CDP Public Cloud / Data Catalog

Cloudera Data Catalog Overview

Date published: 2019-11-14
Date modified: 2025-04-07

https://docs.cloudera.com/

Legal Notice

Contents

About Cloudera Data Catalog

Cloudera Data Catalog is a service within Cloudera that enables you to understand, manage, secure, and govern data assets across the enterprise.
Cloudera Data Catalog helps you understand data lying in your data lake. You can locate relevant data of interest based on various parameters. Using Cloudera 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.

Note: Cloudera Data Catalog is supported in the EU and APAC regional Control Plane. You can ensure that you manage, secure, collaborate, and govern data assets across multiple clusters and environments within the EU region or APAC region where your organization operates as per the data protection regulatory requirements.

Cloudera 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

Cloudera Data Catalog Terminology

Profiler architecture in VM-based environments

In a VM-based environment, the Cloudera Data Catalog Profiler architecture uses a Cloudera Data Hub workload cluster.

Note:

The VM-based architecture (using the Cluster) is deprecated from the 3.0.0 release but remains available until 7.2.18 is supported (Sept 2025). Therefore, based profilers will also not be available in versions after 7.2.18. Only Compute Cluster enabled environment will be able to run profilers after version 7.2.18.

For more information, see Cloudera Support lifecycle policy.

Figure 1: VM-based profiler architecture

After registering a VM-based environment, you have to launch a Cloudera Data Hub cluster for each data lake to provide the resources and services required for a profiler workload. This can be handled by Cloudera Data Catalog. For more information, see Launch profiler Cluster.

Note: In comparison to a Kubernetes pod in a Compute Cluster enabled environment, a Cloudera Data Hub workload cluster reserves compute resources even when a profiler task is not running. Also, more services are required to be included in the Cloudera Data Hub cluster template in contrast to the default compute cluster:

Zookeeper

For configuration information, naming, synchronization and group services over large clusters in distributed systems

Yarn

For resource management

1. Cloudera Data Hub uses the internal service called Cloudbreak to start the necessary services in the Profiler Cloudera Data Hub cluster. It is also used to access data about profilers and the data lake. In comparison, the Cluster Proxy provides the connection between the Cloudera Data Hub UI service and the rest of the Cloudera Data Catalog services.
2. An additional Amazon Relational Database (PostgreSQL) is used to store data required for the profiling process, such as, Custom Sensitivity Profiler Rules, profiler-data lake mappings and datasets.
3. Knox is used to authenticate services between your and Cloudera's environment
4. Livy is used together with a dedicated Scheduler Service to start the individual profiler instances with Spark jobs.
5. The Cloudera Data Hub cluster manages the different services responsible for the profiling.
   a. Profiler Admin service is similar to an interface for Profilers. It allows Cloudera Data Hub to fetch information from the workload Cloudera Data Hub about scheduled jobs, profiler configurations and so on.

      Profiler Metrics is responsible for the metrics calculation and synching it to Cloudera Data Hub database and Atlas.
   b. The profilers use a cloud storage called Profiler output bucket as a temporary storage to aggregate all their collected data, such as profiler snapshots, which help to continue the profiling by saving interim data.
6. The final profiler results are stored in an attached cloud storage.

Related Information

Cloudera Data Hub

Introduction to Data Lakes

Cloudera Management Console

Profiler architecture in Compute Cluster enabled environment

Next to the Cloudera Data Hub based profiler cluster, Cloudera Data Catalog offers the possibility to run profilers as a containerized service in a standardized Kubernetes base cluster called Externalized Compute Cluster. This consumes far less resources and provides auto-scaling.

Note:

The VM-based architecture (using the Cluster) is deprecated from the 3.0.0 release but remains available until 7.2.18 is supported (Sept 2025). Therefore, based profilers will also not be available in versions after 7.2.18. Only Compute Cluster enabled environment will be able to run profilers after version 7.2.18.

For more information, see Cloudera Support lifecycle policy.

Figure 2: Profiler architecture in Compute Cluster enabled environment

1. Once the container-ready environment is set up, a default Kubernetes cluster (Externalized Compute Cluster) is also created in this environment.

   Note: The Kubernetes jobs and API server offers the same API and UI interface capabilities for you as the VM-based Cloudera Data Hub, therefore, there is no difference in use.

2. The Profiler Launcher Service (PLS) internal to Cloudera Data Catalog schedules Kubernetes jobs, cron jobs in the compute cluster using HTTP API calls. Each type of a profiler has its own Kubernetes cron-jobs for handling scheduled profilers.
3. Once the time of the schedule is reached the Kubernetes job will launch a pod that will start profiling a data lake or ranger audit logs. The configuration for the jobs are received via the Cloudera Data Catalog API.
4. Using these settings the profiler connects to a data lake, identifies all the assets present in the data lake then starts profiling.
5. The results will be synced to Atlas and Cloudera Data Catalog using their respective APIs.

Related Information

Cloudera Data Hub
Cloudera Management Console
Using Compute Clusters in AWS environments
Using Compute Clusters in Azure environments

Before you start

Before you access Cloudera Data Catalog on premises, you must deploy the service.

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

Cloudera Data Catalog is supported on both Openshift and ECS clusters. For more information, see the related documents.

Prerequisites to access Cloudera Data Catalog

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

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

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

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

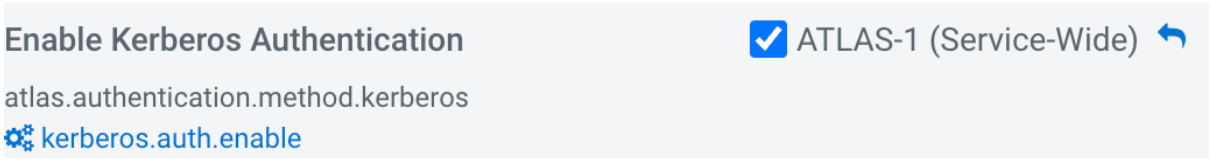
You need the following roles to use Cloudera Data Catalog:

- DataCatalogCspRuleManager

  Using this role, you can create, deploy new Cluster Sensitivity Profiler rules, create new regex, and run validations on newly created rules.
- DataCatalogCspRuleViewer

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

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

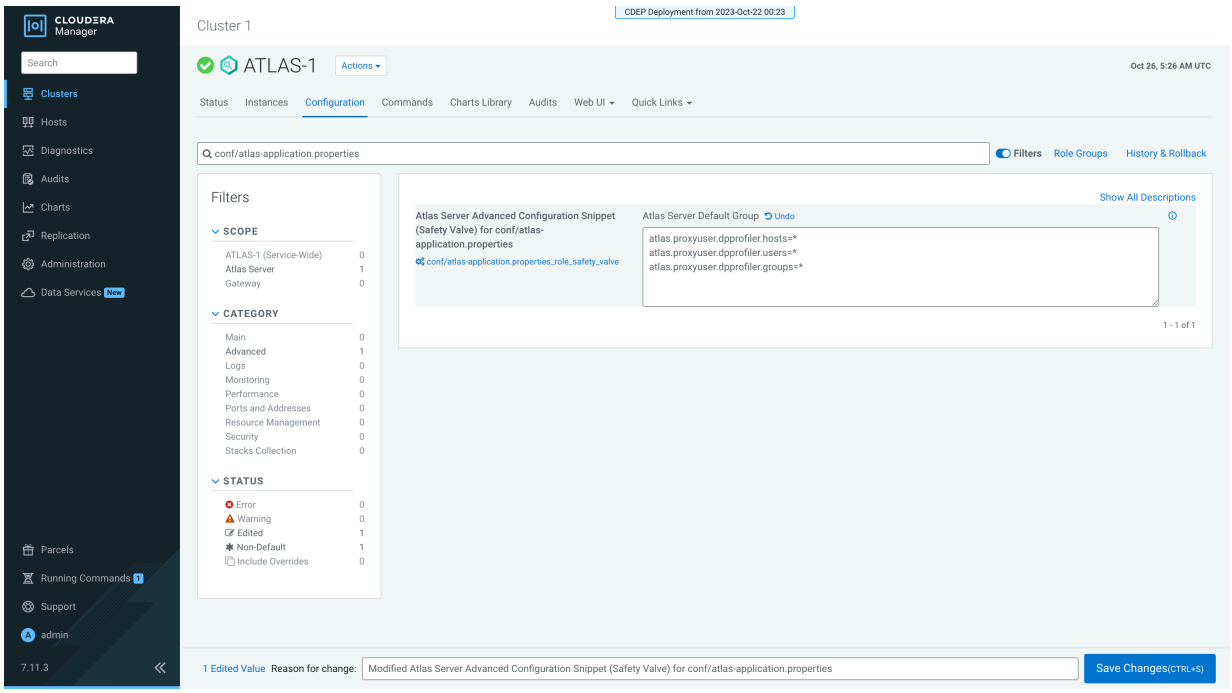


3. In the search bar, look up the following property:conf/atlas-application.properties_role_safety_valve and enter the values for the Atlas service:

   atlas.proxyuser.dpprofiler.hosts=*

   atlas.proxyuser.dpprofiler.users=*

   atlas.proxyuser.dpprofiler.groups=*



4. For the Apache Ranger configuration updates, go to Cloudera Manager Clusters Ranger Configuration .
5. 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: *

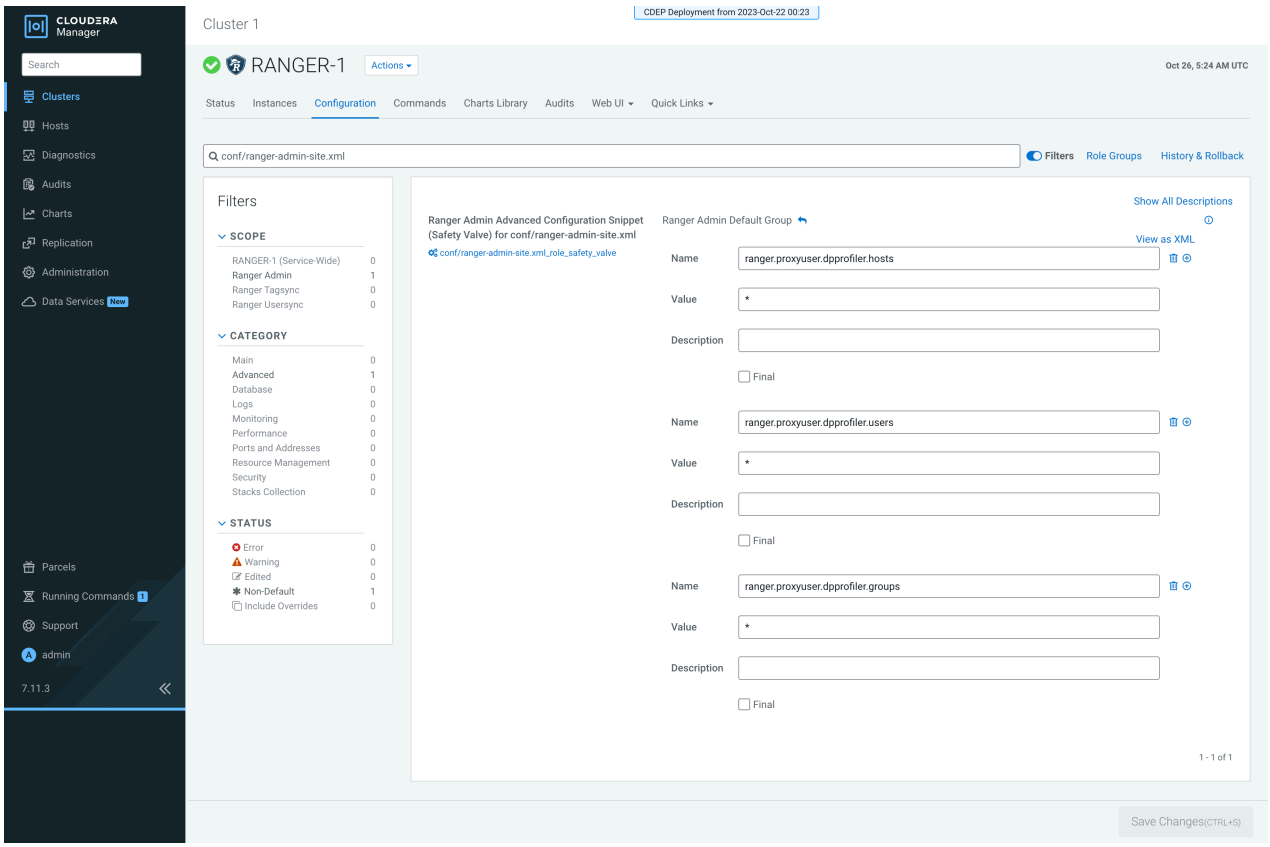
   Name: ranger.proxyuser.dpprofiler.users

   Value: *

   Name: ranger.proxyuser.dpprofiler.groups

   Value: *

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



Later, restart Atlas and Ranger services respectively.

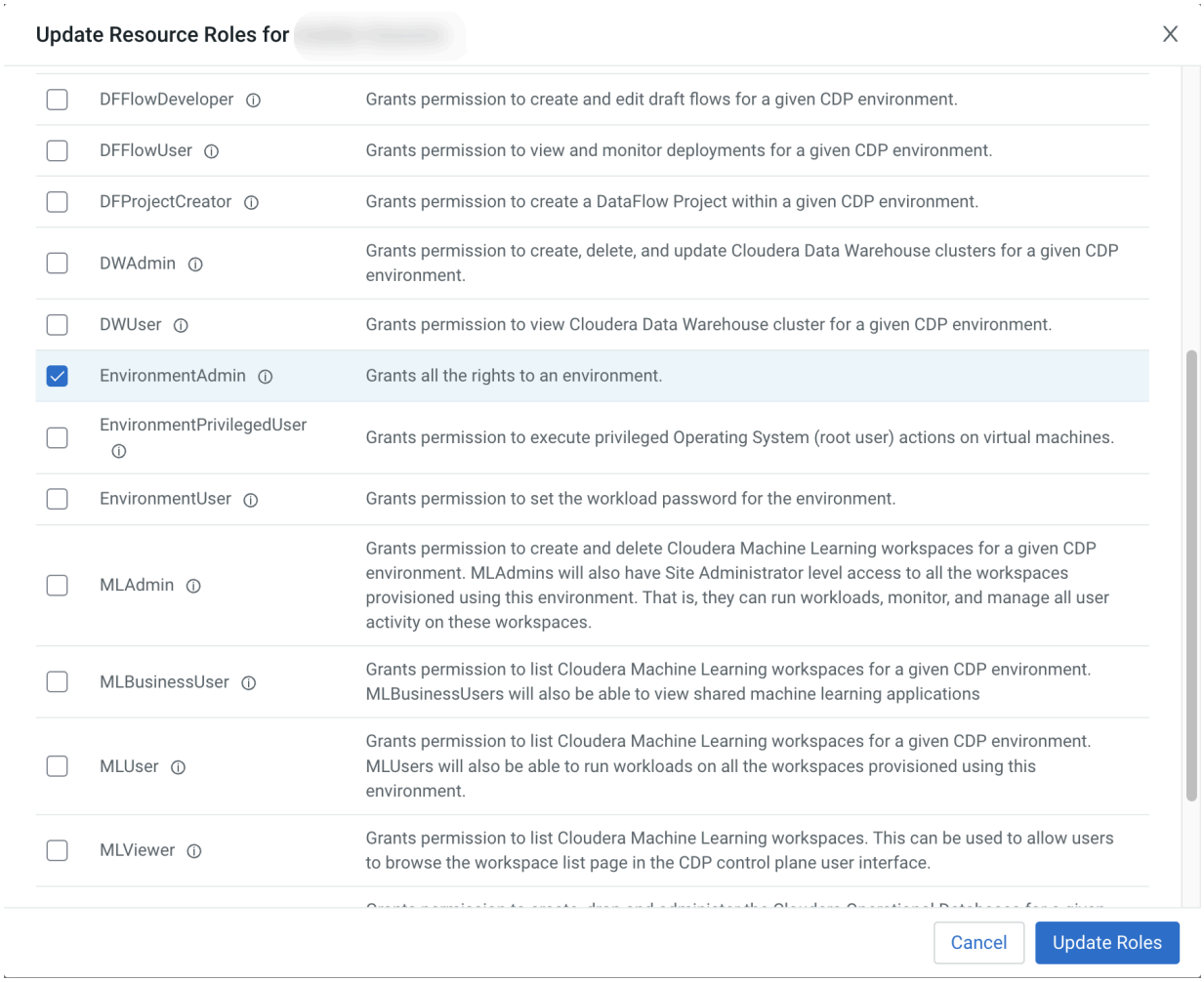
Providing role access

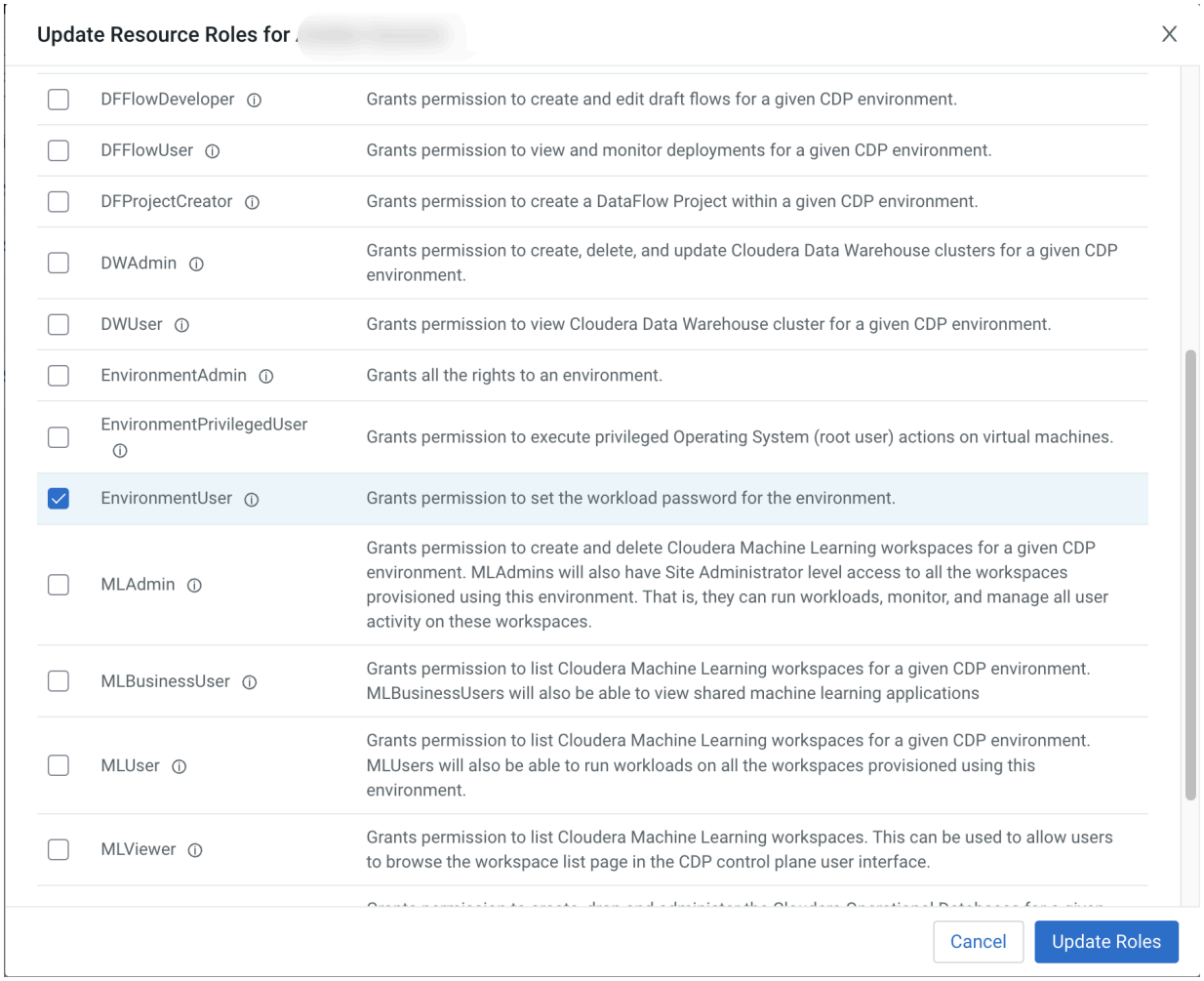
You must provide the required role access to use Cloudera Data Catalog.

Procedure

1. From Cloudera Management Console Environments > Select an environment > select the Actions drop-down > Manage Access.

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





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

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

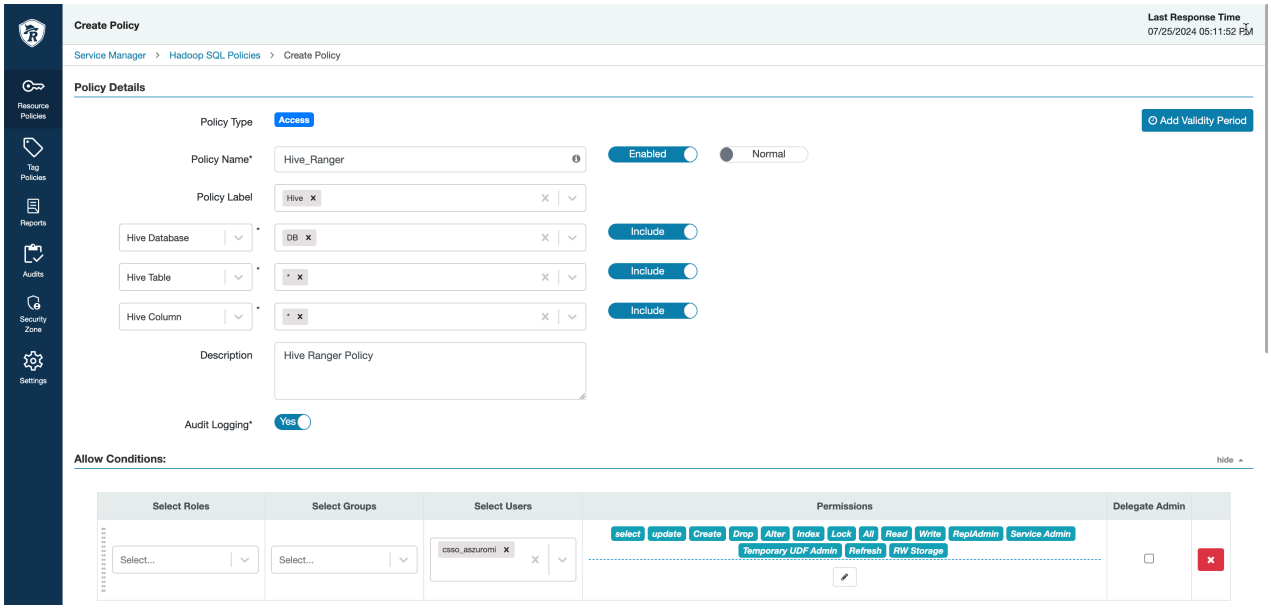
Related Information
Authorization for viewing Assets

Authorization for viewing assets

Cloudera Data Catalog users must have appropriate authorization set up in Apache Ranger to view certain asset types.

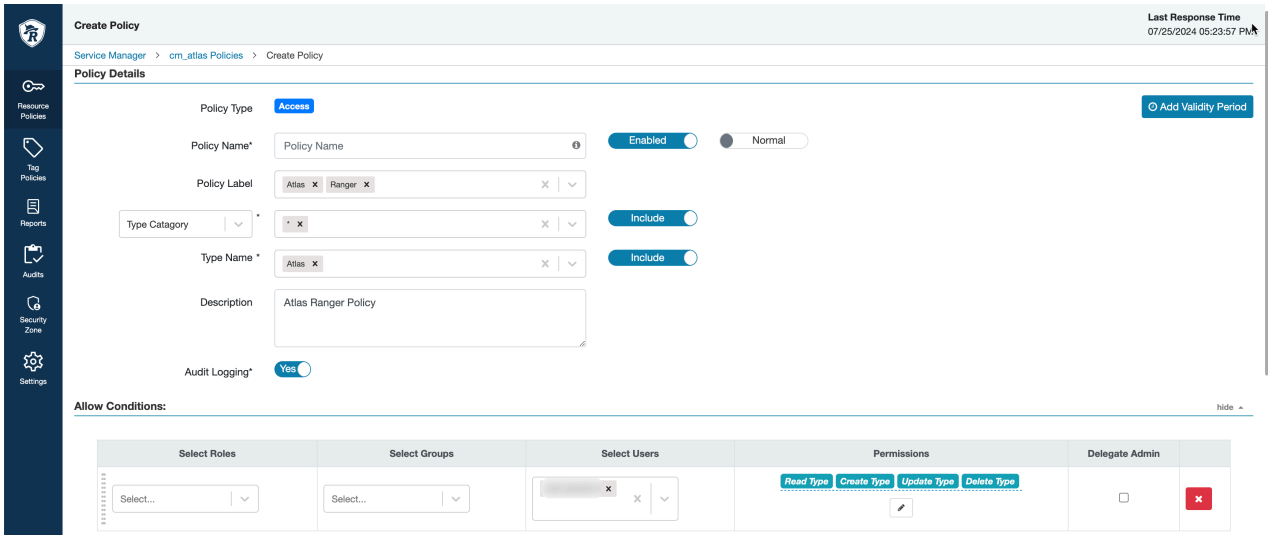
Hive Ranger Policy - You must set up Hive Ranger resource-based policies as per your work requirements for Hive assets in Cloudera Data Catalog.

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



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

Ranger Policies Overview

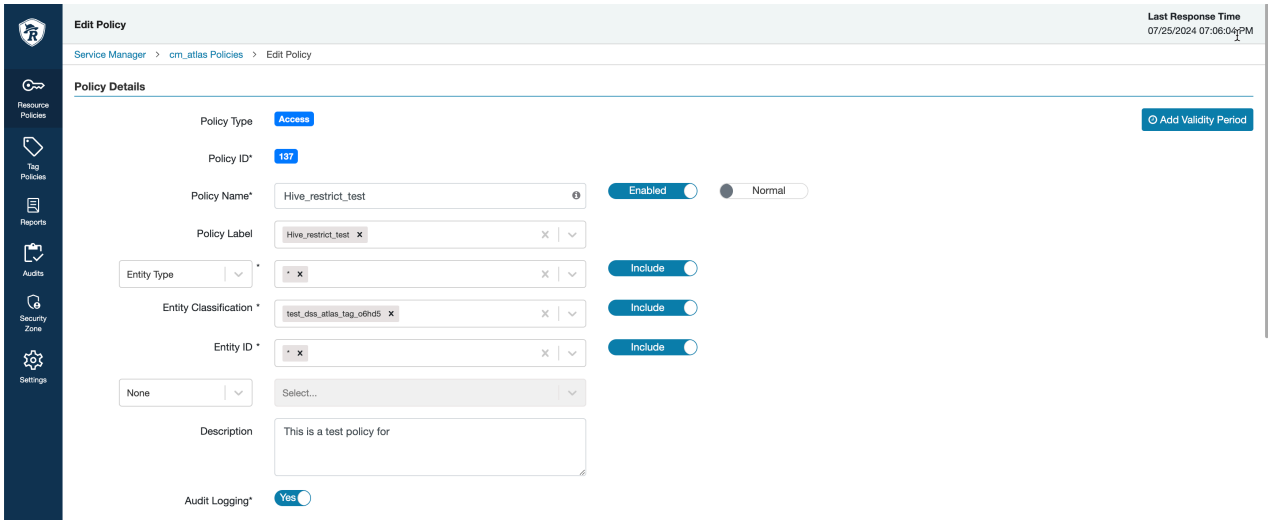
Configuring resource-based services

Restricting access for certain users of Cloudera Data Catalog

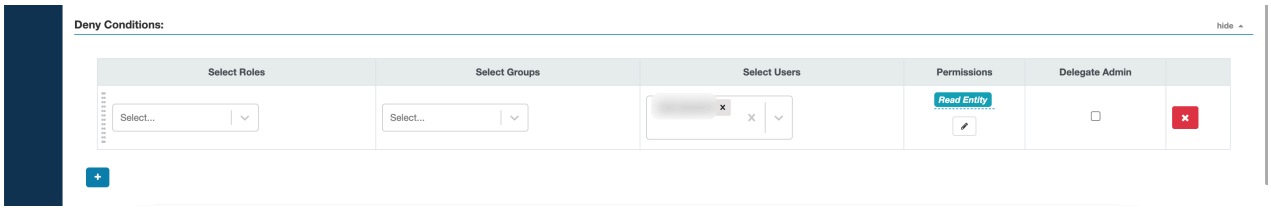
Restricting access for certain users of Cloudera Data Catalog

To have a fine-grained access to the same user from accessing the assets in Cloudera 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.

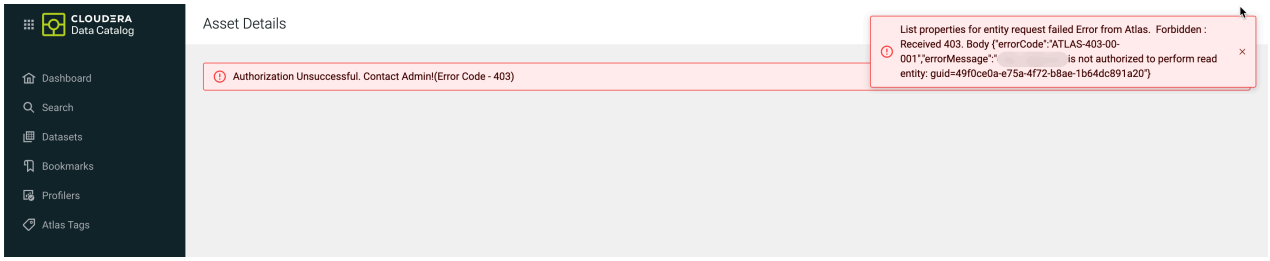
To create the Ranger policy to restrict users from accessing asset details, for example, with a specific classification, refer to the following images:



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



The result is depicted in the following image, where the user has no permissions to access the specified dataset.



Reducing resource consumption with restricted users

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

- Audit summarization 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 the Hive Column Profiler respectively.

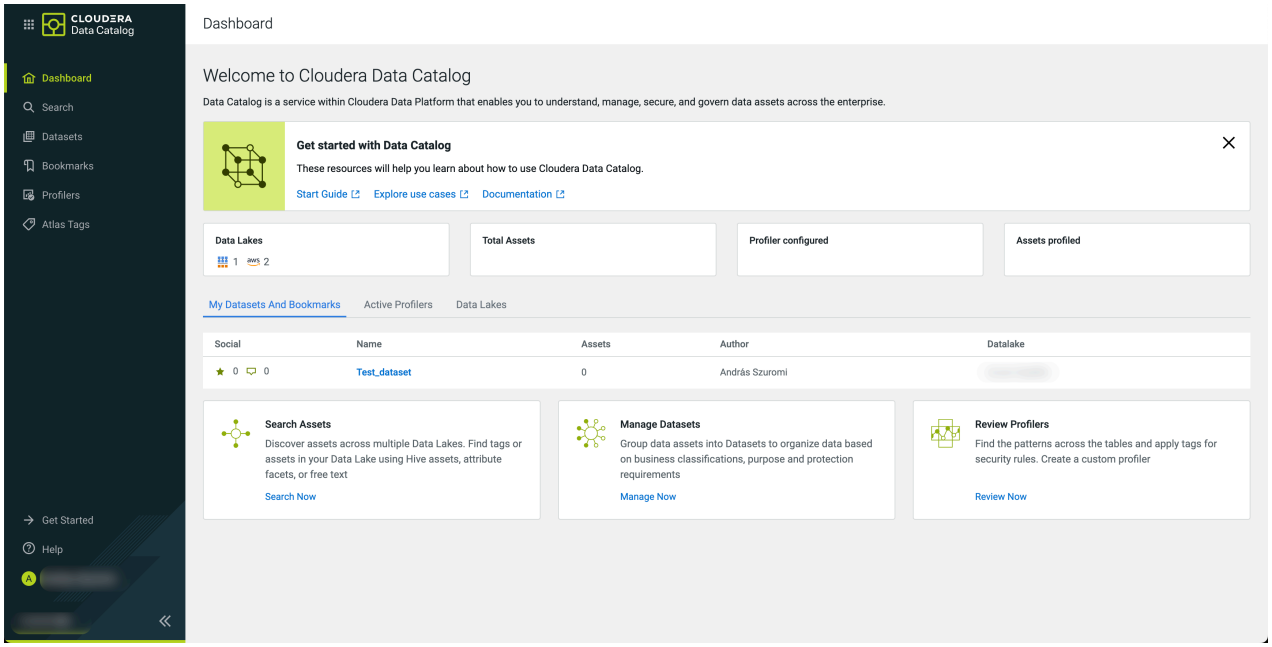
Related Information

Authorization for viewing Assets

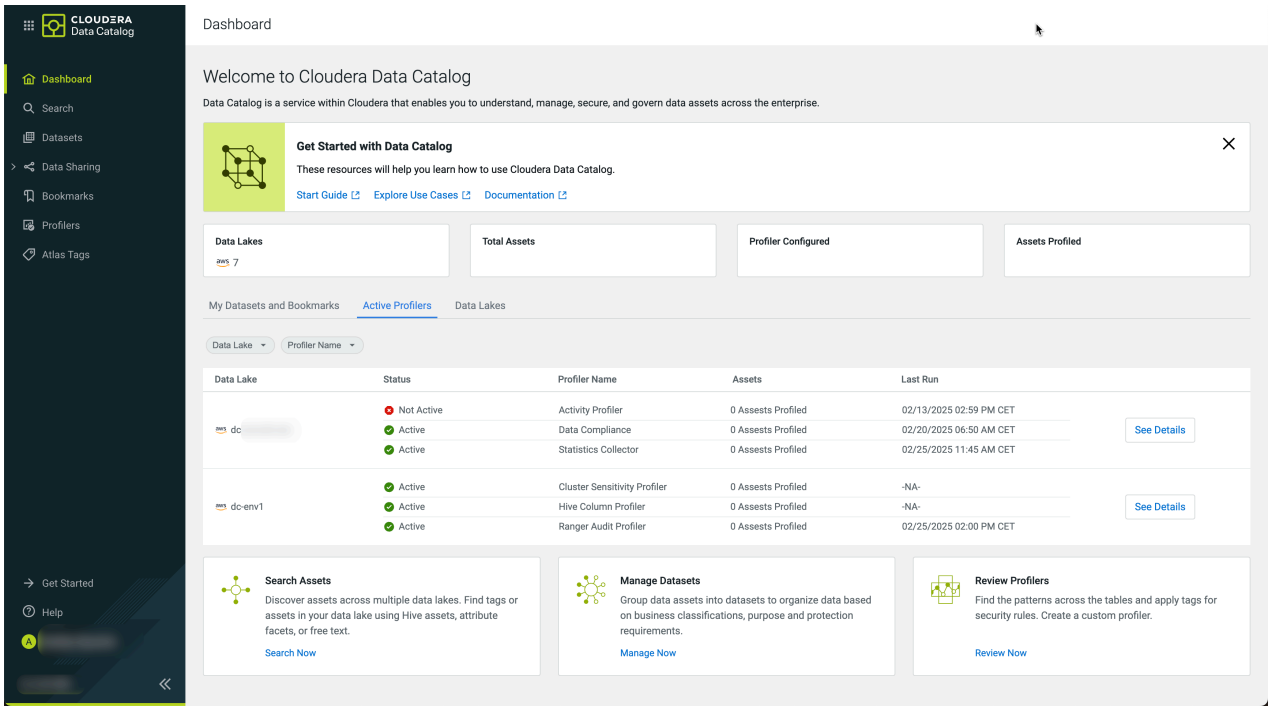
Cloudera Data Catalog Dashboard

The Dashboard 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.

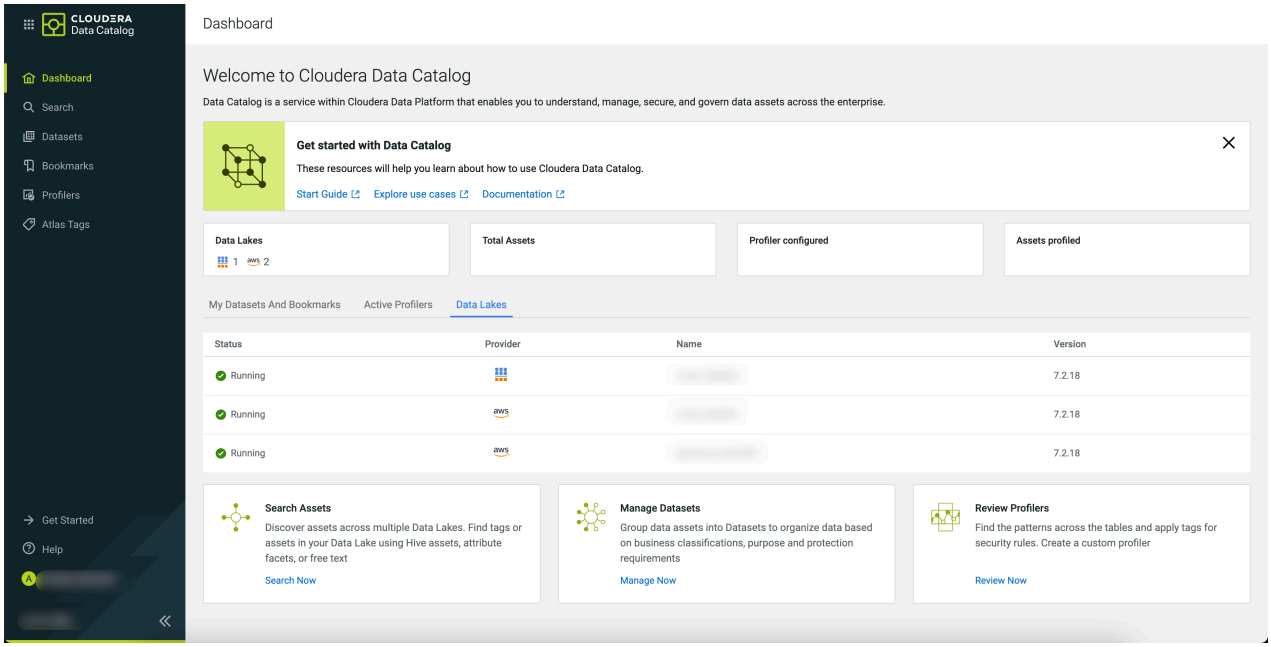
Data Stewards can view the **Dashboard** at a glance, and also focus on the most important tasks, enabling faster decision making as well as immediate action. The application lets you perform multiple actions for different types of content that helps in visualizing information with ease.



The displayed cards and tabs 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 **Dashboard** page contains information on the data lakes, the total number of assets that are profiled, along with the assets that are scanned for data.



Additionally, you can manage the datasets created and bookmarked by you, check the status of profilers, and search or discover assets.

Cloudera Data Catalog terminology

An overview of terminology used in Cloudera Data Catalog.

Profiler

Enables the Cloudera 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, a Data Lake can be realized in practice with an Cloudera Manager enabled Cloudera 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 Cloudera Data Services on premises 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 a Embedded Container Service cluster, which creates and manages an embedded Kubernetes infrastructure for use with Cloudera Data Services on premises.

Openshift Container (OCP)

OpenShift is an enterprise platform for container orchestration.

Data Asset

A data asset is a physical asset located in the Cloudera 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 Cloudera 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.

Related Information

About Cloudera Data Catalog

Introduction to Data Lakes