CDP Public Cloud Management Console

Azure Environments

Date published: 2019-08-22
Date modified:

https://docs.cloudera.com/

Legal Notice

Contents

Working with Azure environments

Refer to the following documentation to learn about creating and managing Azure environments in CDP:

Related Information

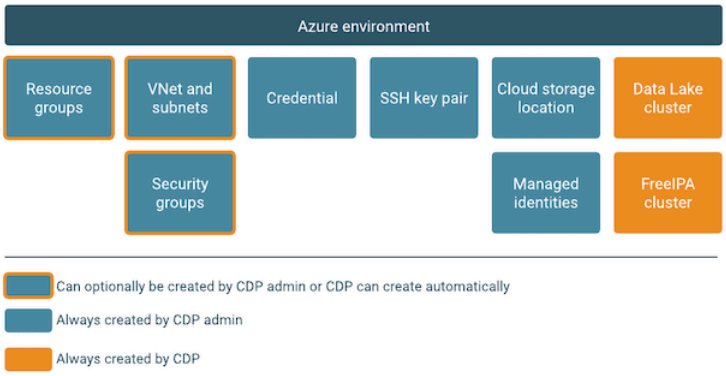
Managing provisioning credentials for Azure
Managing Data Lakes
Managing FreeIPA

Introduction to Azure environments

In CDP, an environment is a logical subset of your cloud provider account including a specific virtual private network. You can register as many environments as you require.

The “environment” concept of CDP is closely related to the virtual private network in your cloud provider account. Registering an environment provides CDP with access to your cloud provider account and identifies the resources in your cloud provider account that CDP services can access or provision. A single environment is contained within a single cloud provider region, so all resources deployed by CDP are deployed within that region within one specific virtual network. Once you’ve registered an environment in CDP, you can start provisioning CDP resources such as clusters, which run on the physical infrastructure in an Azure data center.

The following diagram enumerates the components of an environment:



The diagram illustrates all major user-created and CDP-created components of an environment:

- The items in dark blue boxes with orange outlines can either be automatically provisioned by CDP on your Azure account, or you can optionally pre-create them and specify them when registering an environment.
- The items in dark blue boxes must be pre-created by your CDP administrator prior to environment registration and then specified when registering an environment.
- The items in orange boxes are automatically provisioned on Azure by CDP as part of environment provisioning.

Note: The items that are user-created don’t get terminated during environment deletion.

As shown in the diagram, an environment consists of the following resources:

Environment component	Description
Virtual network with subnets	An environment corresponds to one specific virtual network (called VNet on Azure) and subnets in which CDP resources are provisioned.

Environment component	Description
Security groups	Security groups (called NSGs on Azure) act as a virtual firewall for your instances to control inbound and outbound traffic. All VM instances provisioned within an environment use your specified security access settings allowing inbound access to your instances from your organization's computers.
Credential	CDP uses the credential for authorization to provision resources (such as compute instances) within your cloud provider account. On Azure, credential creation involves creating an application and a service principal within the Azure Active Directory manually (app-based credential).
SSH public key	When registering an environment on a public cloud, a CDP administrator provides an SSH public key. This way, the administrator has root-level access to the Data Lake instance and Data Hub cluster instances.
Cloud storage location and managed identities	When registering an environment, you must provide an ADLS Gen2 location for storing: • All workload cluster data • Cluster service logs and Ranger audits Furthermore, you must create and assign managed identities so that CDP can access the storage location.
Data Lake	A data lake is automatically provisioned when an environment is created. It provides a mechanism for storing, accessing, organizing, securing, and managing data.
FreeIPA	A FreeIPA server is automatically provisioned when an environment is created. It is responsible for synchronizing your users and making them available to CDP services, Kerberos service principal management, and more.
Resource groups	CDP creates multiple new resource groups for resources that it deploys on your Azure account.

You may want to register multiple environments corresponding to different regions that your organization would like to use. Once your environment is running, you can provision Data Hub clusters, Data Warehouses, and other resources in it.

Register an Azure environment from CDP UI

Once you've met the Azure cloud provider requirements, register your Azure environment.

Before you begin

This assumes that you have already fulfilled the environment prerequisites described in Azure requirements.

Required role: EnvironmentCreator

Steps

1. Navigate to the Management Console > Environments > Register environment:
2. On the Register Environment page, provide the following information:

Parameter	Description
General Information	

Parameter	Description
Environment Name (Required)	Enter a name for your environment. The name: • Must be between 5 and 28 characters long. • Can only include lowercase letters, numbers, and hyphens. • Must start with a lowercase letter.
Description	Enter a description for your environment.
Select Cloud Provider (Required)	Select Azure.
Microsoft Azure Credential (Required)	
Select Credential	Select an existing credential or select Create new credential. For instructions on how to create a credential, refer to Create an app-based credential.

3. Click Next.
4. On the Data Access and Data Lake Scaling page, provide the following information:

Parameter	Description
Data Lake Settings	
Data Lake Name (Required)	Enter a name for the Data Lake cluster that will be created for this environment. The name: • Must be between 5 and 100 characters long • Must contain lowercase letters • Cannot contain uppercase letters • Must start with a letter • Can only include the following accepted characters are: a-z, 0-9, -.
Data Lake Version (Required)	Select Cloudera Runtime version that should be deployed for your Data Lake. The latest stable version is used by default. All Data Hub clusters provisioned within this Data Lake will be using the same Runtime version.
Fine-grained access control on ADLS Gen2	
Enable Ranger authorization for ADLS Gen2 Identity	If you would like to use Fine-grained access control, enable this option and then select the Ranger RAZ managed identity created in the Minimal setup for cloud storage.
Data Access and Audit	
Assumer Identity (Required)	Select the Assumer managed identity created in Minimal setup for cloud storage.
Storage Location Base (Required)	Provide the ADLS Gen2 location created for data storage in Minimal setup for cloud storage.
Data Access Identity (Required)	Select the Data Lake Admin managed identity created in Minimal setup for cloud storage.
Ranger Audit Identity (Required)	Select the Ranger Audit managed identity created in Minimal setup for cloud storage.
IDBroker Mappings	We recommend that you leave this out and set it up after registering your environment as part of Onboarding CDP users and groups for cloud storage. **Note:** If you are using Fine-grained access control, this option is disabled, because you should onboard your users and groups via Ranger instead of using IDBroker mappings.
Scale (Required)	Select Data Lake scale. By default, “Light Duty” is used. For more information on data lake scale, refer to Data Lake scale.

5. Click on Advanced Options to make additional configurations for your Data Lake. The following options are available:

Parameter	Description
Hardware and Storage	For each host group you can specify an instance type. For more information on instance types, see Sizes for virtual machines in Azure.
Cluster Extensions	
Recipes	You can optionally select and attach previously registered recipes to run on a specific Data Lake host group. For more information, see Recipes.

6. Click Next.
7. On the Region, Networking and Security page, provide the following information:

Parameter	Description
Region	
Select Region (Required)	Select the region that you would like to use for accessing and provisioning resources from CDP. If you would like to use a specific existing virtual network, the virtual network must be located in the selected region.
Resource Group	
Select Resource Group (Required)	You have two options: • Select one existing resource group. If you select this, all CDP resources will be provisioned into that resource group. • Select Create new resource groups to have CDP create multiple resource groups.
Customer Managed Encryption Keys	
Enable Customer-Managed Keys	Enable this if you would like to provide a Customer-Managed Key (CMK) to encrypt environment's disks and databases. For more information, refer to Customer managed encryption keys.
Select Encryption Key Resource Group	Select the resource group where the CMK is located.
Encryption key URL	Provide the URL of the key value where the CMK resides. This is the same as the key identifier that you can copy directly from Azure Portal.
Managed identity for encryption	If using Azure Database for PostgreSQL Flexible Server, you can optionally select a managed identity created for encrypting it. For more information, refer to Managed identity for encrypting Azure Database for PostgreSQL Flexible Server.
Network	
Select Network (Required)	You have two options: • Select the existing virtual network where you would like to provision all CDP resources. Refer to VNet and subnets. • Select Create new network to have a new network with three subnets created.
Select Subnets (Required)	This option is only available if you choose to use an existing network. Multiple subnets must be selected and CDP distributes resources evenly within the subnets.
Network CIDR (Required)	This option is only available if you select to create a new network. If you selected to create a new network, provide Network CIDR that determines the range of private IPs that VMs will use. This must be a valid private IP CIDR IP in IPv4 range. For example 10.10.0.0/16 are valid IPs. /16 is required to allow for enough IP addresses.
Create Private Subnets	This option is only available if you select to have a new network and subnets created. Is is turned on by default so that private subnets are created in addition to public subnets. If you disable it, only public subnets will be created. **Important:** For production deployments, Cloudera recommends that you use private subnets. Work with your internal IT teams to ensure that users can access the browser interfaces for cluster services.

Parameter	Description
Enable Public Endpoint Access Gateway	When CCM is enabled, you can optionally enable Public Endpoint Access Gateway to provide secure connectivity to UIs and APIs in Data Lake and Data Hub clusters deployed using private networking. If you are using your existing VPC, under Select Endpoint Access Gateway Subnets, select the public subnets for which you would like to use the gateway. The number of subnets must be the same as under Select Subnets and the availability zones must match. For more information, refer to Public Endpoint Access Gateway documentation.
Create Private Endpoints	By default, the PostgreSQL Azure database provisioned for your Data Lake is reachable via a service endpoint (public IP address). To increase security, you can optionally select to have it reachable via a private endpoint instead of a service endpoint. **Note:** This option is only available if an existing resource group is selected. **Note:** Only the subnets that have Azure private endpoint network policies turned off are eligible for private endpoint creation. At least one such subnet is required. If you select to create a private endpoint and you are using your own VNet, you have two options: • Select “Create new private DNS zone” and CDP creates and manages a private DNS zone for you in the provided existing resource group. • Select your existing private DNS zone. If you select to create a private endpoint and you would like for CDP to create a new VNet, CDP creates a private DNS zone for you. For more information, refer to Private endpoint for Azure Postgres.
Create Public IPs	This option is disabled by default when CCM is enabled and enabled by default when CCM is disabled.
Flexible Server	During environment registration in CDP, the Flexible Server in public service mode is used by default, but you can specify to use the Flexible Server in private service mode (“Private Flexible Server”). For more information, refer to Using Azure Database for PostgreSQL Flexible Server.
Proxies	
Select Proxy Configuration	Select a proxy configuration if previously registered. For more information refer to Setting up a proxy server.
Security Access Settings	
Select Security Access Type (Required)	This determines inbound security group settings that allow connections to the Data Lake and Data Hub clusters from your organization’s computers. You have two options: • Create new security groups - Allows you to provide custom CIDR IP range for all new security groups that will be created for the Data Lake and Data Hub clusters so that users from your organization can access cluster UIs and SSH to the nodes. This must be a valid CIDR IP in IPv4 range. For example: 192.168.27.0/24 allows access from 192.168.27.0 through 192.168.27.255. You can specify multiple CIDR IP ranges separated with a comma. For example: 192.168.27.0/24,192.168.28.0/24. If you use this setting, several security groups will get created: one for each Data Lake host group the Data Lake and one for each host group), one for each FreeIPA host group, and one for RDS; Furthermore, the security group settings specified will be automatically used for Data Hub, Data Warehouse, and Machine Learning clusters created as part of the environment. • Provide existing security groups (Only available for an existing VPC) - Allows you to select two existing security groups, one for Knox-installed nodes and another for all other nodes. If you select this option, refer to Security groups to ensure that you open all ports required for your users to access environment resources.
SSH Settings	

Parameter	Description
New SSH public key (Required)	Upload a public key directly from your computer. **Note:** CDP does not use this SSH key. The matching private key can be used by your CDP administrator for root-level access to the instances provisioned for the Data Lake and Data Hub.
Add tags	You can optionally add tags to be created for your resources on Azure. Refer to Defining custom tags.

8. Click on Advanced Options to make additional configurations for FreeIPA. The following options are available:

Parameter	Description
Hardware and Storage	For each host group you can specify an instance type. For more information on instance types, see Sizes for virtual machines in Azure.
Cluster Extensions	
Recipes	You can optionally select and attach previously registered recipes to run on FreeIPA nodes. For more information, see Recipes.

9. Click Next.
10. On the Storage page, provide the following information:

Parameter	Description
Logs	
Logger Identity (Required)	Select the Logger managed identity created in Minimal setup for cloud storage.
Logs Location Base (Required)	Provide the ADLS Gen2 location created for log storage in Minimal setup for cloud storage.
Backup Location Base	Provide the ADLS Gen2 location created for FreeIPA and Data Lake backups in Minimal setup for cloud storage. If not provided, the default Backup Location Base uses the Logs Location Base.
Telemetry	
Enable Workload Analytics	Enables Cloudera Observability support for workload clusters created within this environment. When this setting is enabled, diagnostic information about job and query execution is sent to Cloudera Observability. For more information, refer to Enabling workload analytics and logs collection.
Enable Deployment Cluster Logs Collection	When this option is enabled. the logs generated during deployments will be automatically sent to Cloudera. For more information, refer to Enabling workload analytics and logs collection.

11. Click on Register Environment to trigger environment registration.
12. The environment creation takes about 60 minutes. The creation of the FreeIPA server and Data Lake cluster is triggered. You can monitor the progress from the web UI. Once the environment creation has been completed, its status will change to “Running”.

After you finish

After your environment is running, perform the following steps:

- You must assign roles to specific users and groups for the environment so that selected users or user groups can access the environment. Next, you need to perform user sync. For steps, refer to Enabling admin and user access to environments.
- You must onboard your users and/or groups for cloud storage. For steps, refer to Onboarding CDP users and groups for cloud storage.
- You must create Ranger policies for your users. For instructions on how to access your Data Lake, refer to Accessing Data Lake services. Once you've accessed Ranger, create Ranger policies to determine which users have access to which databases and tables.

Register an Azure environment from CDP CLI

Once you've met the Azure cloud provider requirements, register your Azure environment.

Before you begin

This assumes that you have already fulfilled the environment prerequisites described in Azure requirements.

Required role: EnvironmentCreator

Steps

Unlike in the CDP web interface, in CDP CLI environment creation is a three-step process with environment creation, setting IDBroker mappings and Data Lake creation being three separate steps. The easiest way to obtain the correct commands is to provide all parameters in CDP web interface and then generate the CDP CLI commands on the last page of the wizard. For detailed steps, refer to Obtain CLI commands for registering an environment.

After you finish

After your environment is running, perform the following steps:

- You must assign roles to specific users and groups for the environment so that selected users or user groups can access the environment. Next, you need to perform user sync. For steps, refer to Enabling admin and user access to environments.
- You must onboard your users and/or groups for cloud storage. For steps, refer to Onboarding CDP users and groups for cloud storage.
- You must create Ranger policies for your users. For instructions on how to access your Data Lake, refer to Accessing Data Lake services. Once you've accessed Ranger, create Ranger policies to determine which users have access to which databases and tables.

Enabling admin and user access to environments

In order to grant admin and user access to an environment that you registered in CDP, you should assign the required roles.

You need to be an EnvironmentCreator in order to register an environment. Once an environment is running, the following roles can be assigned:

- EnvironmentAdmin - Grants all rights to the environment and Data Hub clusters running in it, except the ability to delete the environment. The user who registers the environment automatically becomes its EnvironmentAdmin.
- EnvironmentUser - Grants permission to view Data Hub clusters and set the workload password for the environment. This role should be used in conjunction with service-specific roles such as DataHubAdmin, DWAdmin, DWUser, MLAdmin, MLUser, and so on. When assigning one of these service-specific roles to users, make sure to also assign the EnvironmentUser role.
- DataSteward - Grants permission to perform user/group management functions in Ranger and Atlas Admin, manage ID Broker mappings, and start user sync for the environment.
- Owner - Grants the ability to mange the environment in CDP, including deleting the environment. The user who registers the environment automatically becomes its Owner. The Owner role does not grant access the environment's clusters (Data Lakes, Data Hubs).

The roles are described in detail in Resource roles. The steps for assigning the roles are described in Assigning resource roles to users and Assigning resource roles to groups.

Related Information

Resource roles

Assigning resource roles to users

Assigning resource roles to groups

Obtain CLI commands for registering an environment

Although you can obtain CDP CLI commands for environment creation from CDP CLI help, the easiest way to obtain them is from the CDP web interface.

You can quickly obtain CDP CLI commands for creating an environment:

- From details of an existing environment
- From the register environment wizard

Create an environment from an existing environment

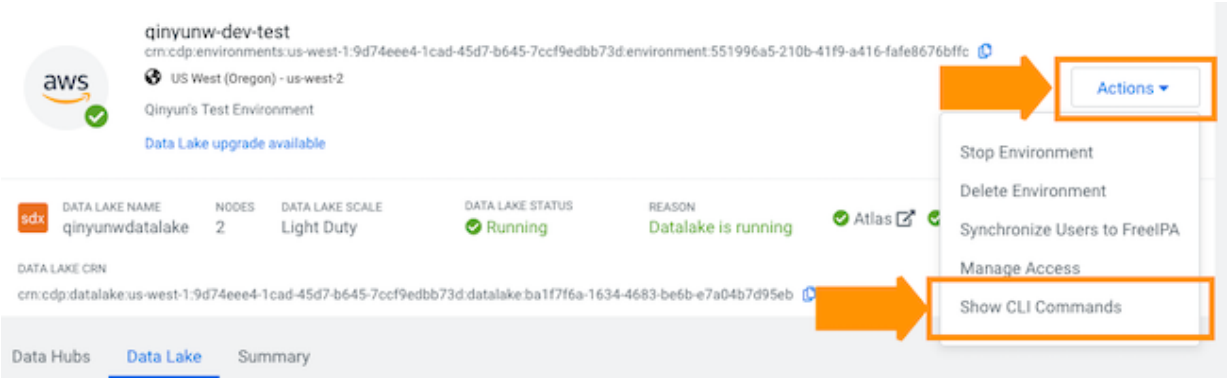
Obtain an environment template from an existing environment to create an environment with the exact same settings.

Since creating an environment, setting IDBroker mappings, and creating a Data Lake are separate actions in CDP CLI, you need to obtain three commands.

Required role: EnvironmentUser, EnvironmentAdmin, or Owner

Steps

1. Log in to the CDP web interface and navigate to the Management Console.
2. Navigate to environment details > Actions > Show CLI commands:



3. Click COPY three times to copy the three commands. These commands allow you to:
   a. Create an environment with the same settings as the existing environment.
   b. Set the same IDBroker mappings as in the original environment.
   c. Create a Data Lake with the same settings.
4. Before you can use these commands, make sure to update the following:
   a. In cdp environments create-<cloud-platform>-environment, update the value of --environment-name. It should be unique within CDP.
   b. In cdp environments set-id-broker-mappings, update the value of --environment-name. It should reference the name of the new environment that you are planning to create.
   c. In cdp datalake create-<cloud-platform>-datalake, update the value of --datalake-name. It should be unique within CDP.
   d. In cdp datalake create-<cloud-platform>-datalake, update the value of --environment-name. It should reference the name of the new environment that you are planning to create.
5. Run the three commands to:
   a. Register an environment.
   b. Set IDBroker mappings.
   c. Create a Data Lake.

Obtain CLI commands from the register environment wizard

Provide environment parameters in the environment wizard, and then on the last page of the wizard generate a CDP CLI template to create an environment and Data Lake with the parameters specified in the wizard. The obtained cluster template can be used to create an environment with the same settings via CDP CLI.

Since creating an environment, setting IDBroker mappings, and creating a Data Lake are separate actions in CDP CLI, you need to obtain three commands.

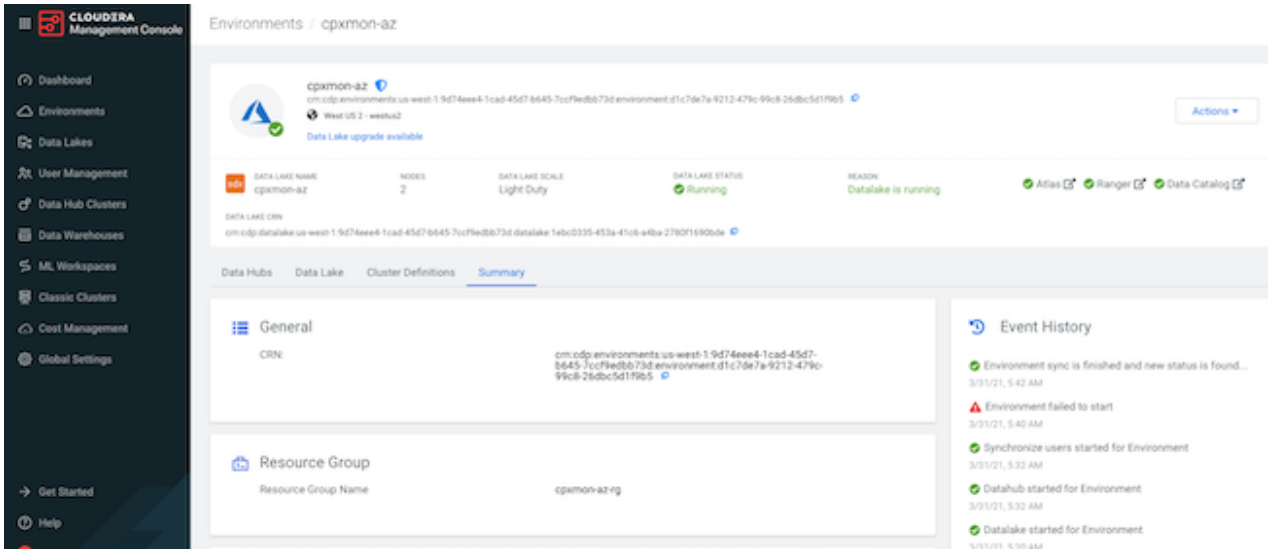
Required role: EnvironmentCreator

Steps

1. Log in to the CDP web interface and navigate to the Management Console.
2. Navigate to Environments > Register Environment.
3. Provide all the parameters for your environment.
4. On the last page, click SHOW CLI COMMAND in the bottom of the page.
5. Copy the three commands (for creating an environment, setting IDBroker mappings, and creating a Data Lake).
6. Run the commands:
   a. First run the command that creates the environment.
   b. Once the command finishes and the environment is running, run the command that sets IDBroker mappings.
   c. Next, run the command that creates the Data Lake.

Understanding environment UI options

To access information related to your environment, navigate to the Management Console service > Environments and click on your environment.



You need to have the EnvironmentUser role or higher for the environment in order to access details of that environment.

From environments details, you can access the following:

- From the Data Hub tab, you can create, manage, and access Data Hub clusters within the environment.
- From the Data Lake tab, you can monitor, manage, and access the Data Lake cluster.
- From the Cluster Definitions tab, you can access all cluster definitions that can be used with the environment.
- From the Summary tab, you can manage and monitor your environment.

The Summary includes the following information:

Option	Description
General	This includes your environment's CRN. CRN is an ID that CDP uses to identify a resource.
Resource group	This lists the resource group(s) that your environment resources are associated with. During environment registration, you either provide an existing resource group or CDP creates multiple resource groups; There is no way to change this once an environment is running.
Credential	This links the provisioning credential associated with the environment and includes the option to change the credential.
Region	This lists the region in which your environment is deployed.
Network	This lists the networking resources used by your environment, provided by you or created by CDP during environment registration. You can add additional subnets for Data Hub clusters deployed in the future.
Security Access	This lists the security groups used by your environment, provided by you or created by CDP during environment registration. You can provide new security groups for Data Hub clusters deployed in the future.
FreeIPA	This includes details of a FreeIPA server running in the environment and includes an Actions menu with FreeIPA management options.
Log Storage and Audits	This lists the cloud storage location used for logs and audits that you provided during environment registration. There is no way to update this location once your environment is running.
Telemetry	This includes your environment's telemetry settings. You can change them for any Data Hub clusters created in the future.
Advanced	This lists the name of your root SSH key.

Related Information

Understanding Data Hub details

Understanding Data Lake details

Monitoring an environment

Once an environment exists, you can access it from the Management Console.

Required role: EnvironmentUser, EnvironmentAdmin, or Owner

Steps

For CDP UI

1. To access an existing environment, navigate to Management Console > Environments and click on your environment.
2. Click on the Summary tab to access environment details.
3. You can monitor the status of your environment from this page.

For CDP CLI

You can also list available environments from CDP CLI using the cdp environments list-environments command. For example:

```
cdp environments list-environments
{
    "environments": [
        {
            "environmentName": "cdp-demo",
```

```
            "crn": "crn:altus:environments:us-west-1:c8dbde4b-ccce-4f8d-
a581-830970ba4908:environment:d3361b40-39ab-4d87-bd5b-abc15f16b90c",
            "status": "DELETE_FAILED",
            "region": "us-east-2",
            "cloudPlatform": "AWS",
            "credentialName": "cdp-demo",
            "description": "Cdp demo"
        },
        {
            "environmentName": "cdp-new",
            "crn": "crn:altus:environments:us-west-1:c8dbde4b-ccce-4f8d-a5
81-830970ba4908:environment:1d2bacde-5c96-47c1-a597-9f276b824028",
            "status": "AVAILABLE",
            "region": "us-east-2",
            "cloudPlatform": "AWS",
            "credentialName": "cdp-demo",
            "description": ""
        }
}
```

To get more information about a specific environment, you can use the following commands:

```
cdp environments describe-environment --environment-name <value>
```

```
cdp environments get-id-broker-mappings --environment-name <value>
```

Related Information

Accessing Data Lake services
Understanding Data Lake details
Understanding Data Hub cluster details
Managing FreeIPA

Environment status options

This topic lists all possible environment status options for the UI and CLI and explains what they mean.

Environment status	Description
Environment creation	
CREATION_INITIATED	Environment creation request was registered in the database and CDP is starting the environment creation flow.
ENVIRONMENT_INITIALIZATION_IN_PROGRESS	Setting up the region and network metadata (public/private and cidr).
ENVIRONMENT_VALIDATION_IN_PROGRESS	Setting up the region and network metadata (public/private and cidr).
NETWORK_CREATION_IN_PROGRESS	If the user chose the create new network option, then CDP creates the network on cloud provider side.
PUBLICKEY_CREATE_IN_PROGRESS	If the user choose the create new SSH key option, then CDP creates the SSH key on cloud provider side.
FREEIPA_CREATION_IN_PROGRESS	Creating the FreeIPA resources for an environment.
Environment update	
UPDATE_INITIATED	Environment update was requested and CDP is starting the update flow (network update, load balancer update, SSH key update).
Environment deletion	
DELETE_INITIATED	Environment deletion request was registered and CDP is starting the deletion flow.

Environment status	Description
NETWORK_DELETE_IN_PROGRESS	If the user chose the create new network option, then CDP deletes the network on cloud provider side.
PUBLICKEY_DELETE_IN_PROGRESS	If the user choosing the create new SSH key option, then CDP deletes the SSH key on cloud provider side.
FREEIPA_DELETE_IN_PROGRESS	Deleting the FreeIPA resources for an environment.
EXPERIENCE_DELETE_IN_PROGRESS	Deleting all the attached clusters (CDW, CML, and so on).
RDBMS_DELETE_IN_PROGRESS	Deleting all the provisioned RDS instances that are related to an environment.
CLUSTER_DEFINITION_DELETE_PROGRESS	Deleting all the cluster definitions that are created for an environment.
UMS_RESOURCE_DELETE_IN_PROGRESS	Deleting all the related UMS resources for an environment.
IDBROKER_MAPPINGS_DELETE_IN_PROGRESS	Deleting all the IBroker mapping for an environment.
S3GUARD_TABLE_DELETE_IN_PROGRESS	Deleting all the Dynamo DB tables for an environment.
DATAHUB_CLUSTERS_DELETE_IN_PROGRESS	Deleting all the attached Data Hub clusters.
DATALAKE_CLUSTERS_DELETE_IN_PROGRESS	Deleting the attached Data Lake cluster.
ARCHIVED	Environment has been deleted (not shown on the UI).
Environment is running	
AVAILABLE	Environment is available (ready to use).
Environment process failed	
CREATE_FAILED	Environment creation failed (Detailed message in the statusReason).
DELETE_FAILED	Environment deletion failed (Detailed message in the statusReason).
UPDATE_FAILED	Environment update failed (Detailed message in the statusReason).
Environment stop	
STOP_DATAHUB_STARTED	Stopping all the Data Hub clusters in an environment.
STOP_DATAHUB_FAILED	Stopping all the Data Hub clusters in an environment failed (Detailed message in the statusReason).
STOP_DATALAKE_STARTED	Stopping the Data Lake cluster in an environment.
STOP_DATALAKE_FAILED	Stopping the Data Lake cluster in an environment failed (Detailed message in the statusReason).
STOP_FREEIPA_STARTED	Stopping the FreeIPA instances in an environment.
STOP_FREEIPA_FAILED	Stopping the FreeIPA instances in an environment failed (Detailed message in the statusReason).
ENV_STOPPED	Environment was successfully stopped.
Environment start	
START_DATAHUB_STARTED	Starting all the Data Hub clusters in an environment.
START_DATAHUB_FAILED	Starting all the Data Hub clusters in an environment failed (Detailed message in the statusReason).
START_DATALAKE_STARTED	Starting the Data Lake cluster in an environment.
START_DATALAKE_FAILED	Starting the Data Lake cluster in an environment failed (Detailed message in the statusReason).
START_FREEIPA_STARTED	Starting all the FreeIPA instances in an environment.
START_FREEIPA_FAILED	Starting all the FreeIPA instances failed in an environment (Detailed message in the statusReason).
START_SYNCHRONIZE_USERS_STARTED	Starting user sync for all the clusters in an environment.

Environment status	Description
START_SYNCHRONIZE_USERS_FAILED	Starting user sync for all the clusters in an environment failed (Detailed message in the statusReason).
FreeIPA instance deletion	
FREEIPA_DELETED_ON_PROVIDER_SIDE	The FreeIPA instance has been deleted on cloud provider side.
Load balancer	
LOAD_BALANCER_ENV_UPDATE_STARTED	Start updating the LoadBalancer on Data Lake in an environment.
LOAD_BALANCER_ENV_UPDATE_FAILED	Failed to update the LoadBalancer on Data Lake in an environment (Detailed message in the statusReason).
LOAD_BALANCER_STACK_UPDATE_STARTED	Start updating the LoadBalancer on Data Hubs in an environment.
LOAD_BALANCER_STACK_UPDATE_FAILED	Failed to update the LoadBalancer on Data Hubs in an environment (Detailed message in the statusReason).

Stop and restart an environment

You can stop an environment if you need to suspend but not terminate the resources within the environment. When you stop an environment, all of the resources within the environment are also stopped, including Data Lakes and Data Hubs. You can also restart the environment.

Warning:

The Machine Learning service does not support environment stop and restart. This means that if ML workspaces are running or expected to be provisioned within an environment, then the environment should not be stopped. If done, this will disrupt running CML workspaces and prevent successful provisioning of ML workspaces in the environment.

Required role: EnvironmentAdmin or Owner

Steps

For CDP UI

1. Navigate to the environment in Management Console > Environments.
2. Click Actions Stop Environment and confirm the action.
3. To restart the environment, click Actions Start Environment .

Warning: We have not tested or certified restarting the environment while Cloudera Data Engineering (CDE) is running.

For CDP CLI

Use the following command to stop an environment:

```
cdp environments stop-environment --environment-name <ENVIRONMENT_NAME>
```

Use the following commands to start an environment:

```
cdp environments start-environment --environment-name <ENVIRONMENT_NAME>
 [--with-datahub-start]
```

Use the following commands to start an environment and all Data Hubs running in it:

```
cdp environments start-environment --environment-name <ENVIRONMENT_NAME>
 --with-datahub-start
```

Delete an environment

Deleting an environment terminates all resources within the environment including the Data Lake.

Before you begin

To delete an environment, you should first terminate all clusters running in that environment.

Required role: Owner or PowerUser

Steps

For CDP UI

1. In Management Console, navigate to Environments.
2. Click on your environment.
3. Click Actions Delete and confirm the deletion.
   - Check the box next to "I would like to delete all connected resources" if you have Data Lake and Data Hub clusters running within the environment. This will delete the Data Lake and Data Hub clusters together with the rest of the environment.

     Note: The "I would like to delete all connected resources" option does not delete any Data Warehouse or Machine Learning clusters running within the environment, so these always need to be terminated prior to environment termination.

     Note: To speed up future environment deployment, the cloudbreak-images resource group, which stores VHDs used for VM deployment, and its content are not deleted during environment deletion.
   - Check the box next to "I understand this action cannot be undone". This is required.
4. Click Delete.

For CDP CLI

When terminating an environment from the CDP CLI, you need to first terminate the Data Lake:

1. Terminate the Data Lake using the following command:

   ```
   cdp datalake delete-datalake --datalake-name <value>
   ```
2. Wait until the Data Lake terminates before proceeding. Use the following commands to check on the status of Data Lake:

   ```
   cdp datalake get-cluster-host-status --cluster-name <value>
   ```

   ```
   cdp datalake list-datalakes
   ```
3. Delete the environment using the following command:

   ```
   cdp environments delete-environment --environment-name <value> --cascadi
   ng
   ```

   The --cascading option deletes or Data Hubs running in the environment.

If environment deletion fails, you can:

- Repeat the environment deletion steps, but also check "I would like to force delete the selected environment". Force deletion removes CDP resources from CDP, but leaves cloud provider resources running.
- Clean up cloud resources that were left on your cloud provider account. See Cleaning up a failed AWS environment.

Only the resources that were provisioned as part of the environment are deleted. For example, if a new network was created by CDP for the environment, the network will be deleted; But if you provided your existing network, it will not be deleted as part of environment deletion.

Cleaning up a failed Azure environment

When environment creation fails, you should delete the environment. If environment termination fails, you should clean up any resources that might have already been created on your AWS account.

When environment creation fails, you should delete the environment by using the steps described in Delete an Environment.

If environment termination fails, you should clean up any resources that might have already been created on your Azure account. To do this, log in to your Azure Portal and do one of the following:

- If CDP created all resource groups for your CDP resources, find it and delete these resource groups. The names of these resource groups will include the name of the environment.
- If you used your own existing resource group, navigate to it and delete all the resources that were created by CDP. The names of these resources will include the name of the environment.

Add subnets to an environment

You can add additional subnets to an existing environment. These subnets will only be used for all Data Hub clusters created within the environment in the future.

Before you begin

These steps assume that you have already created the subnets that you want to add to the environment.

Required role: EnvironmentAdmin or Owner

Steps

1. Navigate to Management Console > Environments and select the environment you want to modify:
2. Click the Summary tab.
3. Scroll down to the Network section.
4. Click the (edit) button, then click the Select Subnets pull down menu and select the subnet you want to add to the Environment.
5. Click the (save) button.

   You should see the new subnet listed in the Network section.

   Note:

   The newly added subnets will not be used for any CDP services other than Data Hub. The newly added subnets will only be used for the Data Hub clusters created within the environment after the new subnets were added. All the existing environment resources such as the Data Lake, FreeIPA, and any existing Data Hub clusters will remain within the subnets originally defined during environment creation.

Add security groups to an environment

You can add additional security groups to an existing environment. These security groups will be used for all Data Hub clusters created within the environment in the future.

Before you begin

These steps assume that you have already created the security groups that you want to add to the environment.

Required role: EnvironmentAdmin or Owner

Steps

1. Navigate to Management Console > Environments and click on the environment you want to modify.
2. Click the Summary tab.
3. Scroll down to the Security Access section.
4. Click the (edit) button, then under Select Security Access Type select the Provide Existing Security Groups option and select the security groups that you want to add to the Environment.
5. Click the (save) button.

   You should see the new security groups listed in the Security Access section.

   Note:

   The newly added security groups will only be used for the Data Hub clusters created within the environment after the new security groups were added. All the existing environment resources such as the Data Lake, FreeIPA, and any existing Data Hub clusters will remain within the security groups originally defined during environment creation.

Change environment's credential

You can change the credential attached to an environment as long as the new credential provides the required level of access to the same Azure subscription as the old credential.

Required roles:

- EnvironmentAdmin or Owner of the environment
- SharedResourceUser or Owner of the credential

Steps

For CDP UI

1. Log in to the CDP web interface.
2. Navigate to the Management Console.
3. Select Environments from the navigation pane.
4. Click on a specific environment.
5. Navigate to the Summary tab.
6. Scroll down to the Credential section.
7. Click to access credential edit options.
8. Select the new credential that you would like to use.
9. Click to save the changes.

For CDP CLI

If you would like to delete a credential from the CDP CLI, use:

```
cdp environments update-environment-credential --environment-name <value>
 --credential-name <value>
```

Enabling environment telemetry

You can optionally enable workload analytics so that diagnostic information about job and query execution is sent to Cloudera Observability for Data Hub clusters. Similarly, you can optionally enable logs collection so that logs generated during deployments will be automatically sent to Cloudera.

Required role:

- PowerUser can set environment telemetry settings for the whole tenant.
- EnvironmentCreator can set environment telemetry settings during environment registration.
- EnvironmentAdmin or Owner can set environment telemetry settings for an existing environment.

Enabling workload analytics

If you enable workload analytics, diagnostic information about job and query execution is sent to Cloudera Observability for Data Hub clusters created within all environments. This is disabled by default and can be enabled:

- For the whole tenant:
  - From the CDP web interface by navigating to Management Console>Global Settings>Telemetry, by turning on the Enable Workload Analytics.
  - Or from the CDP CLI using the following command:

    ```
    cdp environments set-account-telemetry --workload-analytics
    ```
- For a specific environment only:
  - During environment creation from the CDP web interface, by turning on the Enable Workload Analytics option under Logs Storage and Audits in the environment creation wizard.
  - For an existing environment, from environment details > Telemetry by turning on the Enable Workload Analytics.
  - For an existing environment, from the CDP CLI using the following command:

    ```
    cdp environments set-telemetry-features --environment-name <some-name> -
    -workload-analytics
    ```

The environment-level setting overrides the tenant-level setting.

Note:

Only Data Hubs created after enabling workload analytics on an environment will send data to Cloudera Observability. Data Hubs created before workload analytics was enabled will not start sending data to Cloudera Observability if workload analytics is enabled for their parent environment.

Enabling cluster logs collection

If you enable cluster logs collection, logs generated during deployments will be automatically sent to Cloudera. This is disabled by default and can be enabled:

- For the whole tenant:
  - From the CDP web interface by navigating to Management Console>Global Settings>Telemetry, by turning on Enable Cluster Logs Collection.
  - Or from the CDP CLI using the following command:

    ```
    cdp environments set-account-telemetry --report-deployment-logs
    ```
- For a specific environment only:
  - During environment creation from the CDP web interface, by turning on the Enable Cluster Logs Collection option under Logs Storage and Audits in the environment creation wizard.
  - For an existing environment, from environment details > Summary >Telemetry by turning on the Enable Cluster Logs Collection.
  - For an existing environment, from the CDP CLI using the following command:

    ```
    cdp environments set-telemetry-features --environment-name <some-name> -
    -report-deployment-logs
    ```

The environment-level setting overrides the tenant-level setting.

Disable cloud storage logging for an existing environment

By default, CDP sends collected service logs from VM nodes to the cloud storage that you provided for logs during environment registration. In some cases, you may want to disable this for an existing environment.

You can disable this option from environment details > Summary > Telemetry by turning off Enable Cloud Storage Logging.

Note:

Disabling this option will affect only Data Hub clusters created after the option was disabled.

Related Information

Configure lifecycle management for logs on Azure

Defining anonymization rules for CDP logs

CDP includes a set of default anonymization rules and allows you to define custom anonymization rules in order to remove sensitive information from CDP logs.

Use PCRE convention for writing custom anonymization rule patterns.

Anonymization rules are applied to the following logs:

- Cluster logs collected during deployments and automatically sent to Cloudera. See Enabling environment telemetry.
- Diagnostics logs that can be collected for troubleshooting and sent to Cloudera Support. See Generating a VM-based diagnostic bundle.

Note: Anonymization rules are only applied to environments created after the rules were added in CDP.

Default anonymization rules

CDP includes a set of default anonymization rules that anonymize the following:

Anonymization rule (PCRE)	Replacement	Description
\b([A-Za-z0-9]|[A-Za-z0-9][A-Za-z0-9\-\._]*[A-Za-z0-9])@(([A-Za-z0-9]|[A-Za-z][A-Za-z0-9\-]*[A-Za-z0-9])\.)+([A-Za-z0-9]|[A-Za-z0-9][A-Za-z0-9\-]*[A-Za-z0-9])\b	email@redacted.host	Email addresses
\d{4}[^\w]\d{4}[^\w]\d{4}[^\w]\d{4}	XXXX-XXXX-XXXX-XXXX	Credit card numbers
\d{3}[^\w]\d{2}[^\w]\d{4}	XXX-XX-XXXX	SSN
FPW\:\s+[\w|\W].*	FPW: [REDACTED]	FreeIPA (workload) password
cdpHashedPassword=.*[']	[CDP PWD ATTRS REDACTED]	Hashed FreeIPA (workload) password.

Creating anonymization rule patterns

Use PCRE convention for writing anonymization rule patterns. For each pattern, come up with a replacement string.

Define custom anonymization rules

You can define custom anonymization rules in CDP. The anonymization rules are only applied to environments created after the rules were added in CDP.

Required role: PowerUser

Steps

For CDP UI

1. Once you have created the rules, navigate to CDP web interface > Management Console > Global Settings > Telemetry > Anonymization rules.
2. Default rules are pre-populated.
3. Click on New rule and add a pattern and replacement string for your rule. Repeat for multiple rules.
4. Test the rules from the same page on the UI under Test rules:
   a. Under Input text paste an example text with sensitive content that should get anonymized by the rules that you added.
   b. Click Test all rules.
   c. The sensitive content should be removed nad replaced in the output printed in the Anonymized result text box.
5. Click Save Changes.

Note:

You can use the Set defaults button if you would like to revert to the default rules.

For CDP CLI

1. If you would like to add new rules, you should first prepare the patterns and replacement strings, and then test them with the following command:

```
cdp environments test-account-telemetry-rules --cli-input-json {
    "testInput": "Email: myemail@cloudera.com",
    "rules": [
        {
            "value": "\\b([A-Za-z0-9]|[A-Za-z0-9][A-Za-z0-9\\-\\._]*[A-Za-z0-9])@(([A-Za-z0-9]|[A-Za-z][A-Za-z0-9\\-]*[A-Za-z0-9])\\.)+([A-Za-z0-9]|[A-Za-z0-9][A-Za-z0-9\\-]*[A-Za-z0-9])\\b",
            "replacement": "email@redacted.host"
        }
    ]
}
```

2. Run the following command to get your current telemetry settings in JSON format:

```
cdp environments get-account-telemetry
```

3. Copy the JSON file that you obtained in the output of this command and paste it into a text editor.
4. Update the JSON file, updating the settings or adding new rules.

   Note:

   Make sure to preserve all the existing rules, or else they will be deleted. Also, make sure to pass the workloadAnalytics and cloudStorageLogging parameters. If you don't pass all of the parameters, the parameters that are not passed will get reset to their default values.
5. Once you have the JSON file updated, run the cdp environments set-account-telemetry command. For example:

```
cdp environments set-account-telemetry --cli-input-json {
    "workloadAnalytics": true,
    "cloudStorageLogging": true,
    "rules": [
        {
            "value": "\\b([A-Za-z0-9]|[A-Za-z0-9][A-Za-z0-9\\-\\._]*[A-
Za-z0-9])@(([A-Za-z0-9]|[A-Za-z][A-Za-z0-9\\-]*[A-Za-z0-9])\\.)+([A-Za-
z0-9]|[A-Za-z0-9][A-Za-z0-9\\-]*[A-Za-z0-9])\\b",
            "replacement": "email@redacted.host"
        }
    ]
}
```

Adding a customer managed encryption key to a CDP environment running on Azure

By default, local Data Lake, FreeIPA, and Data Hub disks attached to Azure VMs and the PostgreSQL server instance used by the Data Lake and Data Hubs are encrypted with server-side encryption (SSE) using Platform Managed Keys (PMK), but you can optionally configure SSE with Customer Managed Keys (CMK).

The CMK can be specified during environment registration and, if present, is used for encrypting Data Lake, FreeIPA, and Data Hub disks and PostgreSQL server instances. Alternatively, it is possible to specify it once the environment is running, in which case the CMK will only be used for Data Hubs created after the CMK was added.

The disks that are attached to the VMs of the Data Lake, FreeIPA, and Data Hub clusters will be associated with a Disk Encryption Set (DES) that is created with the key URL as the underlying encryption key version. The DES dedicated to the CDP environment will be created in the resource group of the environment before the FreeIPA launch at the beginning of the environment creation process.

This documentation covers the following topics:

- Azure prerequisites for using a CMK
- Registering a new Azure environment and specifying a CMK
- Creating a Data Hub with the same CMK
- Checking whether a CMK exists for an Azure environment
- Adding a CMK to an existing Azure environment

Azure prerequisites for using a CMK

You can use your existing Azure vault and vault key or create a new Azure vault and vault key.

For detailed requirements, refer to Azure Prerequisites: Customer managed encryption keys.

Register an Azure environment with a CMK

You can specify an existing customer managed key (CMK) during Azure environment registration and the encryption key will be used to encrypt the VMs and databases running in the environment.

Steps

For CDP UI

You can register your environment as described in Register an Azure environment from CDP UI, just make sure that on the Region, Networking and Security page you enable the following:

1. Under Customer-Managed Keys, click Enable Customer-Managed Keys.
2. In the same section, under Select Resource group select the resource group where the CMK is located.
3. Provide the URL of the key value where the CMK resides. This is the same as the key identifier that you can copy directly from Azure Portal.

For CDP CLI

Use the following CDP CLI command to create an environment with the encryption key created earlier. Replace the placeholders with actual values. For example <ENVIRONMENT-NAME> should be replaced with an actual name.

The parameters important for this feature are highlighted:

```
cdp environments create-azure-environment \
--environment-name <ENVIRONMENT_NAME> \
--credential-name <CREDENTIAL_NAME> \
--region <REGION> \
--security-access cidr=<YOUR-ORGANIZATION’S_CIDR> \
--public-key '<SSH_PUBLIC_KEY>' \
--log-storage <STORAGE_LOCATION> \
--description '<DESCRIPTION>' \
--resource-group-name <ENV_RESOUCE_GROUP_NAME> \
--encryption-key-resouce-group-name <CMK_RESOUCE_GROUP_NAME> \
--encryption-key-url <KEY_RESOUCE_ID>

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

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

The encryption key can be present in a separate resource group than in which the environment is being created. The resource group in which encryption key is present can be provided using --encryption-key-resouce-group-name. Either --encryption-key-resouce-group-name or --resource-group-name must be present. Depending on the resource group where you environment and encryption key reside, there are three possible use cases:

- If your encryption key is in the same resource group as you wish to create the environment in, you can provide a common resource group with --resource-group-name parameter. In this case, you do not need to provide the --encryption-key-resouce-group-name parameter.
- If your encryption key is in a different resource group than you wish to create the environment in and you wish to use some other existing resource group for the environment, you can provide encryption key’s resource group with --encryption-key-resouce-group-name and the environment’s resource group with --resource-group -name.

If you don't want to specify an existing resource group for the environment, you just need to provide the encryption key's resource group using --encryption-key-resouce-group-name and the environment's resource groups will be created by CDP.

You can obtain more complete commands using the instructions in Obtain CLI commands for registering an environment.

If you are using Azure Database for PostgreSQL Flexible Server in CDP, you can also specify a managed identity and thus use the same CMK for encrypting the Azure Flexible Server database instance used by CDP. For more information, see Configuring a CMEK for data encryption in Azure Database for PostgreSQL Flexible Server .

Create a Data Hub on Azure with a CMK

Use CDP web interface or CDP CLI to create a Data Hub cluster. Note that this doesn't require any extra steps, so you can refer to the Data Hub documentation.

The disks that are attached to the VMs of the Data Hub cluster will be associated with a Disk Encryption Set (DES) that is created with the key URL as the underlying encryption key version. The DES dedicated to the CDP environment will be created in the resource group of the environment at the beginning of the environment creation process (before the FreeIPA launch, specifically).

Check Azure environment's CMK

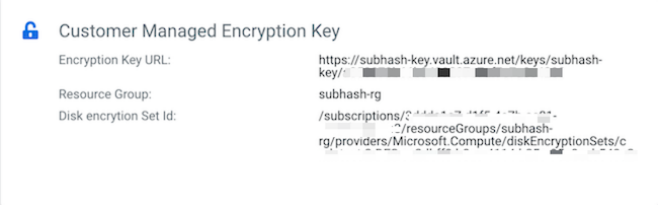
You can check in an Azure environment's summary whether a CMK is used to encrypt a given environment.

Steps

For CDP UI

1. In the Management Console, navigate to Environments and click on the environment for which you would like to set a CMK.
2. Click on the Summary tab.
3. Scroll down to the Customer Managed Encryption Key section.

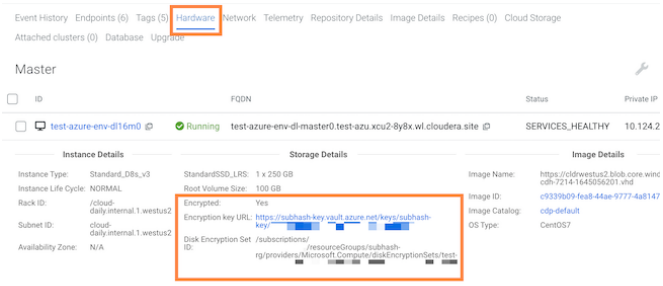
   If a CMK exists, the entry will look similar to:

4. You can also find the CMK information by clicking on the Data Lake tab and then navigating to the Hardware tab.
5. Click on >> on the right side of the screen next to any of the nodes to expand the node related information.

6. Under Storage Details, the following information is included about the CMK:
   - Whether the node has been encrypted with a CMK
   - CMK encryption key URL
   - Disk Encryption Set ID

   For example:



For CDP CLI

You can use the cdp environments update-azure-encryption-resources command to check environment's CMK encryption status.

If encryption key is set, the cdp environments update-azure-encryption-resources command will return the disk encryption set, encryption key, and encryption key's resource group details.

Check Data Hub's CMK

You can check in a Data Hub's hardware tab whether a CMK is used to encrypt it.

Note: A Data Hub's CMK is the same as the one used for the environment in which the Data Hub is running, so you can also find this information in the environment's details.

Steps

1. Navigate to the Data Hub details > Hardware tab.
2. Click on >> on the right side of the screen next to any of the nodes to expand the node related information.
3. Under Storage Details, Under Storage Details, the following information is included about the CMK:
   - Whether the node has been encrypted with a CMK
   - CMK encryption key URL
   - Disk Encryption Set ID

Set a CMK for an existing Azure environment

You can set a CMK for an existing environment. The CMK will be only used for encrypting disks of Data Hubs created after the CMK was added.

Note:

The CMK added to an existing environment will only be used for encrypting disks of Data Hubs created after the CMK was added. Data Lake and FreeIPA disks and databases are not encrypted with the CMK.

Steps

For CDP UI

You can add the encryption key to an existing environment that does not have an encryption key set by navigating to the Summary page of the environment.

1. In the Management Console, navigate to Environments and click on the environment for which you would like to set a CMK.
2. Click on the Summary tab.
3. Scroll down to the Customer Managed Encryption Key section.
4. If no CMK has been set, you will see a message stating that there is no customer-managed key enabled.
5. Click on the edit button in the top right corner of the tab.
6. Click on the toggle button next to Enable Customer-Managed Keys to enable adding a CMK.
7. Provide the following:
   a. Provide the encryption key URL. This is the same as the key identifier that you can copy directly from Azure Portal.
   b. If your encryption key resource group is different from the environment's resource group, provide the name of the resource group.
   c. Click Save.

For CDP CLI

You can add an encryption key for an existing environment that does not yet have encryption enabled.

If your encryption key resource group is same as environment resource group, use:

```
cdp environments update-azure-encryption-resources \
  --environment-name <ENVIRONMENT_NAME> \
  --encryption-key-url <CMK_URL> \
```

If your encryption key resource group is different from environment resource group, use:

```
cdp environments update-azure-encryption-resources \
  --environment-name <ENVIRONMENT_NAME> \
  --encryption-key-url <CMK_URL> \
  --encryption-key-resource-group-name <CMK_RESOURCE_GROUP>
```

If you are using Azure Database for PostgreSQL Flexible Server in CDP, you can also specify a managed identity and thus use the same CMK for encrypting the Azure Flexible Server database instance used by CDP. For more information, see Configuring a CMEK for data encryption in Azure Database for PostgreSQL Flexible Server .

Defining custom tags

In the Management Console user interface, you can define tenant-level or environment-level custom tags across all instances and resources provisioned in your organization's cloud provider account.

Resource tagging

When you create an environment or other resources shared across your cloud provider account, CDP automatically adds default tags to the Cloudera-created resources in your cloud provider account. You can also define additional custom tags that CDP applies to the cluster-related resources in your account.

You can use tags to protect the cloud resources of your CDP environment. Using the tags, you can exclude the resources that should not be removed during housekeeping or security deletion activities that can result in data corruption and data loss.

Default tags

By default, CDP applies certain tags to cloud provider resources whenever you create the resource, for example an environment.

CDP applies the following tags by default:

- Cloudera-Resource-Name: the workload-appropriate Cloudera resource name. For example, an IPA CRN for an IPA, a data lake CRN for a data lake, or a Data Hub CRN for a Data Hub cluster. This CRN serves as a unique identifier for the resource over time.
- Cloudera-Creator-Resource-Name: Cloudera resource name of the CDP user that created the resource.
- Cloudera-Environment-Resource-Name: name of the environment with which the resource is associated.

Custom tags

There are two types of custom tags that you can define in the Management Console: tenant-level tags that apply to Cloudera-created resources across your organization's entire cloud provider account, and environment-level tags.

In the Management Console user interface, you can define tenant-level tags across all instances and resources provisioned in your organization's cloud provider account. These resources include not only provisioned instances, but disks, networks, and other resources as well. In your cloud provider account you can search or filter on either the tag key or value. Tenant-level tags cannot be overridden during environment creation.

In addition to tenant-level tags, you can also define environment-level tags. Environment-level tags are inherited by the resources specific to an environment. For example, environment-level tags are inherited by the following resources:

- FreeIPA
- Data lakes
- Data Hubs
- Data Warehouses
- Operational Databases

As with tenant-level tags, you can search or filter on the key tag or key value in your cloud provider account.

Note: CDP applies custom tags during creation of the resources. For example, you can only define environment-level tags during environment registration. If you want to add or update cloud provider resource tags, you must do so through the cloud provider API.

For more information about using tags on cloud provider resources, consult AWS, Azure, or Google Cloud documentation. It is your responsibility to ensure that your tags meet your cloud provider requirements.

Supported services

While some CDP services such as Data Hub inherit environment-level tags, others require that you add tags when provisioning or enabling the data service. The following table shows how tags can be added for various CDP services:

CDP service	How to add tags
Data Lake	Inherits tenant or environment level tags.
FreeIPA	Inherits tenant or environment level tags.
Data Engineering	Does not inherit tenant or environment level tags but you can define tags when enabling CDE.
Data Hub	Inherits tenant or environment level tags and you can add more tags when creating a Data Hub.
Data Warehouse	Inherits tenant or environment level tags.
DataFlow	Inherits tenant level tags and you can add more tags when enabling CDF.
Machine Learning	Does not inherit tenant or environment level tags but you can define tags when creating a CML workspace.
Operational Database	Inherits tenant or environment level tags and you can add more tags when creating a COD database via CLI.

Defining tenant-level tags

Required roles: PowerUser can define tags for the whole tenant.

- EnvironmentAdmin or Owner can set environment telemetry settings for a specific environment.

Steps

1. In the Management Console, click Global Settings Tags .
2. Click Add.
3. Define both a key and a value for the tag. Both the key and the value must be between 4- 255 characters, with the following restrictions:

   Key

   Allowed characters are hyphens (-), underscores (_), lowercase letters, and numbers. Keys must not start with the following words: 'azure', 'microsoft', and 'windows'. Keys must start with a lowercase letter and must not start or end with spaces.

   Value

   Allowed characters are hyphens (-), underscores (_), lowercase letters, and numbers. Values must not start or end with spaces. You can configure variables in the *{{{variableName}}}* format. The following variables are supported for tenant-level tags:

   - *{{{cloudPlatform}}}* = AWS, AZURE or GCP
   - *{{{userName}}}* = CDP username
   - *{{{userCrn}}}* = Customer Resource Number (CRN) of CDP user
   - *{{{creatorCrn}}}* = CRN of CDP resource creator
   - *{{{time}}}* = Actual time
   - *{{{accountId}}}* = CDP account ID
   - *{{{resourceCrn}}}* = Generated string of CDP resource CRN
4. Click Add, and if necessary repeat the process for additional tags.

Note: Tenant-level tags are applied only to resources created after you define the tag. Any changes to the tag value do not propagate to existing resources that have the tag.

Defining environment-level tags

You define environment-level tags during environment registration.

Required roles: EnvironmentCreator can set tags for a specific environment during environment registration.

Steps

1. In the Management Console, click Environments Register Environment .
2. Proceed through the environment registration steps.
3. After you define data access, add any environment-level tags by clicking Add and defining the tag key and value.

Related Information

Use Tags to Organize Azure Resources (Azure)

Using Azure Database for PostgreSQL Flexible Server

CDP uses Azure Database for PostgreSQL Flexible Server. The Flexible Server allows a highly available database to be deployed for Data Lake and Data Hub clusters. You can create Flexible Server instances with public access, where the Azure Database for PostgreSQL server is accessed through a public endpoint, or with private access (“Private Flexible Server”), where the flexible server has no public endpoint accessible through the internet. The latter option requires a private DNS zone to be specified, and a delegated subnet to be created and added to your CDP Azure environment beforehand.

Using the Flexible Server offers the following benefits to CDP customers:

- Flexible Server allows you to deploy PostgreSQL version 14 and above. See Supported PostgreSQL major versions in Azure Database for PostgreSQL - Flexible Server.
- Unlike the previously used Single Server database instances, Flexible Server database instances can be stopped and restarted during Data Lake and Data Hub cluster stop and restart. This offers a great cost-saving opportunity.

  Note: After 7 days of stopping the Flexible Server database instance, Azure automatically restarts the database. See Limits in Azure Database for PostgreSQL - Flexible Server.

- Flexible Server is multi-AZ capable and offers zone-redundant High Availability. With the Flexible Server, Data Lakes are backed with a highly-available PostgreSQL configuration of two instances. When using a multi-AZ deployment, the Flexible Server instances are deployed in multiple availability zones for additional fault tolerance.

For a detailed comparison of Single Server and Flexible Server offerings, refer to the Comparison table in the Azure documentation.

Note:

On April 10, 2024, Microsoft has announced the general availability of the Flexible Server networking option with Azure Private Link for the Azure Database for PostgreSQL service. This new capability enables simplified creation of PostgreSQL instances with private network access, which previously has only been possible using a delegated subnet. Cloudera is considering adding support for this newly available option and enabling Cloudera customers to upgrade their current clusters that currently rely on Azure Database for PostgreSQL Single Server instances.

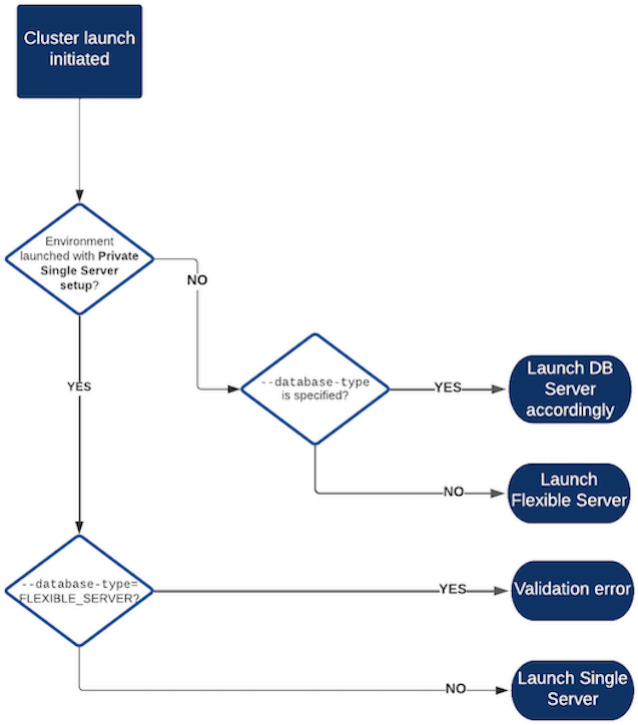
Database server options

You can create Flexible Server instances with public access, where the Azure Database for PostgreSQL server is accessed through a public endpoint, or with private access (“Private Flexible Server”), where the Flexible Server has no public endpoint accessible through the internet. The latter option requires a private DNS zone to be specified, and a delegated subnet to be created and added to your CDP Azure environment beforehand.

The Flexible Server with public access is used by default (as it does not require any special networking setup), but during environment creation you can specify to use the Flexible Server with private access. It is also possible to specify to use the Single Server.

New CDP environments on Azure automatically use Flexible Server with public endpoints and Data Hubs automatically inherit the settings from the environment they run in, but you can also enable Flexible Server when creating a Data Hub.

In general, when registering an Azure environment and creating a Data Hub, you can choose to use Flexible Server, Private Flexible Server, or Single Server, but the exact options vary depending on the environment settings. The logic that CDP uses is illustrated in the following flow chart:



1. When Data Lake or Data Hub creation is initiated, CDP checks if the parent environment has been launched with the private Single Server setup. That is, it checks if one of the following parameters are configured:

```
--flexible-server-subnet-ids   is empty
              --create-private-endpoints   is specified
```

or

```
--flexible-server-subnet-ids  is empty
              --existing-network-params.databasePrivateDnsZoneId  is spec
ified
```

2. If the environment has not been configured for Private Single Server, CDP checks if --database-type=FLEXIBLE_SERVER parameter is specified.
   a. If it is specified, CDP launches the specified database type (Flexible Server or Single Server).
   b. If it is not specified, CDP launches Flexible Server, with public or private setup, based on the environment level settings.
3. If the parameters have been configured, CDP checks if --database-type=FLEXIBLE_SERVER parameter is set.
   a. If --database-type=FLEXIBLE_SERVER parameter is set, a validation error is returned. The validation error states the following:

```
Your environment was created with Azure Private Single Server database s
etup. If you would like to start your cluster with private flexible serv
er database, you have to change your environment network setup.
```

   b. If --database-type=FLEXIBLE_SERVER parameter is not set, CDP launches a cluster with a Single Server.

Limitations

The following limitations currently apply:

- Environments deployed prior to the release of this feature use Single Server with PostgreSQL 11. Upgrading these environments to Flexible Server with PostgreSQL 14 is currently not supported.

- CDP does not support Private Link with Azure Database for PostgreSQL - Flexible Server, which is currently in preview.

In order to set up this feature, you should review Azure prerequisites and then you can enable Flexible Server during Azure environment registration in CDP as described in the linked documentation:

Azure prerequisites for Flexible Server

The following Azure prerequisites must be in place in order to use the Flexible Server with CDP:

- If you would like to use zone redundant-HA, select a region that supports it.
- The policy provided for the Azure credential must have the required permissions mentioned below.
- When deployed in “private service” mode (without public endpoints), Flexible Server instances need to be deployed in a delegated subnet within the VNet. This is not required when public endpoints are used.
- When deployed in “private service” mode, providing the private DNS zone information is mandatory in order to be able to perform DNS resolution. This is not required when public endpoints are used.

Read on to learn how to meet these prerequisites.

Supported regions

Ensure that your selected region has compute availability for Flexible Server. See Flexible Server Azure Regions.

Credential permissions

The Prerequisites for the provisioning credential documentation lists role definitions that includes sufficient permissions for Flexible Server. The role definitions include the following additional permissions that are required for Flexible Server:

Permission	Description
Microsoft.DBforPostgreSQL/flexibleServers/write	Creates a server with the specified parameters or updates the properties or tags for the server.
Microsoft.DBforPostgreSQL/flexibleServers/delete	Deletes an existing server.
Microsoft.DBforPostgreSQL/flexibleServers/start/action	Starts an existing server.
Microsoft.DBforPostgreSQL/flexibleServers/stop/action	Stops an existing server.
Microsoft.DBforPostgreSQL/flexibleServers/firewallRules/write	Restrict which networks can access the database instance.

Delegated subnet

When deployed in private service mode (without public endpoints), Flexible Server instances need to be deployed in a delegated subnet within the VNet.

As mentioned in Azure documentation, to be able to utilize private access with VNet integration, it is a prerequisite to delegate a subnet to Microsoft.DBforPostgreSQL/flexibleServers. This delegation means that only Azure Database for PostgreSQL Flexible Servers can use that subnet. No other Azure resource types can be in the delegated subnet.

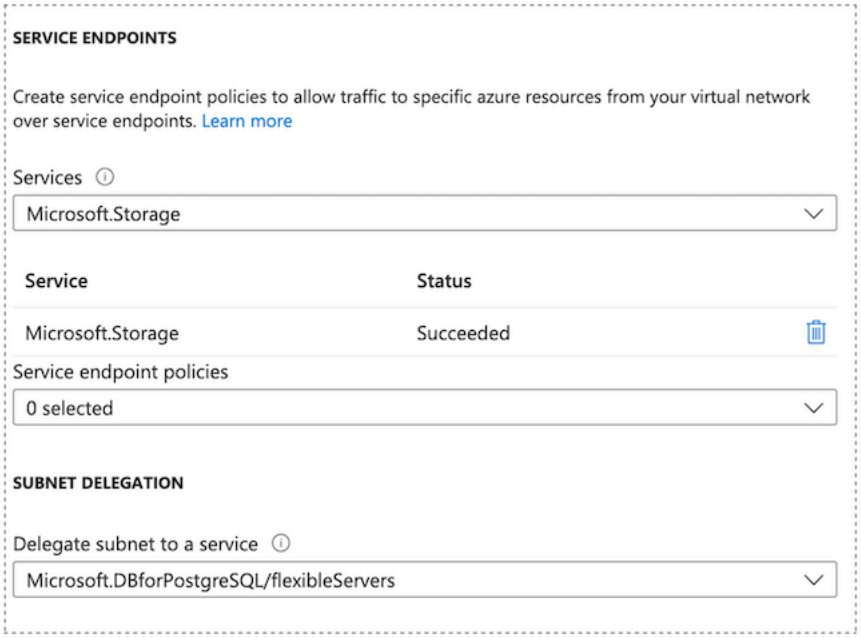
You need to create such a delegated subnet and provide it to CDP during environment registration. This delegated subnet will be used by Azure Database for PostgreSQL instances. The delegated subnet provided during environment registration will be used by default for all Azure Database for PostgreSQL instances used in CDP.

Note: Although you can currently select multiple subnets, the larger subnet is always used by CDP. That is, if there are two subnets provided, one with 128 available IPs and another with 256 available IPs, the second one will be picked. Even though you can select multiple delegated subnets, we recommend that you provide only one.

Creating a delegated subnet

For a step-by-step official guide on how to perform the delegation, see Delegate a subnet to an Azure service. For considerations on the delegated subnet's sizing, see Virtual network concepts.

Here is a screenshot from Azure Portal showing the desired setting:



The Microsoft.Storage service endpoint is set automatically during deployment by Azure.

After the subnet has successfully been delegated, don't forget to record the full subnet ID or the name of the subnet for later use as an input (referred to as <delegated-subnet-id>). For example: /subscriptions/3ddda1c7-d1f5-4e7b-ac81-abcdefgh/resourceGroups/rg/providers/Microsoft.Network/virtualNetworks/vnet/subnets/subnet

Private DNS Zone

When using "private service" mode with Azure virtual network, a Private DNS Zone is mandatory in order to be able to perform DNS resolution.

As mentioned in Using a Private DNS Zone, a Private DNS Zone is used for creating a new Azure Database for PostgreSQL Flexible Server using private network access. Specifically, a Private DNS Zone that ends with .postgres.database.azure.com should be created.

If you do not provide the Private DNS Zone, CDP creates a new Private DNS Zone in your resource group with the naming convention <resource-group-name>.flexible.postgres.database.azure.com

Note: If you are editing an existing environment to enable Flexible Server on it, you must provide a Private DNS Zone (unlike during environment creation where CDP can create it for you).

Note: Only one Private DNS Zone is created per resource group.

Creating a Private DNS Zone

When selecting a name for the Private DNS zone, use one of the following forms:

[name1].[name2].postgres.database.azure.com

or

[name].postgres.database.azure.com

After the Private DNS zone has been created, don't forget to record the full Private DNS Zone ID, referred as <dns-zone-id>, for later use as an input. For example: /subscriptions/3ddda1c7-d1f5-4e7b-ac81-abcdefgh/resourceGroups/drg/providers/Microsoft.Network/privateDnsZones/flexible.private.postgres.database.azure.com

For instructions on how to create a Private DNS zone, see Quickstart: Create an Azure private DNS zone using the Azure portal.

Virtual Network Link

After the Private DNS Zone has been created in Azure, you need to link the VNet that you would like to use for your CDP environment to the Private DNS Zone. Once linked, resources hosted in that VNet can access the Private DNS Zone.

If you let CDP create the Private DNS Zone, then CDP creates the required Virtual Network Link connecting the VNet of your deployment with the created Private DNS Zone.

To learn more about virtual network links, see What is a virtual network link?.

Linking the virtual network

For instructions on how to link the VNet to the previously created Private DNS Zone, see Link the virtual network.

Enable Flexible Server during Azure environment creation

During environment registration in CDP, the Flexible Server in public service mode is used by default, but you can specify to use the Flexible Server in private service mode ("Private Flexible Server").

When CDP is deployed in "private service" mode (without public endpoints), during environment creation you can provide:

- An ID of the delegated subnet
- A private DNS zone ID (optional parameter).

The virtual network link does not need to be specified as input. If you do not provide these, CDP creates them for you.

The steps below show you how to enable "Private Flexible Server". If you would like to enable Flexible Server instances with public access, you do not need to do anything special, as this option is used by default.

Prerequisites

See Azure prerequisites for Flexible Server.

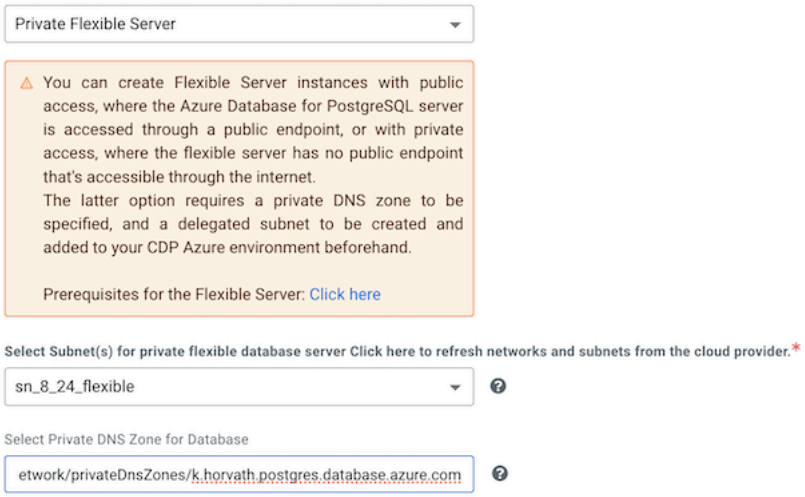
Steps

For CDP UI

1. In the Management Console > Environments, click on Register environment and start registering an Azure environment as usual.
2. In the Network and Availability section, enable the Private Flexible Server by selecting "Private Flexible Server" from the dropdown. The "Flexible Server" option is pre-selected by default. The other two options are "Private Flexible Server" and "Single Server".
3. Select a delegated subnet for the Private Flexible Server.

   Note: Although you can currently select multiple subnets, the larger subnet is always used by CDP. That is, if there are two subnets provided, one with 128 available IPs and another with 256 available IPs, the second one will be picked. Even though you can select multiple delegated subnets, we recommend that you provide only one.

4. Select a Private DNS Zone for the Private Flexible Server. If you do not select one, it will be created automatically.



5. Finish registering your Azure environment in CDP.

For CDP CLI

1. Register an Azure environment using the cdp create-azure-environment CDP CLI command including --existing-network-params with a reference to the Private DNS Zone ID and --flexible-server-subnet-ids with a reference to the delegated subnet ID. The virtual network link does not need to be specified as input.

   Note: Although you can currently provide multiple subnets, the larger subnet is always used by CDP. That is, if there are two subnets provided, one with 128 available IPs and another with 256 available IPs, the second one will be picked. Even though you can select multiple delegated subnets, we recommend that you provide only one.

   For example:

```
cdp environments create-azure-environment
  --environment-name <env-name>
...
  --existing-network-params networkId=dp-rg-test-vnet,resourceGroupName=
dp-rg,subnetIds=/subscriptions/3ddda1c7-d1f5-4e7b-ac81-0523f483b3b3/reso
urceGroups/dp-rg/providers/Microsoft.Network/virtualNetworks/dp-rg-vnet/
subnets/a,/subscriptions/3ddda1c7-d1f5-4e7b-ac81-0523f483b3b3/resourceGr
oups/dp-rg/providers/Microsoft.Network/virtualNetworks/dp-rg-vnet/subnet
s/2,default,databasePrivateDnsZoneId=<dns-zone-id>
  –-flexible-server-subnet-ids <delegated-subnet-id>
```

   The following table explains the required parameters:

Parameter name	Description	Possible values
existing-network-params.databasePrivateDnsZoneId (string)	The ID of an existing private DNS zone used for the database.	Full resource reference

flexible-server-subnet-ids (array)	Comma separated list of the subnet names or full resource IDs delegated for flexible server. This can be specified in two formats: subnet1,subnet2 or /subscriptions/3ddda1c7-d1f5-4e7b-ac81-0523f483b3b3/resourceGroups/dp-rg/providers/Microsoft.Network/virtualNetworks/dp-rg-vnet/subnets/1,/subscriptions/3ddda1c7-d1f5-4e7b-ac81-0523f483b3b3/resourceGroups/dp-rg/providers/Microsoft.Network/virtualNetworks/dp-rg-vnet/subnets/2 The parameter takes a list of subnet IDs (or creates the IDs in case subnet names are provided), validates if they are indeed delegated, and takes the subnet with the largest CIDR range.	List of full resource reference(s). Although this parameter accepts a comma-separated list of subnets, the subnet with the largest CIDR range is always used. Therefore, you should provide only one subnet.

2. Set IDBroker mappings as usual using the cdp environments set-id-broker-mappings command.
3. Create a Data Lake as usual using the cdp create-azure-datalake CDP CLI command, including a reference to the database HA type and the database PostgreSQL engine version.

Related Information

Register an Azure environment from CDP UI
Register an Azure environment from CDP CLI
Enabling admin and user access to environments

Enable Flexible Server during Data Hub creation

A Data Hub uses the same Flexible Server or Single Server settings as the environment in which it runs, but you can choose to enable a Flexible Server on Data Hubs running in an environment that uses a Single Server. This option can be specified during Data Hub creation in the Advanced Options under Network and Availability > Select Azure Database Server.

There are three possible scenarios:

- If you choose to use Private Flexible Server and specify a delegated subnet and a Private DNS Zone on the environment level (either during environment creation or later by editing the environment), then from the Select Azure Database Server dropdown you can select to use Private Flexible Server or Single Server for your Data Hub. You cannot select to use (Public) Flexible Server in this case. If you choose to use Private Flexible Server for your Data Hub, it will use the delegated subnet and private DNS zone specified on the environment level.
- If you choose to use (Public) Flexible Server and do not specify a delegated subnet and a private DNS zone on the environment level, then from the Select Azure Database Server dropdown you can select to use Flexible Server or Single Server for your Data Hub. You cannot select to use Private Flexible Server in this case.
- If you choose to use Single Server on the environment level, then from the Select Azure Database Server dropdown you can select to use Flexible Server or Single Server for your Data Hub. You cannot select to use Private Flexible Server in this case.

If you do not explicitly specify any Azure Database Server option, CDP follows the logic described in Private Flexible Server setup > Database Server Options.

Steps

For CDP UI

This option can be specified during Data Hub creation in the Advanced Options under Network and Availability > Select Azure Database Server.

For CDP CLI

Create a Data Hub as usual using the cdp create-azure-cluster CDP CLI command and specifying the —-database-type.

For example:

```
create-azure-cluster —-database-type FLEXIBLE_SERVER
```

The following table explains the required parameters:

Parameter	Description	Possible values
database-type (string)	The type of the azure database. FLEXIBLE_SERVER is the next generation managed PostgreSQL service in Azure that provides maximum flexibility over your database, built-in cost-optimizations. SINGLE_SERVER is a fully managed database service with minimal requirements for customizations of the database.	• FLEXIBLE_SERVER • SINGLE_SERVER

Related Information

Network and availability

Enable Private Flexible Server on an existing environment

You can change an existing environment to launch the new Data Hubs with a Private Flexible Server. You can achieve this by editing the Network settings of your environment.

Prerequisites

See Azure prerequisites for Flexible Server.

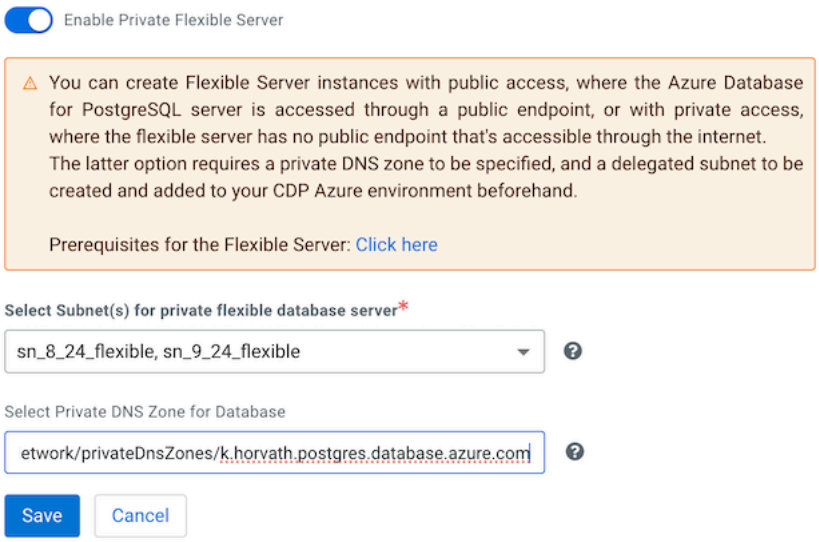
Steps

For CDP UI

1. In the Management Console, navigate to Environments and then navigate to a specific environment.
2. Click on the Summary tab.
3. Scroll down to Network and click the (edit) button.
4. Click on the toggle button next to Enable Private Flexible Server to enable the feature.
5. Under Select subnet(s) for private flexible database server, select a delegated subnet.

6. Select a Private DNS Zone for the Private Flexible Server. This is required.

Note: If you are editing an existing environment to enable a Flexible Server on it, you must provide a Private DNS Zone (unlike during environment creation where CDP can create it for you).



7. Click Save.

For CDP CLI

Use the following command:

```
cdp environments update-azure-database-resources
 --environment <PASTE-CRN>
 --database-private-dns-zone-id <SPECIFY-DNS-ZONE-ID>
 --flexible-server-subnet-ids <SPECIFY-SUBNET-IDs>
```

For example:

```
cdp environments update-azure-database-resources
  --environment crn:cdp:environments:us-west-1:071ea970-ac52-4a7a-a68f-75d
64cf59809:environment:n7aa8c0d-d220-4459-bd47-f19e5291d911
  --database-private-dns-zone-id /subscriptions/1e0c1142-6256-61b9-
b80a-c5888d6e1e22/resourceGroups/b3x-at50645-weu-bde/providers/
Microsoft.Network/privateDnsZones/b3xeng.postgres.database.azure.com
  --flexible-server-subnet-ids /subscriptions/1e0c1142-6256-61b9-b80a-
c5888d6e1e22/resourceGroups/spoke-network-at50645-eng-300361-az01/
providers/Microsoft.Network/virtualNetworks/vnt-300361-az01/subnets/
USER-0303
```

Configuring a CMK for data encryption in Azure Database for PostgreSQL Flexible Server

You can optionally use a customer managed key (CMK) for encrypting data in the Azure Flexible Server database instance used by CDP.

As described in Adding a customer managed encryption key to a CDP environment running on Azure, by default CDP clusters are encrypted with server-side encryption (SSE) using Platform Managed Keys (PMK) and CDP allows you to provide an existing CMK for encrypting CDP clusters.

When using the CMK for encrypting CDP clusters, you can also use that same CMK for encrypting the Azure Flexible Server database instance used by CDP. If you would like to do this, in addition to the typical CMK prerequisites, you should create a managed identity with specific permissions and then after providing the other CMK-related parameters (CMK resource group and URL) during CDP environment registration on Azure provide that managed identity during Azure environment registration in CDP.

Azure prerequisites

You should first meet the CMK-related prerequisites described in Azure Requirements: Customer managed encryption keys (add additional credential permissions and create a key vault and vault key).

In addition to that, you should create a managed identity as described in Managed identity for encrypting Azure Database for PostgreSQL Flexible Server.

Create an environment with a CMEK for encrypting Flexible Server

Steps

For CDP UI

Follow the usual steps for creating a CDP environment on Azure and make sure to do the following:

1. In the Register Environment wizard, on the Region, Networking and Security page find the Customer-Managed Keys section.
2. Click Enable Customer-Managed Keys.
3. In the same section, under Select Resource group select the resource group where your CMK is located.
4. Provide the URL of the key value where the CMK resides. This is the same as the key identifier that you can copy directly from Azure Portal.
5. Under Managed identity for encryption, select the managed identity created as a prerequisite.

For CDP CLI

Add the following CDP CLI parameters to the cdp environments create-azure-environment command:

```
--encryption-key-resource-group-name <CMK_RESOUCE_GROUP_NAME>
--encryption-key-url <KEY_RESOUCE_ID>
--user-managed-identity <EXISTING_MANAGED_IDENTITY>
```

The --encryption-key-resource-group-name is not needed in some cases, as described in CDP CLI steps listed in Adding a customer managed encryption key to a CDP environment running on Azure.

While the first two parameters are required for using CMK for CDP in general the third parameter must be added in addition for encrypting a Flexible Server. It should be specified as in the following example:

```
--user-managed-identity /subscriptions/3ddda1c7-d1f5-4e4b-ac81-0523f483b
3b1/resourcegroups/test-daily-rg/providers/Microsoft.ManagedIdentity/use
rAssignedIdentities/test-adminIdentity
```

Update an existing environment to add the managed identity

Steps

For CDP UI

The steps for adding the managed identity are similar as those described in Set a CMK for an existing Azure environment, just in the Managed identity for encryption field, a managed identity should be provided in addition to the Encryption Key Resource Group and the Encryption Key URL.

For CDP CLI

The steps for adding the managed identity are similar as those described in Set a CMK for an existing Azure environment, just the --user-managed-identity should be specified in addition.

Use the following CDP CLI command:

```
cdp environments update-azure-encryption-resources \
 --environment <ENVIRONMENT_NAME> \
 --encryption-key-url <VAUL_KEY_URL> \
 --encryption-key-resource-group-name <RG_NAME> \
 --user-managed-identity <EXISTING_MANAGED_IDENTITY>
```

Enable Single Server

During environment registration in CDP, the Flexible Server in public service mode is used by default, but you can specify to use the Single Server.

Steps

For CDP UI

During environment registration in CDP, in the Network and Availability section, enable the Single Server by selecting “Single Server” from the dropdown. The “Flexible Server” option is pre-selected by default.

For CDP CLI

1. Register an Azure environment using the cdp create-azure-environment CDP CLI command as usual.
2. Set IDBroker mappings as usual using the cdp environments set-id-broker-mappings command.
3. Create a Data Lake using the cdp create-azure-datalake CDP CLI command, including a reference to the database HA type. For example:

```
cdp datalake create-azure-datalake
  --datalake-name <datalake-name>
...
  --database-type SINGLE_SERVER
```

The following table explains the required parameters:

Parameter	Description	Possible values
database-type (string)	The type of the azure database. FLEXIBLE_SERVER (recommended) is the next generation managed PostgreSQL service in Azure that provides maximum flexibility over your database, built-in cost-optimizations. SINGLE_SERVER is a fully managed database service with minimal requirements for customizations of the database.	• FLEXIBLE_SERVER • SINGLE_SERVER

4. Create a Data Hub using the create-azure-cluster CDP CLI command, including a reference to the database HA type and the database PostgreSQL engine version. For example:

```
cdp datahub create-azure-cluster
 --cluster-name <marketplace-data-hub>
...
 --datahub-database HA
```

```
--database-type SINGLE_SERVER
```

The following table explains the required parameters:

Parameter	Description	Possible values
datahub-database (string)	Represents the database availability type.	• HA (This means, same zone HA) • NON_HA
database-type (string)	The type of the azure database. FLEXIBLE_SERVER (default value) is the next generation managed PostgreSQL service in Azure that provides maximum flexibility over your database, built-in cost-optimizations. SINGLE_SERVER is a fully managed database service with minimal requirements for customizations of the database. If you do not specify this parameter, FLEXIBLE_SERVER is used by default.	• FLEXIBLE_SERVER • SINGLE_SERVER

Troubleshooting Flexible Server

Refer to this documentation for troubleshooting Azure environments using Flexible Server.

Missing default outbound access

Problem:

Data Lake provisioning failed due to missing outbound access.

Solution:

By default, Single Server utilizes service endpoints (Azure Database for PostgreSQL server - Microsoft.Sql) that provide secure and direct connectivity to DB service over an optimized route over the Azure backbone network from the Virtual Machines, but that is not the case with Flexible Server. This means that there are certain cases when you have to explicitly ensure that outbound (i.e. egress) network connectivity from the selected Virtual Network is set up using NAT-Gateway, UDR, or similar.

This requirement is applicable in the following conditions:

- An existing network is being used.
- Create Public IP is disabled.
- Public Endpoint Gateway is disabled.

The lack of default outbound access can be mitigated by using private setup or using public IPs instead.

For more information, see Default outbound access in Azure.

Enabling a private endpoint for Azure Postgres

By default CDP uses service endpoints, but you can select to use private endpoints instead. During environment registration you can optionally select the “Create Private Endpoint” option to use private endpoints instead of using a service endpoint. Currently, only one service or private endpoint is used, for Azure Postgres.

The endpoint type - either service or private endpoints - and whether to use your own private DNS zones or have CDP create them are decided at environment creation. All services capable of that endpoint type will use the selected endpoint type - this currently means only the Postgres server.

Note:

The endpoint type cannot be changed after the environment was created.

By default service endpoints are used, and in order to use private endpoints the user has to explicitly enable the “Create Private Endpoints“ option from either the CDP UI or CDP CLI. There is no option to not use either. Once you have turned on private endpoints then you can specify your own private DNS zone if you wish.

Prerequisites

You should also decide if you would like to use your own Azure private DNS zone or have CDP create an Azure private DNS zone for you, and meet the related prerequisites. Review the requirements described in Private endpoint for Azure Postgres.

Enabling private endpoints

You can enable private endpoints and select to use a CDP-managed Azure private DNS zone or your own Azure private DNS zone during Azure environment registration in the Management Console.

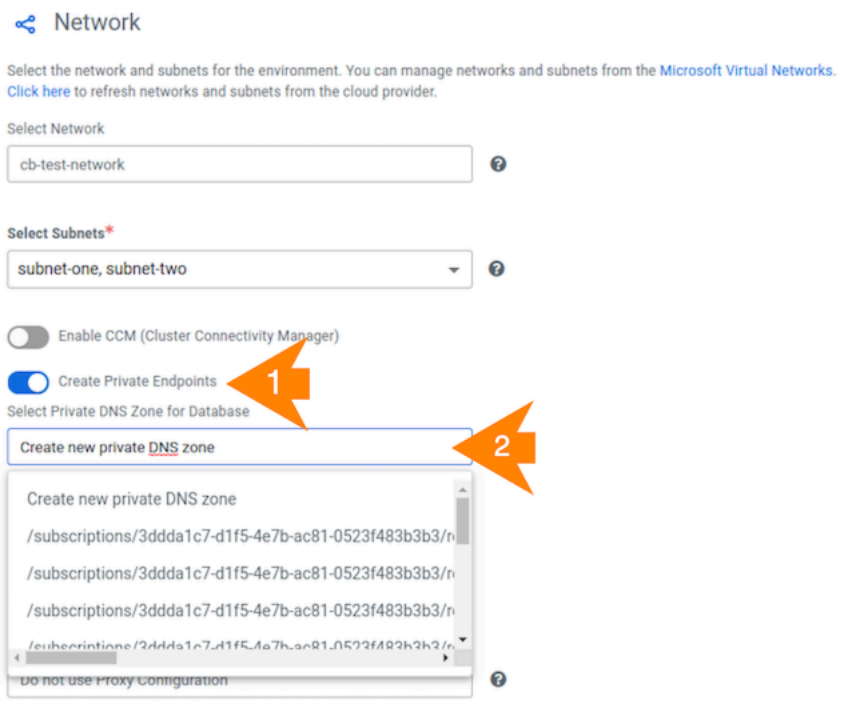
Required role: EnvironmentCreator

Steps

For CDP UI

When registering an environment specify the following:

1. In the Resource group section, select to use your pre-created resource group.
2. In the Network section of the register environment wizard, enable the "Create Private Endpoints" option.
3. A new drop down option appears. You have two options:
   a. Select an existing private DNS zone.
   b. If you would like CDP to create the private DNS zone, select “Create new private DNS zone”.



4. Provide other Azure environment parameters as usual and complete environment registration.

For CDP CLI

The parameter createPrivateEndpoints defines that Azure Postgres will be configured with a Private Endpoint and a Private DNS Zone. When this option is disabled, Azure Service Endpoints are created. This option is disabled by default, so Azure Service Endpoints are used by default.

This is how to specify this via CDP CLI when running the cdp environments create-azure-environment command:

```
--create-private-endpoints
--no-create-private-endpoints
```

If you would like to specify your own Azure private DNS zone, you should additionally include the databasePrivateDnsZoneId as part of existing-network-params:

```
--existing-network-params networkId=<NETWORK_ID>,resourceGroupName=<NETWORK_RG_NAME>,subnetIds=<SUBNET_1>,<SUBNET_2>,databasePrivateDnsZoneId=<PRIVATE_DNS_ZONE_RESOURCE_ID>
```

The <PRIVATE_DNS_ZONE_RESOURCE_ID> that you need to provide should look similar to:

```
/subscriptions/<SUBSCRIPTION_ID>/resourceGroups/<RESOURCE_GROUP_NAME>/providers/Microsoft.Network/privateDnsZones/privatelink.postgres.database.azure.com
```

For example:

```
/subscriptions/a9a10f9c-5323-4fd2-8s14-747b2d68784c/resourceGroups/my-resource-group/providers/Microsoft.Network/privateDnsZones/contoso.com
```

Related Information

Register an Azure environment from CDP UI

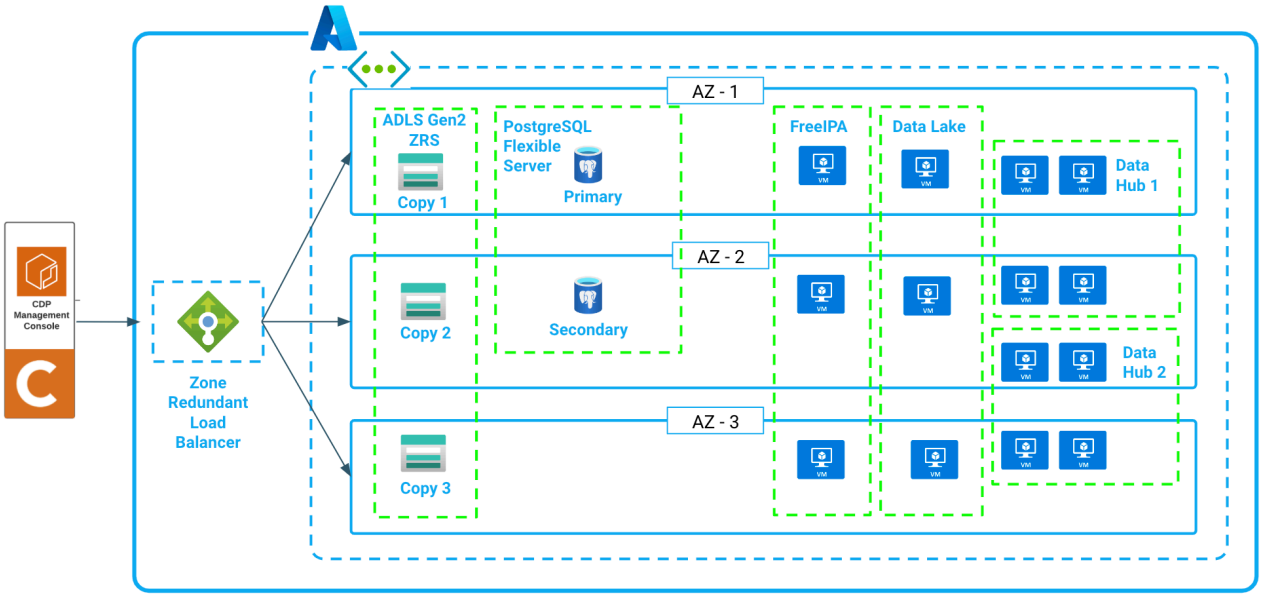
Register an Azure environment from CDP CLI

Deploying CDP in multiple Azure availability zones

You can optionally choose to deploy Data Lake, FreeIPA, and Data Hubs across multiple availability zones (multi-AZ). With multi-AZ support, newly created Azure environments, enterprise Data Lakes and Data Hubs using HA templates can be deployed across multiple availability zones of the selected Azure region. This provides fault tolerance during the extreme event of an availability zone outage.

Note: Only Enterprise Data Lakes can use multi-AZ; other Data Lake shapes do not support it. Enterprise Data Lake is only available in Runtime 7.2.17 and newer; therefore, multi-AZ for CDP on Azure is only available in Runtime 7.2.17 and newer.

Each Azure region has multiple availability zones, which act as failure domains, preventing small outages from affecting entire regions. If you choose to deploy your CDP environment (FreeIPA and Data Lake) and Data Hubs across multiple availability zones, each of these components is spread across three availability zones, providing high availability and fault tolerance. This is illustrated in the following diagram:



With the multi-AZ option enabled, your services are deployed in the following way:

- Azure environments are always created with three FreeIPA servers, deployed on virtual machines spread across three available zones.
- In an Azure Enterprise Data Lake each host group is configured so that virtual machines of all critical services are spread across three available zones.
- In HA Data Hubs, virtual machines of each host group are evenly spread across three availability zones, following a round-robin logic.

When a zone failure happens and a cluster needs to be repaired, the replacement VMs are always provisioned in the same subnet and availability zone as the old ones since the detached disks can only be reattached to a VM in the same availability zone. This means that if there is an availability zone outage, cluster repair is not possible.

By default, If you do not enable multi-AZ, CDP and CDP customers do not have visibility into how Azure distributes VMs across availability zones, because the Azure Portal or CLI do not provide this information.

When creating Data Hubs via CDP CLI, you have the option to specify the AZ, which, in addition to allowing you to select the AZs that should be used, allows you to set up AZ targeting, where all nodes of the cluster are placed on the same AZ. This enables creating disaster recovery scenarios, where a primary and secondary cluster are running in different AZs. If an AZ outage occurs and the primary cluster is lost, it is guaranteed that the secondary cluster is not impacted.

Use cases

A multi-AZ Data Lake and FreeIPA constitute a resilient environment that provides a solid basis for multi-AZ Data Hubs and CDP data services. Data Hubs and CDP data services depend on the FreeIPA instance in the Data Lake to provide DNS resolution. Deploying FreeIPA across multiple availability zones ensures that critical DNS resolution is available in the event of an availability zone outage. Furthermore, a medium duty or enterprise Data Lake provides high availability, and additional compute and memory resources for key SDX services and is recommended for production workloads.

Deploying your Data Hubs across multiple availability zones is key if your mission-critical applications depend on HBase and Kafka. Multiple availability zone deployment for operational workloads is considered best practice by the cloud vendors. It ensures that your applications can continue to run in the event of an availability zone outage.

When an entire availability zone fails, HBase automatically rebalances regions among the remaining instances in the cluster to maintain availability. The write-ahead log (WAL), which is replicated across the three availability zones is automatically replayed by the newly assigned region servers in other availability zones to ensure writes to the database are not lost.

When using the multi availability zone feature, CDP ensures that Kafka replicates partitions across brokers in different availability zones. During an availability zone failure this ensures that no data is lost and applications can continue to access the data they need. Cruise Control, which is deployed alongside every Kafka cluster in CDP Public Cloud, detects that topics need to be rebalanced to the remaining brokers. Once the availability zone is back online, you can repair your Kafka cluster, restoring the initial broker distribution across availability zones. Afterwards Cruise Control kicks in and ensures that all topic partitions are balanced across the cluster.

Limitations

The following limitations apply when deploying a multi-AZ CDP:

- When an AZ is down, you cannot create a new Data Hub, and create or activate CDP data services within the environment. Existing workloads will continue to work.
- When an AZ is down, you cannot resize, stop, or restart Data Hubs.
- Non-AZ environments or clusters cannot be converted to multi-AZ.

Azure requirements

In order to use multi-AZ, you should meet the following Azure requirements:

1. The Azure region that you select should support setting up Azure PostgreSQL Flexible Server in Zone-Redundant HA mode and also the instance types to be used. See Flexible Server Azure Regions.
2. The ADLS Gen2 storage account should be created as zone-redundant storage (ZRS). To specify ZRS via Azure CLI during storage account creation, the --sku option should be set to Standard_ZRS. Below is a sample Azure CLI command:

```
azure % az storage account create \
 --name test-storage \
 --resource-group rg-test-rg \
 --access-tier Cool \
 --allow-blob-public-access false \
 --allow-cross-tenant-replication false \
 --allow-shared-key-access true \
 --enable-hierarchical-namespace true \
 --skuStandard_ZRS
```

Register a multi-AZ environment

You can register a multi-AZ AWS environment via CDP UI or CDP CLI. You may choose to enable multi-AZ for Data Lake only or for FreeIPA only. There is no requirement to enable both.

Steps

For CDP UI

Register your environment as usual, just make sure to do the following:

1. On the Data Access and Data Lake Scaling page:
   a. Select to use the Enterprise Data Lake.
   b. On the same page, scroll down and in the bottom of the page enable the Advanced Options.
   c. In the Network and Availability section enable the Enable Multiple Availability Zones for Data Lake toggle button in order to enable multi-AZ for Data Lake. The option is disabled by default. The option only appears when the Enterprise Data Lake is selected.
2. On the Region, Networking, and Security page:
   a. Scroll down and in the bottom of the page enable the Advanced Options.
   b. In the Network and Availability section enable the Enable Multiple Availability Zones for Data Lake toggle button in order to enable multi-AZ for FreeIPA. The option is disabled by default.
3. Finish registering your environment as usual.

For CDP CLI

Use the following CDP CLI commands to register an environment with a multi-AZ Data Lake and FreeIPA:

1. Register an Azure environment using thecdp environments create-azure-environment command and include multiAz=true in the--free-ipa parameter as shown in this example:

   ```
   cdp environments create-azure-environment \
   --environment-name test-env \
   ...
   --free-ipa instanceCountByGroup=3,multiAz=true \
   ```

   If you do not include the multiAz=true, the default AZ distribution will be used.

   You can also optionally include the--availability-zonesparameter to select the specific availability zones that should be used. Valid values for availability zones are 1,2 and 3. If this parameter is not provided, all AZs are used. For example:

   ```
   cdp environments create-azure-environment \
   --environment-name test-env \
   ...
   --free-ipa instanceCountByGroup=3,multiAz=true \
   --availability-zones 1 2
   ```

2. Set IDBroker mappings as usual using thecdp environments set-id-broker-mappingscommand.
3. Create a Data Lake using the cdp datalake create-azure-datalake command and adding the --multi-az parameter. For example:

   ```
   cdp datalake create-azure-datalake \
   --datalake-name test-dl \
   --environment-name test-env \
   ...
   --scale ENTERPRISE \
   --runtime 7.2.17 \
   --multi-az
   ```

Create a multi-AZ Data Hub

You can create multi-AZ Data Hubs within any existing environment. Detailed steps are provided below.

Prerequisites

You can create a multi-AZ Data Hub in a multi-AZ environment only. If you are trying to create a multi-AZ Data Hub in an environment that uses the default AZ distribution, you need to first edit that environment and add AZs to it.

Steps

For CDP UI

To enable multi-AZ when creating a Data Hub on Azure, navigate to the Advanced Options > Network And Availability and in the “Azure Availability Zones” section click the toggle button next to Enable using multiple availability zones.

For CDP CLI

You can create a multi-AZ Data Hub by adding the --multi-az option to the Data Hub creation command.

In the --instance-groups parameter, you can optionally include theavailabilityZones to select the specific availability zones that should be used. If this parameter is not provided, all three AZs are used. For example:

```
cdp datahub create-azure-cluster \
 --cluster-name test-cluster1 \
```

```
 --environment-name test-env \
 --cluster-template-name "7.2.17 - Data Engineering: Apache Spark, Apache
Hive, Apache Oozie" \
 --multi-az \
```

```
cdp datahub create-azure-cluster \
 --cluster-name test-cluster1 \
 --environment-name test-env \
 --cluster-template-name "7.2.17 - Data Engineering: Apache Spark, Apache
Hive, Apache Oozie" \
 --multi-az \
 --instance-groups
          nodeCount=1,instanceGroupName=compute,instanceGroupType=CORE,ins
tanceType=Standard_D5_v2,rootVolumeSize=100,attachedVolumeConfiguration=
\[\{volumeSize=100,volumeCount=0,volumeType=StandardSSD_LRS\}\],recovery
Mode=MANUAL,availabilityZones=\[1,2\]
          nodeCount=0,instanceGroupName=gateway,instanceGroupType=CORE,
instanceType=Standard_D8_v3,rootVolumeSize=100,attachedVolumeConfigurati
on=\[\{volumeSize=100,volumeCount=1,volumeType=StandardSSD_LRS\}\],recov
eryMode=MANUAL,availabilityZones=\[2,3\]
          nodeCount=1,instanceGroupName=master,instanceGroupType=GATEWA
Y,instanceType=Standard_D16_v3,rootVolumeSize=100,attachedVolumeConfigur
ation=\[\{volumeSize=100,volumeCount=1,volumeType=StandardSSD_LRS\}\],re
coveryMode=MANUAL,availabilityZones=\[1,2,3\]
          nodeCount=3,instanceGroupName=worker,instanceGroupType=CORE,inst
anceType=Standard_D5_v2,rootVolumeSize=100,attachedVolumeConfiguration=\
[\{volumeSize=100,volumeCount=1,volumeType=StandardSSD_LRS\}\],recoveryM
ode=MANUAL,availabilityZones=\[1,3\]
```

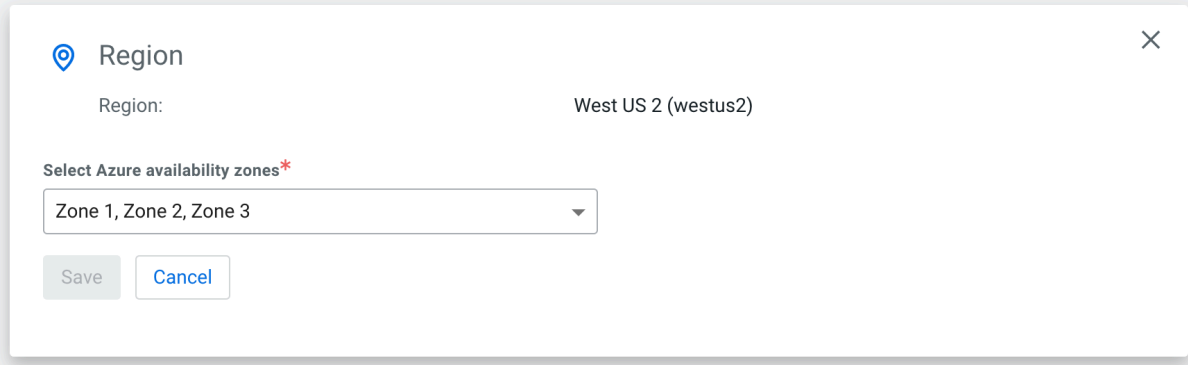
Modify environment's AZs for new Data Hubs

You can modify an environment's AZs. In this case, the added AZs will only be used for new Data Hub clusters. The environment clusters and existing Data Hubs running in it will continue to use the original AZs.

Steps

For CDP UI

1. In the Management Console, navigate to environment details > Summary.
2. Scroll down to the Region section.
3. Under Select Azure availability zones, select the availability zones:



4. Click Save.

For CDP CLI

Below is the CLI command for updating AZs of an existing environment:

```
cdp environments update-azure-availability-zones \
--environment-name test-env \
--availability-zones12 3
```

Restricting access for CDP services that create their own security groups on Azure

The security groups that you select to use during environment registration are only used for the Data Lake, FreeIPA, Data Hubs, and Operational Databases running in that environment. The Kubernetes-based CDP services (Data Warehouse and Machine Learning) create their own security groups with rules that should be restricted separately.

The following table explains where and when you can restrict these rules:

Note:

If you do not restrict these endpoints, CDP defaults to opening access to all (0.0.0.0/0).

CDP service	Type of access that can be restricted	When and where to restrict	Link to related documentation
Data Engineering	Admin access to Kubenetes endpoints can be restricted.	Restrict admin access to Kubernetes endpoints during enabling Data Engineering via the Whitelist IPs parameter.	Enabling Cloudera Data Engineering
Data Warehouse	Both admin access to Kubernetes endpoints and end user access are always set to the same range that can be set in environment activation settings. While the access to the Kubernetes endpoints is a combination of the Cloudera Control Plane's CIDR and your CIDR provided in environment activation settings, the access to the end user access points (JDBC, UI) is only your CIDR provided in environment activation settings.	In Data Warehouse environment's activation settings.	Editing environment details
Machine Learning	There are two separate options, one for admin access to Kubernetes endpoints and another for end user access.	During ML workspace provisioning, under Network Settings: • The Load Balancer Source Ranges parameter can be used to restrict end user access. • Selecting the checkmark Restrict access to Kubernetes API server to authorized IP ranges allows you to restrict admin access to Kubernetes endpoints.	Provisioning ML Workspaces

Configure lifecycle management for logs on Azure

To avoid unnecessary costs related to ALS Gen2 cloud storage, you should create lifecycle management rules for your cloud storage container used by CDP for storing logs so that these logs get deleted once they are no longer useful.

Some examples of CDP logs stored in cloud storage are: cloudera server logs, cloudera agent logs, autossh logs, freeipa logs, ranger audit logs, datahub services logs, datalake logs, cm management services logs, and so on. These logs are mostly useful for troubleshooting, so they can be periodically deleted.

Azure allows you to set up lifecycle management rules for your ADLS cloud storage. For example, you can set a specified expiration period for a cloud storage location so that files in that location get deleted automatically on a scheduled basis. Cloudera recommends that you do this for the cloud storage location that you provided to CDP for log storage.

Consider the following when setting up lifecycle management rules:

- As logs and data locations may overlap with each other (in case the same bucket or container is used for both), ensure to use the correct path prefixes in order to delete only the logs. The prefixes are listed below.
- When setting an expiration period, consider how long you would like to keep the logs to allow enough time for troubleshooting. For example, in case your Data Lake, FreeIPA or Data Hub cluster is ever down, you should be able to access the logs for troubleshooting.

To set up lifecycle management:

- Review the prefixes listed below.
- Follow the instructions in Optimize costs by automating Azure Blob Storage access tiers.

Prefixes based on Azure environment's logs location base

Prior to creating lifecycle management rules, review this information to ensure that you use the correct path.

Note: Path logic changed in February 2021. Starting in February 2021, the path automatically contains the cluster-logs folder as a peer of the cluster-backups folder, creating a better structural separation between logs and backups.

	The "cluster-logs" prefix is automatically generated if a bucket name without any subdirectories is used as logs location	The "cluster-logs" prefix is automatically generated if subdirectories are provided	If your environment was registered prior to February 2021: If you defined a sub-directory, then that subdirectory is used instead of "cluster-logs"
Logs location provided during environment registration	**abfs:// mycontainer@myaccount.dfs.core.windows.n**	**abfs://mycontainer/my-dl@myaccount.dfs.core.window**	**abfs://mycontainer/my-dl@myaccount.dfs.core.window**
FreeIPA prefix for lifecycle rule	mycontainer/cluster-logs/freeipa	mycontainer/my-dl/cluster-logs/freeipa	mycontainer/my-dl/freeipa
DataLake prefix for lifecycle rule	mycontainer/cluster-logs/datalake	mycontainer/my-dl/cluster-logs/datalake	mycontainer/my-dl/datalake
DataHub prefix for lifecycle rule	mycontainer/cluster-logs/datahub	mycontainer/my-dl/cluster-logs/datahub	mycontainer/my-dl/datahub

Related Information

Enabling environment telemetry