Managed Storage Access for AWS (Preview)

Date published: 2021-10-21

Date modified: 2021-10-21

Legal Notice

© Cloudera Inc. 2021. 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.

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

Introduction

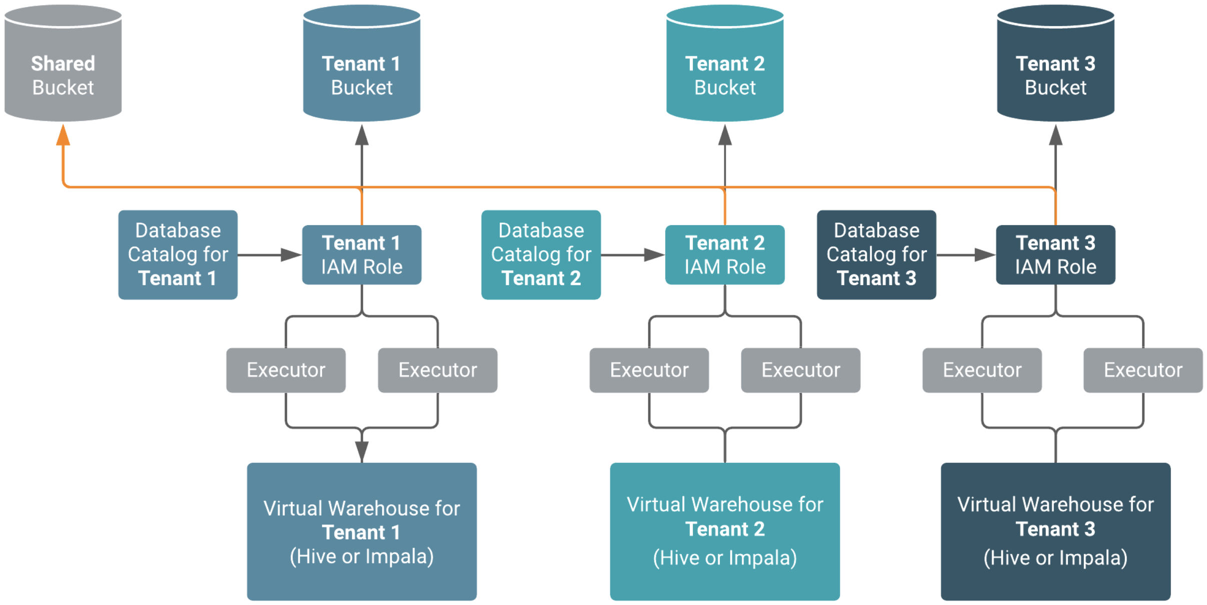
Understanding how Cloudera Data Warehouse (CDW) stores data for multiple tenants and a high-level overview of the configuration tasks prepares you as DWAdmin to set up a managed storage warehouse.

The multitenant storage technique in Cloudera Data Warehouse (CDW) offers increased security over the storage method used in earlier releases that based all storage access in CDW on a single EC2 instance role. The old method listing all S3 buckets that need to be accessible by Hive or Impala made the impact of a leaked access token for a role unacceptably wide, even though the access token is normally very time-constrained. The new technique requires a separate Database Catalog plus at least one Virtual Warehouse per tenant.

You configure a dedicated IAM role to give only the tenant-specific default database catalog instance and its associated virtual warehouse data access to the following buckets:

- The tenant-specific bucket(s)
- Potentially shared bucket(s)
- The SDX Data Lake bucket

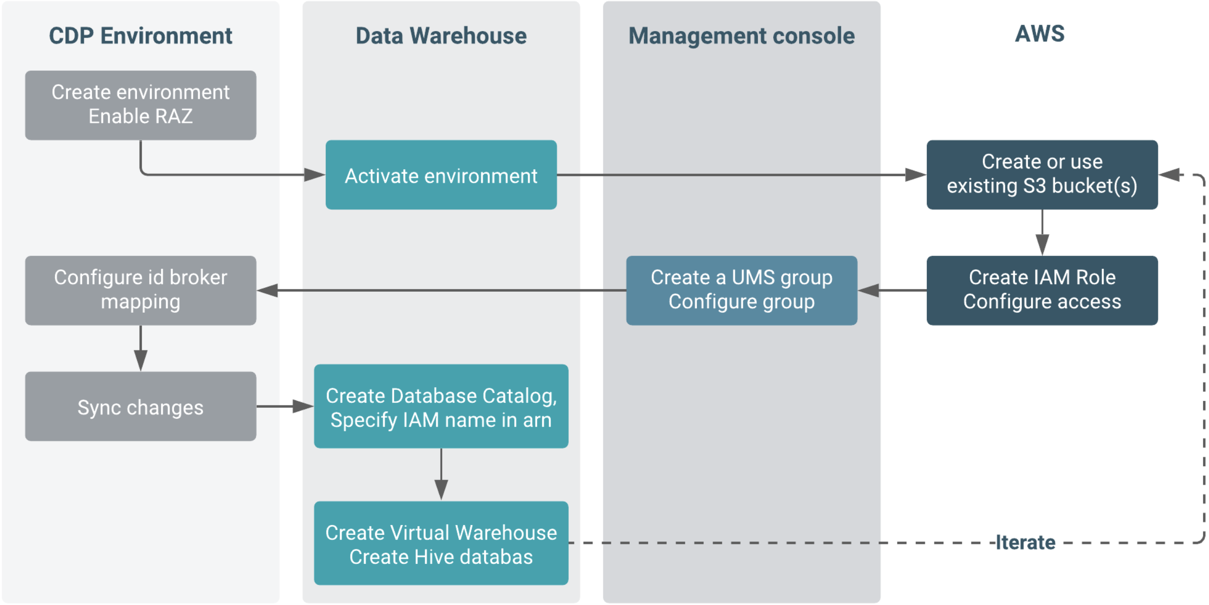
If the access token for a role is leaked, the data of others is not compromised; only one tenant's own, and shared, data is vulnerable. A custom identity broker IDBroker approach to accessing the Virtual Warehouse is depicted in the following graphic:


Shared Bucket
Tenant 1 Bucket
Tenant 2 Bucket
Tenant 3 Bucket
Database Catalog for Tenant 1
Tenant 1 IAM Role
Database Catalog for Tenant 2
Tenant 2 IAM Role
Database Catalog for Tenant 3
Tenant 3 IAM Role
Executor
Executor
Executor
Executor
Executor
Executor
Virtual Warehouse for Tenant 1 (Hive or Impala)
Virtual Warehouse for Tenant 2 (Hive or Impala)
Virtual Warehouse for Tenant 3 (Hive or Impala)

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.

Using a Ranger Remote Authorization (RAZ)-enabled environment, you can control access to data based on user roles and classifications. As a CDP Admin, you can apply Ranger fine-grained access control policies to S3 buckets, directories, and files.

The following diagram shows the high-level steps for configuring the managed storage access:


CDP Environment
Data Warehouse
Management console
AWS
Create environment Enable RAZ
Activate environment
Create or use existing S3 bucket(s)
Configure id broker mapping
Create a UMS group Configure group
Create IAM Role Configure access
Sync changes
Create Database Catalog, Specify IAM name in arn
Create Virtual Warehouse Create Hive databas
Iterate

After you obtain the entitlements required for this feature CDW_ALLOW_MULTI_DEFAULT_DBC and CDW_STORAGE_ROLES, you can configure storage as shown in this diagram:

- In CDP, register an environment that enables RAZ and uses the SDX Data Lake.
- Activate the environment in CDW.
- Manually create S3 bucket(s), or use existing ones, for access by the IAM tenant role.
- In AWS, create an IAM role for the tenant.
- Give the IAM role certain permissions.
- Configure a trust relationship with the IDBROKER_ROLE.
- Give the Instance Profile (idbroker assume role) the permission to assume the role of the tenant.
- In the CDP Management Console, create a UMS group.
- Add the UMS machine user (created when you activated the CDW environment) to the UMS group.
- Add the id broker mapping to the environment.
- Sync user group changes for the RAZ-enabled environment with FreeIPA.
- Create a separate Database Catalog with a unique IAM role.

- Create a tenant-specific Virtual Warehouse based on the Database Catalog, SDX Data Lake, and RAZ-enabled environment.
- Create a tenant-specific Hive or Impala database that points to tenant-specific buckets.

  As the metadata is shared across all tenants, Ranger grants access to tenant users' data via a group at the database level.

- For each tenant, repeat the actions above, starting from the step after activating the environment for CDW.

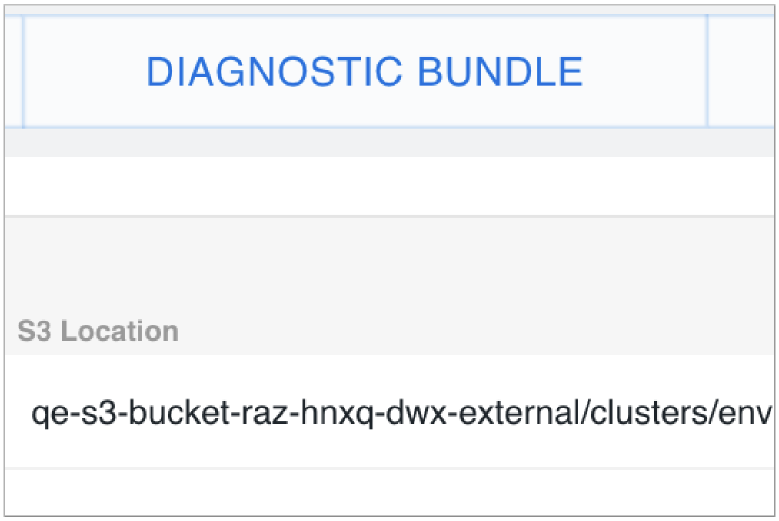
The IAM role accesses a tenant specific bucket s3-tenant-1 and SDX Data Lake bucket. The SDX Data Lake bucket is needed for shared data and other data used by Database Catalog and to integrate WXM. WXM writes Impala query data to the shared bucket. You need the UMS group to add the IDBroker mapping, as a single UMS machine user cannot have multiple ID broker mappings to different IAM roles.

The following topics describe step-by-step how to set up your environment, IAM roles, Database Catalog, and Virtual Warehouse for managed storage access.

Limitations of managed storage access

The diagnostic bundle for a Virtual Warehouse is correctly located in an SDX Data Lake bucket. However, there is a known UI issue with the location of the diagnostic bundle.

In the Diagnostic Bundles > General Information, the first component of the path has an external bucket name instead of an SDX Data Lake bucket name.



DIAGNOSTIC BUNDLE

S3 Location

qe-s3-bucket-raz-hnxq-dwx-external/clusters/env

The correct path would look something like this:
<SDX-DL_BUCKET>/clusters/<cluster-id>/<warehouse-id>/warehouse/debug-artifacts/hive/<compute-id>/<compute-id>-0723024748-0723144748-01234.zip.

Setting up managed storage access

About this task: The RAZ-controlled warehouse supports only multiple Shared Data Experience (SDX) Database Catalogs instead of CDW-specific, isolated database catalogs. In a RAZ-controlled environment, the Elastic Kubernetes Service (EKS) ec2 executor will not have read/write permissions to the S3 bucket. Consequently, after activating the CDW environment, you cannot remove RAZ control. RAZ-control of CDW continues during upgrades of the Database Catalog and Virtual Warehouse.

Before you begin:

- Request activation of the following entitlements:
  - CDW_ALLOW_MULTI_DEFAULT_DBC
  - CDW_STORAGE_ROLES
- You meet the requirements described in the AWS requirements documentation.

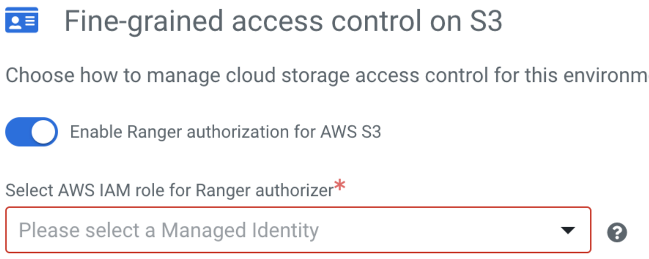
Required role: PowerUser

Creating the CDP environment and IAM roles

You need to register and activate a new environment to enable RAZ in CDW.

1. Register an environment with RAZ enabled using the CDP web interface or CDP CLI.

   In the web interface, in Fine-grained access control on S3, select Enable Ranger Authorization for AWS S3.



Fine-grained access control on S3

Choose how to manage cloud storage access control for this environment

Enable Ranger authorization for AWS S3

Select AWS IAM role for Ranger authorizer*

Please select a Managed Identity

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.

2. In Cloudera Data Warehouse Overview, locate the RAZ-enabled environment, and click Activate ⚡ to activate the environment for CDW.

   This action disables the standard default Database Catalog automatically created after activation. The UMS machine user is created and attached to the environment when you activate the CDW environment. Later, you see how to add this same UMS machine user to a different UMS group for each tenant.

   Repeat the following steps for each tenant.

3. In AWS, manually create one or more S3 buckets for the tenant, or use existing buckets. Use bucket-level encryption and a key that is accessible by the IAM role of the tenant.

4. In AWS, create an IAM role that has significant privileges to read/write to all tenant-specific buckets as well as across all tenants' shared SDX Data Lake bucket.

   For example, create a role named role-tenant-1 and a bucket called s3-tenant-1.

5. In the IAM role, configure a trust relationship with IDBROKER_ROLE.

   For example:{

```
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": [
          "s3.amazonaws.com",
          "ec2.amazonaws.com"
        ],
        "AWS":
"arn:aws:iam::<AWS_ACCOUNT_ID>:IDBROKER_ROLE"
      },
      "Action": "sts:AssumeRole"
    }
  ]

}
```

   The IDBROKER_ROLE needs the trust relationship to assume the role role-tenant-1.

6. Make a note of the IAM role ARN.

7. Attach the following AWS policies to the IAM role: AmazonEKSWorkerNodePolicy, AmazonEKS_CNI_Policy, AmazonEC2ContainerRegistryReadOnly.

   The AmazonEKSWorkerNodePolicy is shown in the next topic.

8. Give the AWS Instance Profile (idbroker assume role), the permission to assume role of role-tenant-1.

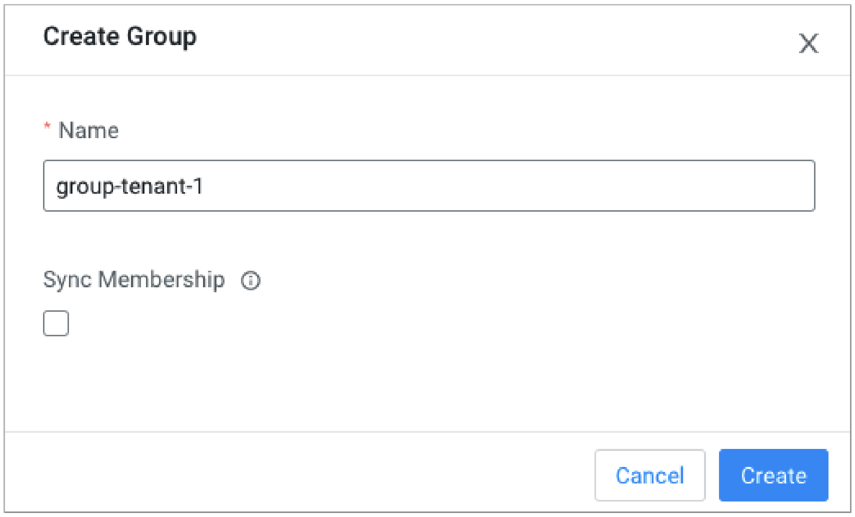
```
 {

    "Version": "2012-10-17",

    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": "sts:AssumeRole",
            "Resource": [

"arn:aws:iam::<AWS_ACCOUNT_ID>:role/idbroker-data-access-role",

"arn:aws:iam::<AWS_ACCOUNT_ID>:role/mock-idbroker-admin-role",
             "arn:aws:iam::<AWS_ACCOUNT_ID>:role/role-tenant-1"
            ]
        }
    ]
 }
```

Creating a UMS group and machine users

This procedure ensures that the CDP machine user gets permission to access the tenant bucket.

Repeat the following steps for each tenant:

1. In Management Console, > User Management > Groups, click CREATE GROUP and create a User Management Service (UMS) group, for example group-tenant-1.



Create Group

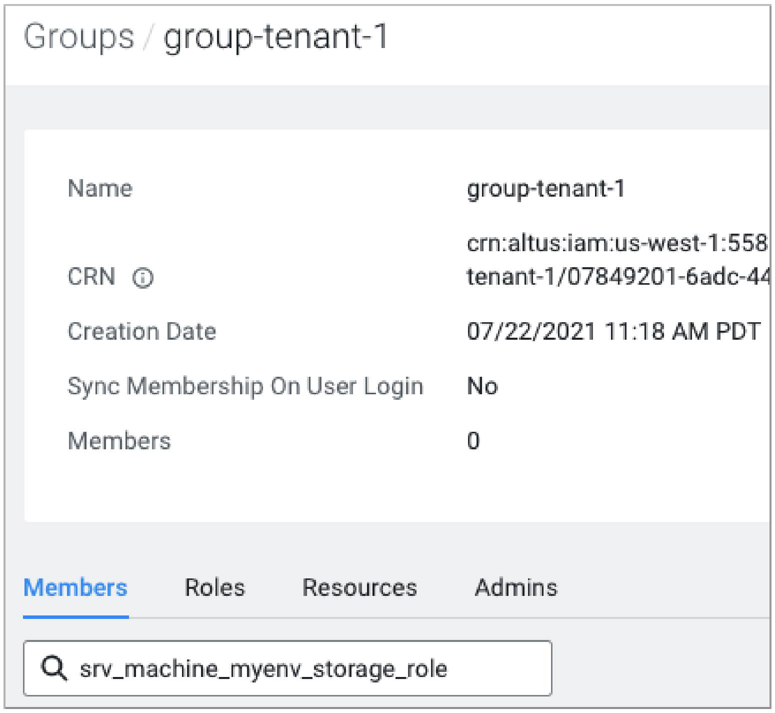
* Name

group-tenant-1

Sync Membership ⓘ

☐

2. In Groups > Members, search for and select your srv_machine_<env id>_storage_role to add this UMS machine user to group-tenant-1.



Groups / group-tenant-1

Name	group-tenant-1
CRN ⓘ	crn:altus:iam:us-west-1:558 tenant-1/07849201-6adc-44
Creation Date	07/22/2021 11:18 AM PDT
Sync Membership On User Login	No
Members	0

srv_machine_myenv_storage_role

3. In Management Console >Environments, select an environment , and click Actions > Manage Access > IDBroker Mappings > Edit.
4. Click + to add a mapping, select the Group-tenant-1 and Role-tenant-1, and specify the role ARN (copied from the IAM role page on AWS).
5. Sync your group changes with FreeIPA by performing a user sync per environment: In the RAZ-enabled environment, click Actions > Synchronize Users to FreeIPA.

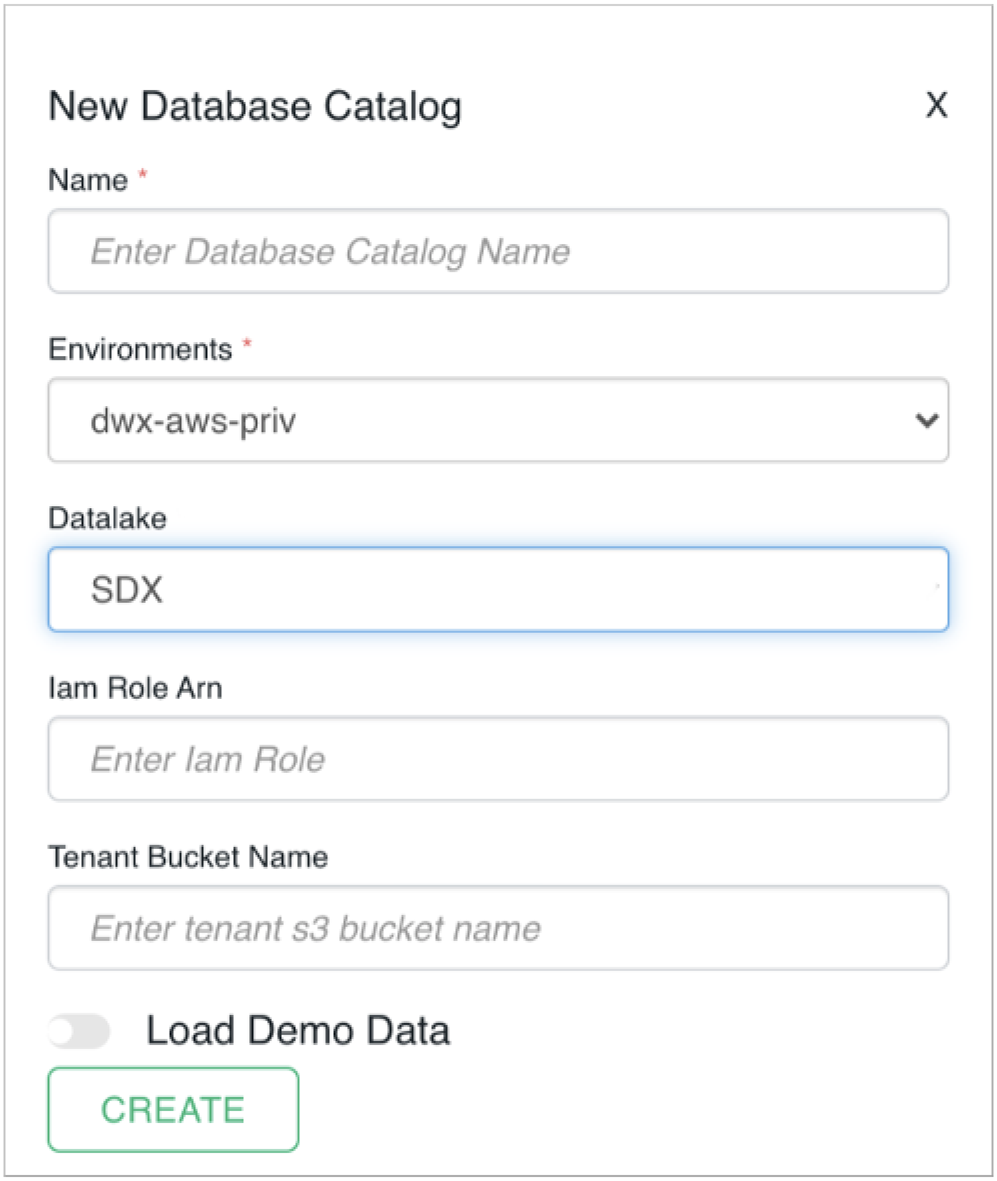
   The UMS machine user gets the permission to access the s3-tenant-1 bucket.

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.

Creating a new Database Catalog

Repeat the following steps for each tenant.

1. Click Data Warehouse > Database Catalog > ADD NEW.



New Database Catalog

Name *

Enter Database Catalog Name

Environments *

dwx-aws-priv

Datalake

SDX

Iam Role Arn

Enter Iam Role

Tenant Bucket Name

Enter tenant s3 bucket name

Load Demo Data

2. In New Database Catalog, enter a role name, role-tenant-1 for example.
3. Select the RAZ-enabled environment.
   In Data Lake, SDX is the required value. The backend Data Lake and Database Catalog database must be the same.
4. Enter the Iam Role ARN you copied earlier.
   Take care to enter the ARN correctly because CDW does not validate your ARN.
5. Enter the tenant bucket name, for example s3-tenant-1, and click CREATE.

Creating a tenant-specific Virtual Warehouse

You need to create a Virtual Warehouse based on the Database Catalog and SDX Data Lake you created in the RAZ-enabled environment. The Database Catalog and Data Lake point to the same backend database. In the tenant-specific Virtual Warehouse, you create a tenant-specific Hive or Impala database that points to tenant-specific buckets. As the metadata is shared across all tenants, Ranger grants access to tenant data at the table level. One or more tenant-specific databases alongside databases for shared data can run in the same HMS instance.

Repeat the following steps for each tenant.

1. In the Data Warehouse service, click Virtual Warehouses > Add New.
2. Specify a name, select the Hive or Impala type, and select the Database Catalog you created for the tenant.
3. On the Overview page under Virtual Warehouses, click options ⋮ > Open Hue.
4. Create a tenant-specific Hive or Impala database where the location for external and managed tables are pointing to the tenant-specific buckets.

   ```
   CREATE (DATABASE|SCHEMA) [IF NOT EXISTS] database_name
   [COMMENT database_comment]
   [LOCATION external_table_path]
   [MANAGEDLOCATION managed_table_directory_path]
   [WITH DBPROPERTIES (property_name=property_value, ...)];
   ```

   Do not set LOCATION and MANAGEDLOCATION to the same path. For more information, see Create a default directory for managed tables.

5. In Ranger, grant the tenant users access to the tenant-specific Hive or Impala database.

Amazon EKS worker node policy example

The JSON code for the EKS worker node policy describes the policy. The policy is representative of one of the three policies you add to the IAM role.

eks-ec2-workernode-policy.json

```
{
 "Version": "2012-10-17",
  "Statement": [
  {
    "Sid": "VisualEditor0",
    "Effect": "Allow",
```

```
      "Action": [
        "kms:Decrypt",
        "s3:ListAllMyBuckets",
        "kms:Encrypt",
        "kms:ListAliases",
        "kms:GenerateDataKey",
        "kms:DescribeKey"
        ],
      "Resource": "*"
    }
    ]
}
```

s3-read-write

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Action": [
                "s3:Get*",
                "s3:Delete*",
                "s3:Put*",
                "s3:ListBucket",
                "s3:ListBucketMultipartUploads",
                "s3:AbortMultipartUpload",
                "s3:GetBucketLocation"
            ],
            "Resource": [
                "arn:aws:s3:::<SDX_BUCKET>",
                "arn:aws:s3:::<SDX_BUCKET>/*",
                "arn:aws:s3:::tenant-1-bucket-1",
                "arn:aws:s3:::tenant-1-bucket-1/*"
            ],
            "Effect": "Allow"
        }
    ]
}
```