Managed Storage Access for Azure (Preview)

Date published: 2021-10-21
Date modified: 2021-10-21

Legal Notice

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.

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

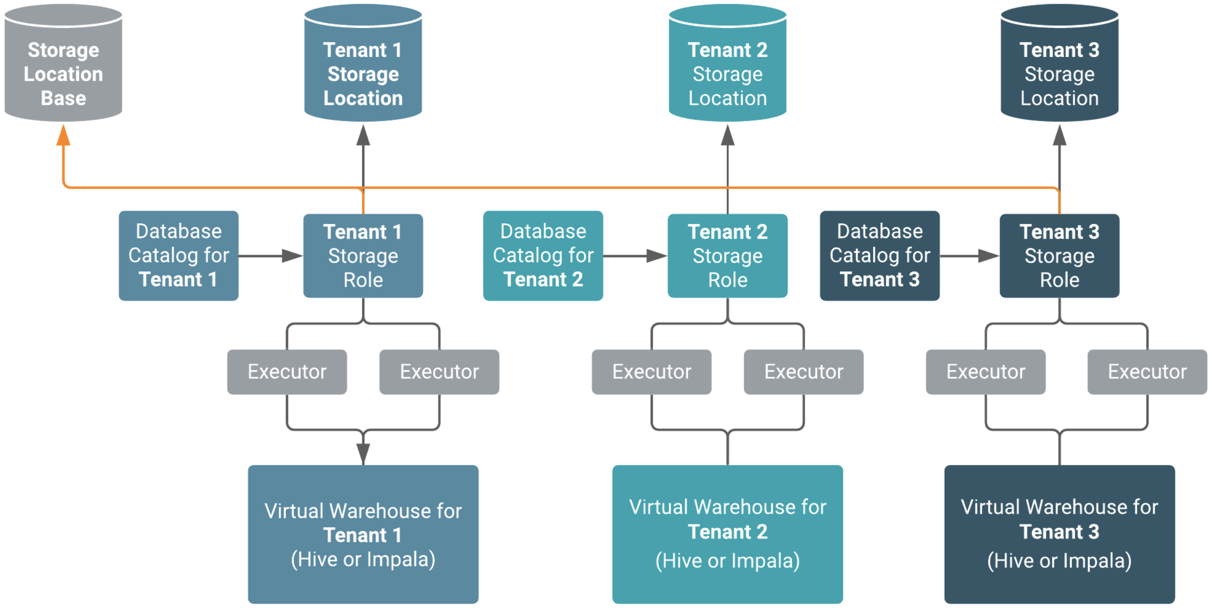
Understanding how Cloudera Data Warehouse (CDW) stores data for multiple tenants and a high-level overview of the configuration tasks prepares you as DWAdmin to set up a RAZ-controlled warehouse.

The multitenant storage technique in CDW requires a separate Database Catalog plus at least one Virtual Warehouse per tenant.

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

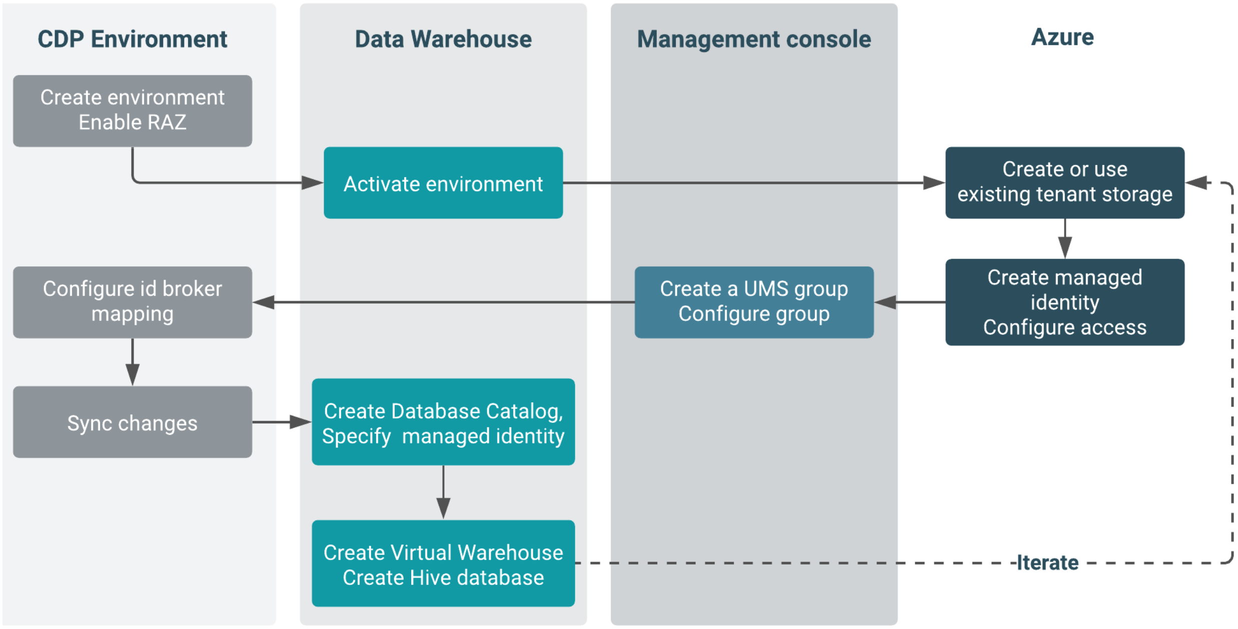
- The tenant storage location
- Potentially shared Storage Location Base
- The SDX Data Lake bucket

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 managed identity IDBroker approach to accessing the Virtual Warehouse is shown in the following graphic:


Storage Location Base
Tenant 1 Storage Location
Tenant 2 Storage Location
Tenant 3 Storage Location
Database Catalog for Tenant 1
Tenant 1 Storage Role
Database Catalog for Tenant 2
Tenant 2 Storage Role
Database Catalog for Tenant 3
Tenant 3 Storage 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)

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 cloud storage.

The following diagram shows the high-level steps for configuring RAZ-enabled storage:


CDP Environment
Data Warehouse
Management console
Azure
Create environment Enable RAZ
Activate environment
Create or use existing tenant storage
Configure id broker mapping
Create a UMS group Configure group
Create managed identity Configure access
Sync changes
Create Database Catalog, Specify managed identity
Create Virtual Warehouse Create Hive database
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 a tenant storage location(s), or use existing ones, for access by the tenant storage role.
- In Azure, create an Azure managed identity for the tenant.
- Assign the tenant specific Azure managed identity as a Storage Blob Data Owner role to the tenant-specific container; assign the tenant specific Azure managed identity as a Storage Blob Data Owner role to Storage Location Base.
- In CDP Management Console, create an UMS group for this tenant.
- 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 managed identity created in the step above.
- 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 storage locations. As the metadata is shared across all tenants, Ranger grants access to tenant 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 managed identity accesses a tenant specific location and SDX Data Lake shared Storage Location Base. The Storage Location Base contains shared data, stores the Directed Acyclic Graph (DAG) data used by the Database Catalog, and provides integration with Cloudera Workload XM. WXM writes Impala query data to the shared Storage Location Base.

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 storing RAZ-enabled data.

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:
    1. CDW_ALLOW_MULTI_DEFAULT_DBC
    2. CDW_STORAGE_ROLES
- You meet the requirements described in the Azure requirements documentation.

Required role: PowerUser

Creating the CDP environment

About this task: You need to register and activate a new environment to enable RAZ in CDW.

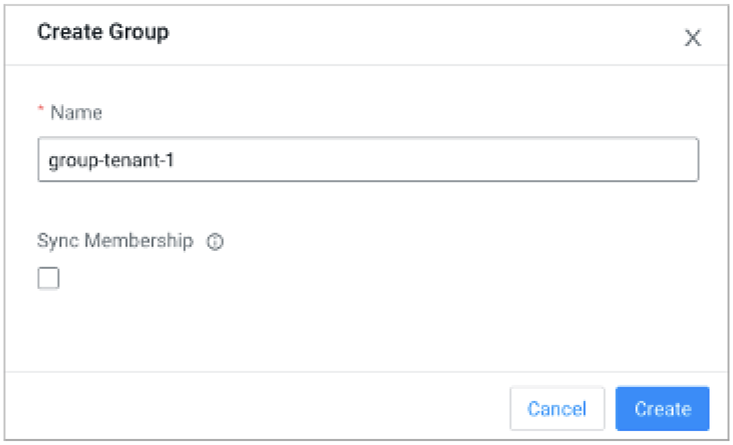
1. Register an environment with RAZ enabled using the CDP web interface or CDP CLI. **Example:** In the web interface, in Fine-grained access control on S3, select Enable Ranger Authorization for AWS S3.
2. In Cloudera Data Warehouse Overview, locate the RAZ-enabled environment, and click Activate to activate the environment for CDW.
   Result: This action disables the standard default Database Catalog that is 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.

3. **Repeat the following steps for each tenant:** In Azure, manually create one or more storage locations for the tenant, or use existing storage locations.
4. In Azure, create a managed identity that has the role **Storage Blob Data Owner** to access a tenant-specific container and across all tenants shared, Storage Location Base in the SDX Data Lake.
   Example: For example, create a managed identity named tenant-1 and a container called container-tenant-1.

Creating a UMS group and machine users

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

1. Repeat the following steps for each tenant:In Management Console, create a User Management Service (UMS) group, for example group-tenant-1.
   Example:



Create Group

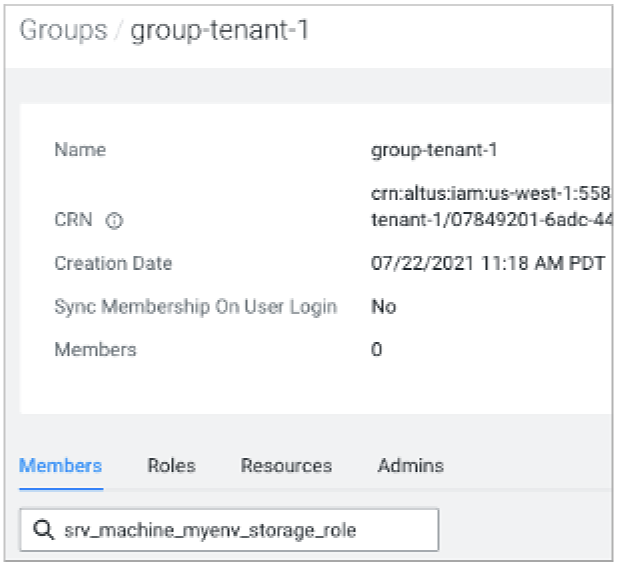
* Name

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.
   Example:



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 complete managed identity.

   For example:

   ```
   /subscriptions/<subsciption_id>/resourcegroups/<resource_group>providers/Microsoft.ManagedIdentity/userAssignedIdentities/<managed-identity-name>
   ```

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.
   Result: The UMS machine user gets the permission to access the tenant specific container.

Creating a new Database Catalog

About this task: When you create a new database catalog, you specify the managed identity you created earlier. In the event of an invalid managed identity specification, the following regular expression appears in the UI.

```
'\/subscriptions\/(.+?)\/resourcegroups\/(.+?)\/providers\/Microsoft.ManagedIdentity\/userAssignedIdentities\/(.+)', 'i'
```

An example that matches the regular expression is:

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.

```
/subscriptions/<subsciption_id>/resourcegroups/<resource_group>provid
ers/Microsoft.ManagedIdentity/userAssignedIdentities/<managed-identit
y-name>
```

- subscriptions: Your subscription ID
- providers: Microsoft.ManagedIdentity (required value)
- resourcegroups: Your resource group
- userAssignedIdentities: Your managed identity name.

Repeat the following steps for each tenant.

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

2. In New Database Catalog, enter the complete description of the managed identity you created earlier.

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.The backend Data Lake and Database Catalog database must be the same.

4. In Tenant Storage Role, enter the complete managed identity you obtained earlier.

   For example:
   ```
   /subscriptions/<subsciption_id>/resourcegroups/<resource_group>
   providers/Microsoft.ManagedIdentity/userAssignedIdentities/<man
   aged-identity-name>
   ```

   CDW attempts to validate your managed identity. If successful, proceed. If validation fails, you see the error message described above. Correct the problem, and try again.

5. In Tenant Storage Location, enter the tenant-specific container **container-tenant-1**for example, and click CREATE.

Creating a tenant-specific Virtual Warehouse

About this task: 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 either 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 , and 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.

   Example:

   ```
    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.

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.