Integrating CDP Data Catalog with AWS Glue Data Catalog (Preview)

Date published: 2021-08-09
Date modified: 2021-12-08

Legal Notice

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.

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

Integrating CDP Data Catalog with AWS Glue Data Catalog

Integrating CDP Data Catalog with AWS Glue Catalog enables the users to browse and discover data as well as register data into SDX (via metadata translation or copy), so that it can be used with Data Hubs and other relevant experiences. While using AWS Glue in Data Catalog, you will be able to experience a complete snapshot metadata view, along with other visible attributes that can power your data governance capabilities.

How integration works

Assuming that the SDX is running in the users' AWS account (that contains the same AWS account which has GlueDataCatalog and the data that has to be discovered), the credentials with the ExternalDataDiscoveryService (which is hosted in SDX) must be shared, so that these two entities can interact with each other. These credentials are used to launch SDX and other workload clusters on the users' AWS account.

Prerequisites:

- You must have full access to AWS Glue Catalog and also have access to the EMR cluster's Hive Metastore instance.
- You must set up the CDP. For more information, see Getting Started with CDP.
- You must have access to your AWS IT Admin and CDP Admin user credentials, which is required to enable CDP to access AWS/EMR managed data in CDP.

Note: AWS policies are managed by AWS IAM and the AWS roles are added to the CDP Management Console. Refer to Using Instance Profile in AWS and Using credentials in Management Console. For more information about the AWS access, see AWS Environments.

Setting up AWS Glue Catalog with CDP Data Catalog

Follow these set up tasks to map your CDP Data Catalog instance with AWS Glue Catalog:

1. Enable the entitlement for your Data Catalog instance by running the following command on your CDP environment:

For example: *$ cdp coreadmin grant-entitlement --entitlement-name DATA_CATALOG_ENABLE_AWS_GLUE --account-id {account_id}*

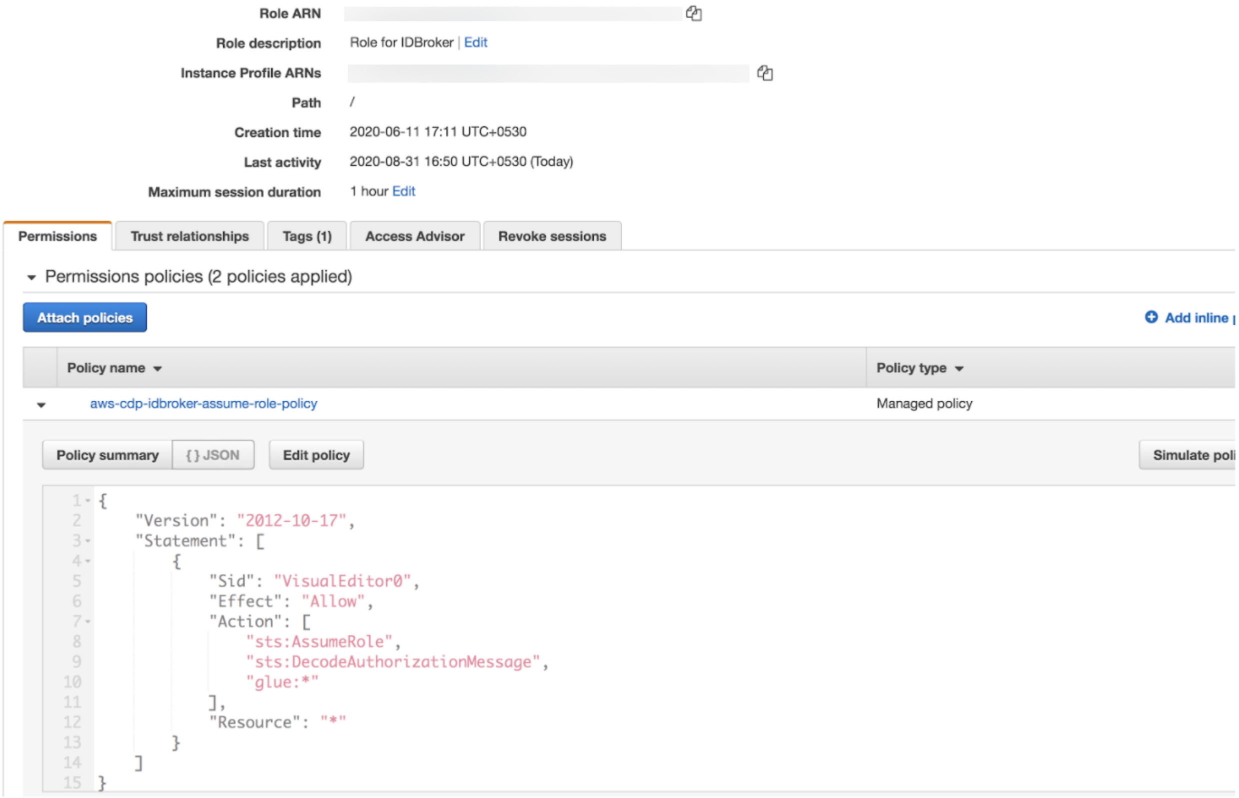
2. You must add relevant permissions in the corresponding AWS account:

    a. Include permission to access Glue Catalog service by editing the policy accordingly.

Make a note of the **Assumer Instance Profile** role that you intend to use and include full access authorization for AWS Glue.

Assumer Instance profile role is the one which you select in the available public cloud documentation listed under **step 10**. For more information, see [Register a CDP environment](#).

Use the following screenshots as a reference to complete the set up.



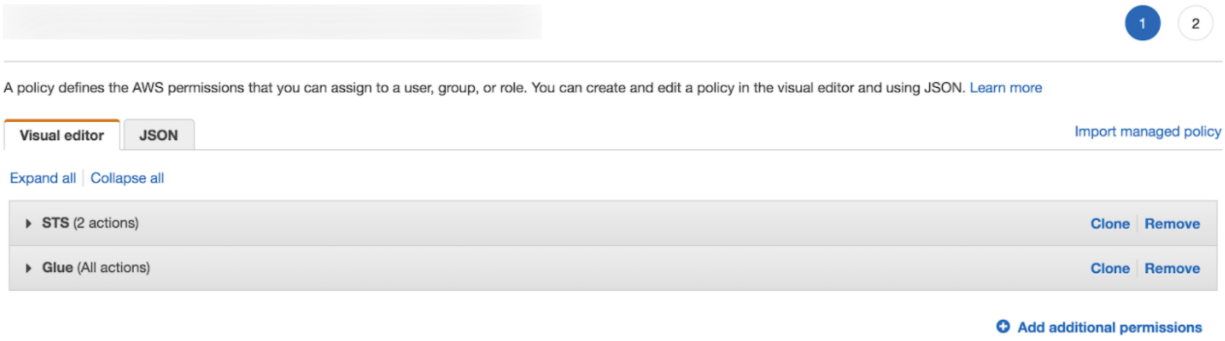
Role ARN [redacted]
Role description Role for IDBroker | Edit
Instance Profile ARNs [redacted]
Path /
Creation time 2020-06-11 17:11 UTC+0530
Last activity 2020-08-31 16:50 UTC+0530 (Today)
Maximum session duration 1 hour Edit

aws-cdp-idbroker-assume-role-policy Managed policy

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": [
                "sts:AssumeRole",
                "sts:DecodeAuthorizationMessage",
                "glue:*"
            ],
            "Resource": "*"
        }
    ]
}
```

Note: For **Role ARN** and **Instance Profile ARNs**, you must include the appropriate account number and role respectively.

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided 'as is' without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.



1 2

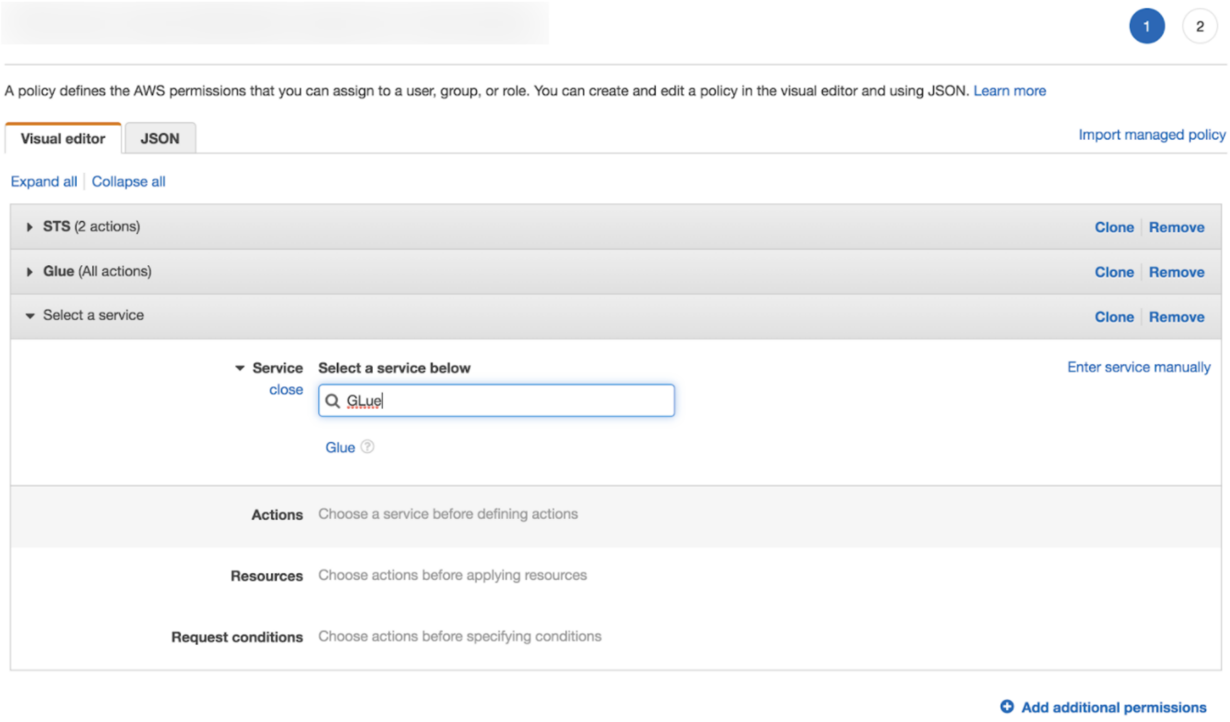
A policy defines the AWS permissions that you can assign to a user, group, or role. You can create and edit a policy in the visual editor and using JSON. Learn more

Visual editor | JSON | Import managed policy

Expand all | Collapse all

- ▸ **STS** (2 actions) — Clone | Remove
- ▸ **Glue** (All actions) — Clone | Remove

Add additional permissions



1 2

A policy defines the AWS permissions that you can assign to a user, group, or role. You can create and edit a policy in the visual editor and using JSON. Learn more

Visual editor | JSON | Import managed policy

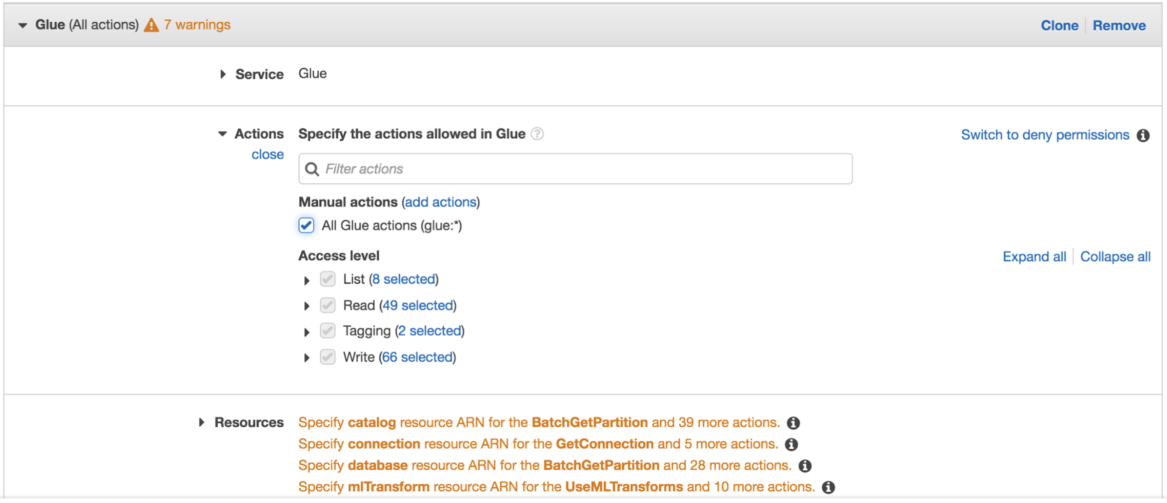
Expand all | Collapse all

- ▸ **STS** (2 actions) — Clone | Remove
- ▸ **Glue** (All actions) — Clone | Remove
- ▾ Select a service — Clone | Remove

▾ **Service** close	**Select a service below** — Enter service manually GLue Glue
Actions	Choose a service before defining actions
Resources	Choose actions before applying resources
Request conditions	Choose actions before specifying conditions

Add additional permissions

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided 'as is' without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.



▾ **Glue** (All actions) ⚠ 7 warnings — Clone | Remove

▸ **Service** Glue

▾ **Actions** close — **Specify the actions allowed in Glue** — Switch to deny permissions

Filter actions

Manual actions (add actions)

☑ All Glue actions (glue:*)

Access level — Expand all | Collapse all

▸ ☑ List (8 selected)

▸ ☑ Read (49 selected)

▸ ☑ Tagging (2 selected)

▸ ☑ Write (66 selected)

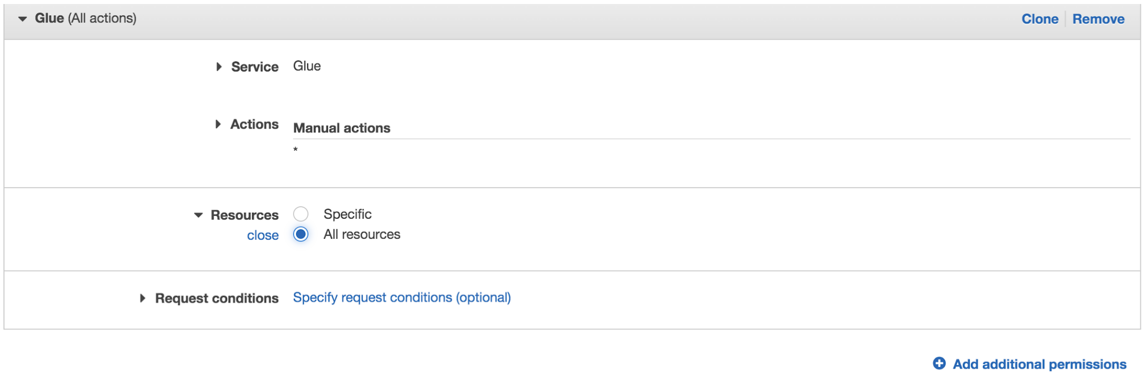
▸ **Resources**

Specify **catalog** resource ARN for the **BatchGetPartition** and 39 more actions.

Specify **connection** resource ARN for the **GetConnection** and 5 more actions.

Specify **database** resource ARN for the **BatchGetPartition** and 28 more actions.

Specify **mlTransform** resource ARN for the **UseMLTransforms** and 10 more actions.



▾ **Glue** (All actions) — Clone | Remove

▸ **Service** Glue

▸ **Actions** **Manual actions**

*

▾ **Resources** close

○ Specific

● All resources

▸ **Request conditions** Specify request conditions (optional)

⊕ Add additional permissions

b. Search for the role attached to the **Instance Profile** of the CDP environment.

Use the **Instance Profile** that you have configured above with Glue related policy in your AWS Environment creation command.

Use the following examples to setup AWS environment and AWS data lake as part of the Glue setup:

```
cdp environments create-aws-environment --profile default --cli-input-json '
{"environmentName”:”ab-ds-cli-7321”,
 "credentialName”:”cd2d-1234”,
 "Region":"us-region-2”,
             "securityAccess":{-insert the value--"},
 "Authentication":{---insert the value---"},

"logStorage":{"storageLocationBase":"s3a://demo-e2e-test-state-bucket/ab-ds-cli-7321/logs","instanceProfile":"arn:aws:iam::<xxxxxxxxxxx>:instance-profile/<role-name>"},
 "vpcId":"vpc-0123456”,
 "subnetIds":["subnet-04fe923b902aa5cf2","subnet-099c7a631f0ebed3c"],
 "s3GuardTableName":"dc-pro-cli-7210",
 "Description":"ab-ds-cli-7321",
 "enableTunnel":false,
 "workloadAnalytics":false,
 "freeIpa":{"instanceCountByGroup":1},
 }'

cdp environments set-id-broker-mappings \
--environment-name "ab-ds-cli-7321" \
--profile default \
--set-empty-mappings \
--data-access-role arn:aws:iam::<xxxxxxxxxxxx>:role/add-role \
--ranger-audit-role arn:aws:iam::<xxxxxxxxxxxx>:role/add-role
```

Similarly, while setting up the data lake use the **Instance Profile** that you configured above with Glue related policy in your data lake creation command:

```
cdp datalake create-aws-datalake --profile default --runtime 7.2.12 --cli-input-json '
{"datalakeName":"ab-ds-cli-7321-sdx",
 "environmentName":"ab-ds-cli-7321",

"cloudProviderConfiguration":{"instanceProfile":"arn:aws:iam::<xxxxxxxxxxx>:instance-profile/<role-name>","storageBucketLocation":"s3a://demo-e2e-test-state-bucket/ab-ds-cli-7321"},
 "scale":"LIGHT_DUTY",
 }'
```

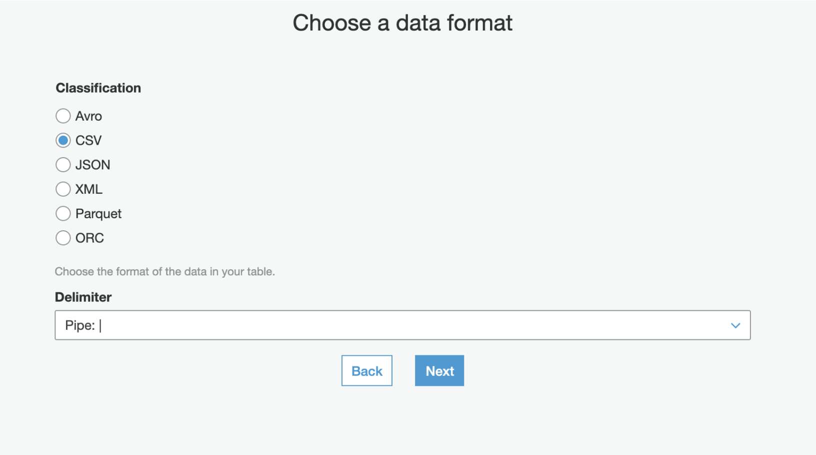
For more information, see Creating an AWS environment with a medium duty data lake using the CLI.

c. Navigate to the attached policy for the role.
d. When you manually create tables in AWS Glue Data Catalog, you must set the **fully qualified path** for the table location.

For example:
“`s3://manowar-e2e-test-state-bucket.s3-us-west-2.amazonaws.com/dc-pro-721-storage/glue/`”

3. You must set up the AWS Glue Data Catalog. For more information, see Populating the Glue Data Catalog. You must select only the CSV format which is currently supported for CDP Data Catalog and the delimiter which is used in the data.

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided ‘as is’ without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.



Choose a data format

Classification

○ Avro
● CSV
○ JSON
○ XML
○ Parquet
○ ORC

Choose the format of the data in your table.

Delimiter

Pipe: |

4. While creating tables in AWS Glue Data Catalog manually, set the fully qualified path for location. For example: ("s3://manowar-e2e-test-state-bucket.s3-us-west-2.amazonaws.com/dc-pro-721-storage/glue/")

Working with AWS Glue in CDP Data Catalog

The AWS Glue metadata must be registered with the CDP Data Catalog. The Glue contains the metadata that is synchronized with Data Catalog. The Glue metadata is accessed using the Data lake option in the Data Catalog service. After setting up AWS Glue with CDP Data Catalog and once the Glue synchronization with CDP is complete, the AWS Glue data lake appears in the Data Catalog instance.

Note: For each AWS environment, there would be a separate listing under the data lake option. The Glue data lake name / identifier in Data Catalog follows the format: <**glue**: **Data Lake Name**>. The Glue assets are of the type: **GLUE EXTERNAL TABLE**.

Accessing AWS Glue in Data Catalog

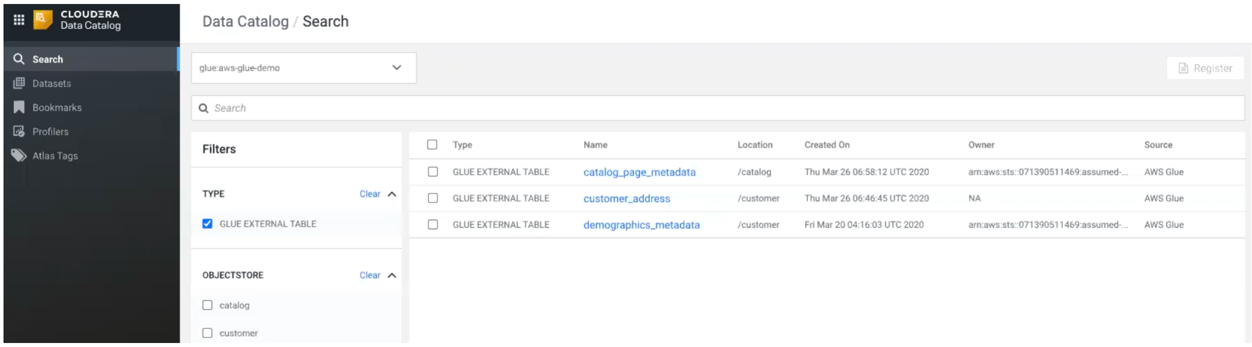
To access the Glue metadata in Data Catalog, you must note the following in your Data Catalog instance.

- List the Glue Metadata by selecting the Glue data lake.
- Select one or more Glue assets and register the same with CDP
- Verify if the registered Glue assets are listed in the Data Catalog owned data lake
- Select the registered Glue asset and click to open the **Asset Details** page

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided 'as is' without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.

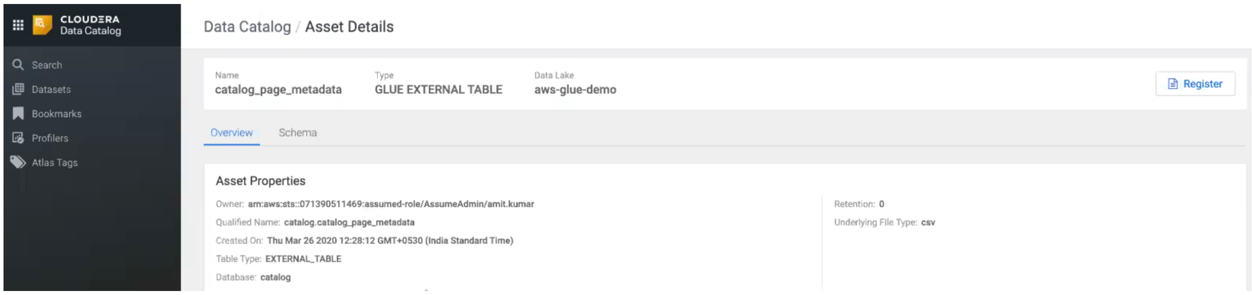
Listing the Glue assets

In Data Catalog, when you select the AWS Glue data lake, you can view the list of Glue assets. These metadata assets are directly sourced from Glue.



□	Type	Name	Location	Created On	Owner	Source
□	GLUE EXTERNAL TABLE	catalog_page_metadata	/catalog	Thu Mar 26 06:58:12 UTC 2020	arn:aws:sts::071390511469:assumed-...	AWS Glue
□	GLUE EXTERNAL TABLE	customer_address	/customer	Thu Mar 26 06:46:45 UTC 2020	NA	AWS Glue
□	GLUE EXTERNAL TABLE	demographics_metadata	/customer	Fri Mar 20 04:16:03 UTC 2020	arn:aws:sts::071390511469:assumed-...	AWS Glue

When you click on one of the assets, the Asset Details page is displayed.



Name: catalog_page_metadata
Type: GLUE EXTERNAL TABLE
Data Lake: aws-glue-demo

Owner: arn:aws:sts::071390511469:assumed-role/AssumeAdmin/amit.kumar
Qualified Name: catalog.catalog_page_metadata
Created On: Thu Mar 26 2020 12:28:12 GMT+0530 (India Standard Time)
Table Type: EXTERNAL_TABLE
Database: catalog

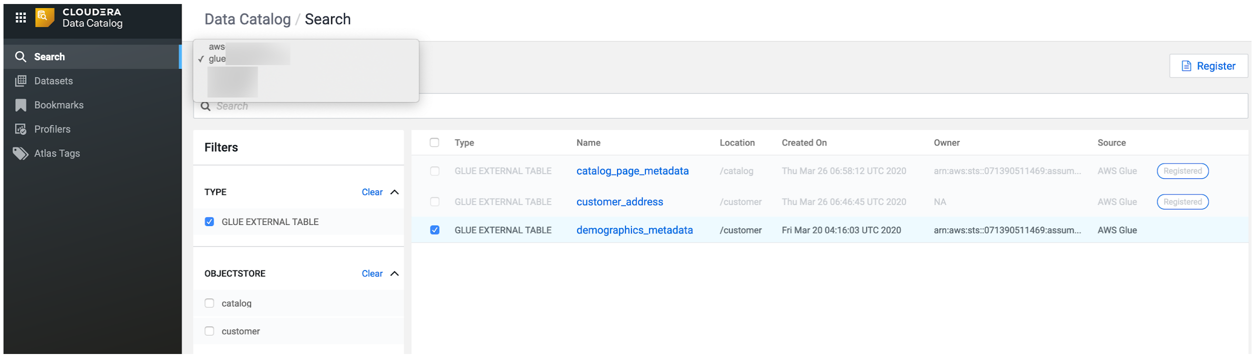
Retention: 0
Underlying File Type: csv

Next, on the main Data Catalog page, you must select the Glue data lake and select one of the Glue assets and register the asset to CDP. Click **Register**.

Optionally, you directly click on the Glue asset and register the asset on the **Asset Details** page.

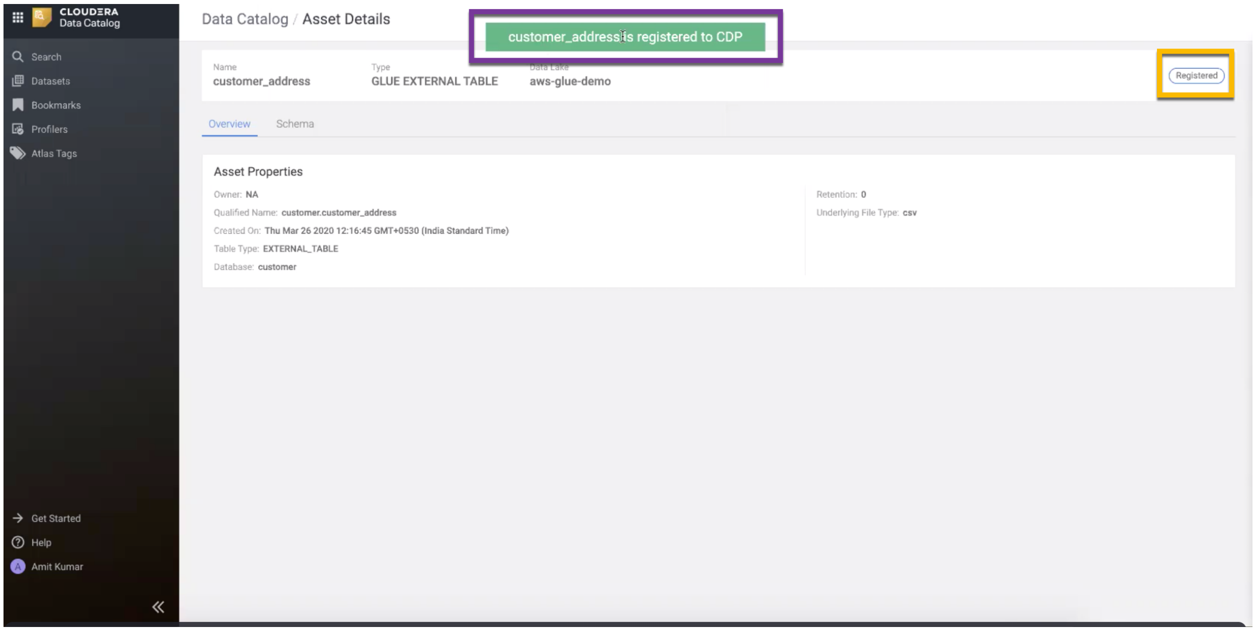
Note: You can select one or more Glue assets on the asset listing page to register them in CDP.

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided 'as is' without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.



	Type	Name	Location	Created On	Owner	Source	
□	GLUE EXTERNAL TABLE	catalog_page_metadata	/catalog	Thu Mar 26 06:58:12 UTC 2020	arn:aws:sts::071390511469:assum...	AWS Glue	Registered
□	GLUE EXTERNAL TABLE	customer_address	/customer	Thu Mar 26 06:46:45 UTC 2020	NA	AWS Glue	Registered
☑	GLUE EXTERNAL TABLE	demographics_metadata	/customer	Fri Mar 20 04:16:03 UTC 2020	arn:aws:sts::071390511469:assum...	AWS Glue	

Once the Glue asset is registered, the asset is imported into CDP.



customer_address is registered to CDP

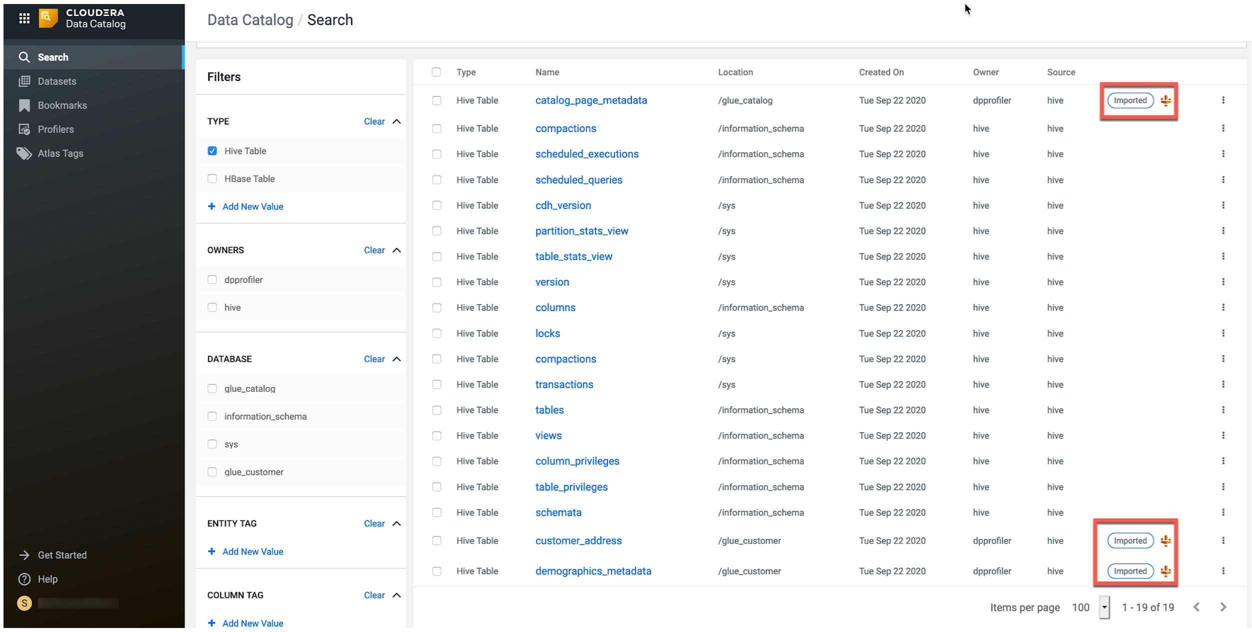
Name: customer_address
Type: GLUE EXTERNAL TABLE
Data Lake: aws-glue-demo

Asset Properties

Owner: NA
Qualified Name: customer.customer_address
Created On: Thu Mar 26 2020 12:16:45 GMT+0530 (India Standard Time)
Table Type: EXTERNAL_TABLE
Database: customer
Retention: 0
Underlying File Type: csv

Next, navigate back to the Data Catalog main page and select the Data Catalog owned data lake and select the type as **Hive Table**. The search results lists all the Hive table assets and you can view the Glue registered asset(s) as well. The registered Glue asset can be identified using the Imported icon.

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided 'as is' without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.



[redacted]

Data Catalog / Search

Filters

TYPE Clear
☑ Hive Table
□ HBase Table
+ Add New Value

OWNERS Clear
□ dpprofiler
□ hive

DATABASE Clear
□ glue_catalog
□ information_schema
□ sys
□ glue_customer

ENTITY TAG Clear
+ Add New Value

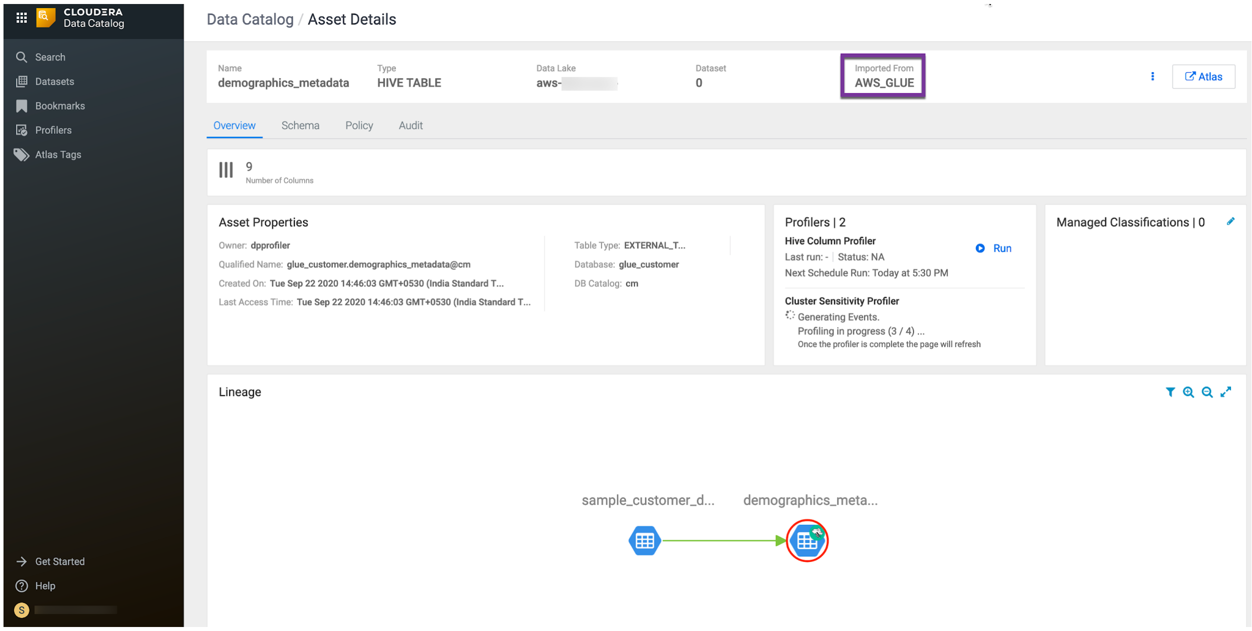
COLUMN TAG Clear
+ Add New Value

□	Type	Name	Location	Created On	Owner	Source		
□	Hive Table	catalog_page_metadata	/glue_catalog	Tue Sep 22 2020	dpprofiler	hive	Imported	⋮
□	Hive Table	compactions	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	scheduled_executions	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	scheduled_queries	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	cdh_version	/sys	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	partition_stats_view	/sys	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	table_stats_view	/sys	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	version	/sys	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	columns	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	locks	/sys	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	compactions	/sys	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	transactions	/sys	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	tables	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	views	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	column_privileges	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	table_privileges	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	schemata	/information_schema	Tue Sep 22 2020	hive	hive		⋮
□	Hive Table	customer_address	/glue_customer	Tue Sep 22 2020	dpprofiler	hive	Imported	⋮
□	Hive Table	demographics_metadata	/glue_customer	Tue Sep 22 2020	dpprofiler	hive	Imported	⋮

Items per page 100 1 - 19 of 19

Note: The entries in the Data Catalog for Glue assets are created in the Hive Metastore.

Click on the Glue registered asset and you can view the **Asset Details** page for the selected Glue asset.



[redacted]

Data Catalog / Asset Details

Name: demographics_metadata
Type: HIVE TABLE
Data Lake: aws-[redacted]
Dataset: 0
Imported From: AWS_GLUE

9
Number of Columns

Asset Properties

Owner: dpprofiler
Qualified Name: glue_customer.demographics_metadata@cm
Created On: Tue Sep 22 2020 14:46:03 GMT+0530 (India Standard T...
Last Access Time: Tue Sep 22 2020 14:46:03 GMT+0530 (India Standard T...
Table Type: EXTERNAL_T...
Database: glue_customer
DB Catalog: cm

Profilers | 2

Hive Column Profiler
Last run: - | Status: NA
Next Schedule Run: Today at 5:30 PM

Cluster Sensitivity Profiler
Generating Events.
Profiling in progress (3 / 4) ...
Once the profiler is complete the page will refresh

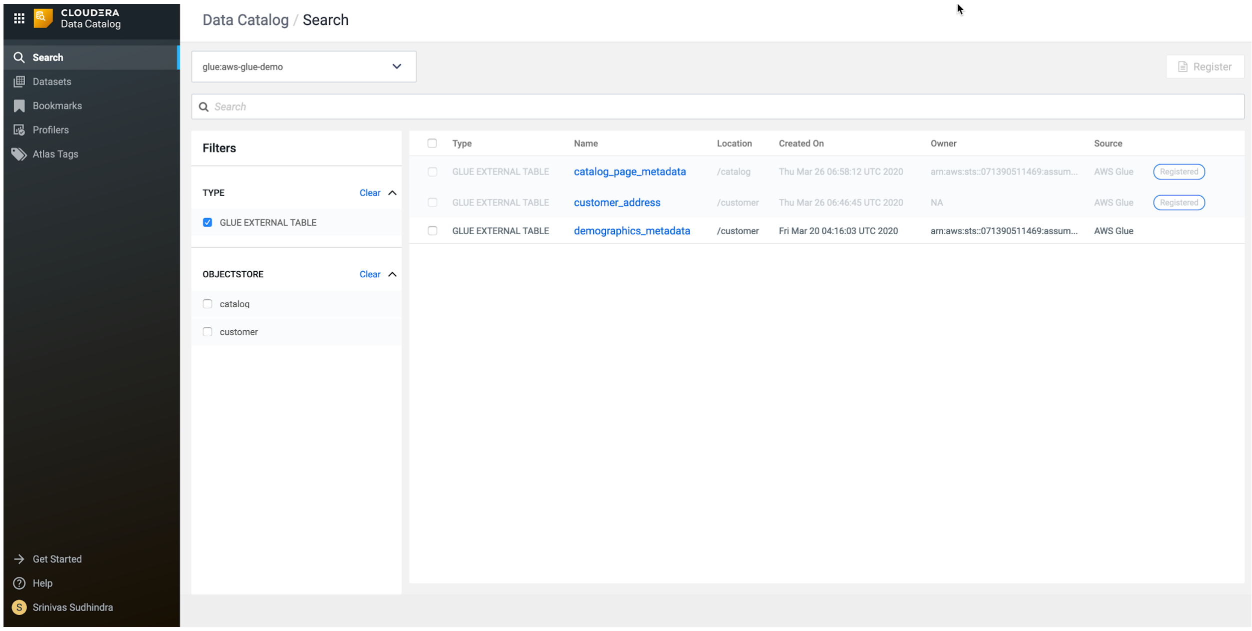
Managed Classifications | 0

Lineage

The Asset Details page for the Glue asset is populated by Atlas. While registering the Glue data, the data was written to the Hive Metastore and later Atlas synchronised the metadata.

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided 'as is' without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.

Go back to the main **Data Catalog** page and select the Glue data lake. Note that the registered Glue asset(s) are greyed out or cannot be selected again.



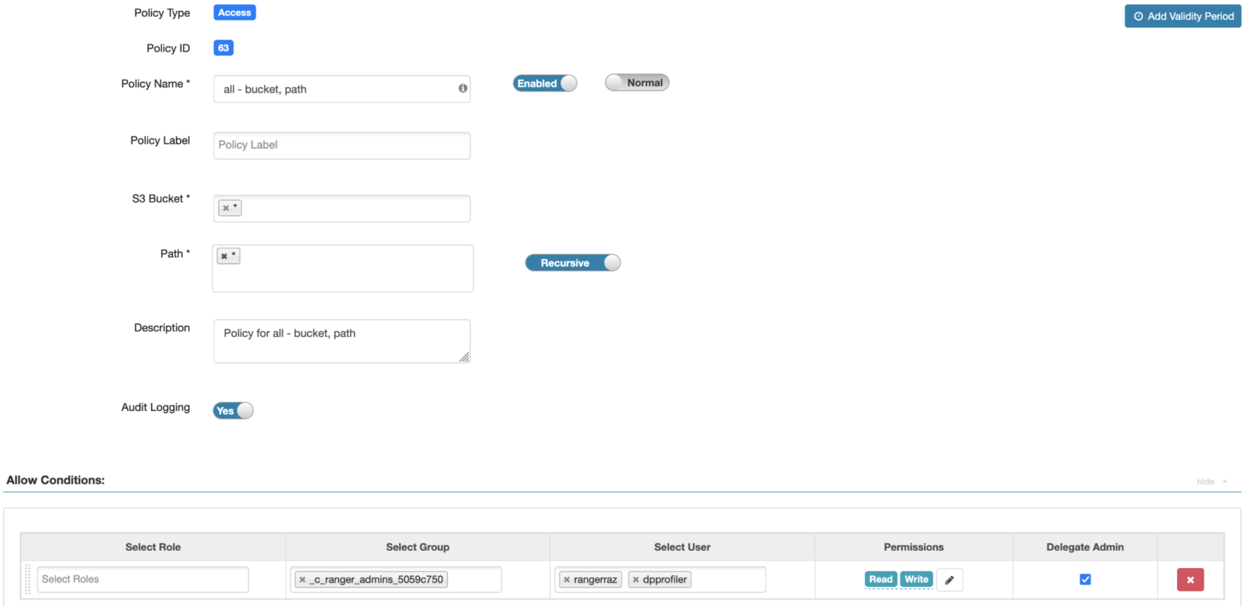
	Type	Name	Location	Created On	Owner	Source	
□	GLUE EXTERNAL TABLE	catalog_page_metadata	/catalog	Thu Mar 26 06:58:12 UTC 2020	arn:aws:sts::071390511469:assum...	AWS Glue	Registered
□	GLUE EXTERNAL TABLE	customer_address	/customer	Thu Mar 26 06:46:45 UTC 2020	NA	AWS Glue	Registered
□	GLUE EXTERNAL TABLE	demographics_metadata	/customer	Fri Mar 20 04:16:03 UTC 2020	arn:aws:sts::071390511469:assum...	AWS Glue	

You can still view the registered Glue assets (powered by Atlas) by clicking on the same and it navigates to the **Asset Details** page as seen above in the image.

Working with Ranger Authorization Service (RAZ) enabled AWS environment

For RAZ enabled AWS environment, you must employ the following permission settings to work with Data Catalog - Glue Integration.

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided 'as is' without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.



Policy Type: Access

Policy ID: 63

Policy Name *: all - bucket, path — Enabled — Normal

Policy Label: Policy Label

S3 Bucket *: × *

Path *: × * — Recursive

Description: Policy for all - bucket, path

Audit Logging: Yes

Allow Conditions:

Select Role	Select Group	Select User	Permissions	Delegate Admin	
Select Roles	× _c_ranger_admins_5059c750	× rangerraz × dpprofiler	Read Write	☑	×

Note: By default “**dpprofiler**” user is not included in the allowed users list. You must manually add the “**dpprofiler**” user in the allowed users list.

Unsupported features

For a Glue data lake, while searching for a specific asset, multiple calls fail with “**400 Bad Request**” which is reflected on the Data Catalog UI in search suggestions. Search for Glue assets in Data Catalog is currently not supported.

Known Issues

CDPDSS-1179: Any Glue registered asset does not fully display the respective Asset Details page when accessed from the Glue Data Lake.

Problem: When any asset in the Data Catalog is Glue registered, it does not display any information in the Asset Details page, when viewed from the Glue Data Lake.

Workaround: Select the same asset in Data Catalog, which is not Glue Data Lake and click on the same to view the Asset Details information.