Fine-grained Access Control for ADLS Gen 2 (Preview)

Date published: 2021-08-10
Date modified: 2021-09-07

Legal Notice

Contents

Introduction

CDP Public Cloud defaults to using cloud storage which introduces new challenges around managing data access across teams and individual users. The Ranger Authorization Service (RAZ) addresses these challenges, enabling users to have the equivalent fine-grained and audit capabilities in Apache Ranger they enjoyed with HDFS files in an on-prem or IaaS deployment.

Many of the use cases that RAZ for ADLS Gen2 enables are cases where access control on files or directories is needed. Some examples include:

- Per-user home directories
- Data engineering (Spark) efforts that require access to cloud storage objects and directories
- Data warehouse queries (Hive/Impala) using external tables.
- Access to Ranger's rich access control policies such as date-based access revocation, user/group/role based controls, along with corresponding audit.
- Tag-based access control via classification propagation originating from directories.

Prior to the introduction of RAZ, controlling access to ADLS Gen2 could be enforced at coarse-grained group level (via IDBroker mappings) and thus would require rearchitecting the implementation of important file-centric activities as well as admin-level access to both the Azure subscription and CDP account.

In HDP and CDH deployments, files and directories would be protected with a combination of HDFS Access Control Lists (ACLs) (in CDH, HDP) and Ranger HDFS policies (in HDP). Similarly, in an AWS CDP Public Cloud environment with RAZ for ADLS Gen2 enabled, Ranger's rich access control policies can be applied to CDP's access to ADLS Gen2 containers, directories, and files and can be controlled with admin-level access to CDP alone.

Limitations

- In order to use RAZ you must create a new environment with RAZ enabled. There is no way to enable RAZ in an existing environment.
- RAZ currently only supports Cloudera Data Hub and Cloudera Machine Learning (CML) and it does not support any other CDP services (such as Data Warehouse, Data Engineering, and so on).
- Only Data Hub and DataLake are tested and supported.
- RAZ has been tested only on the following Data Hub cluster types:
    - Data Engineering
    - Data Engineering HA
    - Data Engineering Spark3
    - Data Mart
    - Operational Database with SQL

  Specifically, Hive, Spark, HBase, and Oozie are supported.
- Data Lake Backup and Restore of RAZ-enabled environments is not supported.

Azure requirements

In addition to the standard Azure requirements described in the Azure requirements documentation, you should meet the following requirements:

Managed identity for RAZ

In addition to creating the managed identities required for the minimal setup for cloud storage, you should create an additional managed identity for RAZ:

Managed Identity	Scope	Roles to assign
`RangerRAZ`	Storage Account	• Storage Blob Data Owner or equivalent Custom role • Storage Blob Delegator

Use the following steps to create the *RangerRAZ* managed identity via Azure Portal:

1. On Azure Portal, navigate to **Managed Identities**.
2. Click **+New**.
3. Select the **Resource group** used for CDP.
4. Select your environment's **Region**.
5. Specify managed identity **Name** (for example *RazIdentity*).
6. Provide tags if required by your organization.
7. Click **Review + create**.

Next, assign the two roles to the *RangerRAZ* managed identity on the scope of the storage account created for CDP:

1. In your Azure Portal, navigate to **Storage accounts** > your storage account > **Access Control (IAM)**.
2. Click **+Add** > Add role assignment.
3. Under **Add role assignment**:
    a. Under **Role**, select *Storage Blob Data Owner*.
    b. Under **Assign access to**, select User assigned managed identity.
    c. Under **Select**, select the *RangerRAZ* managed identity created earlier.
    d. Click **Save**.
4. Repeat steps 1 and 2, but this time select the *Storage Blob Delegator* role.

Custom role to replace Storage Blob Data Owner

You can optionally use the following policy definition to create a custom role that can be used instead of Storage Blob Data Owner:

```
{
   "properties": {
       "roleName": "Cloudera CDP Storage Authorization",
       "description": "Provide privileges that Cloudera CDP requires for storage
access",
       "assignableScopes": [
           "/subscriptions/abce3e07-b32d-4b41-8c78-2bcaffe4ea27"
       ],
       "permissions": [
           {
               "actions": [
"Microsoft.Storage/storageAccounts/blobServices/generateUserDelegationKey/action"
               ],
               "notActions": [],
               "dataActions": [
"Microsoft.Storage/storageAccounts/blobServices/containers/blobs/manageOwnership/acti
on",
"Microsoft.Storage/storageAccounts/blobServices/containers/blobs/modifyPermissions/ac
tion",
"Microsoft.Storage/storageAccounts/blobServices/containers/blobs/read",
"Microsoft.Storage/storageAccounts/blobServices/containers/blobs/write",
"Microsoft.Storage/storageAccounts/blobServices/containers/blobs/delete",
"Microsoft.Storage/storageAccounts/blobServices/containers/blobs/move/action"
               ],
               "notDataActions": []
           }
       ]
   }
}
```

Registering an environment

You can use CDP web interface or CDP CLI to register an environment with RAZ enabled.

In general when registering an environment, you need to perform the following special steps to enable RAZ:

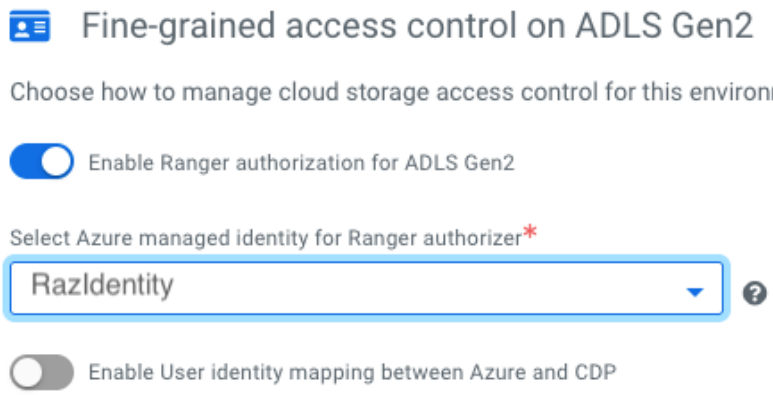
- Select the latest available version of Runtime. Version 7.2.9 or newer can be used.
- Enable **Fine-grained access control on ADLS** and provide the managed identity created earlier.

Register an environment via CDP web interface

Use the following example steps to register an Azure environment with RAZ enabled.

Steps

1. Log in to the CDP web interface.
2. Navigate to the **Management Console > Environments** and click on **Register environment**.
3. Provide an **Environment Name.**
4. Select a provisioning credential.
5. Click **Next**.
6. Provide a **Data Lake Name**.
7. Make sure to select *Runtime 7.2.9* or newer from the **Data Lake version** dropdown.
8. In the **Data Access and Audit** section, provide your data storage location and managed identities created for minimal setup for cloud storage.
9. In the **Fine-grained access control on ADLS** section, click on the toggle button to enable Ranger authorization for ADLS Gen2 and select the managed identity created for RAZ.



10. Click **Next**.
11. Select your region, network, security groups, provide an SSH key, and if needed add tags.
12. Click **Next**.
13. In the Logs section, provide your logs storage location and managed identities created for minimal setup for cloud storage.
14. Click **Register Environment**.

Related links

Register an Azure environment

Register an environment via CDP CLI

Install Beta CDP CLI

In order to register an environment via CDP CLI with RAZ enabled, you need to download and install Beta CDP CLI.

Prerequisites

Do not install both the regular and beta CLIs in the same Python environment, as they use the same entry points and will therefore conflict. Either use a separate virtual environment or uninstall the cdpcli first before installing cdpcli-beta.

Steps

1. Install Python if needed, as described in CDP CLI documentation for your platform. Do not install CDP CLI. Instead, proceed to step 2.
2. Run the following to install Beta CDP CLI:
   pip3 install cdpcli-beta
3. Configure access keys as described in Generating an API access key.

Register an environment

If you already have CDP CLI templates for creating an environment, you can modify them by adding the additional parameters required for RAZ. These additional options are highlighted in the below examples. Note that these examples include “<VALUE>” placeholders, which need to be replaced with actual values:

```
cdp environments create-azure-environment \
--environment-name <ENVIRONMENT_NAME> \
--credential-name <CREDENTIAL_NAME> \
--region <REGION> \
--security-access cidr=<YOUR_CIDR> \
--public-key <SSH_PUBLIC_KEY> \
--log-storage <LOG_STORAGE_CONFIGURATION>\
--use-public-ip

cdp environments set-id-broker-mappings \
--environment-name <ENVIRONMENT_NAME> \
--data-access-role <DATA_ACCESS_IDENTITY>\
--ranger-audit-role <RANGER_AUDIT_IDENTITY> \
--ranger-cloud-access-authorizer-role <RAZ_IDENTITY> \
--set-empty-mappings

cdp datalake create-azure-datalake \
--datalake-name <DATALAKE_NAME> \
--environment-name <ENVIRONMENT_NAME> \
--cloud-provider-configuration <STORAGE_LOCATION_BASE_CONFIGURATION> \
--enable-ranger-raz
```

Note: You can obtain the exact commands from CDP web interface, as described in Obtain CLI commands for registering an environment.

Related links

Register an Azure environment

Creating a Data Hub cluster

Once your environment is running, use the usual steps to create Data Hub clusters. The following cluster templates have been tested and can be used with RAZ:

- Data Engineering
- Data Engineering HA
- Data Engineering Spark3
- Data Mart
- Operational Database with SQL

Custom variants of these templates are expected to work.

Related links

Create a cluster from a definition on Azure

Create a custom cluster on Azure

Ranger policies

Ranger includes a set of preloaded resource-based services and policies, and allows you to create additional policies. No specific additional policies are required.