Data Catalog Overview

Date published: 2019-11-14
Date modified: 2023-03-10

Legal Notice

Contents

Data Catalog Overview

CDP is a hybrid data platform designed for unmatched freedom to choose—any cloud, any analytics, any data. CDP delivers faster and easier data management and data analytics for data anywhere, with optimal performance, scalability, and security.

Data Catalog is a service within Cloudera Data Platform (CDP) that enables you to understand, manage, secure, and govern data assets across enterprise data lakes. Data Catalog helps you understand data across multiple clusters and across multiple environments (on-premises, cloud, and hybrid).

You can access the Data Catalog user interface by logging into Cloudera Data Platform > Select Data Catalog.

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

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 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 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

Data Catalog Overview

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.

Data Catalog users must have either an EnvironmentAdmin or EnvironmentUser role assigned.

The Power User must provide the requisite access to subscribers who plan to use Data Catalog, either as EnvironmentAdmin or EnvironmentUser.

EnvironmentAdmin	EnvironmentUser
Can perform similar actions as EnvironmentUser.	Can create Dataset and related actions (Add assets, remove assets, tag assets, tag asset columns, and few others) for data lakes.
Additionally, in the Management Console, can perform the following: • Delete Environments. • Stop Environments. • Upgrade data lake.	Connect to Atlas and Ranger entities for data lakes for which there is access.
	Search for assets on the search page.
	Bookmark any dataset (even with no data lakes or data lakes to which access is not available).
	Access profilers.
	Create custom tags.
	Launch profilers on a data lake from the search page. Ability to launch the Workload cluster (Must have the Power User role assigned).
	Use Filters on the Search page of Data Catalog to filter results.
	Launch profilers on a data lake from the search page.

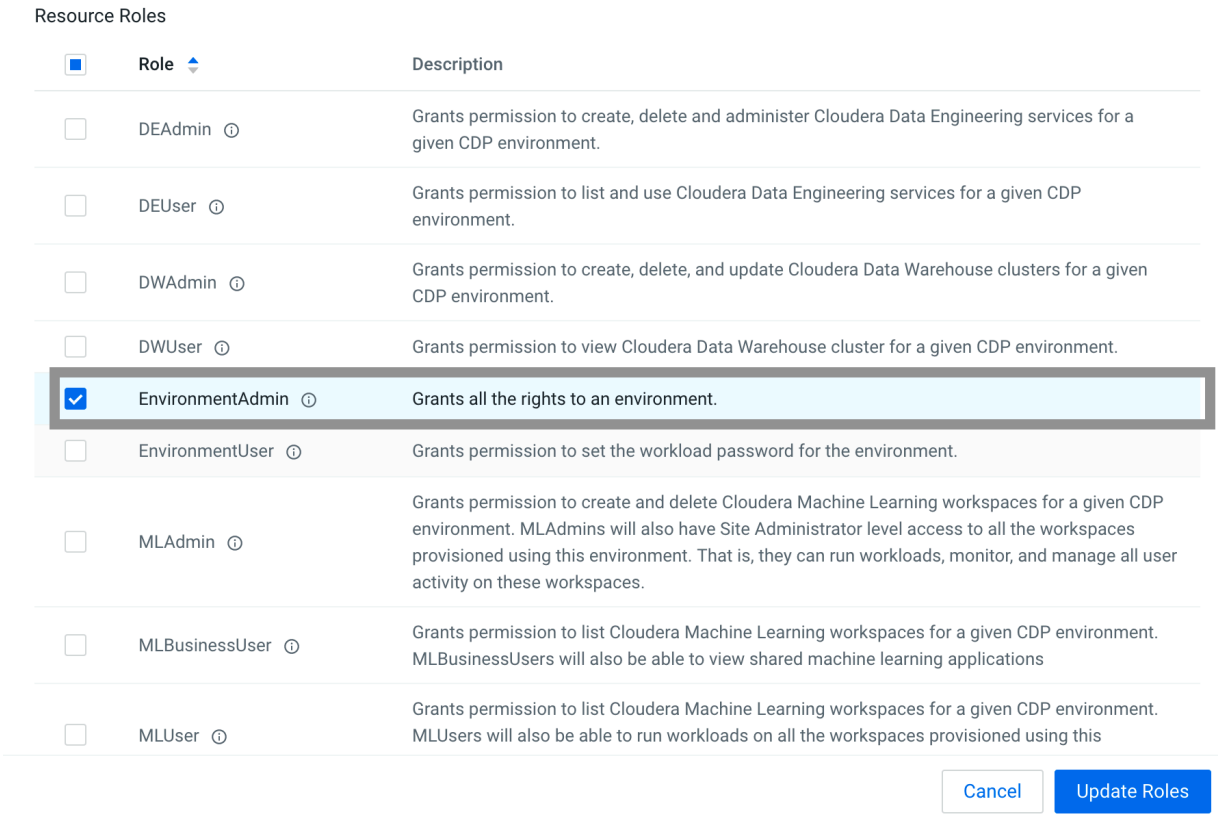
Providing role access

You must provide the required role access to use Data Catalog.

Procedure

1. From Cloudera Data Platform > Management Console > Environments > Select an Environment > select Actions drop-down > Manage Access.

2. Search for the user who requires Data Catalog access > Select the check-box, either EnvironmentAdmin or EnvironmentUser > Click Update Roles.



Resource Roles

Role	Description
DEAdmin	Grants permission to create, delete and administer Cloudera Data Engineering services for a given CDP environment.
DEUser	Grants permission to list and use Cloudera Data Engineering services for a given CDP environment.
DWAdmin	Grants permission to create, delete, and update Cloudera Data Warehouse clusters for a given CDP environment.
DWUser	Grants permission to view Cloudera Data Warehouse cluster for a given CDP environment.
EnvironmentAdmin	Grants all the rights to an environment.
EnvironmentUser	Grants permission to set the workload password for the environment.
MLAdmin	Grants permission to create and delete Cloudera Machine Learning workspaces for a given CDP environment. MLAdmins will also have Site Administrator level access to all the workspaces provisioned using this environment. That is, they can run workloads, monitor, and manage all user activity on these workspaces.
MLBusinessUser	Grants permission to list Cloudera Machine Learning workspaces for a given CDP environment. MLBusinessUsers will also be able to view shared machine learning applications
MLUser	Grants permission to list Cloudera Machine Learning workspaces for a given CDP environment. MLUsers will also be able to run workloads on all the workspaces provisioned using this

Cancel Update Roles

3. Navigate back to the Environment details page and select Actions > select Synchronize Users to FreeIPA
   Allow the sync operation to complete and the changes to take effect.

Attention: For tenant specific roles and related permissions, see Account roles.

Related Information

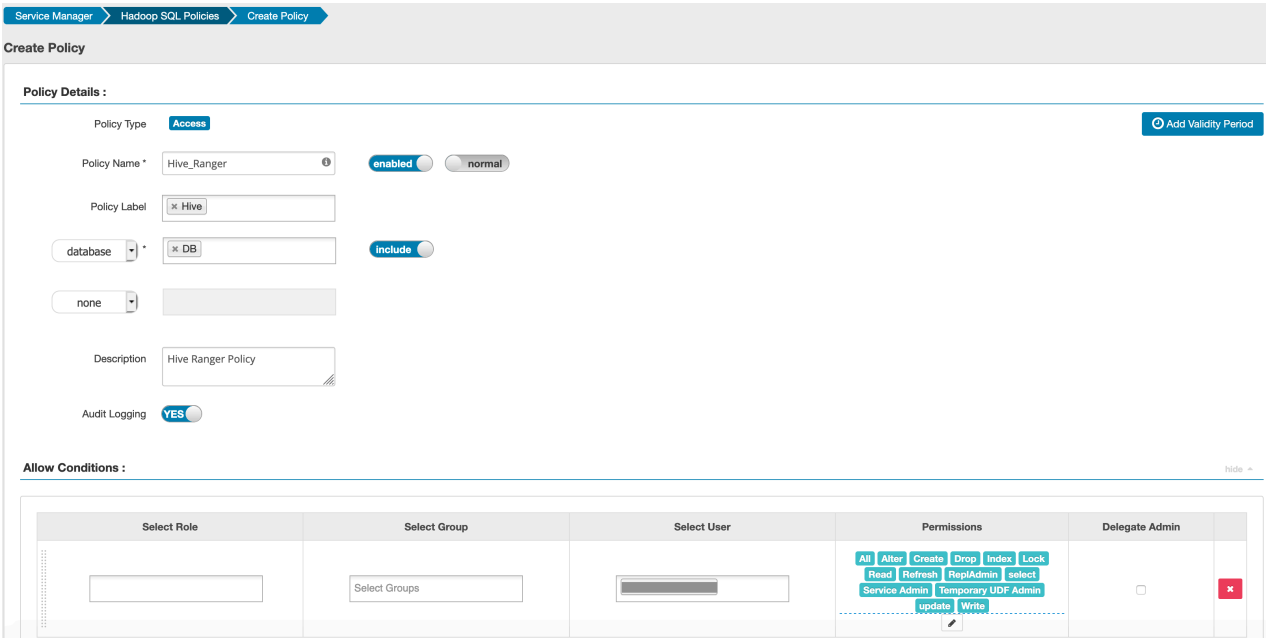
Authorization for viewing Assets

Authorization for viewing Assets

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

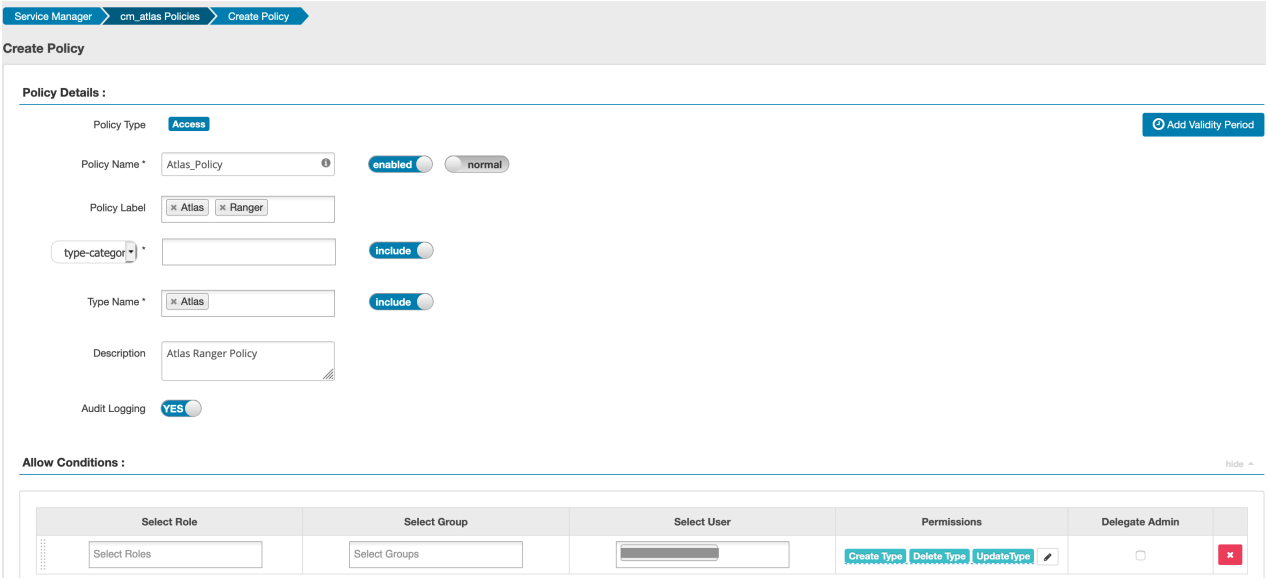
Hive Ranger Policy - You must set-up Hive Ranger policies as per your requirement to work with Hive assets in Data Catalog.

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



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.



Related Information

Providing role access

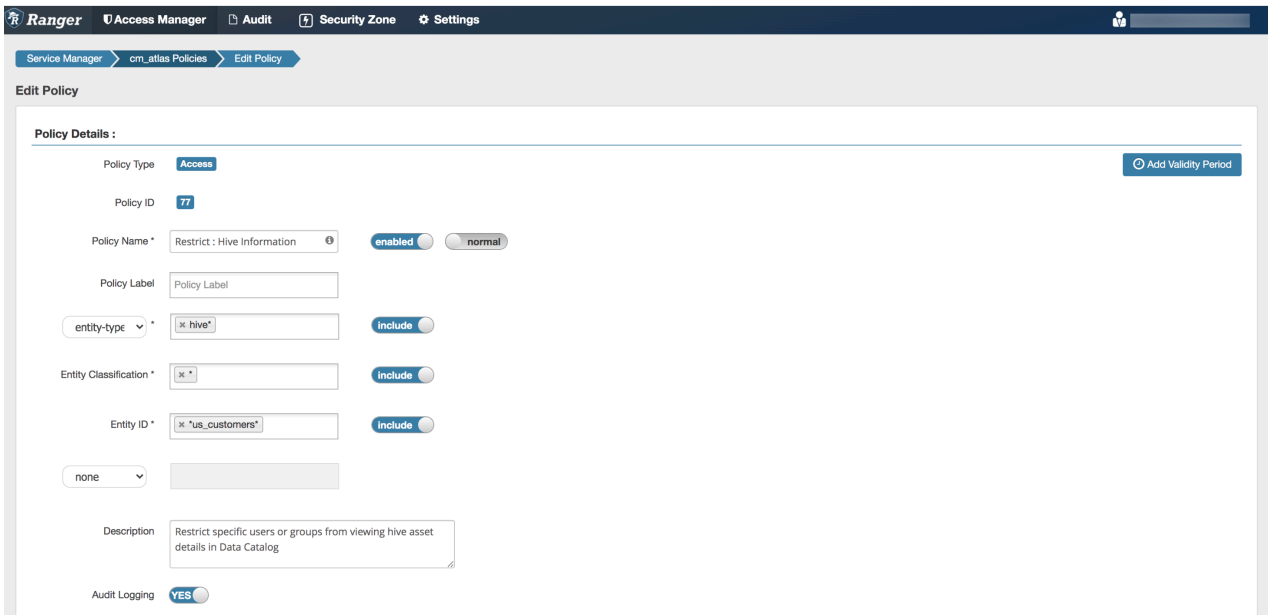
Restricting access for certain users of Data Catalog

Restricting access for certain users of Data Catalog

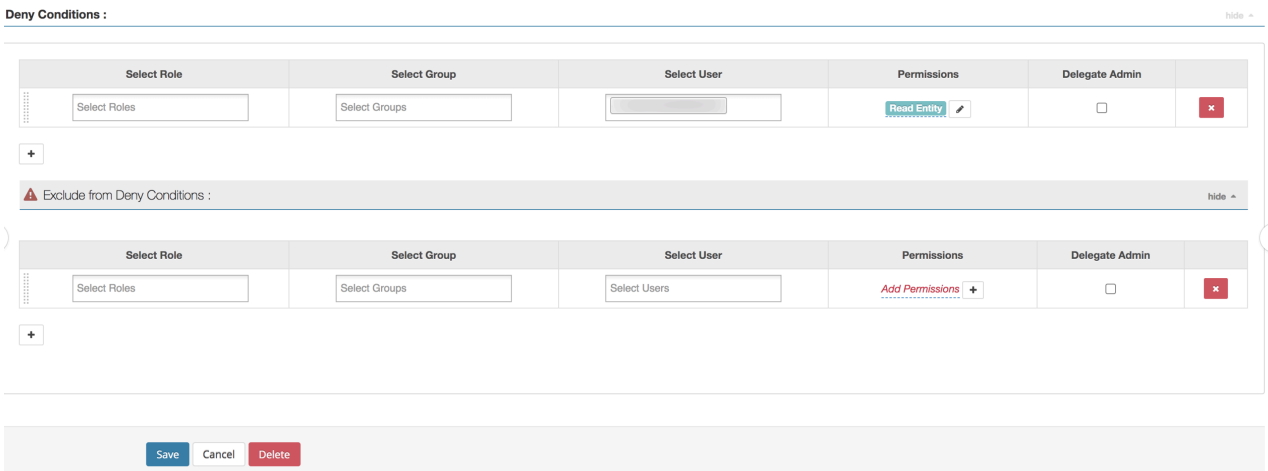
When a subscriber is provided with user access either as EnvironmentAdmin or EnvironmentUser, by default, the user is synchronized with default Atlas policies in Ranger. This implies that the same user has access to all the datasets in the Data Catalog environment.

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.

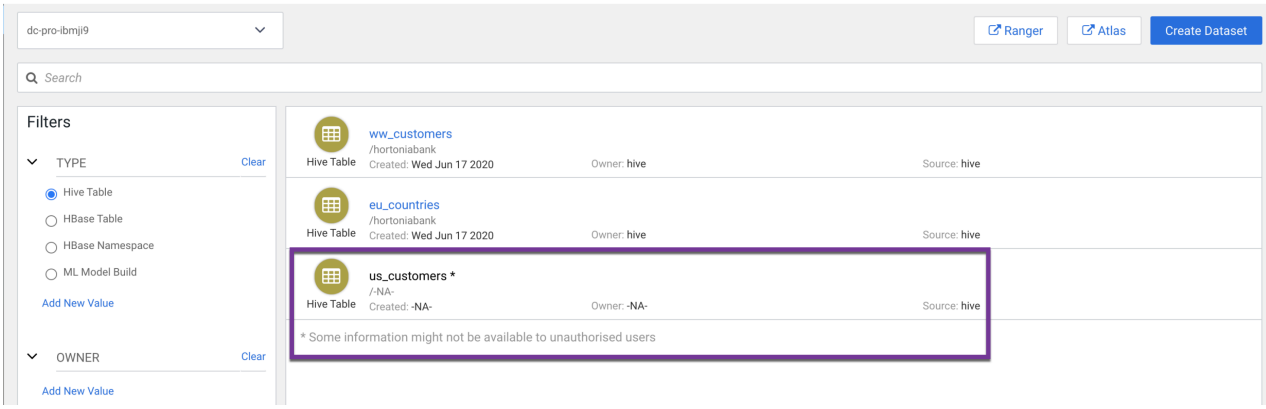
To create the Ranger policy to restrict users from accessing asset details, refer to the following images:



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



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.



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.

Related Information

Authorization for viewing Assets

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.

Related Information

Understanding the Data Catalog Profiler

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.

Understanding the Data Catalog Profiler

Data Catalog includes a profiler engine that can run data profiling operations as a pipeline on data located in multiple data lakes. You can install the profiler agent in a data lake and set up a specific schedule to generate various types of data profiles. Data profilers generate metadata annotations on the assets for various purposes.

Attention: You can launch the profiler engine (DataHub cluster) for a data lake after selecting the data lake and clicking Launch Profilers. Note that, if the selected data lake already has the profiler engine attached, then the Launch Profilers option is not displayed.

Note: Profilers DataHub can be set up only by accessing the Data Catalog UI and is not supported through the Management Console.

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).

Related Information

Understanding Datasets
Understanding the Cluster Sensitivity Profiler
Understanding the Hive Column Profiler
Understanding the Ranger Audit Profiler

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

Related Information

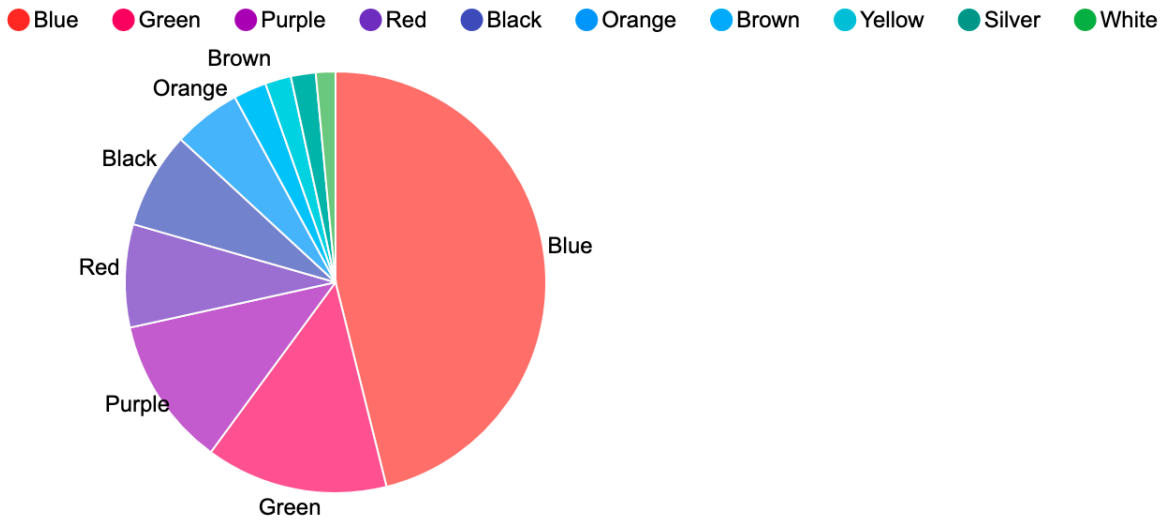
Understanding the Data Catalog Profiler

Understanding the Hive Column Profiler

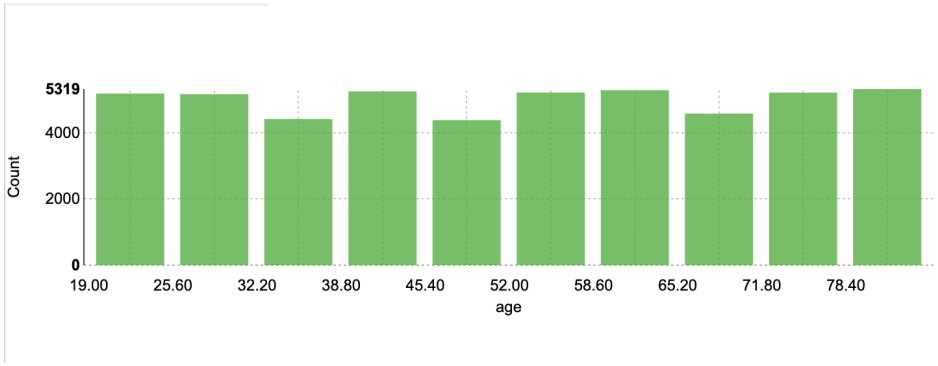
You can view the shape or distribution characteristics of the columnar data within a Hive table based on the Hive Column Profiler.

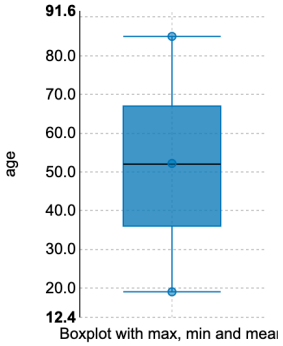
There are different charts available to help visualize the shape and distribution of the data within the column as well as summary statistics (such as means, null count, and cardinality of the data) for a column. The profiler computes column univariate statistics that are displayed using an appropriate 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

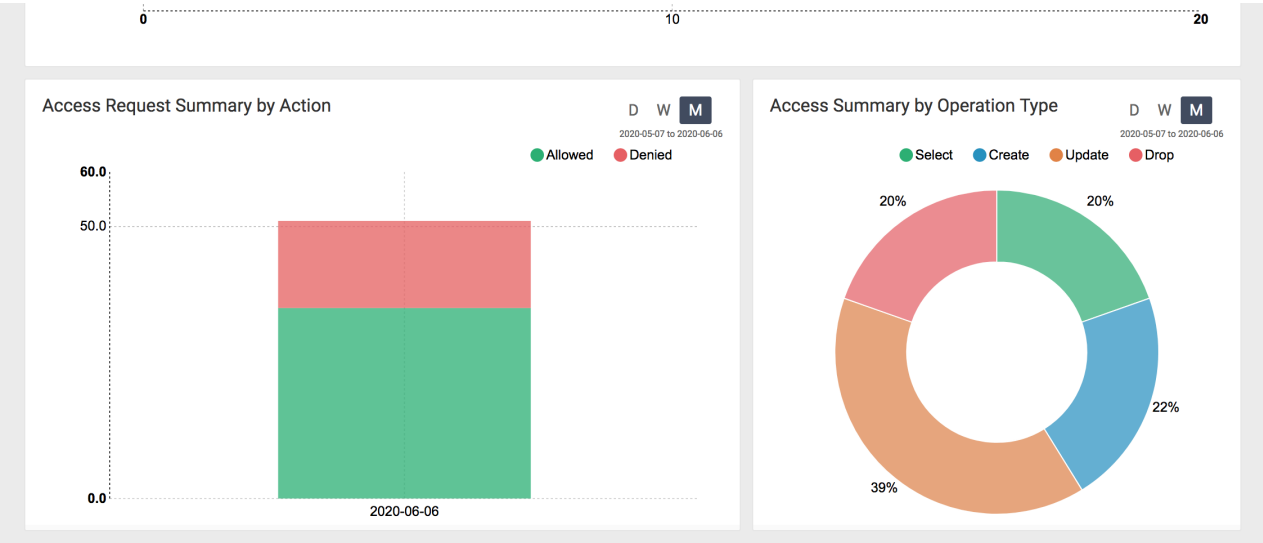
Understanding the Data Catalog Profiler

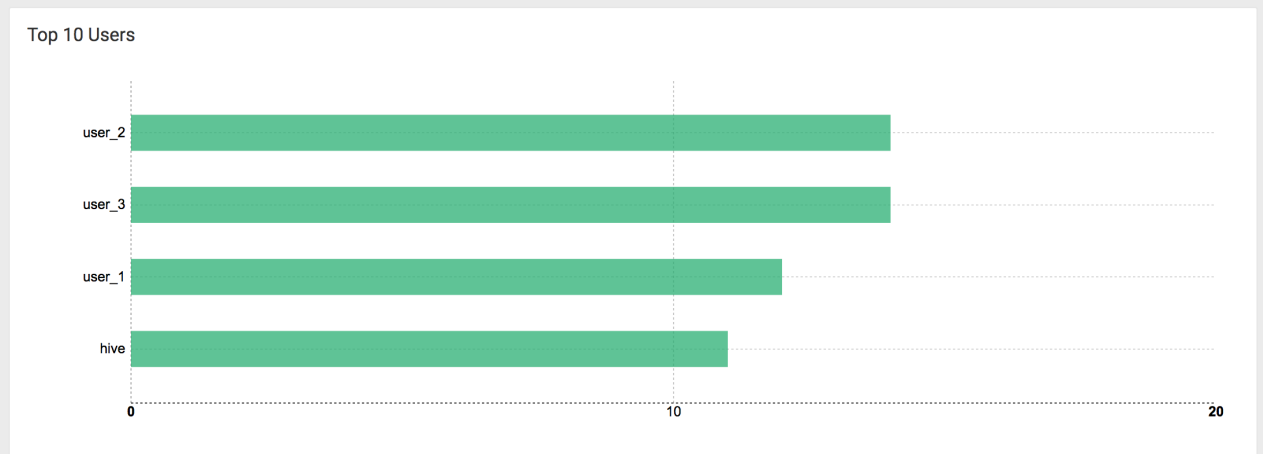
Understanding the Ranger Audit Profiler

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 Data Catalog Profiler

Understanding the Hive Column Profiler