

CDP Public Cloud

AWS Requirements

Date published: 2019-08-22

Date modified:

CLOUDERA

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

AWS account requirements.....	4
AWS permissions.....	4
AWS resources and services.....	4
AWS region.....	4
Supported AWS regions.....	4
VPC and subnets.....	6
Security groups.....	7
Default AWS security groups.....	10
SSH key pair.....	14
EC2 instances.....	14
Custom images.....	15
Cross-account access IAM role.....	15
IAM policy definitions.....	15
Create a cross-account IAM role.....	16
Reduced access policy definition.....	20
AWS cloud storage prerequisites.....	23
Minimal setup for AWS cloud storage.....	23
Onboarding CDP users and groups (RAZ).....	33
Onboarding CDP users and groups (No RAZ).....	33
Policy definitions - minimal.....	39
Policy definitions - onboarding.....	41
Using S3 encryption.....	41
Using S3 Express One Zone for data storage.....	42
Customer managed encryption keys.....	44
AWS limits.....	46
List of AWS resources.....	46
AWS outbound network destinations.....	52
Access to workload UIs.....	57
Supported browsers.....	57
Other resources.....	58
CDP CIDR.....	58

AWS account requirements

Prior to registering your AWS environment in CDP, use this document to verify that your AWS account has all the resources required by CDP and that your CDP administrator has adequate permissions to configure the resources and services in AWS.

Related Information

[AWS permissions](#)

[AWS resources and services](#)

[Overview of AWS resources used by CDP](#)

[AWS outbound network access destinations](#)

[Access to workload UIs](#)

[Supported browsers](#)

[Other resources](#)

AWS permissions

As a CDP administrator, you must be able to perform the following tasks in your AWS account:

- Create policies and roles in IAM.
- Create and administer VPCs, subnets, security groups, and S3 buckets.
- Perform administrative tasks in various AWS services such as EC2 and CloudFormation.

AWS Administrator privileges provides all the permissions you need to create the resources for CDP. For the list of AWS resources that CDP services use, refer to [AWS resources used by CDP](#).

AWS resources and services

CDP uses the following resources in your AWS account.

Use the following guidelines to ensure that CDP has access to the resources in your AWS account and that your AWS account has all the necessary resources required by CDP:

AWS region

Prior to registering an environment, you should decide which AWS region you would like to use.


A single AWS environment registered in CDP corresponds to a single VPC located in a specific region, and all the resources deployed by CDP on AWS are deployed into that VPC.

Typically, you may want to deploy clusters into the same region that contains the S3 buckets that you want to access for input and output data. Therefore, when selecting the region to use, you should consider where your data is located. Furthermore, CDP requires that the S3 storage location provided during environment registration is in the same region as the region selected for the environment.

If you need to use multiple regions, you need to register multiple environments, one per region.

Supported AWS regions

CDP supports the following AWS regions.

Region Name	Region ID	Environment	Data Hub	Data Warehouse	Machine Learning	Data Engineering	DataFlow	Operational Database
US East (Ohio)	us-east-2	##	##	##	##	##	##	##
US East (N. Virginia)	us-east-1  Note: The us-east-1e availability zone is not supported.	##	##	##	##	##	##	##
US West (N. California)	us-west-1	##	##			##	##	##
US West (Oregon)	us-west-2	##	##	##	##	##	##	##
Africa (Cape Town)	af-south-1	##	##		##		##	##
Asia Pacific (Mumbai)	ap-south-1	##	##	##	##	##	##	##
Asia Pacific (Osaka)	ap-northeast-3							
Asia Pacific (Seoul)	ap-northeast-2	##	##	##	##	##	##	##
Asia Pacific (Singapore)	ap-southeast-1	##	##	##	##	##	##	##
Asia Pacific (Sydney)	ap-southeast-2	##	##	##	##	##	##	##
Asia Pacific (Jakarta)	ap-southeast-3	##	##	#	##	##	##	#
Asia Pacific (Hong Kong)	ap-east-1	##	##		##	##	##	#
Asia Pacific (Tokyo)	ap-northeast-1	##	##	##	##	##	##	##
Canada (Central)	ca-central-1	##	##	##	##	##	##	##
China (Beijing)	cn-north-1							
China (Ningxia)	cn-northwest-1							
EU (Frankfurt)	eu-central-1	##	##	##	##	##	##	##
EU (Ireland)	eu-west-1	##	##	##	##	##	##	##
EU (London)	eu-west-2	##	##	##	##	##	##	##
EU (Milan)	eu-south-1	##	##	##	##	##	##	##
EU (Paris)	eu-west-3	##	##	##	##	##	##	##
EU (Spain)	eu-south-2	##	##					##
EU (Stockholm)	eu-north-1	##	##	##	##	##	##	##
Israel (Tel Aviv)	il-central-1	##	##					
Middle East (Bahrain)	me-south-1	##	##				##	##
Middle East (UAE)	me-central-1	##	##				##	##
South America (São Paulo)	sa-east-1	##	##	##	##	##	##	##
AWS GovCloud (US-East)	us-gov-east-1							
AWS GovCloud (US-West)	us-gov-west-1							

Related Information

[Unable to view subnet created in "us-east-1e" in CDP console](#)

VPC and subnets

When registering an AWS environment in CDP, you will be asked to select a VPC and two or more subnets.

You have two options:

- Use your existing VPC and subnets for provisioning CDP resources.
- Have CDP create a new VPC and subnets. All CDP resources will be provisioned into this new VPC and subnets.

Existing VPC and subnets

Verify the limits of the VPC and subnets available in your AWS account to ensure that you have enough resources to create clusters in CDP.

If you would like to use your own AWS VPC, it must meet the following requirements:

- The VPC has at least two subnets, each in a different availability zone.
- If you plan to expose your cluster via a public load balancer, you need to have at least one public and one private subnet in the same availability zone and your cluster must be deployed to one of those subnets.
- The VPC subnets must be connected to an Internet Gateway OR a NAT Gateway. The VPC should be able to make an outbound connection with the internet or set of CIDRs and ports provided by Cloudera.
- CDP supports public subnets and private subnets. For private subnets, you must enable [Cluster Connectivity Manager \(CCM\)](#).



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.



Warning:

Be careful if deploying into an existing subnet that has a non-default Network Access Control List (NACL) applied to it. In such case, a security group at EC2 level and a NACL at subnet level are both present at the same time. This can cause a problem because the two settings are evaluated independently (For incoming traffic, the NACL set at the subnet level is evaluated first, then the security group set at the EC2 level. For outgoing traffic the evaluation is the converse), and if an "allow" rule does not exist at both levels, the traffic will not be admitted. Therefore, if you need to use such a setup, you should ensure that the "allow" rule exists on both levels.

- Ensure the CIDR block for the subnets is sized appropriately. In general there is no way to increase the subnet size without recreating the environment and VPC, although Data Warehouse service allows you to [use overlay networks](#).
- When using CCM on AWS, you have an option to use [Public Endpoint Access Gateway](#). If you would like to use it, you should have at least 2 public subnets in the VPC. The exact number of public subnets should be equal to the number of private subnets. Furthermore, availability zones of the public and private subnets must match.

Depending on the CDP services that you are planning to use, you may also need the following:

- If you are planning to use Data Engineering, DataFlow, Data Warehouse, or Machine Learning you must enable Amazon DNS with the VPC. Corporate DNS is not supported. For guidelines on how verify your DNS settings, refer to sections 1-3 in [AWS environment requirements checklist for the Data Warehouse service](#).
- If you are planning to use Data Engineering, DataFlow, or Machine Learning, you must tag the VPC and the subnets as shared so that Kubernetes can find them. For load balancers to be able to choose the subnets correctly, you are also required to tag private subnets with the `kubernetes.io/role/internal-elb:1` tag, and public subnets with the `kubernetes.io/role/elb:1` tag.
- If you want to use the private network feature with Data Warehouse service, you must have three public subnets and three private subnets in your VPC.
- If you are planning to use the Data Warehouse service, you must verify that the VPC has the settings listed in the [AWS environment requirements checklist for the Data Warehouse service](#).

- If you plan to use the private networking feature in the Data Warehouse service, refer to [Supported deployment modes for private networking in AWS](#) and [Prerequisites for private networking in AWS environments](#).
- If you are planning to use DataFlow, Data Engineering, Data Warehouse, or Machine Learning, you may also want to review the following AWS documentation: [Amazon EKS - Cluster VPC Considerations](#) and [Creating a VPC for your Amazon EKS Cluster](#).

New VPC and subnets

If you choose to allow CDP to create a new VPC, six subnets will be created automatically. One subnet is created for each availability zone assuming three AZs per region; If a region has two AZs instead of three, then still three subnets are created, two in the same AZ.

You will need to specify a valid CIDR in IPv4 range that will be used to define the range of private IPs for EC2 instances provisioned into these subnets. Default is 10.10.0.0/16. Consider changing the IP range to correspond to corporate policies for standardized IP address ranges. The CIDR must match the <network mask>/16 pattern.

By default CDP creates 6 subnets (3 private and 3 public) and divides the address space as follows:

- 3 x /19 private subnets for FreeIPA, Data Lake, Data Hub, Data Warehouse, Machine Learning
- 3 x /24 public subnets

You can disable creating private subnets, in which case only 3 public subnets will be created.

By default, when creating a new network, CDP uses public endpoints. But during environment registration you can optionally select the “Create Private Endpoints” option to use private endpoints instead of public endpoints.

If you choose to use private endpoints, make sure to review [Outbound network access destinations](#).

VPCs can be created and managed from the [VPC console](#) on AWS. For instructions on how to create a new VPC on AWS, refer to [Create and configure your VPC](#) in AWS documentation.

Security groups

Security groups determine the inbound and outbound traffic to and from your CDP environment. That is, you should use security group settings to allow users from your organization access to CDP resources.

You have two options:

- Use your existing security groups (recommended for production)
- Have CDP create new security groups

You should verify the security group limits in your AWS account to ensure that you can create security groups for CDP.

Existing security groups

If you would like to create your own security groups, two security groups must be created: the first security group will be used for all gateway nodes and the second security group will be used for all other nodes. The gateway nodes accept incoming requests for the cluster services and so require an additional port. These security groups will be applied when creating a data lake and FreeIPA during environment creation and when you create Data Hub clusters.

Review the following guidelines prior to adding security groups rules. This describes all the inbound ports that need to be open and provides guidelines for what to enter as a source range:



Note: The communication via TCP/UDP 0-65535 and ICMP is essential for healthy operation of CDP environments, Data Hubs, and data services running within the VPC, so ensure that you open these ports as described below. While some services only need well-known fixed ports, a majority of them depend on ephemeral (i.e. dynamically or randomly allocated) ports; This is why the wildcard 0-65535 TCP/UDP port range is used in the absence of a detailed breakdown of individual ports. Since overall access to the VPC is typically secured by other means, the use of the wildcard rules does not pose a higher risk against external attacks.

Gateway security group

This is used for gateway nodes:

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to cluster hosts. You should open it to your organization's CIDR.
TCP	443	Your CIDR and CDP CIDR	This port is used to access the Data Lake and Data Hub cluster UIs via Knox gateway. You must open this port to your organization's CIDR in order to access cluster UIs. When CCM is enabled, you only need to set this to your CIDR.
TCP	9443	CDP CIDR	This port is used by CDP to maintain management control of clusters and data lakes. By default, when CDP creates the security groups automatically, it opens this port to the correct IP. This port is not needed when CCM is enabled.
TCP, UDP	0-65535	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

Example rules provided in the VPC console on AWS:

Edit inbound rules

Inbound rules control the incoming traffic that's allowed to reach the instance.

Type	Protocol	Port Range	Source	Description
Custom ...	TCP	22	Custom ... 74.217.76.101/32	Set to my CIDR (SSH)
Custom ...	TCP	443	Custom ... 74.217.76.101/32	Set to my CIDR (UI access)
Custom ...	TCP	9443	Custom ... 52.36.110.208/32, 52.40.16...	Set to CDP CIDR
Custom ...	TCP	0-65535	Custom ... 10.10.0.0/16	Open to internal commun...
Custom ...	UDP	0-65535	Custom ... 10.10.0.0/16	Open to internal commun...

[Add Rule](#)

Default security group

This is used for all nodes except Knox gateway nodes:

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to the hosts. You should open it to your organization's CIDR.

Protocol	Port Range	Source	Description
TCP	443	Your CIDR	This port is only required if you are planning to spin up Machine Learning workspaces since HTTPS access to ML workspaces is available over port 443. If you are not planning to use the Machine Learning service, you do not need to open this port.
TCP	9443	CDP CIDR	This port is used by CDP to maintain management control of clusters and data lakes. By default, when CDP creates the security groups automatically, it opens this port to the correct IP. This port is not needed when CCM is enabled.
TCP, UDP	0-65535	Your VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC. TCP port 5432 is used by the Data Lake for communication with its attached database.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

Example rules provided in the VPC console on AWS:

On AWS, you can create security groups and edit their rules from the [VPC console](#) > Security Groups.

To create a security group, click on Create security group and provide the following:

You need to create two security groups: Knox and Default (You will see this terminology in the Management Console UI and CLI, so if you decide to choose different names, make sure that you are able to distinguish between the two security groups).

Use the guidelines and examples provided above when editing rules.

To edit security group rules, select the security group and click on Inbound Rules > Edit rules:

Type	Protocol	Port Range	Source	Description
All traffic	All	All	sg-0005b65421c337f17	Allow node
All traffic	All	All	sg-07986c2d6711f3915	
SSH	TCP	22	192.175.27.0/24	Allow node



Note:

There is a known issue where even if you create and specify your own security groups, the Data Warehouse and Machine Learning services create their own security groups. For instructions on how to restrict access on the security groups created by the Data Warehouse service, refer to [Restricting access to endpoints in AWS environments](#).

New security groups

If you would like CDP to create the security groups for you, you need to provide a CIDR range for inbound traffic to EC2 instances from your organization. CDP creates multiple security groups: one for each Data Lake host group, one for each FreeIPA host group, and one per host group when DataFlow, Data Hub, Data Warehouse, and Machine Learning clusters are created. On these security groups, CDP opens ports as described in [Default security group settings](#) documentation.



Note:

There is a known issue where even if you create and specify your own security groups, the Data Warehouse and Machine Learning services create their own security groups. For instructions on how to restrict access on the security groups created by the Data Warehouse service, refer to [Restricting access to endpoints in AWS environments](#).

Related Information

[Restricting access for CDP services that create their own security groups on AWS](#)

Default security group settings on AWS

Depending on what you chose during environment creation, CDP can create security groups for your environment automatically or you can provide your own security groups.

**Note:**

Even if you create and specify your own security groups, the Data Engineering, Data Warehouse, and Machine Learning services create their own security groups. Refer to [Restricting access for CDP services that create their own security groups on AWS](#) for instructions on how to restrict access.



Note: The communication via TCP/UDP 0-65535 and ICMP is essential for healthy operation of CDP environments, Data Hubs, and data services running within the VPC, so ensure that you open these ports as described below. While some services only need well-known fixed ports, a majority of them depend on ephemeral (i.e. dynamically or randomly allocated) ports; This is why the wildcard 0-65535 TCP/UDP port range is used in the absence of a detailed breakdown of individual ports. Since overall access to the VPC is typically secured by other means, the use of the wildcard rules does not pose a higher risk against external attacks.

Environment security groups

Depending on what you chose during environment creation, CDP can create security groups for your environment automatically or you can provide your own security groups.

- If you choose to use your own security groups, you are asked to create Knox and Default security groups as described in the [Security groups](#) documentation.
- If you choose for CDP to create all security groups required for an environment, the following security groups are created:

Data Lake: master

AWS naming convention: `#{environment-name}-#{random-id}-ClusterNodeSecurityGroupmaster-#{random-id}`

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to cluster hosts. You should open it to your organization's CIDR.
TCP	443	Your CIDR and CDP CIDR	This port is used to access the Data Lake and Data Hub cluster UIs via Knox gateway. You should open it to your organization's CIDR in order to access cluster UIs. This port is also required if you are planning to spin up Machine Learning workspaces since HTTPS access to ML workspaces is available over port 443. If you are not planning to use the Machine Learning service, you do not need to open this port. When CCM is enabled, you only need to set this to your CIDR.
TCP	9443	CDP CIDR	This port is used by CDP to maintain management control of clusters and data lakes. This port is not used when CCM is enabled.
TCP, UDP	0-65535	Your VPC's CIDR (for example 10.10.0.0/16) and your subnet's CIDR (for example 10.0.2.0/24).	This is required for internal communication within the VPC.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

Data Lake: IDBroker

AWS naming convention: `${environment-name}-${random-id}-ClusterNodeSecurityGroupidbroker-${random-id}`

Azure naming convention: `idbroker${dl-name}${numeric-id}sg`

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to cluster hosts.
TCP, UDP	0-65535	Your VPC's CIDR (for example 10.10.0.0/16) and your subnet's CIDR (for example 10.0.2.0/24).	This is required for internal communication within the VPC.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

FreeIPA

AWS naming convention: `${environment-name}-freeipa-${random-id}-ClusterNodeSecurityGroupmaster-${random-id}`

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to cluster hosts. You should open it to your organization's CIDR.
TCP	9443	CDP CIDR	This port is used by CDP to maintain management control of clusters and data lakes. This port is not used when CCM is enabled.
TCP, UDP	0-65535	Your VPC's CIDR (for example 10.10.0.0/16) and your subnet's CIDR (for example 10.0.2.0/24).	This is required for internal communication within the VPC.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

Database

AWS naming convention: `dsecg-dbsvr-${random-id}`

Protocol	Port Range	Source	Description
TCP	5432	Your VPC's CIDR (for example 10.10.0.0/16)	This port is used for communication between the Data Lake and its attached database.

Data Hub security groups

Depending on what you chose during environment creation, CDP can create security groups for your Data Hub clusters automatically or it can use your pre-created security groups:

- If during environment creation, you provided your own security groups, CDP uses these security groups when deploying clusters.
- If during environment creation you chose for CDP to create new security groups, new security groups are created for each Data Hub cluster as follows:

Data Hub: master

AWS naming convention: `${cluster-name}-${random-i}-ClusterNodeSecurityGroupmaster-${random-id}`

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to cluster hosts. You should open it to your organization's CIDR.
TCP	443	Your CIDR and CDP CIDR	This port is used to access the Data Lake and Data Hub cluster UIs via Knox gateway. You should open it to your organization's CIDR in order to access cluster UIs. When CCM is enabled, you only need to set this to your CIDR.
TCP	9443	CDP CIDR	This port is used by CDP to maintain management control of clusters and data lakes. This port is not used when CCM is enabled.
TCP, UDP	0-65535	Your VPC's CIDR (for example 10.10.0.0/16) and your subnet's CIDR (for example 10.0.2.0/24).	This is required for internal communication within the VPC.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

Data Hub: worker

AWS naming convention: `#{cluster-name}-#{random-id}-ClusterNodeSecurityGroupworker-#{random-id}`

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to cluster hosts.
TCP, UDP	0-65535	Your VPC's CIDR (for example 10.10.0.0/16) and your subnet's CIDR (for example 10.0.2.0/24).	This is required for internal communication within the VPC.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

Hub: compute

AWS naming convention: `#{cluster-name}-#{random-id}-ClusterNodeSecurityGroupcompute-#{random-id}`

Protocol	Port Range	Source	Description
TCP	22	Your CIDR	This is an optional port for end user SSH access to cluster hosts.
TCP, UDP	0-65535	Your VPC's CIDR (for example 10.10.0.0/16) and your subnet's CIDR (for example 10.0.2.0/24).	This is required for internal communication within the VPC.
ICMP	N/A	Your internal VPC CIDR (for example 10.10.0.0/16).	This is required for internal communication within the VPC.

Data Warehouse security groups

CDP always creates new security groups when Cloudera Data Warehouses (CDW) are deployed.

Machine Learning security groups

CDP always creates new security groups when Cloudera Machine Learning (CML) workspaces are deployed.

Data Engineering security groups

CDP always creates new security groups when Cloudera Data Engineering (CDE) clusters are deployed.

DataFlow security groups

CDP always creates new security groups when Cloudera DataFlow (CDF) environments are enabled.

SSH key pair

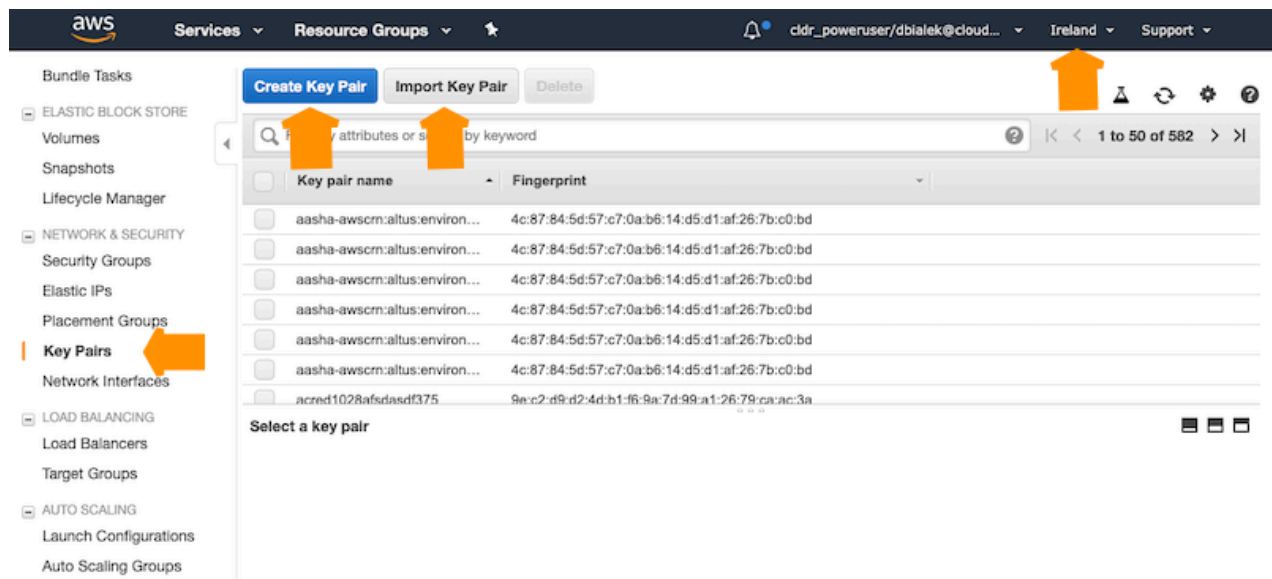
When registering an environment, you will be asked to provide an SSH public key for which you have a matching private key. The minimum SSH key size is 2048 bits.

You have two options:

- Select a public SSH key that is already in EC2 in the region that you would like to use.
- Upload a public SSH key from your computer.

The SSH public key will be used for root-level access to Data Lake and Data Hub instances. Only those users who have the corresponding private key would be able to login as an admin user.

On AWS, you can create or import SSH keys in the [EC2 console](#) > Key Pairs in the specified region by using the Create Key Pair or Import Key Pair options:



For instructions on how to create or import your SSH key on AWS, refer to [Creating a key pair using Amazon EC2](#) and [Importing your own public key to Amazon EC2](#) in AWS documentation.

EC2 instances

CDP provisions EC2 instances as part of environment creation process (for Data Lake and FreeIPA) and for compute clusters.

Therefore, you should verify the limits on the number and type of EC2 instances in your AWS account to ensure that you are able to provision an environment and create clusters in CDP.

CDP supports Amazon EC2 reserved instances; That is, If you have purchased reserved instances, CDP uses them automatically according to [AWS policy](#).

For a list of supported EC2 instance types, refer to [Cloudera Data Platform \(CDP\) Public Cloud service rates](#).

EC2 instances for CDW clusters

On AWS, you can access EC2 instance metadata for your CDW cluster from a running instance using only Instance Metadata Service V2 (IMDSv2). CDP provisions instances for a CDW cluster with IMDSv2. CDP no longer supports IMDSv1 for CDW. Cloudera has removed the capability to access IMDSv1 for a CDW cluster in CDP.

ECS instances for Data Hub and other clusters

On AWS, you can access EC2 instance metadata for Data Hub and clusters other than CDW from a running instance using IMDSv1 or IMDSv2. CDP supports IMDSv1 but does not support IMDSv2 for these clusters, so you should not enable IMDSv2 alone for any EC2 instances used by CDP. For information about configuring IMDS, refer to [Configuring the instance metadata service](#).

Custom images

By default CDP provides a set of default images that are used for all provisioned VMs, but you can optionally use custom images for Data Lake, FreeIPA, and Data Hub.

You might require a custom image for compliance or security reasons (a “hardened” image), or to have your own packages pre-installed on the image, for example monitoring tools or software.

If you would like to use custom images instead of the default images, refer to [Custom images and image catalogs](#).

Cross-account access IAM role

To allow CDP to create resources in your AWS account, you create a cross-account access IAM role in your AWS account and grant CDP access to the role as a trusted principal by specifying a specific AWS account and an external ID.

The policy for the cross-account access IAM role must have the permissions enumerated in the documentation linked below. In addition, the IAM role must reference the specific AWS account ID and external ID provided in the Management Console.

IAM policy definitions

When creating an IAM policy for the provisioning credential, use the following IAM policy definitions.

You have two options:

Option	Description	Policies
Default policy	This policy is appropriate for registering a CDP environment and provisioning all CDP services. This is the same IAM policy as the Default policy provided in the Management Console web interface.	<ul style="list-style-type: none"> Default policy

Option	Description	Policies
Reduced access policies	Instead of the extensive default policy, you can use minimal reduced access policies. Note that the first policy is required for provisioning a CDP environment. The remaining policies are optional and are only needed if you would like to use the specific CDP services.	<ul style="list-style-type: none"> Minimal policy for CDP environment, Data Hub, and Operational Database (COD) Minimal policy for Cloudera Data Engineering (CDE) Minimal policy for Cloudera Data Warehouse (CDW) Minimal policy for Cloudera DataFlow (CDF) AWS IAM restricted roles and policies for compute and CML Minimal policy for using customer managed encryption keys

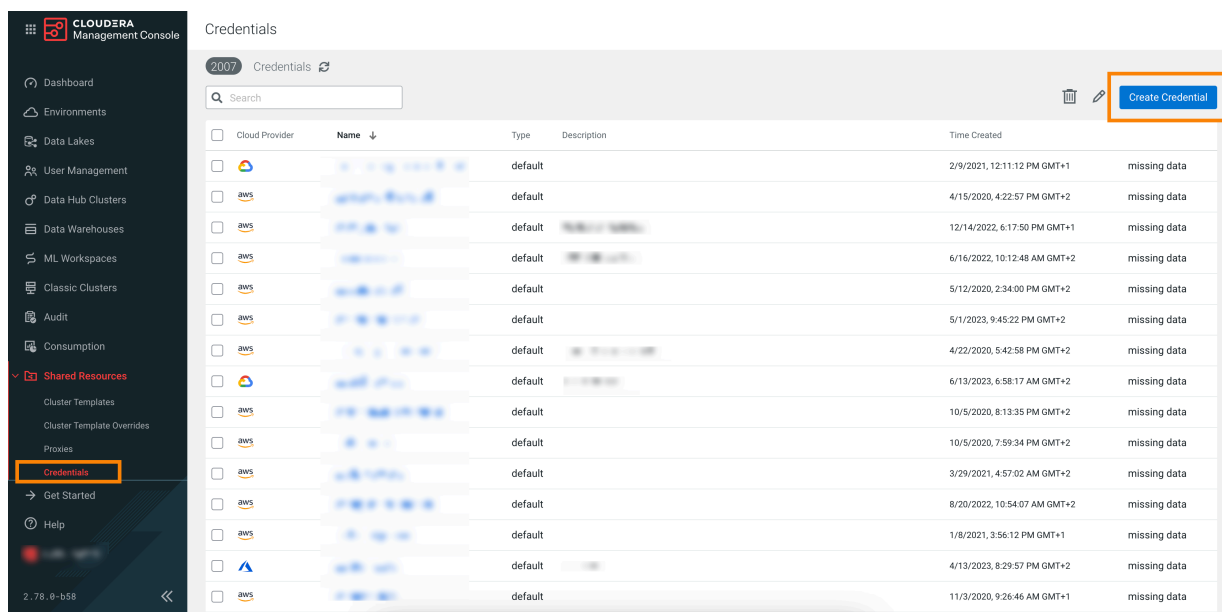
Create a cross-account IAM role

In order to use role-based authentication, you must create an IAM role on AWS.

Before you begin

Prior to creating a cross-account IAM role on AWS, log in to the CDP web interface and obtain the parameters that need to provide for the IAM role:

1. In the Management Console, navigate to Shared Resources Credentials Create Credential :



- Note or copy the Cross-account Access Policy, Account ID, and External ID listed in the UI:

CloudERA Management Console

Credentials / Create Credential

Create Cross-account Access Policy

Copy the following JSON to create an [AWS IAM policy](#)

Default Minimal

The default role allows for the default set of operations including everything that the minimal role allows for.

```
{
  "Statement": [
    {
      "Sid": "CloudFormationFull",
      "Action": [
        "cloudformation:*"
      ],
      "Effect": "Allow",
      "Resource": [
        "*"
      ]
    },
    {
      "Sid": "CloudWatchMetrics"
    }
  ]
}
```

Create Cross-account Access Role

Use Service Manager Account ID and External ID to create an [AWS IAM role](#)

Service Manager Account ID*

External ID*

Cross-account Role ARN*

Enter Cross-account Role ARN

Create > SHOW CLI COMMAND



Note:

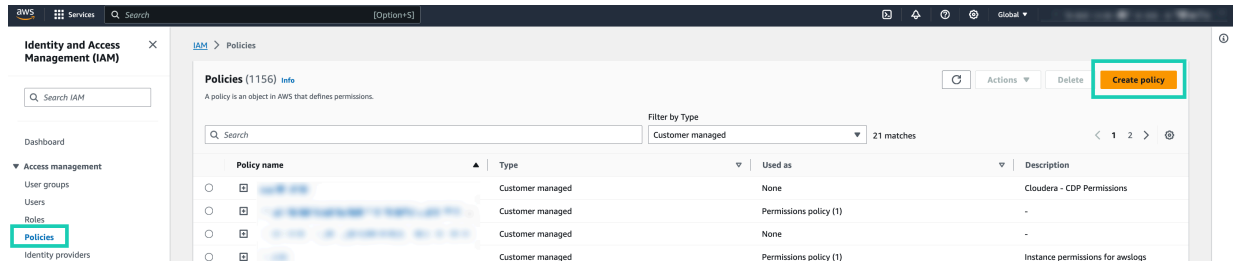
The external ID is tied to the CDP user who obtained it. Therefore, only the user who obtained the external ID is able to complete the credential creation flow in CDP with a given external ID. If a CDP user tries using an IAM role with an external ID obtained by another CDP user, CDP will return an error message.

You will need them to complete the following steps.

Procedure

- Log in to the AWS Management Console.

2. Navigate to the IAM console Policies and click Create policy:

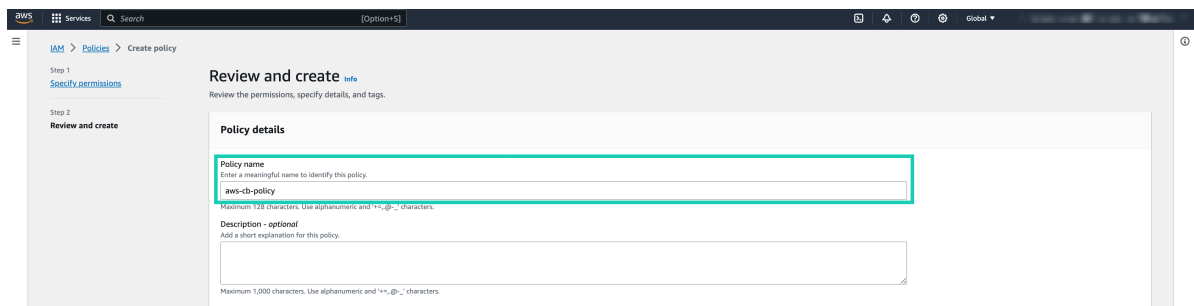


a) Select the JSON tab and paste the Cross-account Access Policy to the Policy editor.



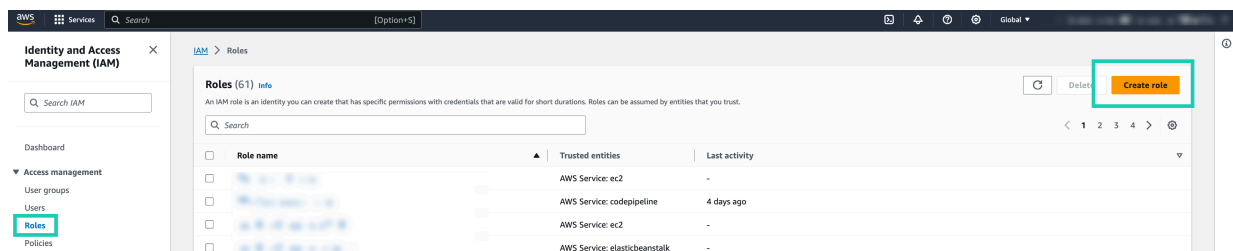
b) Click Next.

c) Enter a name for your policy in the Name field of the Review policy page:



d) Click Create policy when finished.

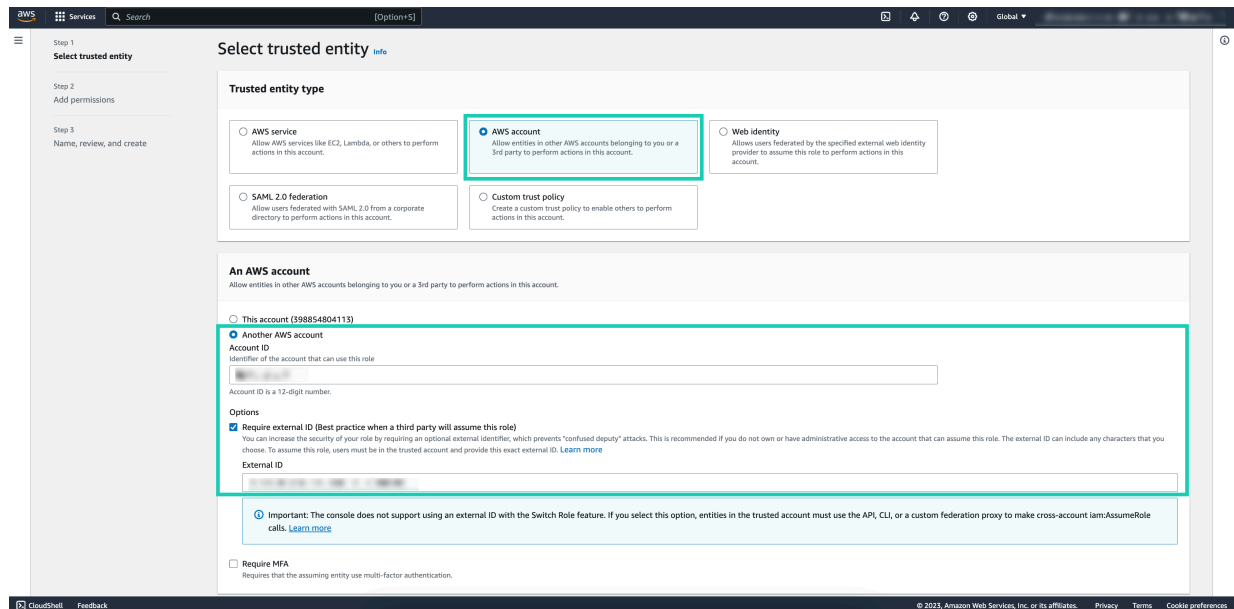
3. Navigate to the IAM console Roles and click Create role:



4. Select AWS account as Trusted Entity type.

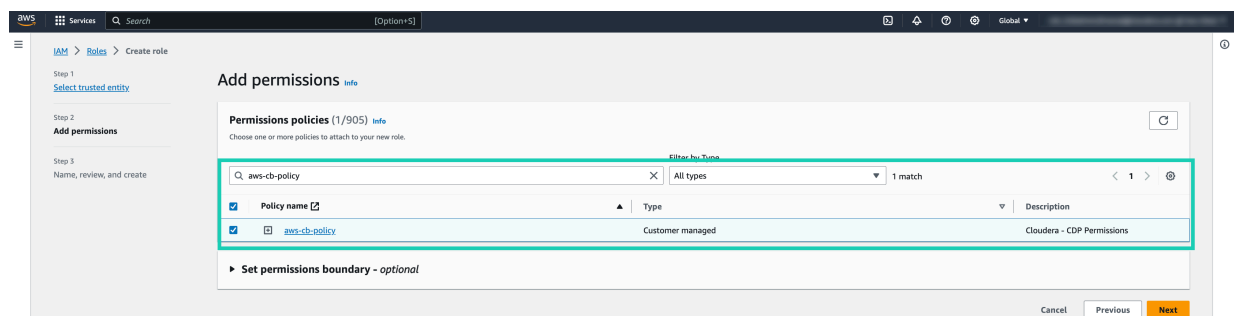
5. Check Another AWS account and provide the following information:

- Copy and paste your Account ID provided in CDP to the Account ID field.
- Check Require external ID under **Options**.
- Copy and paste the External ID from CDP to the External ID field.



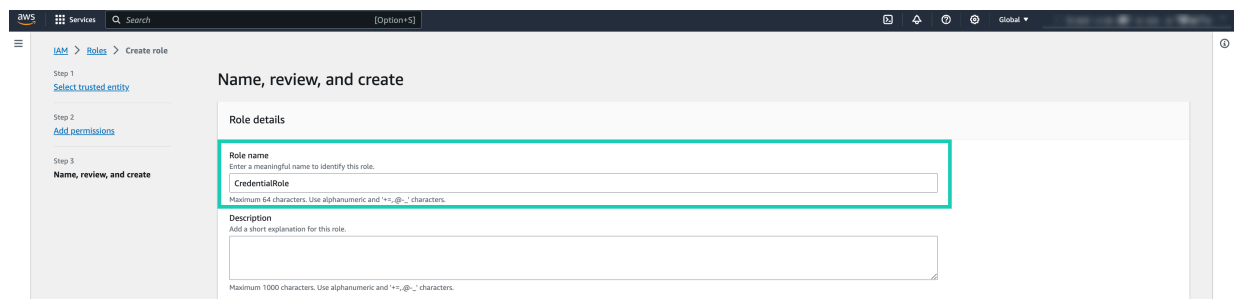
6. Click Next.

7. Search for the previously created policy, and add the permission to the role by checking the box.



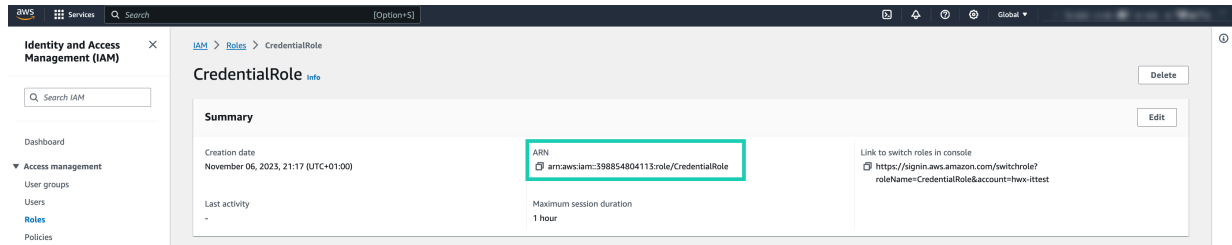
8. Click Next.

9. Enter a role name to the Role name field:



10. Click Create role to finish the role creation process.

11. Search for the created role, and obtain the IAM Role ARN. You will need it to create a role-based credential:



What to do next

Once you are done creating the IAM role on AWS, [create a role-based credential](#) in CDP.

Reduced access policy definition

The following reduced policy definition can optionally be used to register a CDP environment and create Data Hubs and Operational Databases in it.



Note: This policy is only sufficient for registering a CDP environment and creating Data Hubs and Operational Databases in it. In order to use other CDP services, you must add additional service-specific permissions to your cross-account role.

The IAM role used for the provisioning credential can use the [reduced access policy](#) instead of the default policy.

You should:

1. Copy this policy.
2. Update it as described below.
3. Once done, you can create the cross-account role using this policy.

Customize the scope of iam:PassRole

CDP utilizes the iam:PassRole as part of the cloud identity federation setup process via IDBroker on AWS. The default policy uses "*" as the resource scope for this action:

```
{
  "Effect": "Allow",
  "Action": [
    "iam:PassRole"
  ],
  "Resource": "*"
}
```

However the scope can be reduced to just a few roles. To do this, you need to edit the following block and replace the "*" with the ARNs of the following roles used for the environment:

- IDBROKER_ROLE
- LOG_ROLE
- RANGER_AUDIT_ROLE
- DATALAKE_ADMIN_ROLE

You choose the name for these roles when you set up CDP as part of the [Minimal Setup for Cloud Storage](#).

After updating according to the minimal setup for cloud storage, the iam:PassRole policy excerpt should be similar to the following:

```
{
  "Effect": "Allow",
  "Action": [
    "iam:PassRole"
  ],
  "Resource": [
    "arn:aws:iam::398854804113:role/IDBROKER_ROLE",
    "arn:aws:iam::398854804113:role/LOG_ROLE",
    "arn:aws:iam::398854804113:role/RANGER_AUDIT_ROLE",
    "arn:aws:iam::398854804113:role/DATALAKE_ADMIN_ROLE"
  ]
}
```

```

    ],
    "Resource": [
        "arn:aws:iam::${ACCOUNTID}:role/IDBROKER_ROLE",
        "arn:aws:iam::${ACCOUNTID}:role/LOG_ROLE",
        "arn:aws:iam::${ACCOUNTID}:role/RANGER_AUDIT_ROLE",
        "arn:aws:iam::${ACCOUNTID}:role/DATALAKE_ADMIN_ROLE"
    ]
}

```

If you have already created these roles and you don't know the name of these roles, you can obtain them using the following steps:

IDBROKER_ROLE

1. Log in to CDP web interface and navigate to the Management Console service > Data Lakes.
2. Scroll down to where you see Event History and select the Hardware tab.
3. Locate the EC2 instance for the idbroker node (if you are running Medium Duty Data Lake, you will see two IDBroker nodes, you can use either):

Environments / svv-aws / Data Lake / Hardware

The screenshot shows the 'Environment Details' page for 'svv-aws' in the 'us-west-2' region, 'us-west-2b' availability zone. Under the 'Hardware' tab, a table lists EC2 instances. The 'Master' instance (ID: i-07fe5627529cb130a) is in a 'Running' state with FQDN 'svv-aws-datalake-master0.svv-aws.xcu2-8y8x.dev.cldr.work'. The 'Idbroker' instance (ID: i-0011589ce16b3feb3) is also in a 'Running' state with FQDN 'svv-aws-datalake-idbroker0.svv-aws.xcu2-8y8x.dev.cldr.work'. Orange arrows highlight these two instances.

4. Click on the instance id. You will be redirected to the EC2 console.
5. In the Instance Details section for this node, you can see the name of the IAM role being used by this instance. Click on this role to get to the details.
6. Copy the ARN of the instance profile. It follows the following naming convention: `arn:aws:iam::<12-digit-AWS-account-id>:instance-profile/<name-of-idbroker-role>`

LOG_ROLE

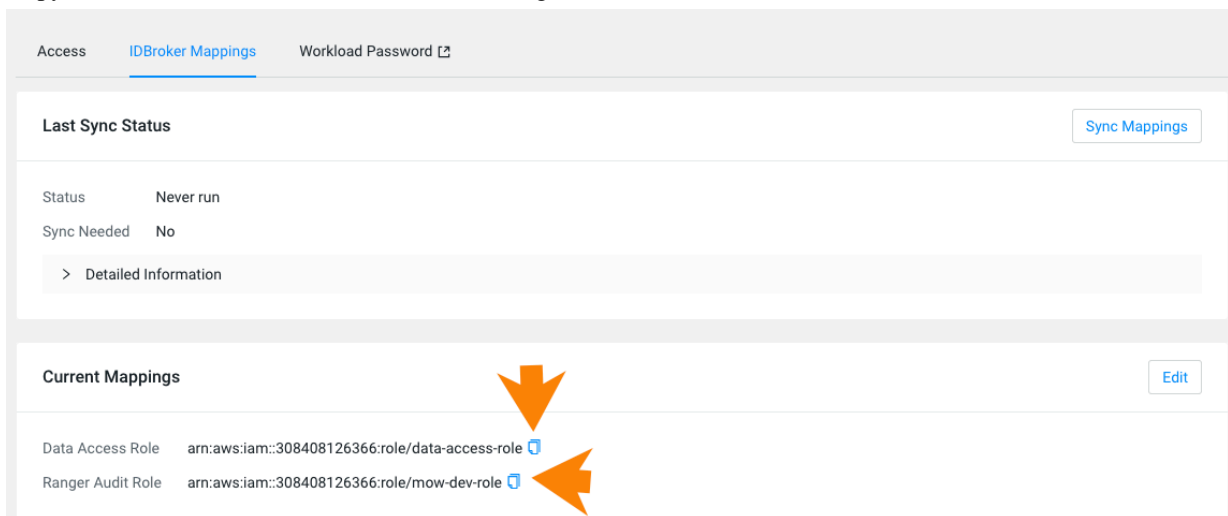
1. Navigate to Environments .
2. Click on the existing environment
3. Click on the Summary tab.
4. Scroll to Logs Storage and Audits and you will see the Instance Profile entry.

5. Copy the ARN of the Instance Profile:



RANGER_AUDIT_ROLE and DATALAKE_ADMIN_ROLE

1. Navigate to Environments.
2. Click on the existing environment.
3. From the Actions menu, select Manage Access.
4. Click on the IDBroker Mappings tab.
5. Under Current Mappings you will see the mappings for the Data Access Role and Ranger Audit Role.
6. Copy the ARNs of the Data Access Role and Ranger Audit Role:



Further permissions reduction (Optional)

You can reduce these policies even further by pre-creating some resources:

- If you set up your own VPC and subnets, then the following actions can be removed from the policy:

```
ec2:CreateVpc
ec2:CreateNatGateway
ec2:CreateRouteTable
ec2:CreateSubnet
ec2:CreateVpcEndpoint
ec2:CreateInternetGateway
ec2>DeleteSubnet
ec2>DeleteInternetGateway
ec2:AttachInternetGateway
ec2:DetachInternetGateway
ec2:DescribePrefixLists
ec2:AllocateAddress
ec2:AssociateRouteTable
ec2:CreateRoute
ec2>DeleteRouteTable
ec2>DeleteVpcEndpoints
ec2:DisassociateRouteTable
ec2:ReleaseAddress
```

```
ec2:DeleteRoute
ec2:DeleteNatGateway
ec2:DeleteVpc
```

- If you use your own security groups, then the following actions can be removed from the policy:

```
ec2:CreateSecurityGroup
ec2:AuthorizeSecurityGroupIngress
ec2:RevokeSecurityGroupEgress
ec2:AuthorizeSecurityGroupEgress
```

- If you use private IPs or set up CCM to communicate with the CDP Control Plane, then the following actions can be removed from the policy:

```
ec2:CreateSecurityGroup
ec2:AuthorizeSecurityGroupIngress
ec2:RevokeSecurityGroupEgress
ec2:AuthorizeSecurityGroupEgress
```

- If you use private IPs or set up CCM to communicate with the CDP Control Plane, then the following actions can be removed from the policy:

```
ec2:AllocateAddress
ec2:ReleaseAddress
```

- If you create your own DynamoDB table for S3Guard, then the following actions can be removed from the policy:

```
dynamodb:DeleteTable
```

S3 bucket, and IAM roles and policies for logs, backups, and data storage

CDP requires that you create and provide at least one S3 bucket for storing workload data and logs. You also need to create and provide multiple IAM roles and policies that allow access to the S3 bucket.



Note:

The Data Warehouse service creates its own S3 buckets, DynamoDB tables, and IAM roles and policies. It does not use the environment's S3 buckets, DynamoDB tables, and IAM roles and policies.

The S3 bucket is used for:

- Storage location base - Workload data storage and Ranger audits
- Logs location base - Service logs, FreeIPA logs
- Backup location base - FreeIPA and Data Lake backups

The S3 bucket must be in the same region as the environment.

For detailed information about the required setup, review the following documentation:

Minimal setup for AWS cloud storage

This minimal secure setup uses one S3 bucket for each Data Lake, and multiple IAM roles and policies.

The example setup includes:

- One S3 bucket with a sub-directory named after your data lake such as `s3a://my-bucket/my-dl`.
- Four IAM roles and two instance profiles:
 - IDBROKER_ROLE role and instance profile
 - LOG_ROLE role and instance profile
 - RANGER_AUDIT_ROLE
 - DATALAKE_ADMIN_ROLE

- Nine IAM policies:
 - One AssumeRole ipolicy (idbroker-assume-role) that can be used by the IDBroker component of the data lake cluster to assume each of the following roles: RANGER_AUDIT_ROLE and DATALAKE_ADMIN_ROLE.
 - A shared policy for accessing S3:
 - aws-cdp-bucket-access-policy
 - Five (or six) policies for specific bucket directory access:
 - Log storage (aws-cdp-log-policy)
 - Ranger audit (aws-cdp-ranger-audit-s3-policy)
 - Data Lake admin and RAZ (aws-cdp-datalake-admin-s3-policy)
 - Data Lake backups (aws-datalake-backup-policy)
 - Data Lake restore (of backups) (aws-datalake-restore-policy)
 - (Optional) FreeIPA backup storage - if using a separate bucket (aws-cdp-backup-policy)
 - Two trust policies:
 - aws-cdp-ec2-role-trust-policy
 - aws-cdp-idbroker-role-trust-policy

**Note:**

S3 Express One Zone buckets should not be used for the S3 cloud storage setup. S3 Express buckets are only supported for use with Data Hub workloads as additional temporary data buckets.

Required cloud storage

One S3 bucket with a sub-directory named after your data lake such as s3a://my-bucket/my-dl is required for this setup.

**Note:**

You can use separate locations for your data, logs, and backups. The Storage Location Base, Logs Location Base, and Backup Location Base, happen to be in the same bucket in this example, but this is not required. If using multiple buckets, the aws-cdp-bucket-access-policy needs to have the additional buckets specified.

**Important:**

Ranger audits are stored under Storage Location Base and not under Logs Location Base. The Logs Location Base is used for Data Lake, Data Hub and FreeIPA logs, and (if no separate location is provided) FreeIPA and Data Lake backups.

During Data Lake creation, CDP automatically creates a location for Ranger audits, Data Lake, Data Hub and FreeIPA logs, and FreeIPA and Data Lake backups. The location for each of these depends on the supplied storage location base and logs location base. The directory structure is created automatically by CDP within these base directories:

Storage Location Base examples

	s3a://my-bucket/	s3a://my-bucket/my-dl
Ranger Audit Logs	s3a://my-bucket/ranger/audit	s3a://my-bucket/my-dl/ranger/audit

Logs Location Base examples

	s3a://my-bucket/	s3a://my-bucket/my-dl
FreeIPA Logs	s3a://my-bucket/cluster-logs/freeipa	s3a://my-bucket/my-dl/cluster-logs/freeipa If your environment was created prior to February 2021, this is s3a://my-bucket/my-dl/freeipa

Backup Location Base examples

	s3a://my-bucket/	s3a://my-bucket/my-dl
FreeIPA Backup	s3a://my-bucket/cluster-backups/freeipa	s3a://my-bucket/my-dl/cluster-backups/freeipa

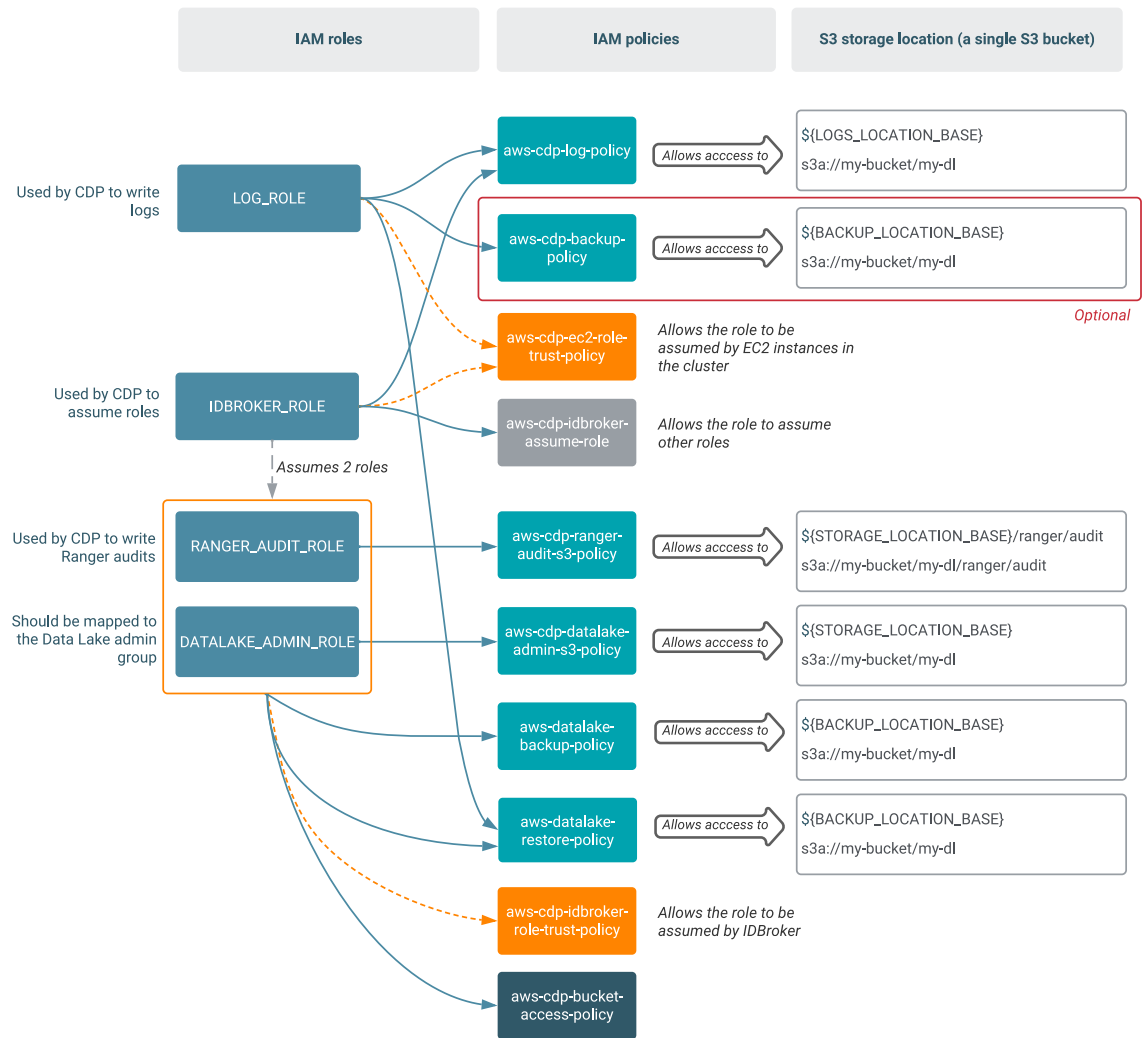
Required IAM resources

The following table lists the IAM roles and IAM policies that need to be created on AWS, and describes which policies should be assigned to which roles (as presented in the diagram, in some cases policies should be assigned to multiple roles). The policy definitions are provided in [IAM policy definitions](#):

Role	Permissions policy	Trust policy	Description
IDBROKER_ROLE	aws-cdp-idbroker-assume-role-policy aws-cdp-log-policy	aws-cdp-ec2-role-trust-policy	<p>The assume role permissions policy must, at a minimum, allow the IDBROKER_ROLE to assume the RANGER_AUDIT_ROLE and the DATALAKE_ADMIN_ROLE. This policy must also allow the IDBROKER_ROLE to assume any other role for which a user or group mapping exists in the IDBroker.</p> <p>Furthermore, the IDBROKER_ROLE needs the same permissions policy as the LOG_ROLE so that it can access the Logs Location Base.</p> <p>The trust policy allows the role to be assumed by the IDBroker EC2 instance.</p>
LOG_ROLE	aws-cdp-log-policy aws-datalake-restore-policy aws-cdp-backup-policy (Optional)	aws-cdp-ec2-role-trust-policy	<p>This role uses the aws-cdp-log-policy permissions policy to provide CDP with access to the specific location called Logs Location Base for logs.</p> <p>If your Backup Location Base is in a separate bucket or folder, you also need to provide the aws-cdp-backup-policy.</p> <p>The trust policy allows the role to be assumed by EC2 instances in the cluster.</p> <p>The aws-datalake-restore-policy is required for both upgrading a Data Lake and restoring a Data Lake backup.</p>
RANGER_AUDIT_ROLE	aws-cdp-ranger-audit-s3-policy aws-cdp-bucket-access-policy aws-datalake-backup-policy aws-datalake-restore-policy	aws-cdp-idbroker-role-trust-policy	<p>This role uses the four permissions policies to provide write access to the Ranger audit sub-directory that CDP creates within the Storage Location Base.</p> <p>The trust policy allows the role to be assumed by IDBroker.</p> <p>The aws-datalake-backup-policy is required for both upgrading a Data Lake and backing up a Data Lake.</p> <p>The aws-datalake-restore-policy is required for both upgrading a Data Lake and restoring a Data Lake backup.</p>

Role	Permissions policy	Trust policy	Description
DATALAKE_ADMIN_ROLE	aws-cdp-datalake-admin-s3-policy aws-cdp-bucket-access-policy aws-datalake-backup-policy aws-datalake-restore-policy	aws-cdp-idbroker-role-trust-policy	<p>This role uses the four permissions policies to provide the Data Lake admin with full access to the whole Storage Location Base.</p> <p>Additionally, if you choose to configure Fine-grained access control, RAZ uses this role to orchestrate the access control.</p> <p>The trust policy allows the role to be assumed by IDBroker.</p> <p>The aws-datalake-backup-policy is required for both upgrading a Data Lake and backing up a Data Lake.</p> <p>The aws-datalake-restore-policy is required for both upgrading a Data Lake and restoring a Data Lake backup.</p>

The following diagram summarizes the roles, policies, and S3 bucket directories in this example setup:





Note: You may choose a different setup, for example one using multiple buckets. Similarly, the IAM role and policy setup and names are just examples and you may choose a different setup. It is possible to have a setup with fewer roles and policies with broader access rights; however, such setup may not be secure for a production environment.

Use the following documentation in order to create the required S3 bucket and IAM resources, and provide them in CDP. The steps involve:

1. Creating an S3 bucket if you don't have one already
2. Creating permissions policies for the minimal setup
3. Creating the required IAM roles and their associated trust policies
4. Providing the parameters in the CDP UI

Create an S3 bucket

You need to provide an S3 location that CDP can use for logs and data storage. If you don't already have an S3 bucket, you can create one from the AWS S3 console or AWS CLI.

Steps

For AWS console

1. In the AWS web interface, navigate to the S3 console.
2. Click on Create bucket.
3. Enter bucket name.
4. Select the region where the bucket should be created.
5. Edit additional settings if needed. For example:
 - Enable encryption
 - Add tags if required by your organization.
6. Click on Create bucket.

For AWS CLI

Use the following command to create an S3 bucket:

```
aws s3api create-bucket --bucket <BUCKET-NAME>
```

See AWS documentation linked below for more information.

Related Information

[Creating a bucket](#)

[AWS CLI Reference: create-bucket](#)

Create permissions policies for the minimal setup

You need to create permissions policies that can be assigned to specific roles required by CDP. You can create these IAM permissions policies from the AWS IAM console or AWS CLI.

Before you begin

Obtain the relevant IAM policy definitions from [IAM policy definitions](#). You need to replace all the placeholders in the policy with actual values.



Note: You only need to create permissions policies. Do not create trust policies in this manner. Trust policies are not created and attached to roles in the same manner as regular IAM policies. Establishing a trust relationship via trust policies is covered as part of a later step described below.

Steps

For AWS console

1. In the AWS web interface, navigate to the IAM console.
2. From the navigation pane, select Policies and then click on Create Policy
3. Choose the JSON tab.
4. Paste the appropriate JSON policy document.



Note: Make sure to replace all the placeholders in the policy with actual values.

5. Click Next.
6. Add tags if required by your organization.
7. Click Next.
8. Type a Name and an optional Description.
9. Click Create policy.

For AWS CLI

Use the following command to create permissions policies:

```
aws iam create-policy --policy-name <NAME> --policy-document <JSON-DOCUMENT>
```

Repeat these steps for all permissions policies that need to be created. For more detailed instructions on how to create IAM policies, refer to the AWS documentation linked below.

Related Information

[Creating IAM policies \(console\)](#)

[AWS CLI Reference: create-policy](#)

Create the IAM roles and establish trust relationships: IDBROKER_ROLE and LOG_ROLE

You can create the IDBROKER_ROLE and LOG_ROLE IAM roles from the IAM console or AWS CLI.

Before you begin

Review the minimal setup for cloud storage outlined above and make sure that you understand which policies you need to attach to which roles.

Steps

For AWS console


1. In the AWS web interface, navigate to the IAM console.
2. From the navigation pane, select Roles.
3. Click on Create Role.
4. Under "Select type of trusted entity" select AWS service.

- Under "Choose a use case" select EC2.


Create role

1 2 3 4


Select type of trusted entity




AWS service
EC2, Lambda and others



Another AWS account
Belonging to you or 3rd party



Web identity
Cognito or any OpenID provider



SAML 2.0 federation
Your corporate directory

Allows AWS services to perform actions on your behalf. [Learn more](#)

Choose a use case

Common use cases

EC2

Allows EC2 instances to call AWS services on your behalf.

Lambda

Allows Lambda functions to call AWS services on your behalf.

Or select a service to view its use cases

- Click Next.
- Attach the permissions policies required for the role that you are creating. Make sure to attach all required policies.
- Click Next.
- Add tags if required by your organization.
- Click Next.
- Provide a name and an optional description for your role, and verify that you have attached all required permissions policies.
- Click on Create role. This creates an IAM role and instance profile.
- Click on the role to see its details.
- Verify that the "Instance Profile ARN" is present:

Roles > test-role

Summary Delete role

Role ARN	arn:aws:iam::[redacted]:role/[redacted]
Role description	Allows EC2 instances to call AWS services on your behalf. Edit
Instance Profile ARNs	arn:aws:iam::[redacted]:instance-profile/test-role
Path	/
Creation time	2021-09-17 14:46 PDT

- Click on the Trust relationships tab.

16. Click on Edit trust relationship.

The screenshot shows the AWS IAM console interface for a role named 'test-role'. The 'Trust relationships' tab is selected, and the 'Edit trust relationship' button is highlighted with an orange arrow. The interface displays various details about the role, including its ARN, description, instance profile ARNs, path, creation time, last activity, and maximum session duration. The 'Trusted entities' section shows one entity: 'The identity provider(s) ec2.amazonaws.com'. The 'Conditions' section indicates that there are no conditions associated with this role.

17. Verify that the JSON declaring a trust policy is the same as described in the minimal setup. If it isn't, replace it with the correct trust policy definition.

18. Click on Update trust policy.

Repeat these steps to create both roles.

For AWS CLI

Use these commands to create the required IAM roles and instance profiles via AWS CLI:

```
aws iam create-role --role name <ROLE-NAME> --assume-role-policy-document
<TRUST-POLICY>
    aws iam attach-role-policy --role name <ROLE-NAME> --policy-arn <P
ERMISSIONS-POLICY-ARN>
    aws iam create-instance-profile --instance-profile-name <IP-NAME>
    aws iam add-role-to-instance-profile --instance-profile-name <IP-NAM
E> --role-name <ROLE-NAME>
```

You should assign the appropriate policies described in the minimal setup. Note that the `aws iam attach-role-policy` command needs to be executed separately for each policy that needs to be attached.

Related Information

[AWS CLI Reference: iam](#)

Create the IAM roles and establish trust relationships: RANGER_AUDIT_ROLE and DATALAKE_ADMIN_ROLE

Use the following steps to create the RANGER_AUDIT_ROLE and DATALAKE_ADMIN_ROLE.

Before you begin

Review the minimal setup for cloud storage outlined above and make sure that you understand which policies you need to attach to which roles.

Steps


For AWS console


1. In the AWS web interface, navigate to the IAM console.
2. From the navigation pane, select Roles.
3. Click on Create Role.
4. Under "Select type of trusted entity" select Another AWS account.


5. Under "Account ID", enter your AWS account ID. You can obtain it by clicking on your user name in the top bar.


Create role 1 2 3 4

Select type of trusted entity

 **AWS service**
EC2, Lambda and others

 **Another AWS account**
Belonging to you or 3rd party

 **Web identity**
Cognito or any OpenID provider

 **SAML 2.0 federation**
Your corporate directory

Allows entities in other accounts to perform actions in this account. [Learn more](#)

Specify accounts that can use this role

Account ID*

Options Require external ID (Best practice when a third party will assume this role)
 Require MFA ⓘ

6. Click Next.
7. Attach the permissions policies required for the role that you are creating. Make sure to attach all required policies described in the minimal setup
8. Click Next.
9. Add tags if required by your organization.
10. Click Next.
11. Provide a name and an optional description for your role, and verify that you have attached all required permissions policies.
12. Click on Create role. This creates an IAM role. No instance profile is created in this case.
13. Click on the role to see its details.
14. Click on the Trust relationships tab.

15. Click on Edit trust relationship.

Roles > test-role2

Summary Delete role

Role ARN arn:aws:iam:::role/test-role2

Role description [Edit](#)

Instance Profile ARNs

Path /

Creation time 2021-09-17 15:02 PDT

Last activity Not accessed in the tracking period

Maximum session duration 1 hour [Edit](#)

Give this link to users who can switch roles in the console <https://signin.aws.amazon.com/switchrole?roleName=test-role2&account=hwxcloudbreak>

Permissions **Trust relationships** **Tags** **Access Advisor** **Revoke sessions**

You can view the trusted entities that can assume the role and the access conditions for the role. [Show policy document](#)

Edit trust relationship

Trusted entities

The following trusted entities can assume this role.

Trusted entities

The account !

Conditions

The following conditions define how and when trusted entities can assume the role.

There are no conditions associated with this role.

16. Paste the JSON declaring a trust policy in place of the default policy. Use the appropriate policy described in the minimal setup.

17. Click on Update trust policy.

Repeat these steps for both roles.

For AWS CLI

Use these commands to create the required IAM roles via AWS CLI:

```
aws iam create-role --role name <ROLE-NAME> --assume-role-policy-document
<TRUST-POLICY>
    aws iam attach-role-policy --role name <ROLE-NAME> --policy-arn <
PERMISSIONS-POLICY-ARN>
```

Note that the `aws iam attach-role-policy` command needs to be repeated if there are multiple policies to attach. You should assign the appropriate policies described in the minimal setup.

Related Information

[AWS CLI Reference: iam](#)

Providing the parameters in the CDP UI

Once you've created the S3 bucket and the required IAM policies, roles, and instance profiles, provide the information related to these resources in the Register Environment wizard.

Parameter	Description	Example
Data Access and Audit		
Assumer Instance Profile	Select the IDBROKER_ROLE instance profile created earlier during IDBROKER_ROLE role creation.	IDBROKER_ROLE
Storage Location Base	Enter the Storage Location Base S3 bucket location created earlier.	my-bucket/my-dl
Data Access Role	Select the DATALAKE_ADMIN_ROLE created earlier.	DATALAKE_ADMIN_ROLE
Ranger Audit Role	Select the RANGER_AUDIT_ROLE created earlier.	RANGER_AUDIT_ROLE
Fine-grained access control on ADLS Gen2		
Enable Ranger authorization for ADLS Gen2	If you would like to use Fine-grained access control , click on "Enable Ranger authorization for ADLS Gen2". Next, select the DATALAKE_ADMIN_ROLE created earlier.	DATALAKE_ADMIN_ROLE
Select Azure managed identity for Ranger authorization		
Logs		
Logger Instance Profile	Select the LOG_ROLE instance profile created earlier during LOG_ROLE role creation.	LOG_ROLE
Logs Location Base	Enter the Logs Location Base S3 bucket location created earlier.	my-bucket-/my-dl
Backup Location Base (Optional)	If you created it, enter the Backup Location Base S3 bucket location. This is optional. If you don't provide this, FreeIPA and Data Lake backups will be stored in the Logs Location Base.	my-bucket-/my-dl

Onboarding CDP users and groups for AWS cloud storage (RAZ environments)

If your AWS environment has [Fine-grained access control](#) enabled, you should onboard your users using Ranger instead of using IDBroker.

For more information, refer to [Ranger policy options for RAZ-enabled AWS environment](#) and [Using Ranger to Provide Authorization in CDP](#).

Onboarding CDP users and groups for AWS cloud storage (no RAZ)

The minimal setup defined earlier spins up a CDP environment and Data Lake with no end user access to cloud storage. Adding users and groups to a CDP environment involves ensuring they are properly mapped to IAM roles to access cloud storage.

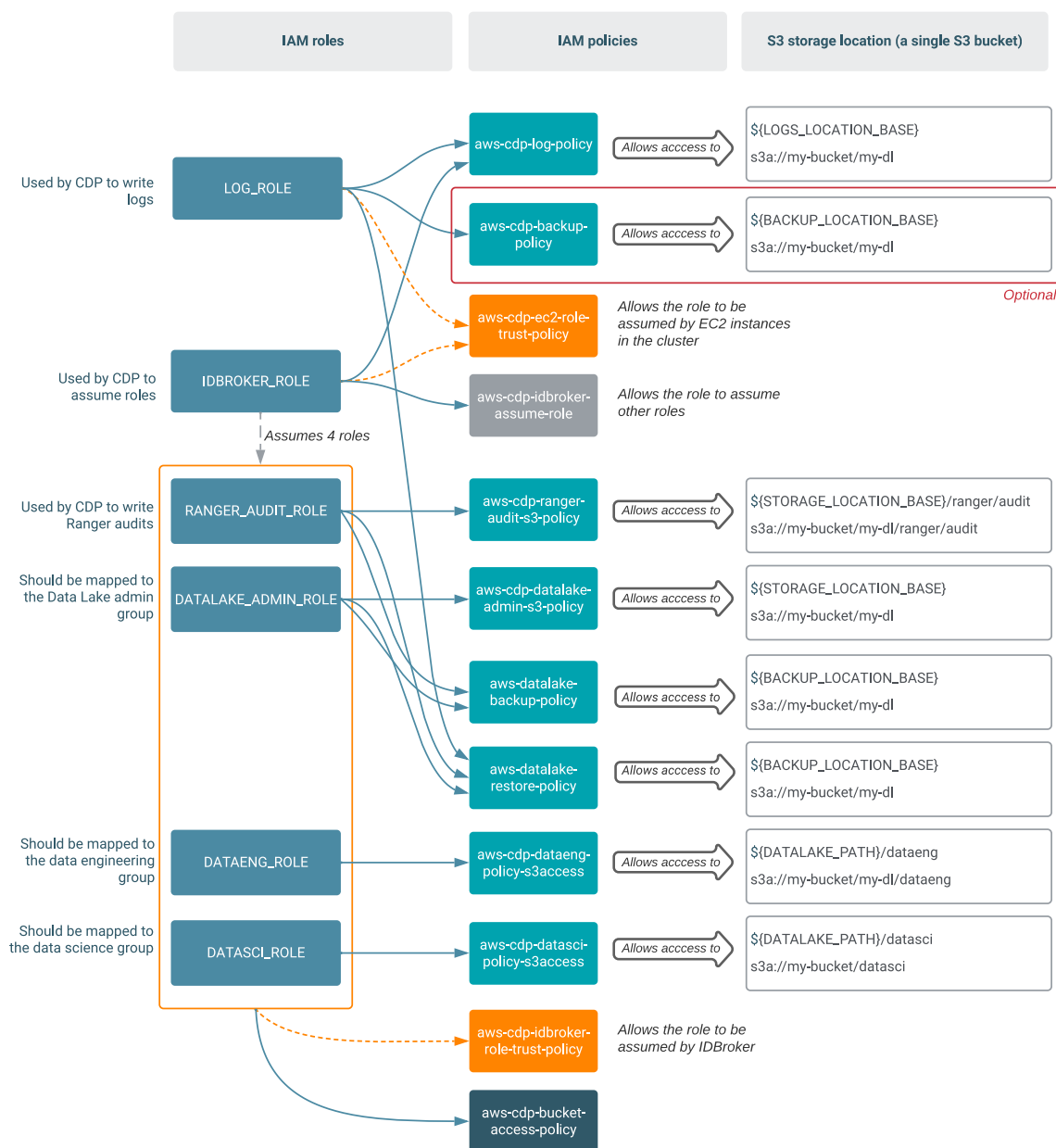


Note: If you are using [Fine-grained access control](#), you should onboard your users using Ranger instead of using IDBroker mappings. Adding IDBroker mappings is disabled for RAZ-enabled environments.

In general, to onboard a new user or group to be onboarded you should have the following IAM roles and policies pre-created in AWS:

- One IAM role for the user/group
- One IAM policy for the user/group role to access the required S3 bucket(s) and path(s)

In the example below, we are adding a data engineering group and a data science group to the CDP environment. The final goal is to have the following that builds on the minimal setup:



In the table below, bolded names are new compared to the minimal setup. The definitions for the policies mentioned in the table can be found in [IAM policy definitions](#).

Role	Permissions policy	Trust policy	Description
DATAENG_ROLE	aws-cdp-dataeng-policy-s3access aws-cdp-bucket-access-policy	aws-cdp-idbroker-role-trust-policy	This role uses the three permissions policies to provide data engineers with access to a specific S3 location (s3a://my-bucket/my-dl/dataeng). The trust policy allows the role to be assumed by IDBroker.

Role	Permissions policy	Trust policy	Description
DATASCI_ROLE	aws-cdp-datasci-policy-s3access aws-cdp-bucket-access-policy	aws-cdp-idbroker-role-trust-policy	This role uses the three permissions policies to provide data scientists with access to a specific S3 location (s3://my-bucket/my-dl/datasci). The trust policy allows the role to be assumed by IDBroker.

Use the following documentation in order to create the required S3 bucket and IAM resources, and provide them in CDP. The steps involve:

1. Creating the additional permissions policies for the minimal setup
2. Creating the required IAM roles and their associated trust policies
3. Creating the mappings in CDP

Create permissions policies for user access

You can create the IAM permissions policies required for user access to the data storage S3 location from the IAM console or AWS CLI.

Before you begin

For IAM policy definitions, refer to [IAM policy definitions](#). Make sure to replace all the placeholders in the policy with actual values.



Note: You only need to create permissions policies. Do not create trust policies in this manner. Trust policies are not created and attached to roles in the same manner as regular IAM policies. Establishing a trust relationship via trust policies is covered as part of a later step described below.

Steps

For AWS console

1. In the AWS web interface, navigate to the IAM console.
2. From the navigation pane, select Policies and then click on Create Policy
3. Choose the JSON tab.
4. Paste the appropriate JSON policy document.



Note: Make sure to replace all the placeholders in the policy with actual values.

5. Click Next.
6. Add tags if required by your organization.
7. Click Next.
8. Type a Name and an optional Description.
9. Click Create policy.

For AWS CLI

Use the following command to create permissions policies:

```
aws iam create-policy --policy-name <NAME> --policy-document <JSON-DOCUMENT>
```

Repeat these steps for all permissions policies that need to be created. For more detailed instructions on how to create IAM policies, refer to the AWS documentation linked below.

Related Information

[Creating IAM policies \(console\)](#)

AWS CLI Reference: create-policy

Create the IAM roles and establish trust relationships: DATAENG_ROLE and DATASCI_ROLE

Use the following steps to create the DATAENG_ROLE and DATASCI_ROLE. These roles allow end users to access specific subsets of S3 cloud storage that you configure in CDP.

Before you begin

Review the setup described above and make sure that you understand which policies you need to attach to which roles.

Steps

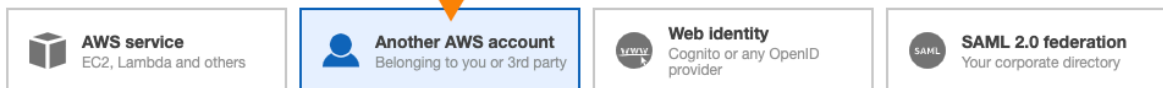
For AWS console

1. In the AWS web interface, navigate to the IAM console.
2. From the navigation pane, select Roles.
3. Click on Create Role.
4. Under "Select type of trusted entity" select Another AWS account.
5. Under "Account ID", enter your AWS account ID. You can obtain it by clicking on your user name in the top bar.

Create role

1 2 3 4

Select type of trusted entity



The screenshot shows four options for selecting a trusted entity:

- AWS service**: EC2, Lambda and others
- Another AWS account**: Belonging to you or 3rd party (Selected)
- Web identity**: Cognito or any OpenID provider
- SAML 2.0 federation**: Your corporate directory

Allows entities in other accounts to perform actions in this account. [Learn more](#)

Specify accounts that can use this role

Account ID*



- Options**
- Require external ID (Best practice when a third party will assume this role)
 - Require MFA ⓘ

6. Click Next.
7. Attach the permissions policies required for the role that you are creating. Make sure to attach all required policies.
8. Click Next.
9. Add tags if required by your organization.
10. Click Next.
11. Provide a name and an optional description for your role, and verify that you have attached all required permissions policies.
12. Click on Create role. This creates an IAM role. No instance profile is created in this case.
13. Click on the role to see its details.
14. Click on the Trust relationships tab.

15. Click on Edit trust relationship.

Roles > test-role2

Summary Delete role

Role ARN arn:aws:iam:::role/test-role2

Role description [Edit](#)

Instance Profile ARNs

Path /

Creation time 2021-09-17 15:02 PDT

Last activity Not accessed in the tracking period

Maximum session duration 1 hour [Edit](#)

Give this link to users who can switch roles in the console <https://signin.aws.amazon.com/switchrole?roleName=test-role2&account=hwxcloudbreak>

Permissions **Trust relationships** **Tags** **Access Advisor** **Revoke sessions**

You can view the trusted entities that can assume the role and the access conditions for the role. [Show policy document](#)

Edit trust relationship

Trusted entities
The following trusted entities can assume this role.

Trusted entities
The account !

Conditions
The following conditions define how and when trusted entities can assume the role.
There are no conditions associated with this role.

16. Paste the JSON declaring a trust policy in place of the default policy. Use the appropriate policy described above.

17. Click on Update trust policy.

Repeat these steps for both roles.

For AWS CLI

Use these commands to create the required IAM roles via AWS CLI:

```
aws iam create-role --role name <ROLE-NAME> --assume-role-policy-document
<TRUST-POLICY>
aws iam attach-role-policy --role name <ROLE-NAME> --policy-arn <PERMIS
SIONS-POLICY-ARN>
```

Note that the `aws iam attach-role-policy` command needs to be repeated if there are multiple policies to attach. You should assign the appropriate policies described above.

Once these resources have been created, you can provide them be editing an existing environment in CDP.

Related Information

[AWS CLI Reference: iam](#)

Adding CDP user/group to IAM role mappings

After creating the two additional IAM roles, one for data engineers and one data scientists, map them to specific user/group in CDP.



Note: If you are using [Fine-grained access control](#), you should onboard your users using Ranger instead of using IDBroker mappings. Adding IDBroker mappings is disabled for RAZ-enabled environments.



Note:

If a user is mapped to multiple roles via group membership, the specific role to be used needs to be provided at runtime. If the user is mapped directly to a role, the direct mapping takes precedence over mapping via group membership. For information on how to specify the role, refer to [Specifying a group when user belongs to multiple groups](#).

Required role: DataSteward, EnvironmentAdmin, or Owner

Steps

For CDP UI

1. The option to add/modify these mappings is available from the Management Console under Environments > click on an environment > Actions > Manage Access > IDBroker Mappings > Edit.
2. Under Current Mappings, click Edit.
3. Click + to display a new field for adding a mapping.
4. Provide the following:
 - a. The User or Group dropdown is pre-populated with CDP users and groups. Select the user or group that you would like to map.
 - b. Under Role, specify the role ARN (copied from the IAM role page on AWS). You should select your DATAENG_ROLE here.
5. Repeat the previous two steps to add additional mapping for the DATASCI_ROLE.
6. For example, in the example setup we created the following roles:
 - DATAENG_ROLE - We created this role while onboarding users and we assume that there is a DataEngineers group that was created in CDP.
 - DATASCI_ROLE - We created this role while onboarding users and we assume that there is a DataScientists group that was created in CDP.

Based on the roles and groups created in this example, the mappings that need to be created are:

IDBroker Mappings ⓘ

User or Group	<input type="text" value="DataEngineers"/>	Role	<input type="text" value="arn:aws:iam:123456789001:role/DATAENG_ROLE"/>	<input type="button" value="⊖"/>	<input type="button" value="⊕"/>
User or Group	<input type="text" value="DataScientists"/>	Role	<input type="text" value="arn:aws:iam:123456789001:role/DATASCI_ROLE"/>	<input type="button" value="⊖"/>	<input type="button" value="⊕"/>

7. Click Save and Sync.

For CDP CLI

If you would like to create the mappings via CDP CLI, you can:

1. Use the `cdp environments get-id-broker-mappings` command to obtain your current mappings.
2. Use the `cdp environments set-id-broker-mappings` command to set additional mappings. The only way to use this command is to:
 - Pass all the current mappings
 - Add the new mappings
3. Next, sync IDBroker mappings. For example:

```
cdp environments sync-id-broker-mappings --environment-name demo3
```

4. Finally, check the sync status. For example:

```
cdp environments get-id-broker-mappings-sync-status --environment-name demo3
```

IAM policy definitions for the minimal cloud storage setup

Use the following IAM policy definitions when defining IAM policies for the minimal cloud storage setup described in the parent topic.

Note that:

- The policy definitions refer to roles by using the convention presented in the table in the parent topic. If the IAM roles that you created use different names, you should update these names in the policy definitions below.
- The policy definitions refer to the example S3 subdirectories presented in the parent topic. If the S3 bucket subdirectories that you created use different names, you should update these names in the policy definitions below.

While creating these IAM policies, make sure to replace the following with actual values:

- `{ARN_PARTITION}` - Replace this with "aws". Or, if you are using one of China regions, replace it with "aws-cn". See [Amazon Resource Names \(ARNs\)](#).
- `{AWS_ACCOUNT_ID}` - Replace this with your our AWS account ID.
- `{DATALAKE_BUCKET}` - Replace this with the name of your S3 bucket. For example my-bucket.
- `{STORAGE_LOCATION_BASE}` - Replace this with the path to your Data Lake directory in the S3 bucket specified as `{DATALAKE_BUCKET}/SOME_PATH`. For example my-bucket/my-dl.
- `{LOGS_BUCKET}` - Replace this with the name of your S3 bucket for logs. For example my-bucket.
- `{LOGS_LOCATION_BASE}` - Replace this with the path to your S3 location for logs. For example my-bucket/my-dl.
- `{BACKUP_BUCKET}` - Replace this with the name of your S3 bucket for backups. For example my-bucket.
- `{BACKUP_LOCATION_BASE}` - Replace this with the path to your S3 location for backups. This location is used for both FreeIPA and Data Lake backups. For example my-bucket/my-dl.

aws-cdp-idbroker-assume-role-policy

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



Note:

The * in the aws-cdp-idbroker-assume-role-policy means that the IDBroker role will be able to assume any role that also has a trust with the IDBroker role. If preferred, the resource can be a list of role ARNs, but this policy would need to be updated each time a new role is mapped. Both a trust from the target role to the IDBroker role and an AssumeRole policy are necessary for roles to be assumed.

aws-cdp-log-policy

Refer to [aws-cdp-log-policy.json](#).

aws-cdp-backup-policy

Refer to [aws-cdp-backup-policy.json](#).



Note: The `aws-cdp-backup-policy` policy is only required if your Backup Location Base is in a separate bucket.

aws-cdp-ranger-audit-s3-policy

Refer to [aws-cdp-ranger-audit-s3-policy.json](#).

aws-cdp-datalake-admin-s3-policy

Refer to [aws-cdp-datalake-admin-s3-policy.json](#).

aws-cdp-bucket-access-policy

Refer to [aws-cdp-bucket-access-policy.json](#).



Note:

If using multiple buckets for Storage Location Base, Logs Location Base, and Backup Location Base, the `aws-cdp-bucket-access-policy` needs to have all the buckets specified.



Note:

The action `"s3:ListAllMyBuckets"` is optional and allows Hue to display a list of buckets in the S3 browser. Without `"s3:ListAllMyBuckets"`, a user must explicitly enter which S3 bucket they want to browse in Hue.

aws-datalake-backup-policy

Refer to [aws-datalake-backup-policy.json](#).

aws-datalake-restore-policy

Refer to [aws-datalake-restore-policy.json](#).

aws-cdp-ec2-role-trust-policy

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "ec2.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

aws-cdp-idbroker-role-trust-policy

```
{
  "Version": "2012-10-17",
  "Statement": [
```



```

    {
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam::${AWS_ACCOUNT_ID}:role/${IDBROKER_ROLE}"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}

```

IAM policy definitions for cloud storage location users

Use the following IAM policy definitions when defining IAM policies for cloud storage location users, such as data scientists and data engineers.

Note that:

- The policy definitions refer to roles by using the convention presented in the table in the parent topic. If the IAM roles that you created use different names, you should update these names in the policy definitions below.
- The policy definitions refer to the example S3 subdirectories presented in the parent topic. If the S3 bucket subdirectories that you created use different names, you should update these names in the policy definitions below.

While creating these IAM policies, make sure to replace the following with actual values:

- `${AWS_ACCOUNT_ID}` - Your AWS account ID.
- `${DATALAKE_PATH}` - Path to your Data Lake directory under the Storage Location Base. For example `my-bucket/my-dl`. This does not have to be under the Storage Location Base, but for simplicity this example assumes it is a subdirectory of the Storage Location Base.

aws-cdp-dataeng-policy-s3access

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor3",
      "Effect": "Allow",
      "Action": "s3:*",
      "Resource": "arn:aws:s3:::${DATALAKE_PATH}/dataeng/*"
    }
  ]
}

```

aws-cdp-datasci-policy-s3access

```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor3",
      "Effect": "Allow",
      "Action": "s3:*",
      "Resource": "arn:aws:s3:::${DATALAKE_PATH}/datasci/*"
    }
  ]
}

```

Using S3 encryption

AWS S3 encryption requires additional permissions added to the IAM roles depending if read only or read/write access is needed.

All IAM roles which need to read data encrypted with SSE-KMS must have the permissions to decrypt using the specific key the data was encrypted with. All IAM roles which need to both read and write data need the encrypt and decrypt permissions (encrypt-only permission is not supported).

IAM role permissions for working with SSE-KMS

Depending on the type of access required, one of the two policies needs to be attached to the roles that access S3. From the examples above, these would be the following roles:

- LOG_ROLE
- RANGER_AUDIT_ROLE
- DATALAKE_ADMIN_ROLE
- DATAENG_ROLE
- DATASCI_ROLE

Use the json below to create the necessary policies and attach them to the correct IAM roles for your environment.

aws-cdp-sse-kms-read-only-policy

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor0",
      "Effect": "Allow",
      "Action": "kms:Decrypt",
      "Resource": [
        "${KEY_ARN}"
      ]
    }
  ]
}
```

aws-cdp-sse-kms-read-write-policy

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor0",
      "Effect": "Allow",
      "Action": [
        "kms:Decrypt",
        "kms:GenerateDataKey"
      ],
      "Resource": [
        "${KEY_ARN}"
      ]
    }
  ]
}
```

Related Information

[Encrypting data on S3](#)

Using S3 Express One Zone for data storage

You can use S3 Express One Zone (S3 Express) with CDP.

If you have additional data buckets that you would like to use with CDP and you do not need zone redundancy, you may use S3 Express buckets, for example for faster processing of temporary data.

The following limitations apply when using S3 Express buckets:

- You can only use S3 Express buckets with Data Hubs running Runtime 7.2.18 or newer. Data services do not currently support it.
- S3 Express buckets may not be used for logs and backups.

If you would like to use an S3 Express bucket for running Data Hub workloads, you should:

1. Add the required permissions in AWS
2. Set bucket region in core-site.xml

Permissions

In order to use an S3Express bucket, you should add the following permissions to the DATALAKE_ADMIN_ROLE role described in the [Minimal setup for AWS cloud storage](#):

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "VisualEditor2",
      "Effect": "Allow",
      "Action": "s3express:CreateSession",
      "Resource": "arn:aws:s3express:region:account-id:bucket/base-bucket-name
--azid--x-s3"
    }
  ]
}
```

Set bucket region in core-site.xml

Set the AWS region of the S3Express bucket in the core-site.xml as follows:

```
<property>
  <name>fs.s3a.bucket.<bucket-name>.endpoint.region</name>
  <value>us-west-2</value>
</property>
```

The following screenshot illustrates how to perform this configuration in Cloudera Manager:

The screenshot shows the Cloudera Manager interface for configuring the hdfs service. The search bar at the top contains the text "core-s". The main content area displays a configuration snippet for the "fs.s3a.bucket.dsp-storage-aws-dev-newwpc.endpoint" property. The value is set to "s3.us-west-2.amazonaws.com". The snippet is titled "Cluster-wide Advanced Configuration Snippet (Safety Valve) for core-site.xml" and is associated with the "hdfs (Service-Wide)" scope. The snippet is marked as "Final" and "Requires additional authorization".

Customer managed encryption keys

By default, Data Lake and FreeIPA's Amazon Elastic Block Store (EBS) volumes and Relational Database Service (RDS) are encrypted using a default key from Amazon's KMS, but you can optionally configure encryption using Customer Managed Keys (CMK). Data Hubs inherit environment's encryption key by default but you have an option to specify a different CMK during Data Hub creation.

Amazon offers the option to encrypt EBS volumes and RDS instances using a default key from Amazon's Key Management System (KMS) or using an external customer-managed KMS. By default, Data Lake and FreeIPA are encrypted using the default key from Amazon's KMS present in the region where the environment is running, but you can provide a customer-managed KMS key instead of the default key. Encryption is configured for block devices and root devices. When encryption is configured for a given cluster, it is automatically applied to all the disk devices of any new VM instances added as a result of cluster scaling or repair.

This document includes the following content:

1. [Environment and Data Hub encryption options](#) on page 44
2. [Permissions for using encryption](#) on page 44
3. [Ensuring that an existing encryption key can be used](#) on page 45
4. [Creating a new encryption key on AWS](#) on page 45

Environment and Data Hub encryption options

By default, Data Hubs use the same default key from Amazon's KMS or CMK as the parent environment but you have an option to pass a different CMK during Data Hub creation.

All possible scenarios are summarized in the following table:

Encryption key during environment registration	Encryption key during Data Hub creation	Result
Absent	Absent	<ul style="list-style-type: none"> EBS and RDS encryption for Data Lake, FreeIPA, and Data Hubs is with the default regional encryption key.
Present	Absent	<ul style="list-style-type: none"> EBS and RDS encryption for Data Lake, FreeIPA, and all Data Hubs is with the CMK provided during environment registration.
Present	Present	<ul style="list-style-type: none"> EBS and RDS encryption for Data Lake and FreeIPA is with the CMK provided during environment registration. If a CMK is provided for a Data Hub, then EBS encryption for the Data Hub is with the CMK provided per host group during the Data Hub creation.
Absent	Present	<ul style="list-style-type: none"> No EBS and RDS encryption is configured for Data Lake and FreeIPA. EBS encryption for the specific Data Hub is with the default or CMK provided per host group during Data Hub creation.

Permissions for using encryption

If you are planning to use encryption, ensure that the [cross-account IAM role](#) used for the provisioning credential includes the following permissions:

EC2 permissions

```
{
```

```

"Version": "2012-10-17",
"Statement": {
  "Effect": "Allow",
  "Action": [
    "ec2:CopyImage",
    "ec2:CreateSnapshot",
    "ec2>DeleteSnapshot",
    "ec2:DescribeSnapshots",
    "ec2:CreateVolume",
    "ec2>DeleteVolume",
    "ec2:DescribeVolumes",
    "ec2:DeregisterImage",
  ],
  "Resource": "*"
}
}

```

KMS permissions

```

{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": [
      "kms:DescribeKey",
      "kms:ListKeys",
      "kms:ListAliases"
    ],
    "Resource": "*"
  }
}

```

If planning to use encryption, ensure that your existing encryption key can be used or create a new encryption key.

Ensuring that an existing encryption key can be used

If you already have an existing encryption key, make sure that the key fulfills the following requirements.

If you have an existing encryption key that you would like to use with Data Hub, make sure that:

- The following are attached as key user:
 - The AWSServiceRoleForAutoScaling built-in role.
 - Your IAM role or IAM user used for the cloud credential.
- To check that these are attached, in the AWS Management Console, navigate to the KMS console > Customer managed keys, select your encryption key, and scroll to Key Users.
- The encryption key is located in the same region where you would like to create clusters with encrypted volumes.

Creating a new encryption key on AWS

If you don't have an existing encryption key, use the following instructions to create one.

1. In the AWS Management Console, navigate to KMS console.
2. Select Customer managed keys.
3. From the Region dropdown, select the region in which you would like to create and use the encryption key.
4. Click Create key.
5. In Step 1: Configure Key:
 - a. Under Key type, choose Symmetric.
 - b. Expand Advanced Options and under Key Material Origin, select "KMS" or "External".

6. In Step 2: Create Alias and Description:
 - a. Enter an alias for your key.
 - b. Defining tags is optional.
7. In Step 3: Define Key Administrative Permissions, select the following:
 - a. Choose your own IAM user/role used for logging into the AWS Management Console. Do not set `AWSServiceRoleForAutoScaling` or the cross-account IAM role as the key admin.
8. In Step 4: Define Key Usage Permissions:
 - a. Select the `AWSServiceRoleForAutoScaling` built-in role.
 - b. Select the cross-account IAM role.
9. In Step 5: Review and edit key policy, you may optionally tweak the key policy as desired, or simply leave it as generated by AWS.
10. Navigate to the last page of the wizard and then click Finish to create an encryption key.

Related Information

[Adding a customer managed encryption key to a CDP environment running on AWS](#)

[Configuring encryption for Data Hub's EBS volumes on AWS](#)

AWS limits

When you create your AWS account, AWS sets limits to the resources available to you. The limits can vary by region.

For example:

- Depending on your AWS account limits, you might only be allowed to provision a certain number of CPU instances, or you might not have default access to GPU instances at all.
- If you are planning to use Endpoint Access Gateway, CDP will create two AWS network load balancers (AWS NLB) per cluster (that is for each Data Lake and Data Hub). Make sure that your AWS NLB limits allow for the load balancer creation.
- If you are planning to use the Machine Learning service, review [Limitations on AWS](#) in ML docs.

Make sure to review your AWS service limits before you proceed.

To view the limits set by Amazon for your account, log in to AWS and go to `EC2 > Limits`. The `EC2 Service Limits` page lists the limits for the resources available to you in your EC2 instance, including limits to the number of instances and hosts. The `Networking Limits` section on the page lists the VPC, subnet, and security group limits for your AWS account. For example, there is typically a limit of 5 VPCs per region.

To review what resources CDP creates on your AWS account during environment registration and later when you deploy specific CDP services, refer to [AWS resources used by CDP](#).

If you require more resources than the limit set by Amazon, you can request Amazon to raise the limit of a resource. On the `EC2 Service Limits` page, click `Request limit increase` for the resource that you want to increase and create an AWS support case for a `Service Limit Increase`.

For more information about AWS limits, refer to [AWS Service Limits](#) in AWS documentation.



Overview of AWS resources used by CDP

The following AWS resources are used by CDP and CDP services.


AWS resources created for a CDP environment


When a CDP environment is created, a FreeIPA cluster and a Data Lake cluster are created.

The following AWS resources are created for FreeIPA (one per environment):

Resource	Description
Virtual Private Cloud (VPC)	If during environment creation you select to have a new VPC and subnets created, then the new VPC and subnets are created on your AWS account. Alternatively, you can provide your own existing VPC and subnets. In both cases, all the resources that CDP provisions for the environment are provisioned into this specific VPC. For example, the EC2 instances provisioned for Data Hub or Data Warehouse are provisioned into that VPC.
Identity and Access Management (IAM)	The cross-account IAM policy that you provided as your credential allows CDP to obtain an access and secret key from AWS, allowing CDP to create resources for your environment and for CDP services such as Data Hub, Data Warehouse, and Machine Learning on your AWS account.
CloudFormation	<p>During environment creation, CloudFormation stack is provisioned for FreeIPA to create required resources. This generates an AWS stack which links and describes the resources of your FreeIPA server. Multi-AZ deployments do not use a CloudFormation template for VM creation. Neither autoscaling groups or launch templates are created. The cluster resources are managed individually using AWS native components (for example, EC2 instances).</p> <p> Caution: Do not attempt to change the configuration of the Auto Scaling groups or the launch template. Any manual modifications (such as modifying suspended processes or subnets) that you make to the Auto Scaling groups or the launch template are not supported and may cause future errors or failures, particularly during repair processes.</p>
Auto Scaling	<p>FreeIPA uses the Auto Scaling service for upscaling FreeIPA, except in a multi-AZ deployment.</p> <p> Caution: After FreeIPA instances have been created, do not attempt to change their configuration of the Auto Scaling groups or the launch template. Any manual modifications (such as modifying suspended processes or subnets) that you make to the Auto Scaling groups or the launch template are not supported and may cause future errors or failures, particularly during repair processes.</p>
Elastic Compute Cloud (EC2)	During environment creation, two or three m5.large EC2 instances are provisioned for the FreeIPA HA server by default. The number of instances depends on the selected Data Lake type. Furthermore, security groups with the rules specified during environment creation are provisioned to define inbound and outbound access to the instances.


In addition, the following resources are created for each Data Lake (one per environment):


Resource	Description
CloudFormation	<p>A CloudFormation stack is provisioned for your Data Lake to create instances, disks, and RDS required. This generates an AWS stack which links and describes the resources of your Data Lake cluster. Multi-AZ deployments do not use a CloudFormation template for VM creation. Neither autoscaling groups nor launch templates are created. In a multi-AZ setup the cluster resources are managed individually using AWS native components (for example, EC2 instances).</p> <p> Caution: Do not attempt to change the configuration of the Auto Scaling groups or the launch template. Any manual modifications (such as modifying suspended processes or subnets) that you make to the Auto Scaling groups or the launch template are not supported and may cause future errors or failures, particularly during repair processes.</p>

Resource	Description
Auto Scaling	<p>The Data Lake uses the Auto Scaling service for upscaling clusters, except in a multi-AZ deployment.</p> <p> Caution: After a Data Lake has been created, do not attempt to change the configuration of the Auto Scaling groups or the launch template. Any manual modifications (such as modifying suspended processes or subnets) that you make to the Auto Scaling groups or the launch template are not supported and may cause future errors or failures, particularly during repair processes.</p>
Elastic Compute Cloud (EC2)	<p>EC2 instances with attached storage are provisioned for the Data Lake nodes:</p> <ul style="list-style-type: none"> • Light duty: Two instances are provisioned: One t3.medium instance (IDBroker) and one m5.2xlarge instance (Data Lake Master node). • Medium duty: Ten instances are provisioned: Two t3.medium instances (IDBroker), three m5.xlarge instances (two Data Lake Master nodes and one Auxiliary node), and five m5.2xlarge instances (three DataLake Core nodes and two Gateway nodes). <p>Furthermore, security groups with the rules specified during environment creation are provisioned to define inbound and outbound access to the instances.</p>
Relational Database Service (RDS)	<p>An RDS instance (db.m5.large) is provisioned for the Data Lake. This RDS instance is used for Cloudera Manager, Ranger, and Hive MetaStore.</p>
Simple Storage Service (S3)	<p>The existing S3 that you provide during environment creation for the Data Lake is used for Data Lake log storage and workload data storage.</p>

AWS resources created for Data Hub

The following AWS resources are created for the Data Hub service:

Resource	Description
CloudFormation	<p>A CloudFormation stack is created for each Data Hub cluster to create instances and disks. This generates an AWS stack which links and describes the resources of your Data Hub cluster. Multi-AZ deployments do not use a CloudFormation template for VM creation. Neither autoscaling groups or launch templates are created. In a multi-AZ setup the cluster resources are managed individually using AWS native components (for example, EC2 instances).</p> <p> Caution: Do not attempt to change the configuration of the Auto Scaling groups or the launch template. Any manual modifications (such as modifying suspended processes or subnets) that you make to the Auto Scaling groups or the launch template are not supported and may cause future errors or failures, particularly during repair processes.</p>
Elastic Compute Cloud (EC2)	<p>An EC2 instance is created for each cluster node. The instance type varies depending on what you selected during Data Hub cluster creation. For each instance, attached storage is provisioned. The storage size and type varies depending on what you selected during cluster creation. Furthermore, security groups with the rules specified during environment creation are provisioned to define inbound and outbound access to the instances. For a list of supported EC2 instance types, refer to Cloudera Data Platform (CDP) Public Cloud service rates.</p>
Relational Database Service (RDS)	<p>Data Hub connects to the Hive MetaStore database on the RDS instance provisioned for the Data Lake.</p>
Simple Storage Service (S3)	<p>The existing S3 bucket that you provided for the Data Lake to use for workload data storage can be accessed from Data Hub clusters via the S3A connector.</p>

Resource	Description
Auto Scaling	Data Hub uses the Auto Scaling service for upscaling clusters, except in a multi-AZ deployment.  Caution: After a Data Hub has been created, do not attempt to change the configuration of the Auto Scaling groups or the launch template. Any manual modifications (such as modifying suspended processes or subnets) that you make to the Auto Scaling groups or the launch template are not supported and may cause future errors or failures, particularly during repair processes.
Key Management Service (KMS)	Data Hub uses KMS for encrypting your disks if during Data Hub cluster creation you select to use disk encryption.

AWS resources created for Data Engineering

The following AWS resources are created for the Cloudera Data Engineering (CDE) service:

Resource	Description
CloudFormation	The initial deployment of services such as the EKS cluster is orchestrated through CloudFormation. This generates an AWS stack which links and describes the resources of your CDE cluster.
Elastic Compute Cloud (EC2)	CDE uses EC2 instances as cluster nodes. For a list of supported EC2 instance types, refer to Cloudera Data Platform (CDP) Public Cloud service rates .
Auto Scaling	CDE uses AWS AutoScaler to add or remove EC2 instances to the kubernetes cluster. Whenever the kubernetes cluster is running low on resources, new EC2 instances are provisioned and jointed into the EKS cluster. Whenever the AutoScaler detects an over-provisioning of resources, it removes and suspends EC2 instances.
Elastic Kubernetes Service (EKS)	EKS is the AWS implementation of the kubernetes stack. All PODs are running within an EKS cluster (one per environment).
ELB Classic Load Balancer	CDE uses Classic Load Balancers for redirecting traffic to EC2 instances.
Key Management Service (KMS)	CDE uses KMS for encrypting your disks if you select to use disk encryption.
Elastic Block Store (EBS)	CDE uses EBS for persistent instance storage.
Elastic File System (EFS)	CDE uses EFS for persistent service and virtual cluster storage.
Relational Database Service (RDS)	CDE uses RDS for provisioning relational databases.

AWS resources created for DataFlow

The following AWS resources are created for the DataFlow (DF) service:

Resource	Description
CloudFormation	The initial deployment of services such as the EKS cluster is orchestrated through CloudFormation. This generates an AWS stack which links and describes the resources of your DataFlow cluster.
Elastic Compute Cloud (EC2)	DataFlow uses EC2 instances as cluster nodes. For a list of supported EC2 instance types, refer to Cloudera Data Platform (CDP) Public Cloud service rates .
Auto Scaling	DataFlow uses AWS AutoScaler to add or remove EC2 instances to the kubernetes cluster. Whenever the kubernetes cluster is running low on resources, new EC2 instances are provisioned and jointed into the EKS cluster. Whenever the AutoScaler detects an over-provisioning of resources, it removes and suspends EC2 instances.

Resource	Description
Elastic Kubernetes Service (EKS)	EKS is the AWS implementation of the kubernetes stack. All PODs are running within an EKS cluster (one per environment).
ELB Classic Load Balancer	DataFlow uses Classic Load Balancers for redirecting traffic to EC2 instances.
Elastic Block Store (EBS)	DataFlow uses EBS for persistent instance storage.
Relational Database Service (RDS)	DataFlow uses RDS for provisioning relational databases.

AWS resources created for Data Warehouse

The following AWS resources are created for the Cloudera Data Warehouse (CDW) service:

Resource	Description
Identity and Access Management (IAM)	During CDW cluster provisioning, the DW service creates an IAM role that defines access to S3 and other provisioned resources. Such role is then attached to the EC2 instance profile to grant PODs within the kubernetes environment access to these resources.
Certificate Manager	CDW creates, stores, and maintains a certificate in the AWS certificate manager. This certificate is used to allow HTTPS connections to the external facing endpoints (i.e. for JDBC or the DAS UI). The certificate is signed by a trusted certificate authority, therefore external consumers and browser can securely connect to DW services without having to deal with untrusted CA or self-signed certificates.
CloudFormation	The initial deployment of services such as the EKS cluster, the CDW-specific RDS database, and S3 buckets is orchestrated through CloudFormation. This generates an AWS stack which links and describes the resources of your DW cluster.
Elastic Compute Cloud (EC2)	CDW uses EC2 instances as cluster nodes. Two different EC2 instance types (through two different auto scaler groups) are used to support shared services and compute requirements within the cluster: m5.2xlarge for always on components, and r5d.4xlarge for compute nodes (Hive and Impala executors). Furthermore, security groups with the rules specified during environment creation are provisioned to define inbound and outbound access to the instances. For a list of supported EC2 instance types, refer to Cloudera Data Platform (CDP) Public Cloud service rates .
Simple Storage Service (S3)	CDW creates its own S3 buckets (separate from the environment's S3 bucket(s)) for storing data and logs.
Auto Scaling	CDW uses AWS AutoScaler to add or remove EC2 instances to the kubernetes cluster. Whenever the kubernetes cluster is running low on resources, new EC2 instances are provisioned and jointed into the EKS cluster. Whenever the AutoScaler detects an over-provisioning of resources, it removes and suspends EC2 instances.
Elastic File System (EFS)	EFS is used as shared filesystem across PODs to persist data (i.e. result cache).
Elastic Load Balancing (ELB)	All inbound traffic is routed through ELB towards the ingress controller of the kubernetes cluster. The ELB is provisioned as a result of the kubernetes ingress controller, which is the single point of entry for services, running in the kubernetes cluster.
Managed Kubernetes Service (EKS)	EKS is the AWS implementation of the kubernetes stack. All DW-deployed PODs are running within an EKS cluster (one per environment).
Key Management Service (KMS)	CDW encrypts data at rest in S3. This requires an encryption key to be generated and stored in KMS. The key is completely under the control of AWS and cannot be exported or otherwise extracted. The S3 buckets are directly referencing the key within KMS, using it to encrypt the stored data.

Resource	Description
Relational Database Service (RDS)	During cluster provisioning, DW provisions an RDS instance to be used as backend database system for metadata, managed and stored by the HMS instances, represented by “DB Catalogs”. Each DB Catalog is implemented as separate database within this single RDS instance.
Security Token Service (STS)	STS is used to generate access tokens (based on roles) to access the resources within the environment's VPC.

AWS resources created for Machine Learning

The following AWS resources are created for the Cloudera Machine Learning (CML) service:

Resource	Description
Identity and Access Management (IAM)	CML creates additional IAM roles and policies for each cluster. Such roles are then attached to the EC2 instance profile.
Elastic Block Store (EBS)	CML uses EBS as block storage.
Elastic Load Balancer (ELB)	CML uses Classic Load Balancers for redirecting traffic to EC2 instances.
Key Management Service (KMS)	CML uses KMS for encrypting your disks if you select to use disk encryption.
Elastic File System (EFS)	EFS is used for project file storage.
Elastic Compute Cloud (EC2)	CML uses EC2 instances as cluster nodes. Three different EC2 instance types (through three different auto scaler groups) are used to support CML infra and compute requirements within the kubernetes cluster. Furthermore, security groups with the rules specified during environment creation are provisioned to define inbound and outbound access to the instances. For a list of supported EC2 instance types, refer to Cloudera Data Platform (CDP) Public Cloud service rates .
Auto Scaling	CML uses AWS AutoScaler to add or remove EC2 instances to the kubernetes cluster. Whenever the kubernetes cluster is running low on resources, new EC2 instances are provisioned and joined into the cluster. Whenever the AutoScaler detects an over-provisioning of resources, it removes and suspends EC2 instances.
Simple Storage Service (S3)	CML uses S3 as the primary store for data and logs.
Security Token Service (STS)	STS is used to generate access tokens (based on roles) to access the resources within the environment's VPC.
Managed Kubernetes Service (EKS)	EKS is the AWS implementation of the kubernetes stack. All PODs are running within an EKS cluster (one per environment).

AWS resources created for Operational Database

The following AWS resources are created for the Cloudera Operational Database (COD) service:

Resource	Description
CloudFormation	A CloudFormation stack is created for each COD database to create instances and disks. This generates an AWS stack which links and describes the resources of your COD database.
Elastic Compute Cloud (EC2)	An EC2 instance is created for each node. The instance type, storage size, and storage type is determined automatically by COD. Furthermore, security groups with the rules specified during environment creation are provisioned to define inbound and outbound access to the instances.
Simple Storage Service (S3)	The existing S3 bucket that you provided for the Data Lake to use for workload data storage can be accessed from COD database via the S3A connector.

Resource	Description
Relational Database Service (RDS)	An RDS instance is provisioned for the COD. This RDS instance is used by Cloudera Manager.

AWS outbound network access destinations

If you have limited outbound internet access (for example due to using a firewall or proxy), review this content to learn which specific outbound destinations must be available in order to register a CDP environment.



Note:

If the cloud provider network that you would like to use for registering a CDP environment uses a custom DNS server that does not allow name resolution for public domain, you should add all the domains listed in the below tables to the DNS forwarder for name resolution.

The following list includes general destinations as well as AWS-specific destinations.

General endpoints

Description/ Usage	CDP service	Destination	Protocol and Authentication	IP Protocol/Port	Comments
Cloudera CCMv1 Persistent Control Plane connection	All services	*.ccm.cdp.cloudera.com 44.234.52.96/27	SSH public/private key authentication	TCP/6000-6049	One connection per cluster configured; persistent
Cloudera CCMv2 Persistent Control Plane connection	All services	US-based Control Plane: *.v2.us-west-1.ccm.cdp.cloudera.com 35.80.24.128/27 35.166.86.177/32 52.36.110.208/32 52.40.165.49/32 EU-based Control Plane: *.v2.ccm.eu-1.cdp.cloudera.com 3.65.246.128/27 AP-based Control Plane: *.v2.ccm.ap-1.cdp.cloudera.com 3.26.127.64/27	HTTPS with mutual authentication	TCP/443	Multiple long-lived/persistent connections

Description/ Usage	CDP service	Destination	Protocol and Authentication	IP Protocol/Port	Comments
Cloudera Databus Telemetry, billing and metering data	All services	US-based Control Plane: dbusapi.us-west-1.sigma.altus.cloudera.com https://cloudera-dbus-prod.s3.amazonaws.com EU-based Control Plane: api.eu-1.cdp.cloudera.com https://mow-prod-eu-central-1-sigmadb-dbus.s3.eu-central-1.amazonaws.com https://mow-prod-eu-central-1-sigmadb-dbus.s3.amazonaws.com AP-based Control Plane: api.ap-1.cdp.cloudera.com https://mow-prod-ap-southeast-2-sigmadb-dbus.s3.ap-southeast-2.amazonaws.com https://mow-prod-ap-southeast-2-sigmadb-dbus.s3.amazonaws.com	HTTPS with Cloudera-generated access key for dbus HTTPS for S3	TCP/443	Regular interval for telemetry, billing, metering services, and used for Cloudera Observability if enabled. Larger payloads are sent to a Cloudera managed S3 bucket.
Cloudera Manager parcels Software distribution	All services	archive.cloudera.com	HTTPS	TCP/443	Cloudera's public software repository. CDN backed service; IP range not predictable.
Control Plane API	All services	US-based Control Plane: api.us-west-1.cdp.cloudera.com EU-based Control Plane: api.eu-1.cdp.cloudera.com AP-based Control Plane: api.ap-1.cdp.cloudera.com	HTTPS with Cloudera-generated access key	TCP/443	Cloudera's control plane REST API.
RPMs Cloudera RPMs for workload agents	All services	cloudera-service-delivery-cache.s3.amazonaws.com	HTTPS	TPC/443	RPM packages for some workload components
Docker Images Software Distribution	Data Engineering DataFlow Machine Learning	container.repository.cloudera.com docker.repository.cloudera.com	HTTPS	TCP/443	Cloudera's public docker registry. CDN backed service; IP range not predictable.

Description/ Usage	CDP service	Destination	Protocol and Authentication	IP Protocol/Port	Comments
Docker Images Software Distribution	Data Engineering DataFlow Data Warehouse Machine Learning	container.repo.cloudera.com US-based Control Plane: prod-us-west-2-starport-layer- bucket.s3.us-west-2.amazonaws.com prod-us-west-2-starport-layer- bucket.s3.amazonaws.com s3-r-w.us-west-2.amazonaws.com *.execute-api.us- west-2.amazonaws.com EU-based Control Plane: prod-eu-west-1-starport-layer- bucket.s3.eu-west-1.amazonaws.com prod-eu-west-1-starport-layer- bucket.s3.amazonaws.com s3-r-w.eu-west-1.amazonaws.com *.execute-api.eu- west-1.amazonaws.com AP-based Control Plane: prod-ap-southeast-1- starport-layer-bucket.s3.ap- southeast-1.amazonaws.com prod-ap-southeast-1-starport-layer- bucket.s3.amazonaws.com s3-r-w.ap-southeast-1.amazonaws.com *.execute-api.ap- southeast-1.amazonaws.com	HTTPS	TCP/443	Moved to container.repo.cloudera.com container.repo.cloudera.com uses ECR which requires S3 URLs.
Docker Images Software Distribution	Data Warehouse	auth.docker.io* cloudera-docker-dev.jfrog.io* docker-images- prod.s3.amazonaws.com* gcr.io* k8s.gcr.io* quay-registry.s3.amazonaws.com* quay.io* quayio-production- s3.s3.amazonaws.com* docker.io* production.cloudflare.docker.com* storage.googleapis.com*	HTTPS	TCP/443	These endpoints are required only for old/existing Data Warehouse environments.
Flow definitions CDP AWS bucket with flow definitions	DataFlow	US-based Control Plane: *.s3.us-west-1.amazonaws.com EU-based Control Plane: *.s3.eu-central-1.amazonaws.com AP-based Control Plane: *.s3.ap-southeast-2.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Outbound internet access to S3 hosts is necessary on all cloud providers when using CDF as the workload needs to query outbound to an S3 location to retrieve the flow definition when creating a deployment.

Description/ Usage	CDP service	Destination	Protocol and Authentication	IP Protocol/Port	Comments
Public Signing Key Retrieval	Data Engineering DataFlow	US-based Control Plane: consoleauth.altus.cloudera.com console.us-west-1.cdp.cloudera.com EU-based Control Plane: console.eu-1.cdp.cloudera.com AP-based Control Plane: console.ap-1.cdp.cloudera.com	HTTPS	TCP/443	Required to allow authentication to CDE virtual Cluster using a CDP Access Key.
SQL Stream Builder PostgreSQL driver install	Data Hub: Streaming Analytics clusters	pypi.org	HTTPS	TCP/443	SQL Stream Builder depends on the python3 PostgreSQL driver. This is only required for Runtime versions 7.2.11, 7.2.12 and 7.2.13.
Control plane IAM API	Machine learning	US-based Control Plane: iamapi.us-west-1.altus.cloudera.com EU-based Control Plane: console.eu-1.cdp.cloudera.com AP-based Control Plane: console.ap-1.cdp.cloudera.com	HTTPS	TCP/443	For connecting to the IAMAPI for fetching the entitlement details.
AMPs Applied ML Prototypes	Machine Learning	https://raw.githubusercontent.com https://github.com	HTTPS	TCP/443	Files for AMPs are hosted on GitHub.
Learning Hub	Machine Learning	https://github.com/cloudera/learning-hub-content	HTTPS	TCP/443	Access Learning Hub in air-gapped environments

AWS-specific endpoints

Description/ Usage	CDP service	Destination	Protocol and Authentication	IP Protocol/Port	Comments
AWS STS	All services	sts.amazonaws.com sts.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	CDP 7.1.1+ required before can be made internal with VPC endpoints.
AWS S3	All services	*.s3.amazonaws.com *.s3.*.amazonaws.com s3.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Can be made internal with VPC endpoints.
AWS RDS	All services	*.*.rds.amazonaws.com	JDBC / Postgres binary protocol / MySQL / RDS CA certs	TCP 5432 / 3306 / 443	VPC Internal. Only Data Engineering uses MySQL and requires port 3306 to be open.

Description/ Usage	CDP service	Destination	Protocol and Authentication	IP Protocol/Port	Comments
AWS EC	DataFlow Data Warehouse Machine Learning	api.ecr.*.amazonaws.com *.dkr.ecr.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Can be made internal with VPC endpoints.
AWS EC2	DataFlow Data Warehouse Machine Learning Operational Database	ec2.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Can be made internal with VPC endpoints.
AWS EKS	Data Engineering DataFlow Data Warehouse Machine Learning	eks.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	AWS does not support EKS VPC endpoints at this time.
AWS Cloudformation	DataFlow Data Warehouse Machine Learning	cloudformation.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Can be made internal with VPC endpoints.
AWS Autoscaling	Data Engineering Data Warehouse Machine Learning	autoscaling.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Can be made internal with VPC endpoints.
AWS EFS	Data Engineering Data Warehouse Machine Learning	elasticfilesystem.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Can be made internal with VPC endpoints.
AWS ELB	Data Engineering DataFlow Data Warehouse	elasticloadbalancing.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Can be made internal with VPC endpoints.
AWS EKS k8s cluster API	Data Warehouse	UNIQUEID.*.eks.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	Optional for new clusters.
AWS RDS API	Data Warehouse	rds.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	AWS does not support RDS API VPC endpoints at this time. This requirement is under further evaluation. Data Warehouse uses Amazon RDS for PostgreSQL.
AWS Service Quotas	Data Warehouse	servicequotas.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	AWS does not support Service Quota via VPC endpoints. Used to check limits and warn prior to hitting the limits.

Description/ Usage	CDP service	Destination	Protocol and Authentication	IP Protocol/Port	Comments
AWS Price List Service	Data Warehouse	pricing.*.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	AWS Price List Service uses us-east-1 or ap-south-1 as the region.
Flow definitions storage	DataFlow	US-based Control Plane: s3.us-west-2.amazonaws.com/dfx-flow-artifacts.mow-prod.mow-prod.cloudera.com EU-based Control Plane: cldr-mow-prod-eu-central-1-dfx-flow-artifacts.s3.eu-central-1.amazonaws.com AP-based Control Plane: cldr-mow-prod-ap-southeast-2-dfx-flow-artifacts.s3.ap-southeast-2.amazonaws.com	HTTPS (one way) IAM authentication	TCP/443	

Access to workload UIs

If you have restricted DNS or networking setup, make sure that *.cloudera.site is resolvable from your network so that members of your organization can access workload UIs.

CDP workloads (including Data Lake) use subdomains under cloudera.site to host various UI endpoints (Cloudera Manager, Ranger, Knox, Hue and so on). CDP automatically provisions these endpoints whenever a Data Lake, Data Hub or another type of workload (for example, Virtual Warehouse in CDW) is created, and routing is set up so that you can access these endpoints from your network.

The subdomains are assigned under cloudera.site using the following convention:

```
<endpoint-name>.<env-truncated-name>.<customer-workload-subdomain>.<regional-subdomain>.cloudera.site
```

Supported browsers

Cloudera validates and tests against the latest version and supports recent versions of the following browsers:

- Google Chrome
- Mozilla Firefox



Note: Mozilla Firefox is not supported by Data Engineering.

- Safari
- Microsoft Edge

Other resources

While this document attempts to provide a complete overview of cloud provider requirements, there is additional documentation that you should review if planning to deploy CDP data services.

The following table includes links to documentation that you should review if planning to deploy the respective CDP data service:

CDP data service	Documentation links
Data Engineering	CDE AWS prerequisites CDE cost management CDE performance management
Data Warehouse	Virtual Warehouse sizing requirements for public cloud environments Virtual Warehouse IP address and cloud resource requirements for public cloud environments Managing costs in the public cloud environments for Cloudera Data Warehouse AWS environment requirements checklist
DataFlow	AWS requirements for DataFlow DataFlow AWS networking DataFlow AWS limits
Machine Learning	CML limitations on AWS CML restricted policies for AWS Use a non-transparent proxy with Cloudera Machine Learning on AWS environments
Operational Database	COD cloud checklist AWS requirements for COD deployment

CDP CIDR

CDP CIDR includes the following IP ranges:

Control Plane Region	IP Ranges
us-west-1	35.80.24.128/27, 35.166.86.177/32, 52.36.110.208/32, 52.40.165.49/32
eu-1	3.65.246.128/27
ap-1	3.26.127.64/27

When creating your own security groups for CDP, you must open required ports to all of these IP ranges.