Atlas in order to work with asset search and Tag flow management.

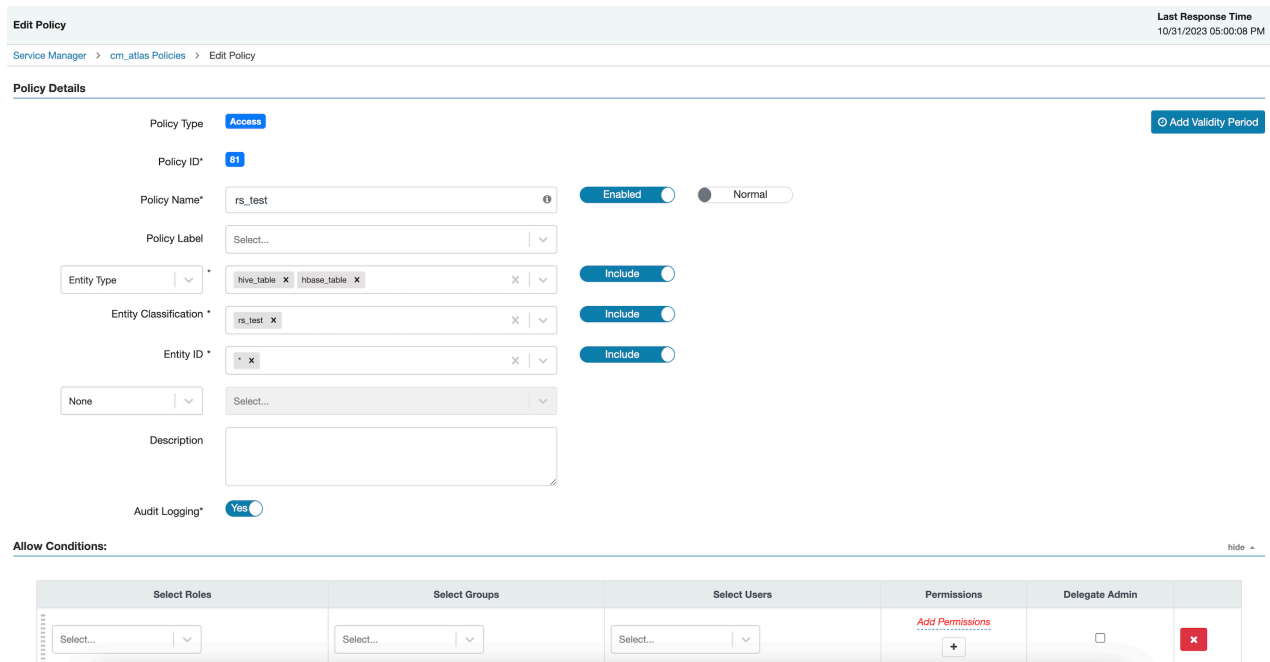
For example, the following diagram provides a sample Atlas Ranger policy.



Restricting access for certain users of Data Catalog

To have a fine-grained access to the same user from accessing the assets in Data Catalog, you can perform some additional changes. For example, if you want to restrict some users from accessing specific table information, you must set-up a Ranger policy such that these users will not have access to the asset details in Data Catalog.

To create the Ranger policy to restrict users from accessing asset details, refer to the following images:



The following image displays the “Deny Conditions” set for the specific user.

Edit Policy Last Response Time
10/31/2023 05:00:08 PM

Service Manager > cm_atlas Policies > Edit Policy

Exclude from Allow Conditions:

Select Roles	Select Groups	Select Users	Permissions	Delegate Admin	
Select...	Select...	Select...	Add Permissions +	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Deny All Other Accesses: False

Deny Conditions:

Select Roles	Select Groups	Select Users	Permissions	Delegate Admin	
Select...	Select...	hrt_e2e_admin X	Read Entity Create Entity Update Entity Delete Entity ✎	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Exclude from Deny Conditions:

Select Roles	Select Groups	Select Users	Permissions	Delegate Admin	
			Add Permissions		<input checked="" type="checkbox"/>

The resultant is depicted in the following image, where the user has no permissions to access the specified dataset. In this example, it is us_customers.

Search

tyik4e-env-datalake | 2

Type	Name	Qualified Name	Created On	Owner	Source
<input type="checkbox"/> Hive Table	us_customers	hortoniabank.us_customers@cm	Sun Oct 29 2023	hive	hive
<input type="checkbox"/> Hive Table	customer_address	test_dss_db.customer_address@cm	Sun Oct 29 2023	hive	hive

Filters:

TYPE

- Hive Table
- HBase Table
- Spark ML Directory
- HDFS Path

OWNERS

- atlas
- cseo_rasharma
- hive
- public

ENTITY TAG

- rs_test

Additionally, when you plan to restrict data access, please note the following:

- Audit summarisation for the asset evolves from the Ranger audit profiler and Metrics service.
- Various Hive Column Statistical metrics for columns of the asset evolves from Atlas as part of the profile_data of a column.

To ensure that the data related to audit summary and Hive Column Statistics are not visible to the subscribers, you must make sure to turn off the audit profiler and Hive Column profiler respectively.

Understanding Datasets

A Dataset is a group of assets that fit search criteria so that you can manage and administer them collectively.

Asset collections enable you to perform the following tasks when working with your data:

- Organize

Group data assets into Datasets 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.

Advanced asset 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`.

- Understand

Audit data asset security and use for anomaly detection, forensic audit and compliance, and proper control mechanisms.

You can edit Datasets after you create them and the assets contained within the collection will be updated. CRUD (Create, Read, Update, Delete) is supported for Datasets.



Note: Datasets must have less than 130 assets.

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 data lake, such as a Hive table, that contains business or technical data.

Data assets are also known as *entities* in Apache Atlas.

About the Data Catalog Profiler

The Data Catalog profiler employs Kubernetes enabled job scheduling and runs profilers jobs on-demand.

These profilers create metadata annotations that summarize the content and shape characteristics of the data assets.



Note: You must be a PowerUser to launch the Profiler.

Profiler Name	Description
Cluster Sensitivity Profiler	A sensitive data profiler- PII, PCI, HIPAA and others.
Ranger Audit Profiler	A Ranger audit log summarizer.
Hive Column Profiler	Provides summary statistics like Maximum, Minimum, Mean, Unique, and Null values at the Hive column level.

For example, data profilers can create summarized information about contents of an asset and also provide annotations that indicate its shape (such as distribution of values in a box plot or histogram).

Understanding the Cluster Sensitivity Profiler

The Cluster Sensitivity Profiler automatically performs context and content inspection to detect various types of sensitive data and suggest suitable classifications or tags based on the type of sensitive content detected or discovered.

Auto-detected data types

Type of data

- Bank account
- Credit card
- Driver number (UK)
- Email
- IBAN number
 - Austria (AUT)
 - Belgium (BEL)
 - Bulgaria (BGR)
 - Switzerland (CHE)
 - Cyprus (CYP)
 - Czech Republic (CZE)
 - Germany (DEU)
 - Denmark (DNK)
 - Spain (ESP)
 - Estonia (EST)
 - Finland (FIN)
 - France (FRA)
 - United Kingdom (GBR)
 - Greece (GRC)
 - Croatia (HRV)
 - Hungary (HUN)
 - Ireland (IRL)
 - Iceland (ISL)
 - Italy (ITA)
 - Liechtenstein (LIE)
 - Lithuania (LTU)
 - Latvia (LVA)
 - Luxembourg (LUX)
 - Malta (MLT)
 - Netherlands (NLD)
 - Norway (NOR)
 - Poland (POL)
 - Portugal (PRT)
 - Romania (ROU)
 - Slovakia (SVK)
 - Slovenia (SVN)
 - Sweden (SWE)
- IP address
- NPI
- Name

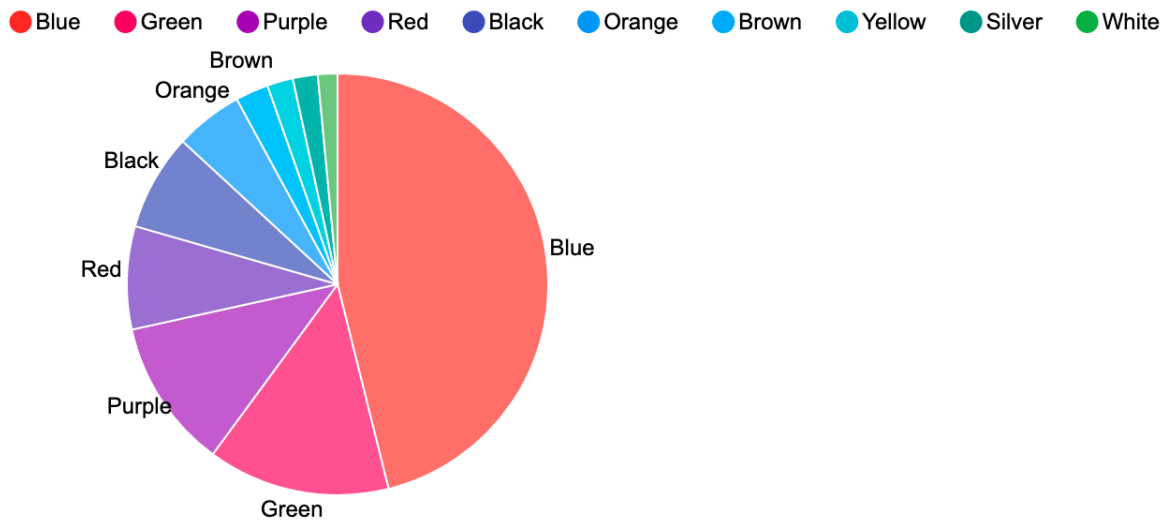
- National ID number
 - Bulgaria (BGR)
 - Switzerland (CHE)
 - Czech Republic (CZE)
 - Denmark (DNK)
 - Spain (ESP)
 - Estonia (EST)
 - Finland (FIN)
 - Greece (GRC)
 - Ireland (IRL)
 - Iceland (ISL)
 - Italy (ITA)
 - Lithuania (LTU)
 - Latvia (LVA)
 - Norway (NOR)
 - Poland (POL)
 - Portugal (PRT)
 - Romania (ROU)
 - Slovakia (SVK)
 - Sweden (SWE)
- National insurance number (UK)
- Passport number
 - Austria (AUT)
 - Belgium (BEL)
 - Switzerland (CHE)
 - Germany (DEU)
 - Spain (ESP)
 - Finland (FIN)
 - France (FRA)
 - Greece (GRC)
 - Ireland (IRL)
 - Italy (ITA)
 - Poland (POL)
 - United Kingdom (UK)
- Bank Routing Number
- US Social Security Number
- Society for Worldwide Interbank Financial Telecommunication (SWIFT)
- Telephone

Understanding the Hive Column Profiler

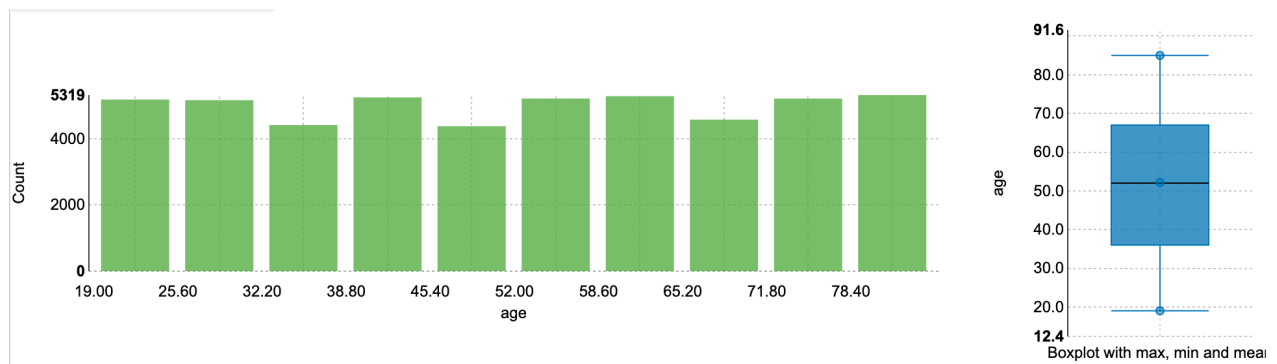
You can view the shape or distribution characteristics of the columnar data within a Hive table based on the Hive Column Profiler.

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, and cardinality of the data) for a column. The profiler computes column univariate statistics that are displayed using an appropriate chart in the Schema tab.

Pie charts are presented for categorical data with limited number of categories or classes. Examples include data such as eye colors that only have a fixed list of values (categories or labels).



When the data within columns is numeric, a histogram of the distribution of values organized into 10 groups (decile frequency histogram) and a box plot with a five-number summary (mean, median, quartiles, maximum, and minimum values) are shown for the column.



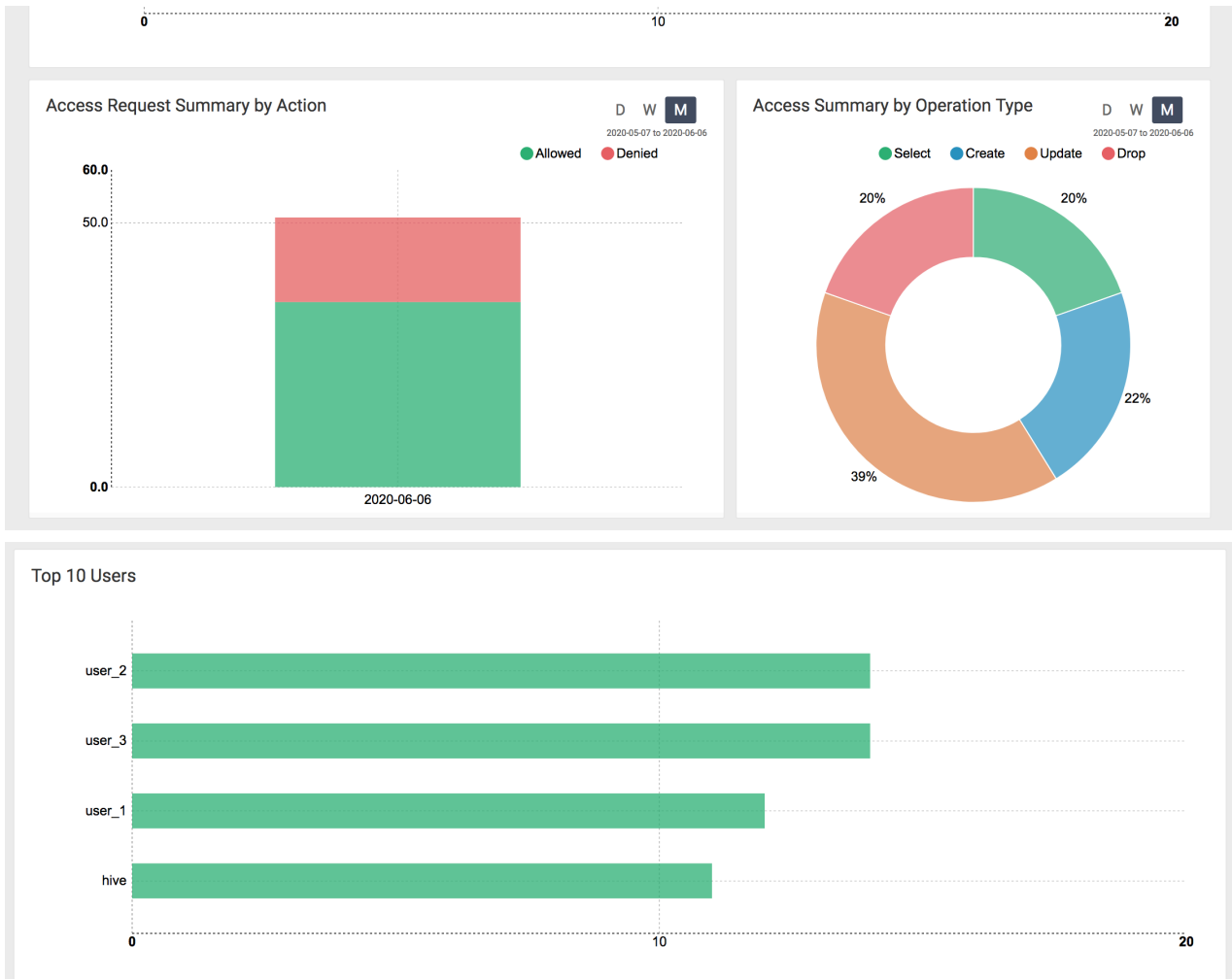
Related Information

[Understanding the Ranger Audit Profiler](#)

Understanding the Ranger Audit Profiler

You can view who has accessed which data from a forensic audit or compliance perspective, visualize access patterns, and identify anomalies in access patterns using the Ranger Audit Profiler.

The audit profiler uses the Apache Ranger audit logs to show the most recent raw audit event data as well as summarized views of audits by type of access and access outcomes (allowed/denied). Such summarized views are obtained by profiling audit records in the data lake with the audit profiler.



Related Information

[Understanding the Hive Column Profiler](#)