

AWS environments

Date published: 2024-01-01

Date modified: 2024-12-05

CLOUDERA

Legal Notice

© Cloudera Inc. 2025. 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 environments overview.....	6
AWS environment requirements checklist.....	7
Activating an AWS environment from Cloudera Data Warehouse.....	10
Load balancers for AWS environments.....	11
AWS restricted policy.....	12
Attaching a managed policy ARN.....	14
Activating your environment in reduced permissions mode.....	16
Minimum set of IAM permissions required for reduced permissions mode.....	19
Reduced permissions mode JSON IAM permissions policy template.....	21
Required tags for CloudFormation stacks created with reduced permissions mode.....	21
Setting up cloud resources for reduced permissions mode.....	22
Deactivating AWS environments created with reduced permissions mode.....	24
Retaining PostgreSQL backups in AWS environments.....	25
Viewing and editing AWS environment details.....	25
Deactivating an AWS environment.....	28
Cloud storage buckets.....	28
Bucket encryption.....	30
Accessing S3 buckets.....	30
Accessing buckets in the same AWS account.....	31
Accessing buckets using a custom key.....	31
Accessing buckets in a different AWS account.....	36
Accessing S3 buckets in a managed policy environment.....	38
Accessing buckets in the same AWS account under a managed policy.....	39
Accessing buckets in a different AWS account under a managed policy.....	41
Accessing S3 buckets in a RAZ environment.....	44
Accessing buckets in the same AWS account under RAZ.....	45
Accessing buckets in a different AWS account under RAZ.....	45
Enabling RAZ manually.....	48

Remote access.....	50
Granting remote access to Kubernetes clusters on Amazon EKS.....	51
Revoking remote access to Kubernetes clusters on Amazon EKS.....	52
Restricting access to endpoints in AWS environments.....	52
Editing the IP CIDRs in the trusted list for endpoints in AWS environments.....	54
Networking.....	54
Overlay networks for AWS environments in Cloudera Data Warehouse service.....	54
Enabling overlay networks in AWS environments.....	56
Use a non-transparent proxy with Cloudera Data Warehouse on AWS environments.....	56
Configure non-transparent proxies for Cloudera Data Warehouse on AWS environments.....	57
Setting up private networking in AWS environments.....	58
Supported deployment modes for private networking in AWS.....	58
Prerequisites for private networking in AWS environments.....	59
Activating an AWS environment with private subnet support.....	60
Architecture for Private Load Balancer, Private Worker Nodes deployment on AWS.....	61
Custom tags in AWS environments.....	62
Upgrades.....	63
Upgrading PostgreSQL 9.6 before EOL.....	63
Validate the upgrade to PostgreSQL 10.16.....	65
Upgrade to PostgreSQL 11.12.....	66
Upgrading Amazon Kubernetes Service (EKS).....	66
Upgrading using your own AMI or reduced permissions.....	67
Dynamically updating the Amazon Machine Image.....	68
Managed storage access.....	69
Setting up managed storage access.....	71
Creating the Cloudera environment and IAM roles.....	71
Creating a UMS group and machine users.....	73
Creating a new Database Catalog.....	75
Creating a tenant-specific Virtual Warehouse.....	75
Tenant IAM role policy.....	75
Using AWS Vault to manage credentials.....	76
Identifying the spill location for Impala temporary data.....	77
Managing temporary scratch files in S3.....	78
Configuring an existing Impala Virtual Warehouse to spill to S3.....	79

Setting the scratch space limit for spilling Impala queries.....80

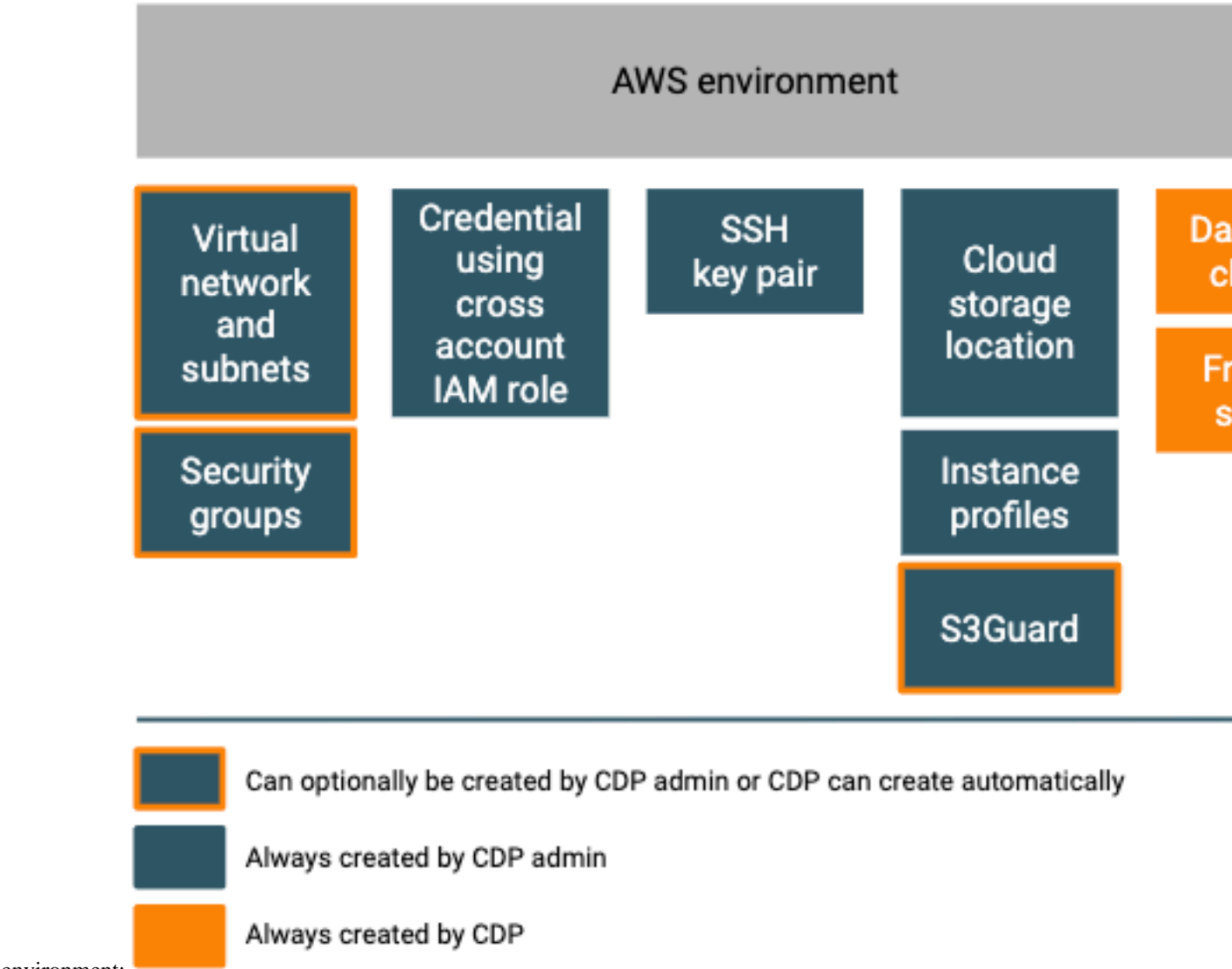
AWS environments overview

Learn about environments on AWS in Cloudera on cloud, including requirements you must meet before activating your environment in Cloudera Data Warehouse.

The Cloudera environment is closely related to the virtual private network in your cloud provider account. Registering an environment with Cloudera Management Console provides Cloudera with access to your cloud provider account and identifies resources in your account that Cloudera services can access, including Cloudera Data Warehouse. A single environment is contained within a single cloud provider region, so all resources deployed by Cloudera are deployed within that region within one specific virtual network. After you have registered an environment with Cloudera Management Console, you can activate the environment in Cloudera Data Warehouse. You create Virtual Warehouses in Cloudera Data Warehouse.

By default, Ranger Authorization is enabled in AWS environments. For an introduction to enabling RAZ, see the [Cloudera Management Console documentation](#).

The following diagram shows the components of an AWS



AWS environment requirements checklist

To successfully activate environments that have been registered with Cloudera on AWS VPCs with Cloudera Data Warehouse service, your AWS VPC must meet these requirements.

1. VPC has DNS resolution and DNS hostnames enabled

Ensure that your AWS VPC has DNS Resolution and DNS Hostnames enabled. For example, in the VPC Dashboard, click Your VPCs in the left navigation menu, and select the VPC you want to use for your Cloudera Data Warehouse service environment on Cloudera. View configuration details to make sure DNS resolution and DNS hostnames are Enabled. The AWS screen looks something like this:

The screenshot shows the AWS VPC Dashboard. In the left navigation menu, 'Your VPCs' is highlighted with an orange arrow. The main panel shows a list of VPCs, with one selected (dw-...). Below the list, the 'Details' section for the selected VPC is shown. Two orange arrows point to the 'DNS hostnames' and 'DNS resolution' fields, both of which are 'Enabled'.

Details			
VPC ID	State	DNS hostnames	DNS resolution
vpc- df8	Available	Enabled	Enabled
1	DHCP options set	Main route table	Main network ACL
Tenancy	dopt-4t	rtb- 0d 8	acl- 0c b
Default	IPv4 CIDR		

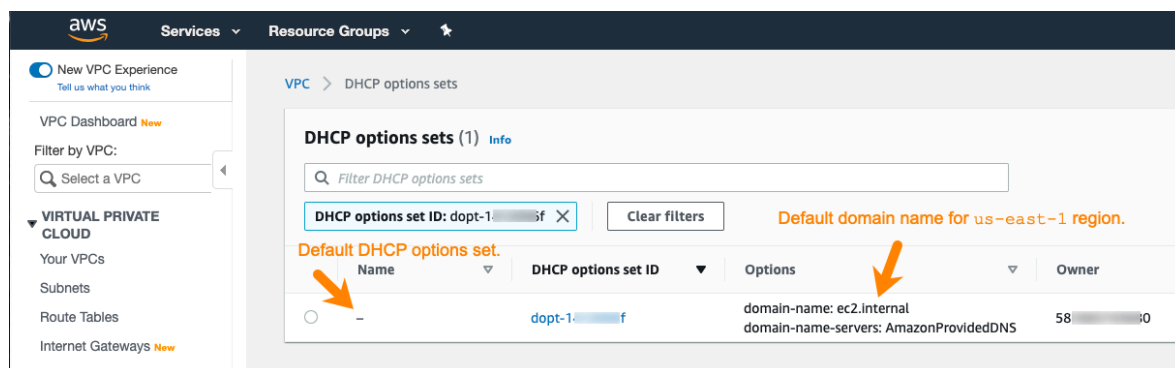
2. DHCP option set uses default domain name with one domain

When you create your VPC to use for the Cloudera Data Warehouse service, ensure that the DHCP option set attached to the VPC uses only one domain and use the default domain name:

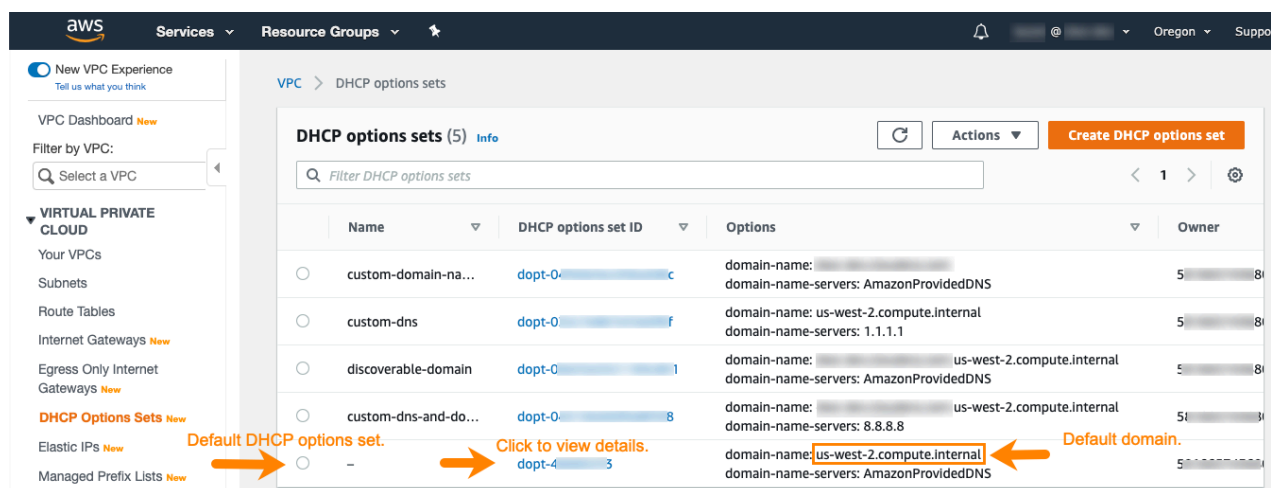
domain-name = <region>.compute.internal;

**Important:**

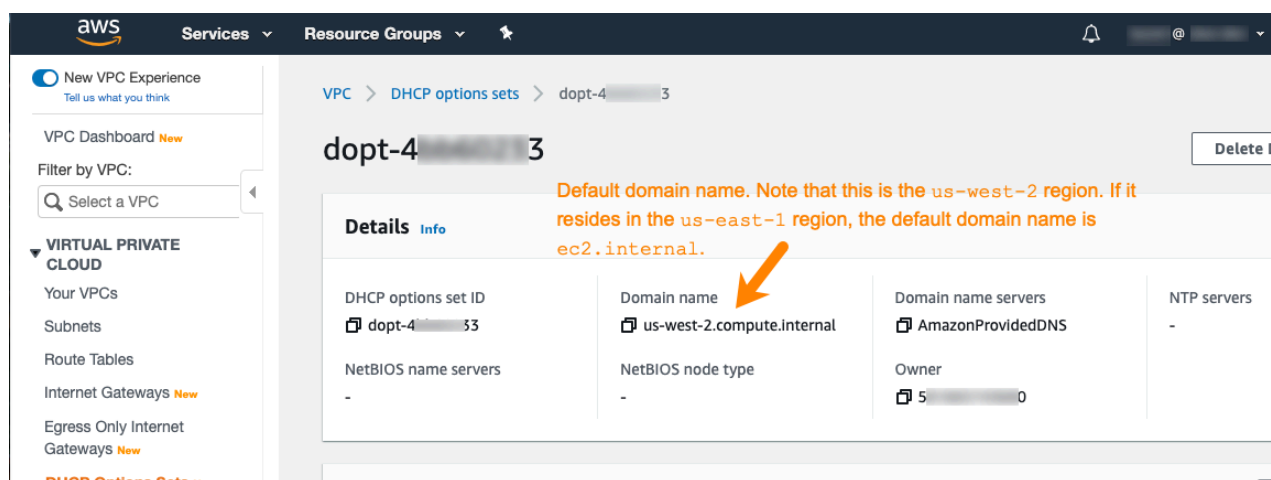
If your VPC is in the us-east-1 region [U.S. East (N. Virginia)], the default domain name is ec2.internal:



You can verify the setting in the VPC Dashboard of the AWS Console. Click the DHCP options set ID of the default DHCP options set (always named "-" by AWS) to view details, including the associated domain:



A details page appears:



3. DHCP option set uses AmazonProvidedDNS

When you create the VPC for the Cloudera Data Warehouse service, AWS automatically creates a set of DHCP options and associates them with the VPC. This set of options specifies the Amazon DNS Server as the default domain name server:

domain-name-server = AmazonProvidedDNS;

Use this setting for VPCs for the Cloudera Data Warehouse service shown in the AWS Console VPC Dashboard above.

4. Ensure the correct subnets in VPC are specified

When you activate an environment for the Cloudera Data Warehouse service, ensure that the subnets are correct. If there are more than three private subnets in the VPC only the top three are selected. However, they may not be the subnets you intend to use for the Cloudera Data Warehouse service.

5. Ensure private subnets have outbound internet connectivity

Your private subnets must have outbound internet connectivity. Check the route tables of private subnets to verify the internet routing. Worker nodes must be able to download Docker images for Kubernetes, billing and metering information, and to perform API server registration.



Note: If you have used only Cloudera Data Hub in your Cloudera environment, you must check and possibly configure connections to additional outbound destinations for Cloudera Data Warehouse.

For more information, see [AWS Outbound Network Access Destinations](#).

6. Ensure the Amazon Security Token Service (STS) is activated

To successfully activate an environment in the Cloudera Data Warehouse service, you must ensure the Amazon STS is activated in your AWS VPC:

1. In the AWS Cloudera Management Console home page, select IAM under Security, Identity, & Compliance.
2. In the Identity and Access Management (IAM) dashboard, select Account settings in the left navigation menu.
3. On the Account settings page, scroll down to the section for Security Token Service (STS).
4. In the Endpoints section, locate the region in which your environment is located and make sure that the STS service is activated.

Prerequisite for enabling a private EKS API server (Preview)

By enabling a private EKS API server, you can ensure that the EKS cluster is setup with only private endpoint enabled, which restricts the public access to your EKS API server from the internet. To set up the Amazon Elastic Kubernetes Service (EKS) cluster in private mode and to enable the private EKS, ensure that the DataLake cluster is created with Cluster Connectivity Manager version 2 enabled.

You must also run the following Cloudera CLI command:

```
cdp dw create-cluster --environment-crn crn:cdp:environments:us-west-1:XXXX
--use-private-load-balancer --aws-options enablePrivateEKS=true,workerSubnet
Ids=privatesubnet-1,privatesubnet-2,privatesubnet-3,lbSubnetIds=privatesubne
t-XXX --profile dev
```



Note: Using a private EKS API server is under technical preview and not recommended for production environments. Cloudera recommends you use this feature in test or development environments.

Related Information

[DHCP Options Sets in the Amazon documentation](#)

[Managing AWS STS in an AWS Region in the Amazon documentation](#)

[Activating environments](#)

Activating an AWS environment from Cloudera Data Warehouse

To use an AWS environment for Cloudera Data Warehouse on cloud you must first activate it.

About this task

When you activate an environment, Cloudera creates an EKS cluster to host Kubernetes-based resources. The underlying compute, network resources are managed by AWS:

- Resource group
- Load balancer(s)
- Public IP address(es)
- Network security group
- Disk(s)

In the Cloudera Data Warehouse environment, instances for shared service components are set up within a Kubernetes (K8s) cluster. The setup begins with three m5.2xlarge instances running the Cloudera Data Warehouse service, but the K8s cluster is capable of autoscaling, automatically adding more instances if necessary to handle increased demand. Additionally, an Amazon Relational Database Service (RDS) (db.r5.large) running PostgreSQL is created to store user metadata for Hue and Data Visualization services. In total, three shared db.r5.large nodes are used for this purpose. [Always active, shared services](#).

Before you begin

- Obtain the DWAdmin role.
- Review the [AWS environment requirements](#).

Procedure

1. In the Cloudera Data Warehouse service, in Environments, locate the environment that you want to activate.
2. Click Activate.
3. In Deployment Mode, select load balancers.

For more information, see [Load balancers for AWS environments](#).



Note: Select Private Load Balancer, Private Executor if you want to enable using a private EKS API server. Using a private EKS API server is under technical preview and not recommended for production environments. Cloudera recommends you use this feature in test or development environments.

To view or configure the public and private subnets that have been specified for your Cloudera environment, click Advanced Settings.

- Private Subnets: Accept the selected subnets you configured during [AWS environment registration](#) for load balancer and workload pods, or deselect subnets. Cloudera recommends three subnets for each load balancer to enable high availability (HA).
- Enable IP CIDR for Kubernetes cluster: Enter the IP Classless Inter-Domain Routing (CIDRs) from which the Kubernetes cluster should accept incoming connections. Connections from other IP ranges are dropped. Obtain your internal network's IP CIDR ranges of IP addresses that need access to endpoints on the Kubernetes cluster. For more information, see [Restricting access to endpoints in AWS](#).
- Enable IP CIDRs for the load balancer: Enter the IP CIDR(s) from which the load balancer should accept incoming connections. Connections from other IP ranges are dropped. Obtain your internal network's IP CIDR ranges of IP addresses that need access to endpoints that are load balanced. For more information, see [Restricting access to endpoints in AWS](#).
- Use Overlay Network: [Overlay Networks for AWS environments](#) can increase the number of available IP addresses for your deployments of Cloudera Data Warehouse if you have an existing Virtual Private Cloud

(VPC). Use this feature if your Virtual Private Cloud subnet has fewer than 1,024 IP addresses. Cloudera recommends that you do not configure more than 200 executor nodes for an overlay network to operate.

- **Attach Managed Policy ARN to Node Role:** If you do not want to provide PutRolePolicy permission in your cross account role, you can [attach a managed policy ARN to a node role](#) to provide the cross account role permissions. You must create a new NodeInstanceRole manually, and provide the ARN during activation of the environment from Cloudera Data Warehouse.
- **Use Reduced Permissions Mode:** If you cannot provide the standard set of IAM permissions required by Cloudera Data Warehouse for environment activation, you can [use reduced permissions mode](#) to activate an AWS environment with fewer than half of these permissions. To use this feature, a [minimum set of IAM permissions](#) are required.
- **Enable CloudWatch Logs:** Enable CloudWatch logs if you use Amazon CloudWatch. In your AWS account, you can then find the logs in /aws/eks/<cluster name>/cluster. Before enabling CloudWatch, you must [add required permissions](#) to your IAM policy to access CloudWatch logs; otherwise, you cannot activate the environment.

4. Click Activate.

Related Information

[Supported deployment modes](#)

[Setting up private networking](#)


[Restricting access to endpoints](#)

[Overlay networks for AWS environments](#)

Load balancers for AWS environments

When you activate an environment, you can select load balancers.

You can select public and private load balancers to evenly distribute the inbound traffic. Deployment Mode options that you can select are as follows:

Option	Brief Description	Requirements and Recommendations
Private Load Balancer, Private Executors	Cloudera Data Warehouse nginx based load balancer runs on private subnets and all workload pods that also run on private subnets.	Requires a jump host or AWS direct connect to access Cloudera Data Warehouse. Cloudera recommends that you use the Private Load Balancer, Private Executors deployment mode if possible for security reasons. Selecting this option is required if you want to use a private EKS API server.  Note: Using a private EKS API server is under technical preview and not recommended for production environments. Cloudera recommends you use this feature in test or development environments.
Public Load Balancer, Private Executors	Cloudera Data Warehouse nginx based load balancer runs on public subnets and all workload pods that also run on private subnets.	Select this option to connect to Cloudera Data Warehouse from anywhere as long as the source CIDR filters allow the connection.
Public Load Balancer, Public Executors	Cloudera Data Warehouse nginx based load balancer runs on public subnets and all workload pods that also run on public subnets.	Select this option to connect to Cloudera Data Warehouse from anywhere as long as the source CIDR filters allow the connection. Cloudera does not recommend selecting this option for security reasons.

AWS restricted policy

As Administrator, you must include the AWS restricted policy in your IAM role to limit access to the environment. You must include this policy before you activate the environment in Cloudera Data Warehouse.

The AWS restricted policy associates a [cross-account role](#) with the environment. If you do not have a Ranger Authorized (RAZ)-enabled Data Lake, simply attach the AWS restricted policy to your cross-account role, and nothing more. If you do have a RAZ-enabled Data Lake, take the following actions:

- Attach the AWS restricted policy described below to your cross-account role.
- Add ARNs to the role or a new sid with appropriate permissions to the AWS restricted policy.

For more information see "[RAZ-enabled Data Lake restricted policy](#)" below.

Attaching the policy to your cross-account role

The AWS restricted policy is split into two files because the content exceeds the AWS file size limit. To attach the policy to your cross-account role, you need to work with both files as described in the following steps:

1. In the AWS management console, find the Cross-account IAM role you created.
2. Go to the following Github links to get the restricted policy files without comments:

- [Restricted policy file 1](#)
- [Restricted policy file 2](#)

For your information, you can get a commented version of each file that explains each line in the policy.

- [Restricted policy file 1 with comments](#)
 - [Restricted policy file 2 with comments](#)
3. Attach the policies to your IAM role.
 4. In the Restricted policy file1 without comments, replace `${DATALAKE_BUCKET}` with the name of your S3 bucket. For example my-bucket.
 5. Attach both restricted policy file 1 and restricted policy file 2 without comments to your IAM role.

Do not attempt to attach the policy files with comments as this would cause an error.

RAZ-enabled Data Lake restricted policy

The AWS restricted policy associates a [cross-account role](#) with the environment, as mentioned above. If you have a Ranger Authorized (RAZ)-enabled Data Lake, one of the following additions to either the cross-account role or cross-account json policy are required:

- File 1 Append ARNs to role: To your cross-account restricted policy, in the existing sid "AttachRole ", append all policy ARNs attached to the RAZ role.
- File 2 Add a new sid: Add "AttachRAZPolicyToNodeInstance" to the cross-account json policy

File 1 append ARNs to role

Append all the cross-account policy ARNs attached to the RAZ role to the sid key "AttachRole" value in your cross-account restricted policy.

```
{
  "Sid": "AttachRole",
  "Effect": "Allow",
  "Action": "iam:AttachRolePolicy",
  "Resource": [
    "arn:aws:iam::*:role/env-*-dwx-stack-EKSServiceRole-*",
    "arn:aws:iam::*:role/env-*-dwx-stack-NodeInstanceRole-*"
  ],
  "Condition": {
    "ForAnyValue:ArnEqualsIfExists": {
```

```

    "iam:PolicyARN": [
      "arn:aws:iam::aws:policy/AmazonEKSClusterPolicy",
      "arn:aws:iam::aws:policy/AmazonEKSServicePolicy",
      "arn:aws:iam::aws:policy/AmazonEC2ContainerRegistryReadOnly",
      "arn:aws:iam::aws:policy/AmazonEKSWorkerNodePolicy",
      "arn:aws:iam::aws:policy/AmazonEKS_CNI_Policy",
      "arn:aws:iam::aws:policy/CloudWatchAgentAdminPolicy"
      ...
    ]
  }
}

```

Replace the ellipsis ... placeholder above with all policy ARNs attached to the RAZ role as shown in the example below. See [IAM policy definitions](#) for more information about these policies.

```

arn:aws:iam::1234567:policy/aws-cdp-datalake-admin-s3-policy
arn:aws:iam::1234567:policy/aws-cdp-bucket-access-policy
arn:aws:iam::1234567:policy/aws-datalake-backup-policy
arn:aws:iam::1234567:policy/aws-datalake-restore-policy

```

File 2 Add sid to policy

Add the new sid "AttachRAZPolicyToNodeInstance" value to the cross-account json policy.

Make sure all the policies attached to RAZ/Data Lake Admin role are in the following regex pattern format:

```

${ANY_WILDCARD_REGEX}

```

Use the pattern as a value for the key "iam:PolicyARN" as shown in the example below:

Example policies attached to the RAZ/Data Lake Admin role

```

arn:aws:iam::1234567:policy/cdp-dev-datalake-admin-s3-policy
arn:aws:iam::1234567:policy/cdp-dev-datalake-bucket-access-policy
arn:aws:iam::1234567:policy/cdp-dev-datalake-backup-policy
arn:aws:iam::1234567:policy/cdp-dev-datalake-restore-policy

```

The regex is "arn:aws:iam::1234567:policy/cdp-dev-datalake*".

```

{
  "Sid": "AttachRAZPolicyToNodeInstance",
  "Effect": "Allow",
  "Action": "iam:AttachRolePolicy",
  "Resource": [
    "arn:aws:iam::*:role/env-*-dwx-stack-NodeInstanceRole-*"
  ],
  "Condition": {
    "ForAnyValue:ArnLikeIfExists": {
      "iam:PolicyARN": "arn:aws:iam::<AWS_ACCOUNT_ID>:policy/
${ANY_WILDCARD_REGEX}"
    }
  }
}

```

For more information about using RAZ, see [fine-grained access control](#) and [Cloudera policies](#) documentation.

Attaching a managed policy ARN

For security reasons, if you do not want to provide PutRolePolicy permission in your cross account role, which would be used later to add an inline policy to the Node instance role, you must create a managed policy. Failure to do so results in an authorization error. You learn how to create the managed policy in a few steps.

About this task

If you are using the [restricted policy with managed policy ARN](#), you need to add the permission to the cross account role from this [Github link](#).



Note: Enable DAS Link will not be supported, since *PutRolePolicy* permission is not available in your cross account role.

Procedure

1. Follow instructions in steps 1-4 to activate your environment in Cloudera Data Warehouse, as described in the "Activating an AWS environment" above.

2. Select Attach Managed policy ARN to Node Role, and pass the ARN.

For example:

When you remove the PutRolePolicy permission, which is one of the standard permissions needed, from your cross account role, the reduced permissions mode UI appears.

For example:

For more information about reduced permissions mode, see the topic below.

3. In Environmental Activations, optionally do not select Reduced Permissions mode.

4. Click Activate.

The [noderole-inline-policy.json](#) is attached to the Node Instance role instead of a inline policy requiring the PutRolePolicy permission in your cross account role.

5. Make the following changes to the [noderole-inline-policy.json](#) file in your cross account role:

- `${DATALAKE_BUCKET}` - Replace this with the name of your S3 bucket. For example my-bucket.
- `${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_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.
- `${BACKUP_BUCKET}` - Replace this with the name of your S3 bucket for backup. For example my-bucket.

Activating your environment in reduced permissions mode

IAM permissions are required by Cloudera Data Warehouse for environment activation. You can choose to provide a reduced set of IAM permissions for environment activation instead of the full set in the AWS restricted permissions policy.

About this task

You can activate an AWS environment for Cloudera Data Warehouse with fewer than half the set of required IAM permissions on your AWS cross-account IAM role. You can choose reduced permissions mode in two ways:

- Select the Use reduced permissions mode when you activate your environment from Cloudera Data Warehouse.
- Alternatively, let the system detect your account does not have the AWS [restricted permissions policy](#).

The following dialog appears. You can select Check to activate environment with reduced permissions mode.



The dialog box is titled "Environment Validations" and has a close button (X) in the top right corner. The main text reads: "You seem to have missing permissions. Permission for {iam:PutRolePolicy} was [implicitDeny]. Do you want to continue activating?". Below this text is a checkbox labeled "Check to activate environment with reduced permissions mode". At the bottom right, there are three buttons: "BACK", "CANCEL", and "ACTIVATE". The "ACTIVATE" button is green and highlighted.

In this task, you activate your environment from Cloudera Data Warehouse in reduced permissions mode. In steps 6 and 7, you manually create the stack and then remove it:

Procedure

1. In the Cloudera Data Warehouse service, in Environments, click the search icon and locate the environment that you want to activate.
2. Activate the environment.

3. In Activation Settings, if you do not have the standard set of required IAM permissions or a restricted policy in place, select Use Reduced Permissions Mode.
For example:

Activate Environment

Do you want to activate the environment "brodia-aws"?

Deployment Mode:*

- ☒ Private Load Balancer, Private Executors
- ☐ Public Load Balancer, Private Executors
- ☐ Public Load Balancer, Public Executors

Advanced Settings



Node Count:



Overprovision compute nodes

Node Count:



- ☐ Use Custom ECR repository
- ☐ Use Overlay Network
- ☐ Attach Managed policy ARN to Node Role
- ☐ Enable Compaction Observability
- ☒ Use Reduced Permissions Mode

Alternatively, if Environment Validations appears, select Check to activate environment with reduced permissions mode. Skip the next step and go to step 6.

4. If you do not want to activate the environment in reduced permissions mode, uncheck the option, and click Activate. Skip the rest of the steps in this procedure. Cloudera Data Warehouse automatically creates the cloud resources in your AWS account for you.
5. Manually create the cloud resources in your AWS account and tag them appropriately, as described in topic, "Setting up cloud resources for reduced permissions mode" below.

Cloudera Data Warehouse pre-populates the required CloudFormation template for you within the AWS console, and you perform the manual steps to create the stack.

6. When you are finished using the stack, manually delete it in the AWS console.

Related Information

[Minimum set of IAM permissions required for reduced permissions mode](#)

Minimum set of IAM permissions required for reduced permissions mode

Review a list of the minimum IAM permissions required to activate AWS environments for Cloudera Data Warehouse in reduced permissions mode.

The following is a list of the minimum permissions that are required for your IAM policy to activate environments for Cloudera Data Warehouse in reduced permissions mode. In this mode you must manually create your CloudFormation stack from a template that Cloudera Data Warehouse pre-populates in the AWS console for you. When you are finished using the stack, you must manually delete its resources in your AWS account.

Table 1: Minimum set of IAM policy permissions required for environment activation in Cloudera Data Warehouse in reduced permissions mode

AWS service	"Allow" actions	Description
Certificate Manager (acm)	DescribeCertificate	Created by Cloud Formation to check the certification status during activation.
	ListCertificates	ACM validation adds DNS records.
CloudFormation (cloudformation)	DescribeStackEvents	Get Cloud Formation stack events, identify cause of failed CF stack creation failure
	DescribeStacks	Check the status of stack--error or completed, then install helm charts
	UpdateStack	Update Custom AMI, upgrade EKS
CloudWatch (logs)	CreateLogGroup	Create/name cloudwatch log group
	CreateLogStream	Create log stream of log group that originates from monitored application or resource
	DescribeLogStreams	List log streams for log groups
	PutLogEvents	Upload log events to log stream
	PutRetentionPolicy	Change number of days Cloudwatch retains
EC2 (ec2)	CreateKeyPair	Create ssh Public key pair, pass to ec2 instances. Not required if passed/set/reused via CloudBreak
	CreateTags	Tag subnets and eks security group. Amazon EKS security group requirements and considerations
	DeleteKeyPair	Delete keypair while deactivating Cloudera Data Warehouse // needed if CB env ssh is not reused
	DeleteTags	

AWS service	"Allow" actions	Description
	DescribeDhcpOptions	See points 2-3 in AWS Requirements Checklist
	DescribeKeyPairs	Validate CloudBreak env ssh key pair exists, not deleted inbetween; check for duplicate keypair in case of Cloudera Data Warehouse created keypair
	DescribeRouteTables	
	DescribeSubNets	See Point 4 in AWS Requirements Checklist
	DescribeVpcAttribute	Validate enableDnsHostnames and enableDnsSupport VPC attributes; see 1 and 3 points in Footnote 3 URL
	DescribeVpcs	Validate ID of set of DHCP options associated with the VPC
EC2 Auto Scaling (autoscaling)	DescribeAutoScalingGroups	Get shared services/compute ASGs, update as part of AZRebalance
	SuspendProcesses	Suspend AZRebalance for autoscaling group; include AZRebalance; cannot suspend AZRebalance in cloudformation; edit/update ASGs with AWS API to avoid AWS re-balancing nodes for AZ (most nodes run in stateful/critical pods)
	UpdateAutoScalingGroup	Calico overlaynetwork option requires no EKS nodes up on installation; with CF stack creation 3 nodes start up, autoscaling group updates desired capacity to Zero via AWS API; need latest SSH key from CloudBreak for EKS node updates; new Launch template passes SSH key and updates in ASG
EKS (eks)	DescribeCluster	Calico overlaynetwork option requires no EKS nodes up on installation; with CF stack creation 3 nodes start up, autoscaling group updates desired capacity to Zero via AWS API; need latest SSH key from CloudBreak for EKS node updates; new Launch template passes SSH key and updates in ASG
	DescribeUpdate	Check status of Updates--enable Private EKS and Cloudwatch on EKS
	TagResource	Tag eks cluster, e.g.: clusterId, envId, clustername, accountId...
	UpdateClusterConfig	Update EKScaluster config Enable Private EKS and Cloudwatch on EKS
	UpdateClusterVersion	Updates an Amazon EKS cluster to the specified Kubernetes version
IAM (iam)	DeleteRolePolicy	
	GetRolePolicy	Delete inline policies like efs, ebs, cluster-autoscaler etc created/attached to Node instance role at deactivation
	ListAttachedRolePolicies*	List policies attached to Ranger RAZ role; attach to NodeInstanceRole for S3 access if RAZ enabled
	PutRolePolicy	Add inline policies like efs, ebs, cluster-autoscaler to Node Instance Role
	SimulatePrincipalPolicy	Simulate Cloud Formation stack formation policies

AWS service	"Allow" actions	Description
RDS (rds)	StartDBInstance	
	StopDBInstance	
	DescribeDBInstances	
S3 (s3)	GetBucketLocation	Needed for external bucket feature via UI, where we validate the VPC and bucket region are the same
	GetObject	Get Cloud Formatoin template while CF stack creation may not be needed for reduced mode
	ListBucket	
	PutObjectAcl	
	PutObject	Put Cloud Formation template in SDX bucket

*Needed only in a Ranger Authorization (RAZ) environment.

Reduced permissions mode JSON IAM permissions policy template

To activate an AWS environment for Cloudera Data Warehouse using reduced permissions mode, you can use this sample JSON template when you register an environment in Cloudera.

In this mode you must manually create your CloudFormation stack from a template that Cloudera Data Warehouse pre-populates in the AWS console for you. When you are finished using the stack, you must manually delete its resources in your AWS account.

Get the JSON [reduced permissions mode policy from Github](#). Use the policy in Step 6 of the procedure to [create your cross-account IAM role](#) for Cloudera. Make the following substitutions in the policy:

- Replace \${ACCOUNT_ID} with your AWS account number.
- Replace \${DATALAKE_BUCKET} with your datalake bucket name.

Related Information

[Create a cross-account IAM role](#)

Required tags for CloudFormation stacks created with reduced permissions mode

This is a list of tags you must manually apply to AWS CloudFormation stack resources when you use the reduced permissions mode to activate environments for Cloudera Data Warehouse.

Table 2: Required tags for CloudFormation stacks created with reduced permissions mode in Cloudera Data Warehouse

Tag key	Tag value
clusterName	Name of the Cloudera environment registered with Cloudera Management Console
data-warehouse-env-owner	Email ID of the account of the user who owns the stack and is a user who has access to the Cloudera environment. This is the email ID of the email address that is listed for Email on the Users page of the User Management module of Cloudera Management Console.
stackName	Name of the Cloudera Data Warehouse CloudFormation stack. In the format: env-<ENVIRONMENT-IDENTIFIER>-dwx-stack For example: env-6g8dsf-dwx-stack

Tag key	Tag value
clusterId	Cloudera Data Warehouse environment ID that is displayed in the environment tile in the Cloudera Data Warehouse UI. For example: env-hmrt2z
Cloudera-Resource-Name	Cloudera environment ID (CRN [Cloudera Resource Name]). In the format: crn:cdp:environments:<REGION>:<ACCOUNT-ID>: environment :<IDENTIFIER> For example: crn:cdp:environments:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:environment:cc8ad776-4704-48f7-a243-97348939becd
actorCrn	CRN (Cloudera Resource Name) from the user's profile in the Cloudera Management Console module of Cloudera. For example: crn:altus:iam:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:user:de370d9f-ebb4-4b75-a89b-5d15306ae143
accountId	Tenant ID. In the above CRN example, the tenant ID is the GUID listed immediately after the AWS region: 9d74eee4-1cad-45d7-b645-7ccf9edbb73d

Setting up cloud resources for reduced permissions mode

Learn how to activate environments on AWS using the reduced permissions mode in Cloudera Data Warehouse. In this mode, you must manually create and delete the CloudFormation stack in the AWS Console.

About this task

Required role: EnvironmentAdmin or PowerUser

When you activate an AWS environment for Cloudera Data Warehouse, if you do not have the standard required IAM permissions, the following message displays in the environment tile of the Cloudera Data Warehouse UI, which provides a link to the AWS Console:

Click the link and perform the following listed steps to navigate to the AWS Console and create the CloudFormation stack.

Before you begin

- Because you need to use the AWS Console to manually create your CloudFormation stack for Cloudera Data Warehouse environment activation, in another browser tab, log into your AWS account before you begin. Make sure that the IAM entity logged in has the two AWS restricted policies described in ["AWS restricted policies"](#).
- You must also have the AWS CLI and the kubectl CLI configured and available on your system to apply the kubeconfig that Cloudera Data Warehouse provides in Step 10 below.



Important:

- Make sure you have the time to complete the task of creating your CloudFormation stack in reduced permissions mode in one sitting. This can take up to 20 minutes. Otherwise, if the creation process outlined below is delayed in the AWS Console for more than an hour, the Cloudera Data Warehouse environment activation times out and the environment will go into an error state.
- Make sure that the IAM entity you use to log into the AWS Console has adequate permissions to create CloudFormation stacks and to run kubectl commands on your AWS environment.
- Make note of the name of the IAM entity you use because you must use it to log in using a terminal window again in Step 10c. of the following procedure.

Procedure

- In the Cloudera Data Warehouse UI Overview page, go to the Environments tab.

2. Locate the environment you want to activate, and click Activate.
3. If the system detects that you do not have the standard required IAM permissions on your AWS account for automatic CloudFormation stack creation by Cloudera Data Warehouse, it displays the following message in the tile:

Step 1 of 2: Insufficient permissions! Visit AWS Console to Create Stack and come back. Creating a stack will take up to 15 minutes.

This message asks you to navigate to the AWS Console to manually create the CloudFormation stack.

4. Click the link Visit AWS Console to Create Stack and the AWS Console opens on the CloudFormationStacksCreate Stack page that is pre-populated with a template in another browser tab.
5. In the Create Stack page, click Next to advance to the Specify stack details page.



Important: Do not change any configurations on the Specify stack details page, including the Stack Name.

6. In the Specify stack details page, click Next to advance to the Configure stack options page where you can specify the required tags for your stack resources. See "Required tags for CloudFormation stacks," which is linked to at the bottom of this page for a list of required tags.
7. After adding the required tags, do not set the remaining options on the page. Scroll down to the bottom of the page and click Next to advance to the Review <ENV-STACK-NAME> page.
8. On the Review <ENV-STACK-NAME> page, scroll down to the bottom, click the I acknowledge that AWS CloudFormation might create IAM resources check box, and then click Create Stack.



Note: Stack creation can take up to 20 minutes depending on network traffic and load. The Cloudera Data Warehouse UI monitors AWS stack creation and displays a "Creating" message in the environment tile.

9. After stack creation has completed, a message displays in the Cloudera Data Warehouse UI environment tile. Click the Open Configurations link and a Configurations dialog box displays.
10. In the Configurations dialog box, perform the following steps:
 - a. Copy the Kubeconfig text to your system clipboard and save it into a text file on your system.
 - b. Copy the Aws Auth text to your system clipboard and save it into a text file on your system. The Aws Auth text provides the IAM cross-account role that is registered in Cloudera to access the EKS cluster on AWS after you perform the kubectl command in the next step.
 - c. In a terminal window, verify that the AWS CLI is configured to use the same IAM entity that you used to create the CloudFormation stack in Step 8. Then, using the kubectl CLI, run the following commands to apply configurations from the two text files that you created in Step 10b:

```
$> export KUBECONFIG=<PATH-TO-THE-KUBECONFIG-TEXT-FILE>
```

```
$> kubectl apply -f <PATH-TO-THE-AWS-AUTH-TEXT-FILE>
```

Look for shell output "configmap/aws-auth created" to confirm the configuration was applied correctly.

If you deployed the CloudFormation stack with a federated user using the AWS console, you need to execute the commands mentioned in 10c (above) in the AWS CloudShell. You might need to manually install the kubectl command in the CloudShell.

11. Back in the Cloudera Data Warehouse UI Configurations dialog box, select Yes, Kubeconfig and AWS Auth configuration are applied checkbox, and then click Finish Activation.

Results

After clicking Finish Activation, the environment is activated and the tile displays a starting message.

Deactivating AWS environments created with reduced permissions mode

Learn how to deactivate an environment that has been activated for use in Cloudera Data Warehouse with the reduced permissions mode. When you deactivate an environment in Cloudera Data Warehouse, the environment registered with Cloudera remains available for use by other applications.

About this task

Required role: EnvironmentAdmin or PowerUser

If a Cloudera environment has been activated for Cloudera Data Warehouse with the reduced permissions mode, then if you deactivate the environment in the Cloudera Data Warehouse UI, you must manually delete the CloudFormation stack in AWS and its associated S3 buckets and DynamoDB table. After you click the deactivation icon in the Cloudera Data Warehouse environment tile, a link displays in the tile that you can use to navigate to the AWS Console to delete these cloud resources.

Procedure

1. In the Cloudera Data Warehouse UI, navigate to the tile for the environment you want to deactivate, and click Deactivate, which launches the Action dialog box.
2. (Optional) In the Action dialog box, you can select one of the following environment deactivation options if appropriate:
 - Choose Drop Data if you want to drop any data Cloudera Data Warehouse created outside of the Data Lake, but retain the Database Catalogs and Virtual Warehouses that are associated with the environment.
 - Choose Force Delete if you want to drop the data and also remove the Database Catalogs and Virtual Warehouses that are associated with the environment.
3. Click Visit AWS Console to Delete Stack, which displays in the Action dialog box. This opens the AWS Console in another browser tab.
4. In the Action dialog box of the Cloudera Data Warehouse UI, click OK, and then wait five minutes so Cloudera Data Warehouse can perform its deletion steps.
5. After waiting five minutes, in the AWS Console, click Delete to delete the CloudFormation stack in your AWS account.
6. In the AWS Console, perform the following tasks:

- Navigate to the S3 service and delete the following S3 buckets that were used by the stack:

- `<S3-BUCKET-NAME>-<LAST-4-DIGITS-OF-ENVIRONMENT-ID>-dwx-managed`

- `<S3-BUCKET-NAME>-<LAST-4-DIGITS-OF-ENVIRONMENT-ID>-dwx-external`

For example, if the bucket name is sales-east and the Cloudera Data Warehouse environment ID is ENV-CK8988, the buckets you should delete are:

```
sales-east-8988-dwx-managed
```

```
sales-east-8988-dwx-external
```

- Navigate to the DynamoDB service and delete the associated DynamoDB table.

Results

After performing these steps, the resources in your Amazon account for the deactivated Cloudera Data Warehouse environment have been deleted.

Retaining PostgreSQL backups in AWS environments

When you create a Cloudera Data Warehouse cluster using the Cloudera CLI create-cluster command, any PostgreSQL backup retention period you set on your Cloud Provider side, is observed by Cloudera.

Procedure

1. In AWX, configure BackupRetentionPeriod.
2. Create a Data Warehouse cluster using the Cloudera CLI create-cluster command.
The Data Warehouse cluster will retain the PostgreSQL backups according to your configuration.

Viewing and editing AWS environment details

This topic describes how to view and edit Cloudera AWS environment details in the Cloudera Data Warehouse service UI.

About this task


You can view Cloudera AWS environment details without leaving the Cloudera Data Warehouse service UI. Accessing the Environment Details page in the Cloudera Data Warehouse UI also enables you to edit the description of the environment and the allowed IP Classless Inter-Domain Routing (CIDRs) that control access to Kubernetes and load balancer service endpoints.

Before you begin

- You must activate an environment before you can view or edit its details. See "Activating AWS environments," which is linked to in the "Related information" section at the bottom of this page.
- Obtain the DWUser role for viewing environment details.
- Obtain the DWAdmin role for editing the environment details.

Procedure

1. In the Cloudera Data Warehouse service, go to the Environments tab.
2. Locate the environment that you want to view.

3. Click  Edit .

In Environment Details, you can view information about the environment, like the Cloudera Data Warehouse release version in which you activated the environment, when you created and last updated the environment, and how many Database Catalogs and Virtual Warehouses use the environment.

In Configurations, you can also make the following changes:

ENVIRONMENT Name: nfqe-aws-7215

	STATUS	VERSION	CREATED BY	DATA
	Running	1.6.1-b220	rbalamohan@cloudera.com	1

GENERAL DETAILS

CONFIGURATIONS

Description:

Please enter the description

Enable IP-CIDR for Kubernetes cluster:

List of allowed IP-CIDR for Kubernetes cluster

Enable IP-CIDR for the load balancer:

List of allowed IP-CIDR for the load balancer

Add External S3 Bucket:

✓ Apply Changes

 Discard Changes

- Add a description for the environment, that makes it easier to identify.
- Add or edit the list of IP CIDR(s) for Kubernetes Cluster to enable access from your internal network to the Kubernetes endpoints.
- Add or edit the list of IP CIDR(s) for the load balancer.
- Add an external S3 bucket to [access data outside your Data Lake](#), such as a CSV File.
- Enable CloudWatch logs if you use Amazon CloudWatch. In your AWS account, you can then find the logs in `/aws/eks/<cluster name>/cluster`.



Note: Before enabling CloudWatch, you must [add required permissions](#) to your IAM policy to access CloudWatch logs; otherwise, your cluster goes into an error state.

4. Click Apply Changes.

Related Information

[Activating environments](#)

[Restricting access to endpoints](#)

Deactivating an AWS environment

Learn how to deactivate an AWS environment for Cloudera on cloud.

Before you begin

Required role: EnvironmentAdmin or PowerUser

Procedure

1. In the Cloudera Data Warehouse service, go to the Environments tab.
2. Locate the environment that you want to deactivate and click Deactivate.
The **Action** modal is displayed.
3. On the Action modal, you can select environment deactivation options:
 - Choose Drop Data to remove the managed buckets that were created by Cloudera Data Warehouse during environment activation. The underlying data in the data lake is untouched. Only the default Database Catalog is retained because it resides in the data lake. Non-default Database Catalogs and Virtual Warehouses are deleted.



Important: All metadata and the Ranger policies for non-default Database Catalogs are deleted.

- Choose Force Delete to drop the data and to remove the Database Catalogs and Virtual Warehouses that are associated with the environment.
4. Click OK to deactivate the environment.

Cloud storage buckets

Cloudera Data Warehouse is integrated with the Data Lake Storage Cloud provider storage, such as AWS S3 or Azure Storage. During Data Lake creation, Cloudera creates storage locations for your data, logs, and backups.


For more information about AWS storage buckets, see [S3 bucket and IAM roles and policies for logs, backup, and data storage](#). For information about which logs are stored in which directories, see [Locations of Impala log files in S3](#).




Typically, during cluster creation, a [managed policy](#) is created automatically by Cloudera Data Warehouse and attached to a node instance role. Alternatively, if you need different permissions than those specified during cluster


creation, you can create or modify the managed policy and attach it to the node instance role. Whether you create the policy manually, or Cloudera Data Warehouse creates the policy automatically, the policy must specify the paths to the log, backup, and data buckets in the Resources array of the `s3readwriteownbuckets` object in the managed policy JSON.

```
"arn:aws:s3:::${LogBucket}/clusters",
"arn:aws:s3:::${LogBucket}/clusters/*",
"arn:aws:s3:::${LogBucket}/<Your configured log path>",
"arn:aws:s3:::${LogBucket}/<Your configured log path>/*",
"arn:aws:s3:::${BackupBucket}/<Your configured backup path>",
"arn:aws:s3:::${BackupBucket}/<Your configured backup path>/*",
"arn:aws:s3:::${DataBucket}/<Your configured data path>",
"arn:aws:s3:::${DataBucket}/<Your configured data path>/*",
"arn:aws:s3:::${DataBucket}/backup",
"arn:aws:s3:::${DataBucket}/backup/*",
```

You get the path and name of the bucket, which was specified during Data Lake creation. To get the paths and names of the buckets, navigate to **Environments Data Lake**. Click **Summary**.



Data Lake Details

NAME	NODES	SCALE
cdw-nfqe-w9aqp2	 2  0  0	Custom


STATUS	STATUS REASON	CRN
 Running	Datalake is running	crn:cdp:datalake

[Data Hubs](#)
[Data Lake](#)
[FreeIPA](#)
[Data Services Clusters](#)
[Cluster Definitions](#)
[Summary](#)

The name and paths of your logs and backup Data Lake buckets appear:


Logs Storage and Audits

Storage Location:	s3a://eng-sdx-longrunning-qe/jubin-aws/audit
Instance Profile:	arn:aws:iam::146617852659:instance-profile/mow-dev


Backup Storage

Storage Location:	s3a://eng-sdx-longrunning-qe/jubin-aws/audit
Instance Profile:	arn:aws:iam::146617852659:instance-profile/mow-dev

One S3 bucket with a sub-directory named after your data lake such as `s3a://my-bucket/my-dl` is created when your Data Lake is created. This bucket is called the Storage Location Base and is intended for your data.

Node instance roles need access to data, logs, and backup buckets in your Data Lake. To configure a node instance role to access these buckets, you [attach the managed policy to the node instance role](#).

Bucket encryption

To allow encryption and decryption of your S3 bucket contents, the environment node instance role must access the [KMS](#) (Key Management System) key. Reading and writing to the bucket is impossible unless you provide this access.

By default, any resource under your account can use and manage keys. However, typically you want to restrict access to keys, authorizing only the node instance role to use the key. To accomplish this, you add code to your [managed policy](#) and attach the policy to the node instance role. The following example policy snippet provides minimal privileges to use the key.

```
{
  "Sid": "Allow use of the key",
  "Effect": "Allow",
  "Principal": {
    "AWS": "arn:aws:iam::555555555555:role/env-id-dwx-stack-NodeInstanceRole-xyz"
  },
  "Action": [
    "kms:Encrypt",
    "kms:Decrypt",
    "kms:GenerateDatakey",
    "kms:ReEncrypt*"
  ],
  "Resource": "*"
},
```

This code contains the following representations:

- 555555555555 is the example account id.
- env-id-dwx-stack-NodeInstanceRole-xyz is the example NodeInstanceRole for the environment.
 - env-id-dwx-stack is the example environment ID prefix.
 - xyz is a random string suffix

Accessing S3 buckets

In Cloudera Data Warehouse on cloudclusters running on AWS environments, you need to configure access to S3 buckets.

The subtopics of this document describe using the Cloudera Data Warehouse UI for configuring access to S3 buckets. This documentation does not apply if you are using Cloudera Data Warehouse as follows:

- In a RAZ (Ranger Authorized) environment
 - See [Accessing buckets in a RAZ environment](#).
- In a managed policy environment with a [managed policy ARN](#)
 - See [Accessing S3 buckets in a managed policy environment](#).
- In [reduced permissions mode](#)

In a RAZ or managed policy environment, or in reduced permissions mode, you use [AWS instance profiles](#) instead of the Cloudera Data Warehouse UI. The UI for configuring an external S3 bucket does not appear in these environments and modes.

Accessing buckets in the same AWS account

This topic explains how to configure read-only or read/write access using default encryption to external S3 buckets that reside in the same AWS account as the Cloudera Data Warehouse on cloud cluster.

About this task




Important: If you configure read-only access to an external S3 bucket, there is no need to restart Virtual Warehouses. However, if you configure read/write access to an external S3 bucket, you must restart Virtual Warehouses by suspending them and starting them again. Alternatively, you can create a new Virtual Warehouse to use the external S3 bucket with read/write access.

Required role: DWAdmin

Before you begin

- Identify and activate the environment you want to configure for access to an external bucket in the same AWS account.
- In the AWS Management Console, identify the external S3 bucket you want to configure access to.

Procedure

1. In the Cloudera Data Warehouse UI Overview, click Environments, choose the environment that is activated for the Virtual Warehouses you want to use with the external AWS bucket, and click  Edit .
2. In **Environment Details**, in the Enter s3 bucket name, type the name of the AWS bucket you want to configure access to.
3. Specify whether Read Only or Read Write access is needed.



Note:

For Read Write access only: if you do not need to use a custom encryption key, leave the ENCRYPTION SETTINGS text box blank. If you do not use a custom key, the system uses AES256 encryption by default. If you need to use a custom encryption key, see the next topic. Read Only access does not involve encryption.

4. Click Add Bucket to save the configuration. A success message displays at the top of the page.

What to do next

If you have configured Read Write access, you must restart the Virtual Warehouses that are associated with this environment for the configuration changes to take effect.

Accessing buckets using a custom key

If you want read/write access from Cloudera Data Warehouse on cloud on AWS to the external S3 bucket using your own custom encryption key, you must configure the encryption key.

About this task

Perform the steps described in this topic to use your own custom key.




Important: If you want to use a custom key, you must perform the configuration described in this topic whether the bucket is in the same AWS account as Cloudera Data Warehouse or if the bucket resides in a different AWS account.

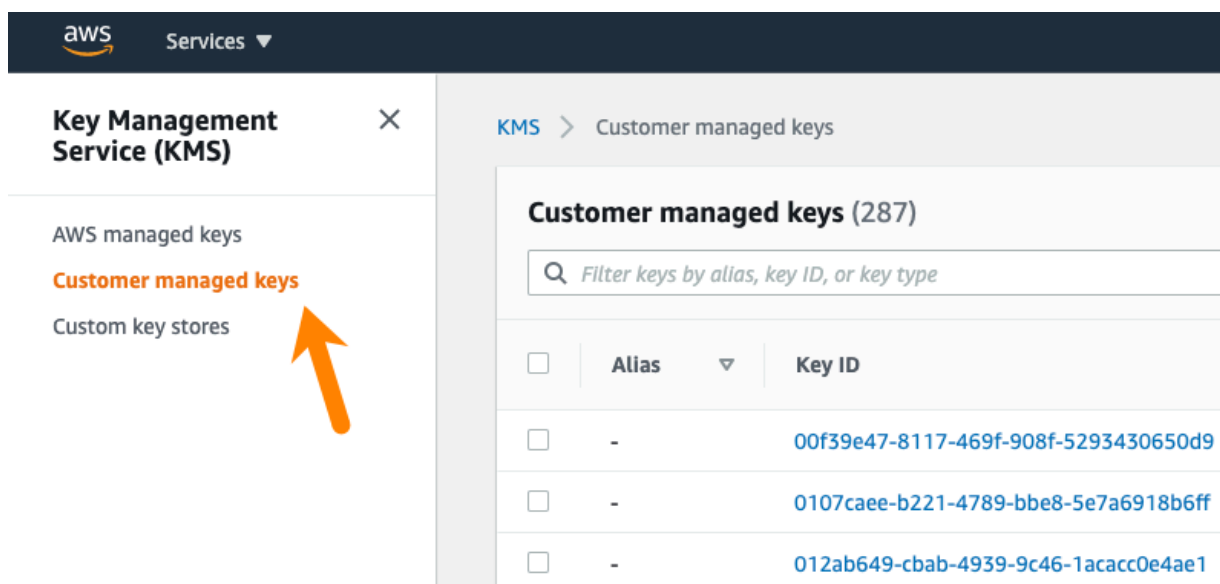
Required role: DWAdmin

Before you begin

- Identify which environment you want to configure for access to an external bucket in another AWS account. In the Cloudera Data Warehouse UI, go to the Environments tab. This causes the Data Catalog and Virtual Warehouses that use this environment to be highlighted in the Cloudera Data Warehouse UI. Choose the environment that is activated for the Virtual Warehouses you want to use with the external AWS bucket.
- In the AWS Management Console for the different account, identify the external S3 bucket you want to configure access to.

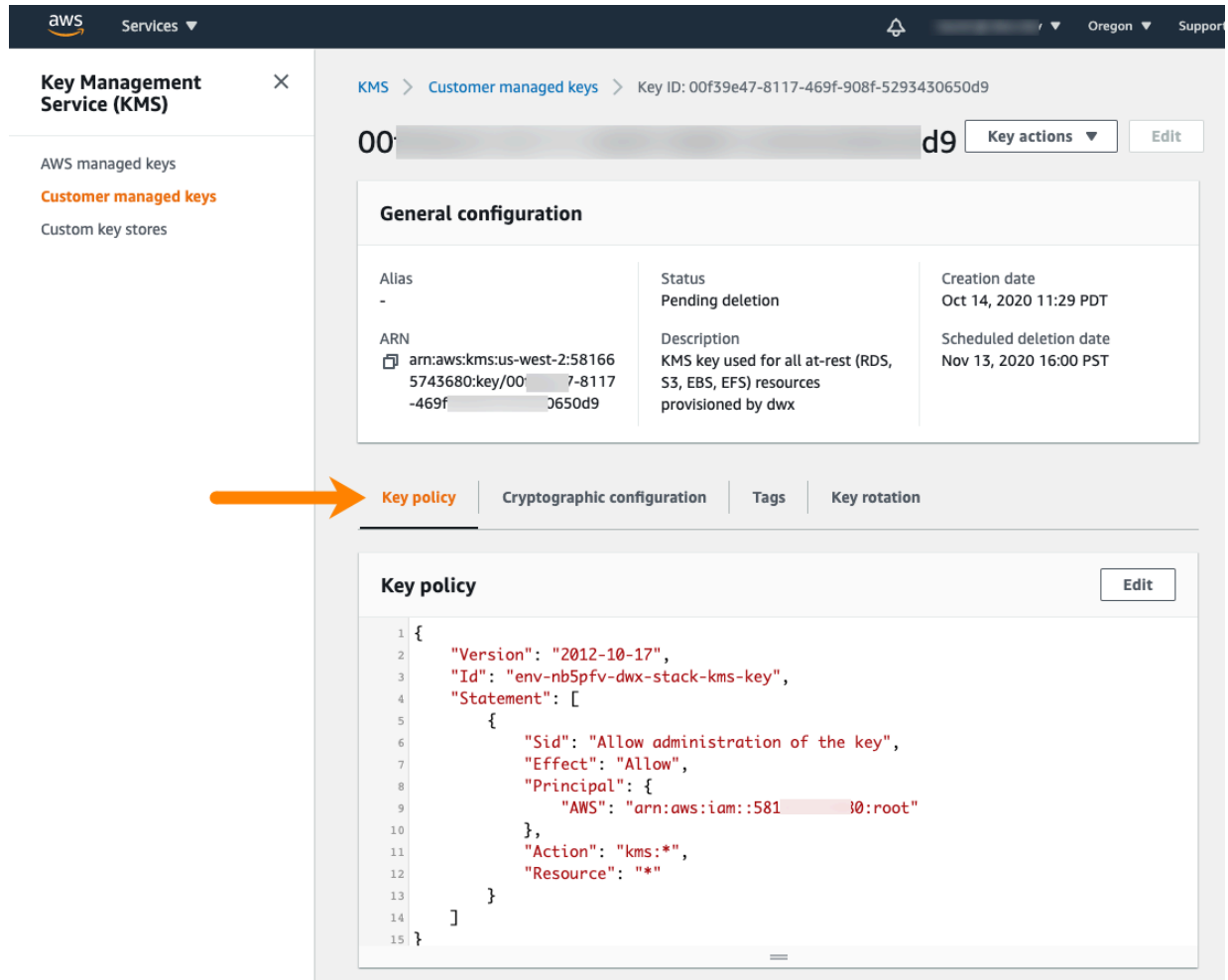
Procedure

- On the Cloudera Data Warehouse UI **Overview** page, go to the Environments tab and locate the environment for which you want to configure access to an external AWS bucket, and then click  **Edit** . This loads the **Environment Details** page.
- Go to the Configuration tab and type the name of the AWS bucket you want to configure access to in the Add External S3 Bucket field.
- (Optional) If you want to configure access to an S3 bucket in a different AWS account, select Bucket belongs to different AWS Account.
- Select Read Write. This causes the ENCRYPTION SETTINGS text box and a key policy to display.
- In the AWS Management Console for the account where the S3 bucket resides, navigate to the Key Management Service, and select Customer Managed Keys in the left navigation menu:



On the Customer managed keys page, select the key you want to use.

6. On the key details page, select the Key policy tab in the center panel of the page:



The screenshot shows the AWS KMS console interface. On the left, the 'Key Management Service (KMS)' sidebar is visible with options for 'AWS managed keys', 'Customer managed keys' (highlighted), and 'Custom key stores'. The main panel displays the details for a specific key (Key ID: 00f39e47-8117-469f-908f-5293430650d9). The 'General configuration' section shows the key's status as 'Pending deletion', creation date as 'Oct 14, 2020 11:29 PDT', and scheduled deletion date as 'Nov 13, 2020 16:00 PST'. Below this, the 'Key policy' tab is selected, indicated by an orange arrow. The 'Key policy' section shows a JSON policy document that allows administration of the key for the principal 'arn:aws:iam::581:root'. The policy document is as follows:

```
1 {
2   "Version": "2012-10-17",
3   "Id": "env-nb5pfv-dwx-stack-kms-key",
4   "Statement": [
5     {
6       "Sid": "Allow administration of the key",
7       "Effect": "Allow",
8       "Principal": {
9         "AWS": "arn:aws:iam::581:root"
10      },
11      "Action": "kms:*",
12      "Resource": "*"
13    }
14  ]
15 }
```

This displays the key policy for the customer managed key.

7. In the Cloudera Data Warehouse Environment Details page, copy the Amazon Resource Names (ARNs) associated with the environment that displays in the panel:

Add External S3 Bucket:

☐ Bucket belongs to different AWS Account

ACCESS MODE:

☐ Read Only ☒ Read Write

ENCRYPTION SETTINGS: ⓘ

Please update source encryption key policy to include the following. Refer this [link](#) for additional instructions.

```
{
  "Sid": "encrypt-decrypt-objects-for-cdw-env-xspjph",
  "Effect": "Allow",
  "Principal": {
    "AWS": [
      "arn:aws:iam::123014800043:role/func-qe-weekly-role",
      "arn:aws:iam::123014800043:role/env-xspjph-dwx-stack-NodeInstanceRole-1BR2OOVUXZH9F"
    ]
  },
  "Action": [
    "kms:Encrypt",
    "kms:Decrypt",
    "kms:ReEncrypt*",
    "kms:GenerateDataKey*",
    "kms:DescribeKey"
  ],
  "Resource": "*"
}
```

Copy these 2 roles that are associated with this environment to your system clipboard.

ADD BUCKET

The actions listed in the above screen capture are the minimum set of actions needed by Cloudera Data Warehouse:

```
"Action": [
  "kms:Encrypt",
  "kms:Decrypt",
  "kms:ReEncrypt*",
  "kms:GenerateDataKey*",
  "kms:DescribeKey"
]
```

The key policy you use should allow at least these actions.

- Return to the key details page in the AWS Management Console, click Edit in the upper right corner of the Key policy tab, paste the two ARNs to append them after the existing ARNs in the key policy, and then click Save changes:

Edit key policy

Key policy

Paste the 2 ARNs under the existing ARNs in the key policy, and then click "Save changes."

```

1 {
2   "Version": "2012-10-17",
3   "Id": "env-nb5pfv-dwx-stack-kms-key",
4   "Statement": [
5     {
6       "Sid": "Allow administration of the key",
7       "Effect": "Allow",
8       "Principal": {
9         "AWS": "arn:aws:iam::581665743680:root",
10        "arn:aws:iam::123014800043:role/func-qe-weekly-role":
11        "arn:aws:iam::123014800043:role/env-xspjph-dwx-stack-NodeInstanceRole-1BR200VUXZH9F"
12      },
13      "Action": "kms:*",
14      "Resource": "*"
15    }
16  ]
17 }

```

Cancel Save changes

The key policy shown in the above screen capture lists "kms:*" in the "Action" section of the policy. This indicates that all actions are allowed. At minimum, Cloudera Data Warehouse requires the "Encrypt", "Decrypt", "ReEncrypt", "GenerateDataKey*", and the "DescribeKey" actions as shown on the screen capture in Step 7.

If there is no Key Policy tab of the key details page, copy and paste the entire key policy in the AWS Management Console from the Cloudera Data Warehouse UI.

- After you save the changes to the key policy in the AWS Management Console, copy the ARN from the General configuration section of the key details page:

General configuration

Alias	ARN	Status
-	arn:aws:kms:us-west-2:581665743680:key/01ac13e1-96da-4b55-94e1-e9c232928bcd	Per

10. In the Cloudera Data Warehouse Environment Details page, add the ARN you copied in Step 9 to the ENCRYPTION SETTINGS text box:

Add External S3 Bucket:

☐ Bucket belongs to dif

ACCESS MODE:
☐ Read Only ☒ Read Write

ENCRYPTION SETTINGS: ⓘ

Please update source encryption key policy to include the following. Refer t

```
{
  "Sid": "encrypt-decrypt-objects-for-cdw-env-xspjph",
  "Effect": "Allow"
```

**Paste the ARN from
the AWS Management
Console to CDW.**



11. Click Add bucket to save the configuration. A success message displays at the top of the page.

What to do next

You must restart the Virtual Warehouses that are associated with this environment for the configuration changes to take effect.

Accessing buckets in a different AWS account

You must configure read-only or read/write access using default encryption to external S3 buckets in AWS accounts that are different from the Cloudera Data Warehouse cluster account.

About this task





Important: If you configure read-only access to an external S3 bucket, there is no need to restart Virtual Warehouses. However, if you configure read/write access to an external S3 bucket, you must restart Virtual Warehouses by suspending them and starting them again. Alternatively, you can create a new Virtual Warehouse to use the external S3 bucket with read/write access.

Required role: DWAdmin

Before you begin

- Identify and activate the environment you want to configure for access to an external bucket in a different AWS account.
- In the AWS Management Console, identify the external S3 bucket you want to configure access to.

Procedure

1. On the Cloudera Data Warehouse UI **Overview** page, go to the Environments tab and locate the environment for which you want to configure access to an external AWS bucket, and then click  Edit .
This loads the **Environment Details** page.
2. Go to the Configuration tab and type the name of the AWS bucket you want to configure access to in the Add External S3 Bucket field.
3. Select Bucket belongs to different AWS Account. The Cloudera Data Warehouse bucket policy appears.
4. Click Copy  .

Add External S3 Bucket:

☒ Bucket belongs to different AWS Account


ACCESS MODE:
☒ Read Only ☐ Read Write

Please update source bucket policy to include the following. Refer this [link](#) for additional instructions.

```

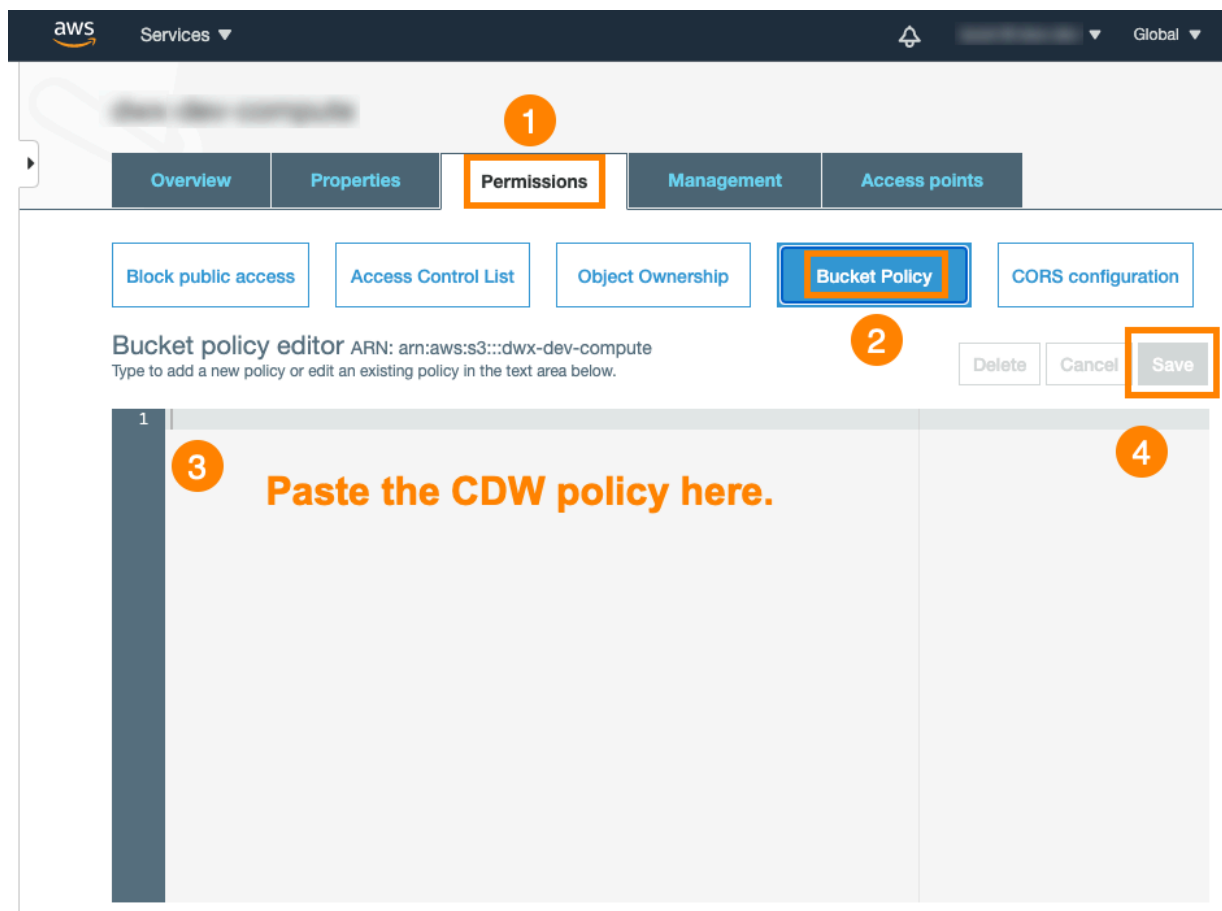
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "read-write-access-for-cdw-env-xspjph",
      "Effect": "Allow",
      "Principal": {
        "AWS": [
          "arn:aws:iam::123014800043:role/func-qe-weekly-role",
          "arn:aws:iam::123014800043:role/env-xspjph-dwx-stack-NodeInstanceRole-1BR2OOVUXZH9F"
        ]
      },
      "Action": [
        "s3:Get*",
        "s3:ListBucket",
        "s3:GetBucketLocation"
      ],
      "Resource": [
        "arn:aws:s3:::foo",
        "arn:aws:s3:::foo/*"
      ]
    }
  ]
}

```



5. Open the AWS Management Console for the different account where the external bucket is located and navigate to the bucket to which you want to configure access.

- On the bucket details page of AWS Management Console, click the Permissions tab, click Bucket Policy, paste the policy from Cloudera Data Warehouse, and click Save:



- In the Cloudera Data Warehouse UI Environment Details page, specify either Read Only or Read Write access for the external bucket.



Note:

For Read Write access only: if you do not need to use a custom encryption key, leave the ENCRYPTION SETTINGS text box blank. If you do not use a custom key, the system uses AES256 encryption by default. If you need to use a custom encryption key, see the previous topic. Read Only access does not involve encryption.

- Click Add Bucket to save the configuration. A success message displays at the top of the page.

What to do next

If you have configured Read Write access, you must restart the Virtual Warehouses that are associated with this environment for the configuration changes to take effect.

Accessing S3 buckets in a managed policy environment

You can add and query data in S3 external buckets you add to Cloudera Data Warehouse service clusters running on AWS environments.

This topic and the following subtopics are applicable if you are using managed policy ARN [feature](#). If you are using RAZ enabled DW, see "Accessing buckets in a RAZ environment".

If you are not using a RAZ-enabled Cloudera Data Warehouse, access to S3 buckets is controlled by the managed policy attached to the [AWS instance profiles](#). To access S3 buckets you add to your Cloudera Data Warehouse service

cluster, you must edit the instance profile to add read/write permissions to the additional buckets as described in the following subtopics.

Related Information

[Accessing S3 buckets in a RAZ environment](#)

Accessing buckets in the same AWS account under a managed policy

In certain scenarios, you might need to interact with data that resides outside of the data lake S3 buckets. You can add a bucket to S3, enable access to the bucket, and then, define external tables based on the data, such as a CSV file, you put into the bucket.

About this task

In this task, you see how to add access to an S3 bucket to [create an external table based on CSV data using Hue](#). You need to enable read/write access to the external S3 bucket before creating the table. From the command line of your cluster, you can run HDFS CLI commands on the S3 bucket. You can also use the S3 bucket for [uploading a UDF jar for registration](#), and then include UDFs in queries from your cluster.



Note: This topic might describe versions of the AWS console that have changed over time.

Before you begin

- The S3 bucket you add to hold the data outside your Data Lake must be in the same AWS account as your Cloudera Data Warehouse service cluster.
- Required role: DWAdmin

About this task

In this task, first you note the Managed Policy ARN attached to the Node Instance Role used while activating the cluster. Next, you edit the managed policy in the JSON file, for example `noderole-inline-policy.json`.

Procedure

1. In your [managed policy](#), locate the sid "putgetmybucketpaths" for editing.

```
"Sid": "putgetmybucketpaths",
"Action": [
    "s3:Get*",
    "s3:Delete*",
    "s3:Put*",
    "s3:ListBucketMultipartUploads",
    "s3:AbortMultipartUpload"
],
"Resource": [
    "arn:aws:s3:::roohi-dwx-priv/clusters",
    "arn:aws:s3:::roohi-dwx-priv/clusters/*",
    "arn:aws:s3:::roohi-dl-bucket/clusters",
    "arn:aws:s3:::roohi-dl-bucket/clusters/*",
    "arn:aws:s3:::roohi-dwx-priv/logs",
    "arn:aws:s3:::roohi-dwx-priv/logs/*",
    "arn:aws:s3:::roohi-backup/backup",
    "arn:aws:s3:::roohi-backup/backup/*",
    "arn:aws:s3:::roohi-dl-bucket/data",
    "arn:aws:s3:::roohi-dl-bucket/data/*",
    "arn:aws:s3:::roohi-dl-bucket/backup",
    "arn:aws:s3:::roohi-dl-bucket/backup/*",
    "arn:aws:s3:::roohi-dl-bucket/tmp",
    "arn:aws:s3:::roohi-dl-bucket/tmp/*"
],
"Effect": "Allow"
```


2. Append resources to the resource section for the buckets you added.
For example, you added a bucket more-sales-data. To enable access to the more-sales-data bucket, you append resources to the end of the "resource" section, as shown in the last two resource names:

```
"Resource" : [
...
"arn:aws:s3:::roohi-dl-bucket/backup/*",
"arn:aws:s3:::more-sales-data",
"arn:aws:s3:::more-sales-data/*"
],
```



Important: There is no comma after the last line of the "Resource" section.

3. Click Review policy in the lower right corner of the page, and then click Save changes.
You can now access the more-sales-data bucket outside your data lake from Hue in your Cloudera Data Warehouse service cluster. For example, you can create external Hive tables that point to the bucket, and join those external tables with tables already in your data lake. You can govern Cloudera Data Warehouse user access to this external S3 bucket using Ranger Hadoop SQL Policies.

Accessing buckets in a different AWS account under a managed policy

You might need to know how to add read/write access to S3 buckets under AWS accounts that are different from the Cloudera Data Warehouse cluster account.

About this task

To enable Cloudera Data Warehouse service cluster access to a bucket you add to S3 under a different AWS account, you must configure the bucket in the different account to access the Cloudera Data Warehouse cluster account. Then, you can configure the Cloudera Data Warehouse service account to access the bucket you added. You perform both of these tasks in the AWS Management Console.

Required role: DWAdmin



Note: This topic might use both "new" and "old" versions of the AWS console.

Before you begin

To configure access to external S3 buckets for your Cloudera Data Warehouse cluster, you must edit the managed policy attached to the [AWS instance profile](#).

You use this cluster ID in Step 5 below.

Procedure

1. In the AWS Console, navigate to AWS Management Console S3, locate the bucket in the other AWS account you added, and then click the bucket name.
2. In the bucket details page, click the Permissions tab, and then click the Bucket Policy sub-tab.

3. In the Bucket Policy sub-tab page, in the Bucket policy editor, add the Cloudera Data Warehouse cluster Id and what permissions you want the Cloudera Data Warehouse service account to have for this bucket:

Amazon S3 > airlines-orc-dwx-dev

airlines-orc-dwx-dev

Overview Properties Permissions Management Access points

Block public access Access Control List **Bucket Policy** CORS configuration


Bucket policy editor ARN: arn:aws:s3::airlines-orc-dwx-dev
Type to add a new policy or edit an existing policy in the text area below. Delete Cancel Save

The block public access settings turned on for this bucket prevent granting public access.

```

1  {
2    "Version": "2012-10-17",
3    "Statement": [
4      {
5        "Sid": "read-access-for-dwx-env-cx57qr",
6        "Effect": "Allow",
7        "Principal": {
8          "AWS": "arn:aws:iam::581665743680:role/env-cx57qr-dwx-stack-NodeInstanceRole-1J37ID398QP8K"
9        },
10       "Action": [
11         "s3:Get*",
12         "s3:ListBucket"
13       ],
14       "Resource": [
15         "arn:aws:s3::airlines-orc-dwx-dev",
16         "arn:aws:s3::airlines-orc-dwx-dev/*"
17       ]
18     }
19   ]
20 }
```

This example policy includes the following specifications:

-  **Tip:** To get the ARN of your Cloudera Data Warehouse cluster account: In the AWS Console, navigate to CloudFormation and locate the stack for your Cloudera Data Warehouse cluster. Click the stack name. In the stack details page, click the Resources tab. In the Resources table, scroll down to the NodeInstanceRole, and then click the hyperlink just to the right of it. At the top of the Summary page, the Role ARN is listed. This is the ARN you must specify for the Principal in the bucket policy.
- The Action section specifies what actions the Principal can perform.
- The Resource section specifies the S3 bucket you added and want your Cloudera Data Warehouse cluster to be able to access.

For details about bucket policies, see [Managing Access to Amazon S3 Buckets Using Bucket Policies](#) in the AWS documentation.

4. Click Save.
5. Note the managed policy ARN attached to the Node Instance Role, used while activating the cluster.
6. Open the managed policy JSON file, for example noderole-inline-policy.json for editing

7. Locate the sid "putgetmybucketpaths" for editing.

```
"Sid": "putgetmybucketpaths",
"Action": [
    "s3:Get*",
    "s3:Delete*",
    "s3:Put*",
    "s3:ListBucketMultipartUploads",
    "s3:AbortMultipartUpload"
],
"Resource": [
    "arn:aws:s3:::roohi-dwx-priv/clusters",
    "arn:aws:s3:::roohi-dwx-priv/clusters/*",
    "arn:aws:s3:::roohi-dl-bucket/clusters",
    "arn:aws:s3:::roohi-dl-bucket/clusters/*",
    "arn:aws:s3:::roohi-dwx-priv/logs",
    "arn:aws:s3:::roohi-dwx-priv/logs/*",
    "arn:aws:s3:::roohi-backup/backup",
    "arn:aws:s3:::roohi-backup/backup/*",
    "arn:aws:s3:::roohi-dl-bucket/data",
    "arn:aws:s3:::roohi-dl-bucket/data/*",
    "arn:aws:s3:::roohi-dl-bucket/backup",
    "arn:aws:s3:::roohi-dl-bucket/backup/*",
    "arn:aws:s3:::roohi-dl-bucket/tmp",
    "arn:aws:s3:::roohi-dl-bucket/tmp/*"
],
"Effect": "Allow"
```

8. Append resources to the resource section for the buckets you added.
For example, if you want to add access to the more-sales-data bucket, you append resources to the end of the "resource" section, as shown in the last two resource names:

```
"Resource" : [
  ...
  "arn:aws:s3:::roohi-dl-bucket/backup/*" ,
  "arn:aws:s3:::more-sales-data" ,
  "arn:aws:s3:::more-sales-data/*"
],
```



Important: There is no comma after the last line of the "Resource" section.

9. Click Review policy in the lower right corner of the page, and then click Save changes. You can access the new bucket from your Cloudera Data Warehouse service cluster now. For example, you can create external Hive tables that point to the bucket.

Accessing S3 buckets in a RAZ environment

In a RAZ (Ranger Authorized) environment, you must configure permissions to access an S3 bucket. The procedures for configuring the permissions differ depending on the AWS account that owns the bucket.

If you have enabled RAZ in your environment, policies attached to the Ranger RAZ Service role control access to external S3 buckets.

Prerequisites

You must meet the following prerequisites before adding access permissions to buckets to the RAZ environment within the same AWS account or in a different account:

- Obtain the DWAdmin role.
- Follow steps similar to the [minimum setup for cloud storage](#) to create the Ranger RAZ role.
- [Register an environment with RAZ](#) using the Cloudera web interface.

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

Accessing buckets in the same AWS account under RAZ

In a Ranger Authorized (RAZ) environment, to access an S3 external bucket in the same AWS account as the cluster, you must configure a policy that manages access to the Cloudera data lake. You modify a JSON file that defines the IAM policy to do this.

About this task

You update the following IAM policy definition for the minimal cloud storage setup: [aws-cdp-datalake-admin-S3-policy](#).

Before you begin

You must meet the prerequisites mentioned in the topic above.

Procedure

Add the ARN of the external bucket to the "Resource" array of values in the [JSON file](#) for the [aws-cdp-datalake-admin-S3-policy](#).

For example, the values in the Resource array give the Cloudera Data Warehouse cluster access to external bucket MY_EXTERNAL_BUCKET.

```
...
    "s3:ListBucketMultipartUploads",
    "s3:ListBucketVersions",
    "s3:ListMultipartUploadParts",
    "s3:PutObject"
  ],
  "Resource": [
    "arn:${ARN_PARTITION}:s3:::${STORAGE_LOCATION_BASE}",
    "arn:${ARN_PARTITION}:s3:::${STORAGE_LOCATION_BASE}/*",
    "arn:${ARN_PARTITION}:s3:::MY_EXTERNAL_BUCKET",
    "arn:${ARN_PARTITION}:s3:::MY_EXTERNAL_BUCKET/*"
  ]
}
```

Accessing buckets in a different AWS account under RAZ

In a Ranger Authorized (RAZ) environment, to access an S3 external bucket in a different AWS account from the Cloudera Data Warehouse cluster, you must configure the bucket policy of the other account. You use the AWS Management Console to do this.

About this task

For information about bucket policies, see [Adding a bucket policy by using the Amazon S3 console](#) in the AWS documentation.

Before you begin

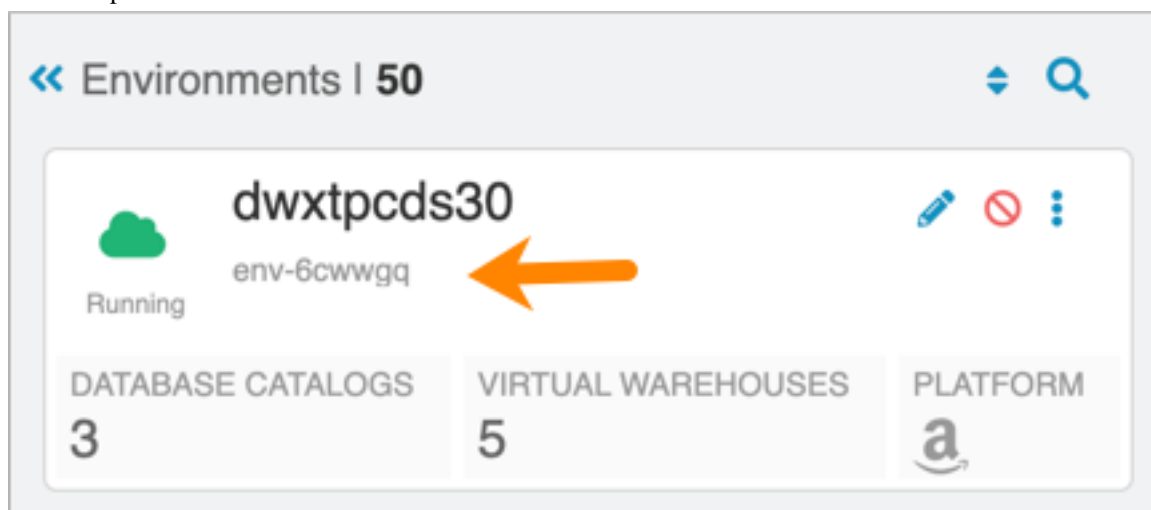
You must meet the prerequisites mentioned in the topic above.

Procedure

1. Add the ARN of the external bucket to the "Resource" array of values in the [JSON file](#) for the aws-cdp-datalake-admin-s3-policy.
For example, the values in the Resource array give the Cloudera Data Warehouse cluster access to external bucket MY_EXTERNAL_BUCKET.

```
...
    "s3:ListBucketMultipartUploads",
    "s3:ListBucketVersions",
    "s3:ListMultipartUploadParts",
    "s3:PutObject"
  ],
  "Resource": [
    "arn:${ARN_PARTITION}:s3:::${STORAGE_LOCATION_BASE}",
    "arn:${ARN_PARTITION}:s3:::${STORAGE_LOCATION_BASE}/*",
    "arn:${ARN_PARTITION}:s3:::MY_EXTERNAL_BUCKET",
    "arn:${ARN_PARTITION}:s3:::MY_EXTERNAL_BUCKET/*"
  ]
}
```

2. Get the cluster ID from the Environments tile in the Cloudera Data Warehouse service UI.
For example:



You use this cluster ID later.

3. In the AWS Console, navigate to AWS Management Console S3, locate the bucket in the other AWS account you added, and then click the bucket name.
4. In the bucket details page, click Permissions, and then click Bucket Policy.

5. In Bucket Policy, in the Bucket policy editor, add the Cloudera Data Warehouse cluster Id and the permissions you want the Cloudera Data Warehouse service account to have to this bucket.

The screenshot shows the AWS Management Console interface for the bucket 'airlines-orc-dwx-dev'. The 'Bucket Policy' tab is selected, and the 'Bucket Policy editor' is open. The policy editor shows a JSON policy document with the following content:

```

1 {
2   "Version": "2012-10-17",
3   "Statement": [
4     {
5       "Sid": "read-access-for-dwx-env-cx57qr",
6       "Effect": "Allow",
7       "Principal": {
8         "AWS": "arn:aws:iam::581665743680:role/env-cx57qr-dwx-stack-NodeInstanceRole-1J37ID398QP8K"
9       },
10      "Action": [
11        "s3:Get*",
12        "s3:ListBucket"
13      ],
14      "Resource": [
15        "arn:aws:s3:::airlines-orc-dwx-dev",
16        "arn:aws:s3:::airlines-orc-dwx-dev/*"
17      ]
18    }
19  ]
20 }
```

This example policy includes the following specifications:

- This first section includes the Sid, which is an optional identifier indicating what the policy does. The Effect specifies that this policy is allowing the Principal to do what is listed below in the Action section. The Principal is where you specify the ARN of the instance role for your Cloudera Data Warehouse cluster account.



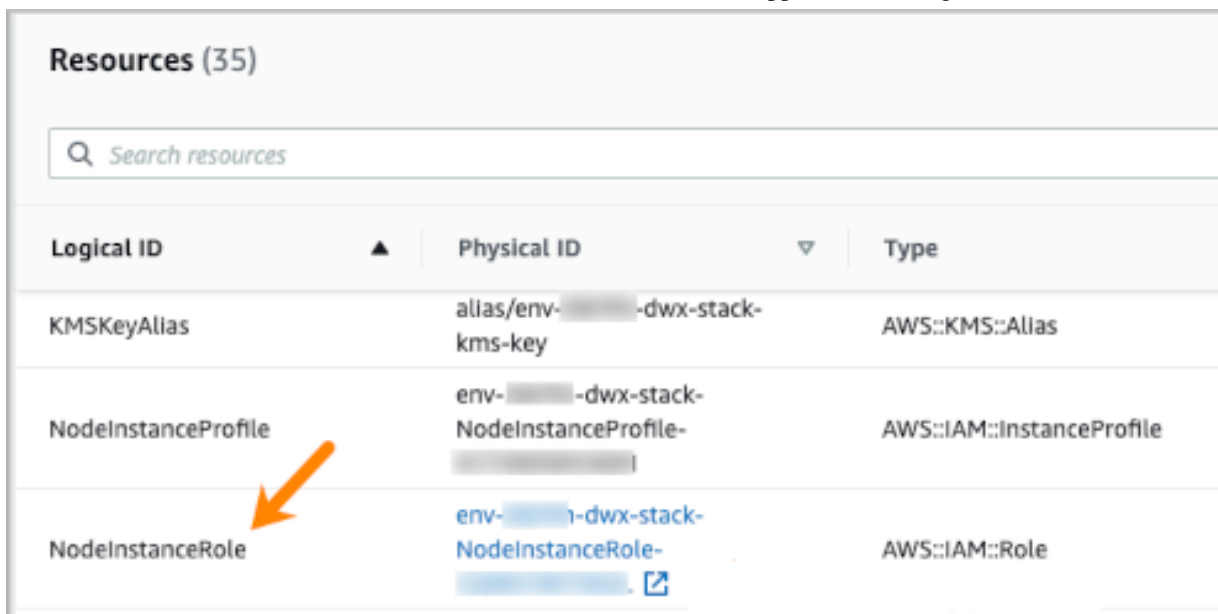
Tip: To get the ARN of your Cloudera Data Warehouse cluster account: Follow steps 6-8 below, then specify the ARM for the Principal in the bucket policy.

- The Action section specifies what actions the Principal can perform.
- The Resource section specifies the S3 bucket you added that your Cloudera Data Warehouse cluster will access.

6. Navigate to AWS Management Console > CloudFormation and locate the stack corresponding to the cluster ID.

- Click the CloudFormation stack name. This stack name is the one in this format: <cluster-ID>-dwx-stack. For example, if the cluster ID is env-6cwwgg, the CloudFormation stack name for this cluster is env-6cwwgg-dwx-stack.

In CloudFormation stack details, in Resources, the NodeInstanceRole appears in the Logical ID column.



Logical ID	Physical ID	Type
KMSKeyAlias	alias/env- -dwx-stack-kms-key	AWS::KMS::Alias
NodeInstanceProfile	env- -dwx-stack-NodeInstanceProfile- 	AWS::IAM::InstanceProfile
NodeInstanceRole	env- -dwx-stack-NodeInstanceRole- 	AWS::IAM::Role

- Click the hyperlink just to the right of it.
At the top of the Summary page, the Role ARN is listed.
- Specify the ARN for the Principal in the bucket policy.
- Click Save.

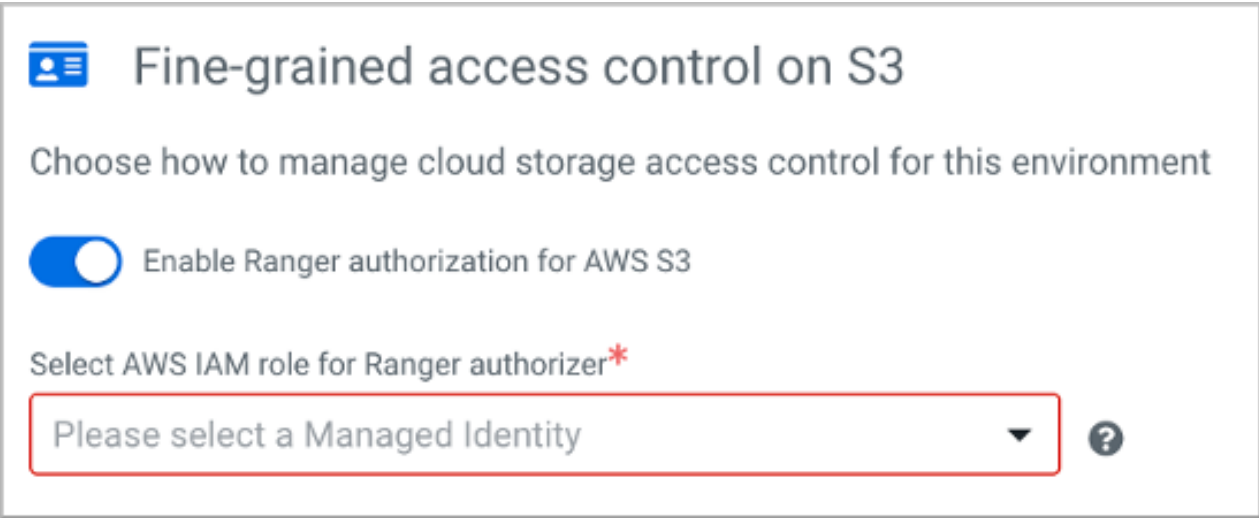
Enabling RAZ manually

You might want to manually enable Ranger Authorization (RAZ) if you have a Cloudera Data Warehouse that predates the capability to enable RAZ for AWS S3. In this case, if you need Ranger authorization in Hue, perform this task.

About this task

For an introduction to enabling RAZ, see the [Cloudera Management Console documentation](#).

When you activate your environment in 1-6.3-b319 (released April 5, 2023) or later, you configure S3 access using the Fine-grained access control on S3 dialog in the Cloudera Management Console UI. If your Data Lake is RAZ-enabled (Enable Ranger authorization for AWS S3), Cloudera Data Warehouse is RAZ-enabled by default, and cannot be turned off. Cloudera Data Warehouse will be RAZ-enabled to provide authorization mainly in Hue. There is nothing more you need to do.



If you activated your Cloudera Data Warehouse environment in 1-6.2-b197 (released Feb 14, 2023) or earlier, your Cloudera Data Warehouse is not RAZ-enabled (Enable Ranger authorization for AWS S3 is not enabled). You can manually enable RAZ for Cloudera Data Warehouse only if the Data Lake is RAZ-enabled. You follow the steps below to manually enable RAZ for AWS S3 mainly for authorization of Hue users.

Before you begin

Your Data Lake must be RAZ-enabled before you can perform these steps.

Procedure

1. Obtain a list of the permissions policies related to the AWS IAM role for Ranger authorizer you created on AWS.

In some cases, this role is interchangeable with the `DATA_LAKE_ADMIN` role for AWS S3. For information about this role and policies, see ["Required IAM resources"](#).

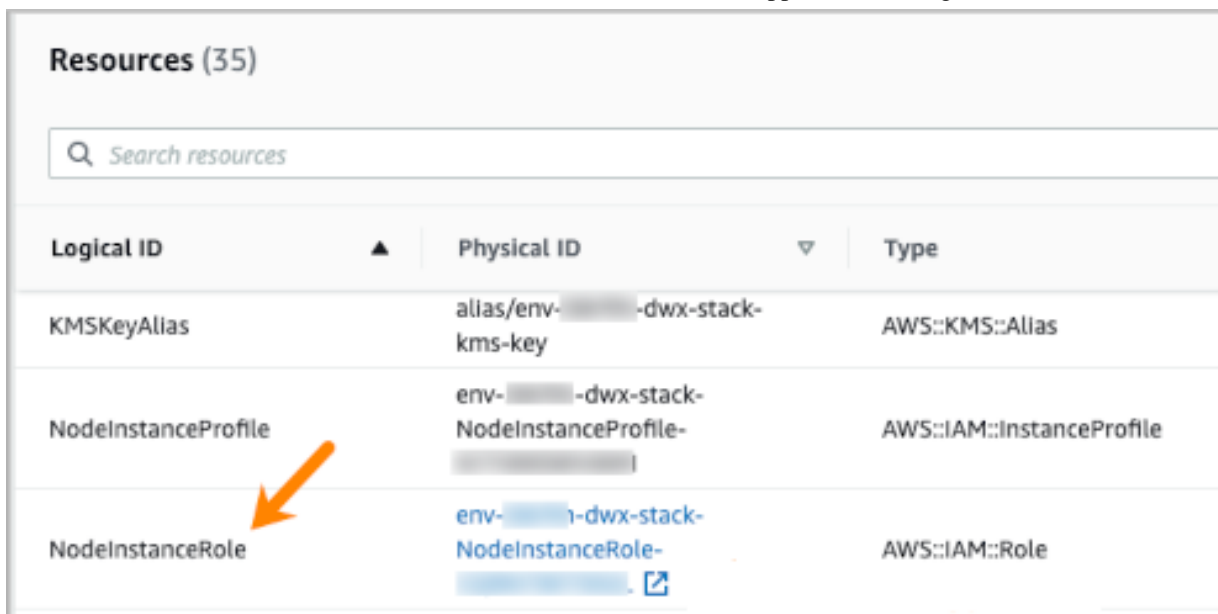
Minimal permissions policies are:

- `aws-cdp-datalake-admin-s3-policy`
- `aws-cdp-bucket-access-policy`
- `aws-datalake-backup-policy`
- `aws-datalake-restore-policy`

2. Get the cluster ID from the Environments tile in the Cloudera Data Warehouse service UI.
3. Navigate to `AWS Management Console CloudFormation` and locate the stack corresponding to the cluster ID.

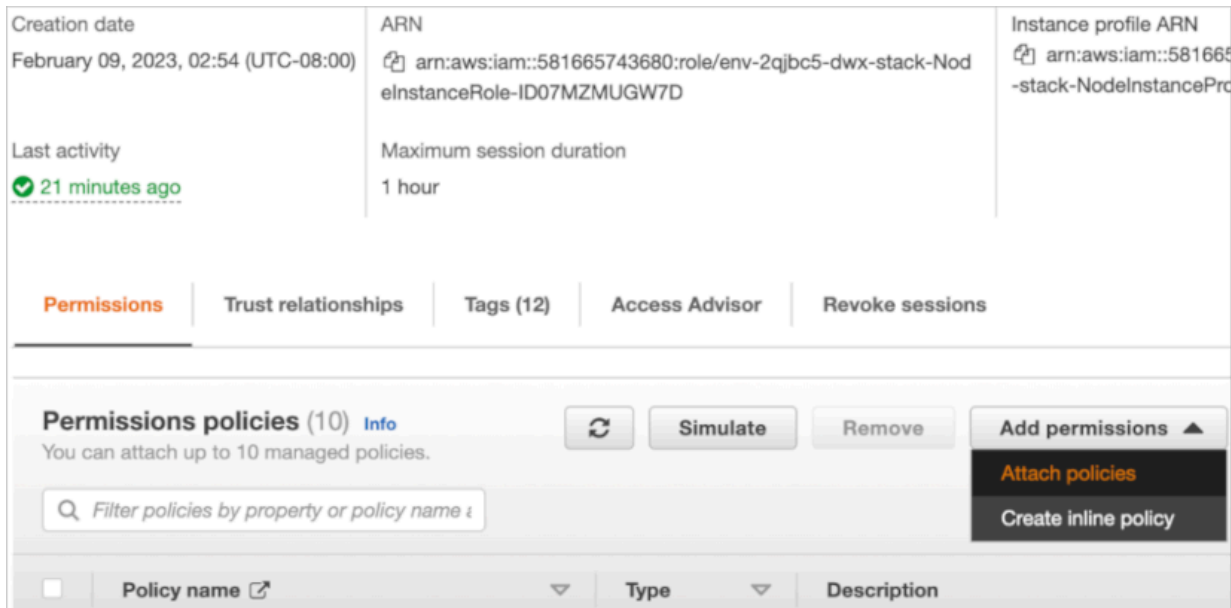
- Click the CloudFormation stack name. This stack name is the one in this format: <cluster-ID>-dwx-stack. For example, if the cluster ID is env-6cwwgg, the CloudFormation stack name for this cluster is env-6cwwgg-dwx-stack.

In CloudFormation stack details, in Resources, the NodeInstanceRole appears in the Logical ID column.



Logical ID	Physical ID	Type
KMSKeyAlias	alias/env- -dwx-stack-kms-key	AWS::KMS::Alias
NodeInstanceProfile	env- -dwx-stack-NodeInstanceProfile- 	AWS::IAM::InstanceProfile
NodeInstanceRole	env- -dwx-stack-NodeInstanceRole- 	AWS::IAM::Role

- Click the NodeInstanceRole link.
The Node instance Role page appears.
- Click **Add Permissions Attach Policies**, and on the next page, in Permissions Policies, select the policies from step 1.



Creation date	ARN	Instance profile ARN
February 09, 2023, 02:54 (UTC-08:00)	arn:aws:iam::581665743680:role/env-2qjbc5-dwx-stack-NodeInstanceRole-ID07MZMUGW7D	arn:aws:iam::581665743680:instance-profile/env-2qjbc5-dwx-stack-NodeInstanceProfile-ID07MZMUGW7D

Last activity: ✔ 21 minutes ago

Maximum session duration: 1 hour

Permissions | Trust relationships | Tags (12) | Access Advisor | Revoke sessions

Permissions policies (10) [Info](#)

You can attach up to 10 managed policies.

[Attach policies](#) | [Create inline policy](#)

Policy name	Type	Description
-------------	------	-------------

Remote access

You can follow a few steps to grant users access to Cloudera Data Warehouse service Kubernetes clusters on Amazon EKS. A step-by-step procedure explains how to restrict access to Kubernetes endpoints and service endpoints of the Kubernetes cluster.

Granting remote access to Kubernetes clusters on Amazon EKS

This topic describes how you can grant users access to Cloudera Data Warehouse service Kubernetes clusters on Amazon EKS.

About this task

To grant remote access to Kubernetes clusters on Amazon EKS, add the Amazon Resource Name identifiers (ARNs) to the list of trusted users.

Required role: DWAdmin

Before you begin

- You must activate an environment before you can grant users access to the Kubernetes cluster.
- Contact your AWS account administrator or the user who is requesting access to the Kubernetes cluster on AWS to get the ARN identifier for their Amazon account:

Using the Amazon Management Console

1. On the Amazon Management Console home page, enter IAM in Find Services, and then select IAM in the search results.
2. On the Identity and Access Management home page, in the left navigation menu, select Users.
3. On the User page, locate the user and click or tap their User name.
4. On the Summary page for the user, their User ARN is listed at the top. Copy it to paste into the Data Warehouse service UI in Step 4 below.

Using the AWS CLI

As an alternative to copying the User ARN from the AWS Management Console, you can also ask the user who is requesting access to enter the following command in the AWS CLI:

```
aws sts get-caller-identity
```


This command evokes the AWS Security Token Service (sts) and returns the following type of information on the requesting user:

```
#Sample output
{
  "UserId": "ABCDE12345FGHIJKLMNO6789",
  "Account": "888888888888",
  "Arn": "arn:aws:iam::888888888888:user/<USERNAME>"
}
```

A link to the Amazon documentation on this command is available at the bottom of this page.

Procedure

1. Log in to the Cloudera Data Warehouse service as DWAdmin.
2. Go to the Environments tab from the **Overview** page.

3. Click  Edit GROUP ACCESS corresponding to the environment for which you want to grant access to ARNs.

Kubeconfig

```
apiVersion: v1
clusters:
- cluster:
    certificate-authority-data:
LS0tLS1CRUdJTiBDRVJUSUZJQ0FURSB0tLS0tCk1JSUN5R
ENDQWJDZ0F3SUJBZ01CQURBTklna3Foa2lHOXcwQkFRc0
ZBREFTVJNd0VRWURWUWFERXdcmRXSmwKY201bGRHVnp
NQjRFRFRjd01ERXlPVEl4TURMElGblhEVE13TURFeU5q
SXhNRE0wTUZvd0ZURVRNQGkVHQTFVRQpBeE1LYTNWaVpYS
nVaWFJsY3pDQ0FTSXdEUVlKS29aSWh2Y05BUUVCQ1FBRG
dnRVRBRENDQVFrO2dnRUlBTDBhaC11SVnnGQ3FCVExn2Hd
```

COPY

HIDE

4. Enter the ARN identifier under Add new group and click Grant Access.

Related Information

[Activating environments](#)

[Amazon documentation on the AWS CLI command 'aws sts get-caller-identity'](#)

Revoking remote access to Kubernetes clusters on Amazon EKS


This topic describes how to revoke remote access to Kubernetes clusters for Cloudera Data Warehouse on cloud.

About this task

Revoke users' access to Cloudera Data Warehouse service Kubernetes clusters on Amazon EKS by deleting their Amazon Resource Name identifiers (ARNs) from the list of trusted users.

Required role: DWAdmin

Procedure

1. In the Cloudera Data Warehouse service, go to the Environments tab.
2. Locate the environment for which you want to revoke user's access and click  Edit GROUP ACCESS .
The Amazon Resource Name (ARN) identifiers that have been granted access to the Kubernetes cluster are displayed.
3. Click Remove Access to revoke user access.
4. Click Apply Changes.

Restricting access to endpoints in AWS environments

This topic describes how to limit access to Hive, Impala, Data Analytics Studio, or Hue endpoints in Cloudera Data Warehouse on cloud.

About this task

You can restrict access to Kubernetes endpoints and load balancer endpoints of the Kubernetes cluster by specifying a list of IP Classless Inter-Domain Routing (CIDRs) that are allowed access. Kubernetes endpoints are used to control the deployment and maintenance of workload components, such as Virtual Warehouses and Database Catalogs. Load balancer endpoints are endpoints of services like Hive, Impala, or Hue. You can specify trusted IP addresses when you activate a on cloud environment to use in the Data Warehouse service or in the Environment Details page.

Otherwise, all external IP addresses can access these endpoints on the Kubernetes cluster that is being used in the Cloudera Data Warehouse service.

Required role: DWAdmin

Before you begin

Contact your network team to get your internal network's IP CIDR ranges of IP addresses that need access to Kubernetes and load balancer service endpoints. All Cloudera IP addresses that need access to these endpoints have already been allowed.

Procedure

1. In the Cloudera Data Warehouse service, in Environments, search for and locate the environment that you want to specify CIDRs.
2. Click Activate to activate the environment.
3. In Activation Settings Advanced Settings, in Enable IP CIDR for Kubernetes cluster specify a comma-separated list of IP CIDRs that you want to be able to access your Kubernetes endpoints:
4. In Activation Settings Advanced Settings, in Enable IP CIDR for the load balancer specify a comma-separated list of IP CIDRs that you want to be able to access your load balancer endpoints:

Advanced Settings

Private Subnets (3 of 6 available selected)*

- subnet-05bbb3a6c02d0a76a (ent-usw2-private)
- subnet-0ac6a804221711fc2 (ent-usw2-private)
- subnet-0ee2ba7b5bb2fcc29 (ent-usw2-private)

Enable IP-CIDR for Kubernetes cluster:

List of allowed IP-CIDR for Kubernetes cluster

Enable IP-CIDR for the load balancer:

List of allowed IP-CIDR for the load balancer

5. After specifying the IP CIDRs, click Activate.

Related Information

[CIDR \(Classless Inter-Domain Routing\)](#)

Editing the IP CIDRs in the trusted list for endpoints in AWS environments

This topic describes how to change access to Hive, Impala, or Hue service endpoints in Cloudera Data Warehouse on cloud.

About this task


After you have added IP CIDRs to the trusted list so they have access to Kubernetes endpoints and the load balancer endpoints of services such as Hive, Impala, or Hue, you can edit the lists of IP addresses to change access to the endpoints.

Required role: DWAdmin

Before you begin

You must activate an environment before you can view its details or edit the IP CIDRs that have been added to the trusted list. See "Activating environments" in the "Related information" section at the bottom of this page.

Procedure

1. In the Cloudera Data Warehouse service, go to the Environments tab.
2. Locate the environment where you want to edit the IP CIDRs that have been specified as part of a trusted list and click  Edit .
This launches the Environment Details page.
3. Go to the Configurations tab, edit the comma-separated list of IP CIDRs: Enable IP CIDR(s) for Kubernetes cluster and Enable IP CIDR(s) for the load balancer.
4. Click Apply Changes.

Related Information

[Activating environments](#)

Networking

Learn how to increase the available IP addresses for Cloudera Data Warehouse on cloud using overlay networks, use a network proxy in Cloudera Management Console, and set up private networking.

Overlay networks for AWS environments in Cloudera Data Warehouse service

This topic explains how you can use overlay networks to increase the number of available IP addresses for your deployments of Cloudera Data Warehouse on cloud.

If you have an existing Virtual Private Cloud (VPC) where you want to deploy a Virtual Warehouse, you must make sure that there is an adequate number of available IP addresses. Otherwise, the Cloudera Data Warehouse service cannot be used efficiently and some features like auto-scaling do not work because the number of IP addresses might become exhausted. For example, each executor node in a Virtual Warehouse uses 8 IP addresses. If you create a SMALL-sized Virtual Warehouse with 2 executor nodes, it uses 16 IP addresses. If you create a LARGE-sized Virtual Warehouse with 40 executor nodes, it uses 320 IP addresses. If your VPC is small, the available IP addresses can be consumed quickly. Using overlay networks solves this issue.

**Important:**

- Use this feature if your VPC subnet has fewer than 1,024 IP addresses.
- Overlay networks are configured to operate up to 200 executor nodes by default. Cloudera recommends that you do not exceed this limit or networking limitation might occur.

About overlay networks

An overlay network is a software-defined layer of network abstraction that is used to run multiple separate, discrete virtualized network layers over the VPC network. In the case of the Cloudera Data Warehouse service, a custom CNI (Container Network Interface) plugin is used to enable the overlay network. It creates two network spaces:

- A node network space, which derives per-node IP addresses from the VPC.
- A Kubernetes pod network space, which derives per-pod IP addresses from the CNI plugin's own network space.

The overlay network is bridged into the node network. As a result, one IP address is required per node instead of one IP address needed per pod. Consequently, there are more available IP addresses and you can use the Cloudera Data Warehouse service efficiently, auto-scaling Virtual Warehouses as needed to meet the demands of your workloads.

When to use overlay networks for Cloudera Data Warehouse service environments

By default, when you create a new network during environment registration with the Cloudera Management Console, Cloudera creates 3 subnets with 8,192 IP addresses per subnet, which means there are 24,576 available IP addresses:

Figure 1: Registering an environment with a new network in Cloudera Management Console

When Cloudera creates your environment, you most likely will have enough IP addresses to use and grow your Virtual Warehouses. However, if you decide to use an existing VPC network on your cloud provider, which might have a limited number of available IP addresses, configuring your environment to use overlay networks for the Cloudera Data Warehouse service avoids IP address exhaustion. Using overlay networks, which can be configured for an environment in the Cloudera Data Warehouse service, uses fewer IP addresses because it uses an "IP per host" model, so your Virtual Warehouse can be used efficiently.

Related Information

[Register an AWS environment](#)

Enabling overlay networks in AWS environments

This topic describes how to enable Cloudera environments on AWS to use overlay networks to increase the number of IP addresses that are available to Cloudera Data Warehouse on cloud.

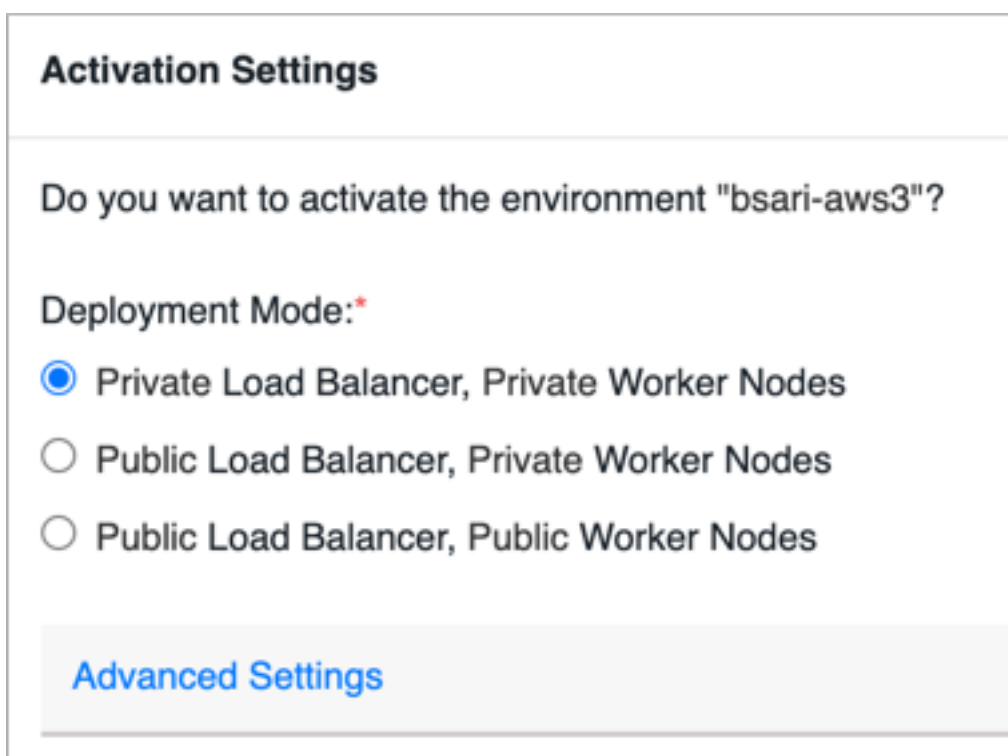
About this task

Required role: EnvironmentAdmin or PowerUser

After you have registered your environment with Cloudera, navigate to the Cloudera Data Warehouse service and perform the following steps to configure your AWS environment to use overlay networks.

Procedure

1. In the Cloudera Data Warehouse service, in the Environments column, click the search icon and locate the environment you registered with Cloudera where you want to enable overlay networks.
2. Click Activate to activate the environment.
3. In Activation Settings, click Advanced Settings to see the options:



Activation Settings

Do you want to activate the environment "bsari-aws3"?

Deployment Mode:*

☒ Private Load Balancer, Private Worker Nodes

☐ Public Load Balancer, Private Worker Nodes

☐ Public Load Balancer, Public Worker Nodes

[Advanced Settings](#)

4. Select Use Overlay Network and click Activate:

Use a non-transparent proxy with Cloudera Data Warehouse on AWS environments

You can configure Cloudera environments to use a network proxy in Cloudera Management Console. Then when you activate the environment for Cloudera Data Warehouse, if you can only use non-transparent proxies with clients that must connect to the Virtual Warehouse that uses the environment, you can configure them.

Difference between transparent and non-transparent network proxies

Transparent proxies are unknown to clients and require no additional client configuration. Usually connections by way of transparent proxies are configured in route tables on your AWS VPC. However, clients are aware of non-transparent proxies and each client must be specifically configured to use the non-transparent proxy connection.

There are benefits to using non-transparent proxies because doing so permits you to pass connection or security information along with the connection request sent by clients. Some organizations' security policies require the use of non-transparent proxies and Cloudera Data Warehouse can support that requirement.

Configure non-transparent proxies for Cloudera Data Warehouse on AWS environments

You can configure an AWS environment to use non-transparent proxy connections when activating environments for Cloudera Data Warehouse.

About this task

This task explains how to configure non-transparent proxies when you activate AWS environments for Cloudera Data Warehouse.

Required role: DWAdmin

Before you begin

- Before you can configure non-transparent proxies during environment activation for Cloudera Data Warehouse, you must make sure that a proxy was configured for the environment when it was registered with Cloudera Management Console. For details about configuring a proxy when registering an environment with Cloudera Management Console, see [Using a non-transparent proxy](#).



Note: When you activate an environment to use with Cloudera Data Warehouse, the option to configure non-transparent proxies is only available if you have added a proxy during environment registration with Cloudera Management Console.

- Before activating the environment that uses a proxy for Cloudera Data Warehouse, set up the following VPC endpoints for your AWS account in the AWS Console:
 - sts.amazonaws.com
 - sts.<REGION>.amazonaws.com
 - .s3.<REGION>.amazonaws.com
 - .s3.amazonaws.com
 - s3.amazonaws.com
 - dynamodb.<REGION>.amazonaws.com
 - api.ecr.<REGION>.amazonaws.com
 - dkr.ecr.<REGION>.amazonaws.com
 - ec2.<REGION>.amazonaws.com
 - cloudformation.<REGION>.amazonaws.com
 - autoscaling.<REGION>.amazonaws.com
 - elasticfilesystem.<REGION>.amazonaws.com
 - elasticloadbalancing.<REGION>.amazonaws.com
- For information about creating VPC endpoints, see the [Amazon documentation](#). If you cannot create a VPC endpoint for one of the required outbound destinations that are listed here, you must delete it from the Bypass proxy settings for these domains text box in Step 5 below, and add it to the proxy allowlist.
- Add other AWS specific outbound destinations to your proxy allowlist because creating VPC endpoints for them is not supported by AWS:
 - eks.<REGION>.amazonaws.com
 - rds.<REGION>.amazonaws.com
 - servicequotas.<REGION>.amazonaws.com
 - pricing.<REGION>.amazonaws.com

For more information about the AWS specific outbound destinations used by Cloudera, see [Outbound network access destinations for AWS](#) in the Cloudera Management Console documentation.

Procedure

1. In the Cloudera Data Warehouse service, go to the Environments tab.
2. Locate the environment you configured a non-transparent proxy for when you registered it with Cloudera Management Console.
3. When you locate the environment, click Activate to launch the Activation Settings dialog box where you can configure non-transparent proxies for the environment.
4. In the Activation Settings dialog box, click Advanced Settings to expand the dialog box options.
5. In the expanded dialog box, select Use default environment Proxy. When you select this option, the Bypass proxy settings for these domains option appears, and a text box listing all destinations that by-pass this proxy also appears. These are the destinations that you created VPC endpoints for in the above [Before you begin](#) section.



Important: If you cannot create a VPC endpoint for one of the required outbound destinations that are listed in the "Before you begin" section, you must delete it from the Bypass proxy settings for these domains text box by clicking the "x" that is adjacent to it, and then add it to the proxy allowlist.

6. Click Activate to complete environment activation for Cloudera Data Warehouse.

Setting up private networking in AWS environments

The topics in this section describe how to enable private subnets in Cloudera Data Warehouse on AWS.

Required role: DWAdmin

In AWS, a *public subnet* is connected to an internet gateway which can send and receive traffic directly to and from the internet. *Private subnets* send outbound traffic from nodes to the internet by using a network address translation (NAT) gateway, and then forwards the traffic to an internet gateway. Private subnets receive no direct inbound connections from the internet. This provides private network connectivity for workload endpoints in Cloudera Data Warehouse service.

Related Information

[Amazon documentation on public and private subnets](#)

Supported deployment modes for private networking in AWS

This topic describes the different types of private networking that you can configure Cloudera Data Warehouse to use on AWS.

The Cloudera Data Warehouse service can be deployed in three different modes on AWS cloud resources. These supported deployment modes determine whether your cluster nodes can receive direct connections from the internet.

The supported deployment modes in AWS are:

1. Public Load Balancer, Public Worker Nodes

Requires that you specify 3 public subnets in the virtual private cloud (VPC) that is registered with the Cloudera Management Console. In this mode, all AWS network components have a publicly visible IP address assigned to them. However, all traffic flows through the Kubernetes ingress controller so services and ports are not directly accessible.

2. Public Load Balancer, Private Worker Nodes

Requires that you specify 3 public subnets and 3 private subnets in the VPC that is registered with the Cloudera Management Console. In this mode, the Amazon EKS nodes are not assigned public IP addresses. All traffic is routed from the load balancer (Amazon ELB), which is located in a public subnet, to the ingress controller and the private subnet containing the private worker nodes.

3. Private Load Balancer, Private Worker Nodes

Requires that you specify 3 private subnets in the VPC that is registered with the Cloudera Management Console. This mode also requires that you set up network routing by way of a VPN between your on-premises network and

the VPC. Both the load balancer and worker nodes reside in private subnets so no Data Warehouse services have publicly visible IP addresses assigned to them.



Note: Selecting private load balancers is required if you want to use a private EKS API server.

Prerequisites for private networking in AWS environments

This topic lists the prerequisite requirements for configuring Cloudera Data Warehouse to use private networks on AWS.

To use the AWS private networking feature in the Cloudera Data Warehouse service, make sure you set up gateway virtual private cloud (VPC) endpoints and that you specify the correct number of public and private subnets for the Cloudera environment you plan to use for the service.

Set up gateway VPC endpoints to improve performance

For the Public Load Balancer/Private Worker Nodes and the Private Load Balancer/Private Worker Nodes deployment modes where worker nodes are running in private subnets, all outbound internet traffic passes through network address translation (NAT) gateways. For example, traffic created by activities such as downloading Docker images or accessing Cloudera services like billing or metering, S3 storage, and Amazon DynamoDB. To improve performance by reducing the number of network hops for accessing S3 and DynamoDB, set gateway VPC endpoints. A gateway endpoint is a gateway that you specify as a target for a route in your route table for traffic that is destined to either Amazon S3 or DynamoDB, the two supported AWS services.

If your VPC does not contain VPC endpoints that target S3 or DynamoDB, use the Amazon documentation, which is linked to at the bottom of the page, to set them up.

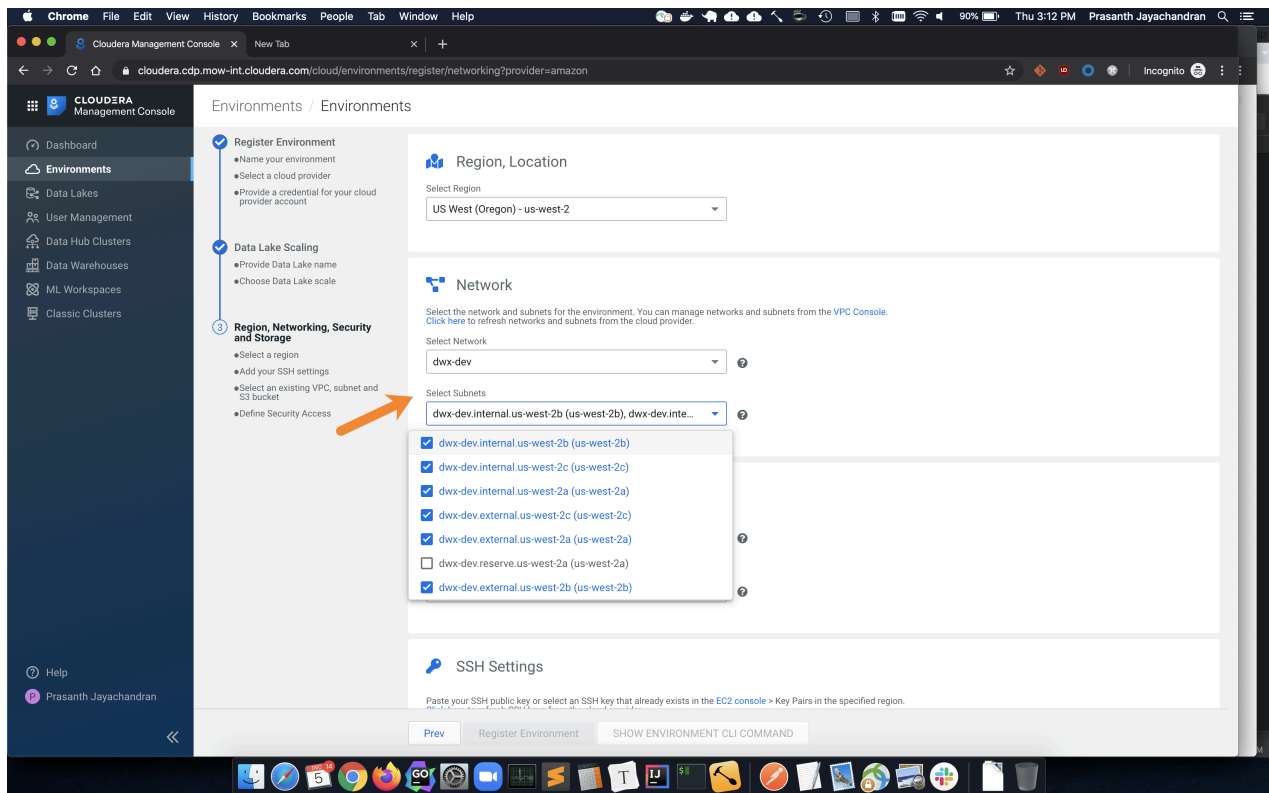
Specify the correct number of public and private subnets for your Cloudera environment

Depending on which deployment mode you choose, select the correct number of private and public subnets in the VPC environment that you register with the Cloudera Management Console. Check the documentation on [Supported deployment modes](#) to make sure what combination of public and private VPC subnets are needed before you register your environment.



Note: If you can only use private subnets in your AWS VPC environment, you can specify three private subnets when you register the environment with Cloudera Management Console. However, if you set up private networking with a Cloudera environment that includes three private subnets and no public subnets, network administrators must make sure there is outbound internet access from private subnets by way of the transit gateway or by another means.

General instructions for registering an environment are included in the Cloudera Management Console documentation. A link to this documentation is provided below. When you are on Step 6 where you specify information on the Region, Networking, Security and Storage page of Cloudera Management Console, specify the appropriate number of public or private subnets in your VPC for the Select Subnets prompt on the page:



Then proceed through the instructions to complete registering your environment with Cloudera.

Related Information

[Endpoints for Amazon S3](#)

[Endpoints for Amazon DynamoDB](#)

[Register an AWS environment with Management Console](#)

Activating an AWS environment with private subnet support

This topic describes the procedure for activating an AWS environment to use private subnets in Cloudera Data Warehouse on cloud.

About this task

After you have registered your environment with Cloudera, navigate to Cloudera Data Warehouse service and perform the following steps to activate the environment.

Required role: DWAdmin

Procedure

1. In the Cloudera Data Warehouse service, go to the Environments tab.
2. Locate the environment for which you want to configure private networking.
3. Click Activate to activate the environment.

4. Specify the Deployment Mode:



Important: Cloudera recommends that you use the Private Load Balancer, Private Worker Nodes deployment mode if possible because it is the most secure.

Activation Settings

Do you want to activate the environment "prakashsdx73"?

Deployment Mode:^{*}

☒ Private Load Balancer, Private Worker Nodes

☐ Public Load Balancer, Private Worker Nodes

☐ Public Load Balancer, Public Worker Nodes

5. To view the public and private subnets specified for your Cloudera environment, click Advanced Settings.

6. Click Activate.

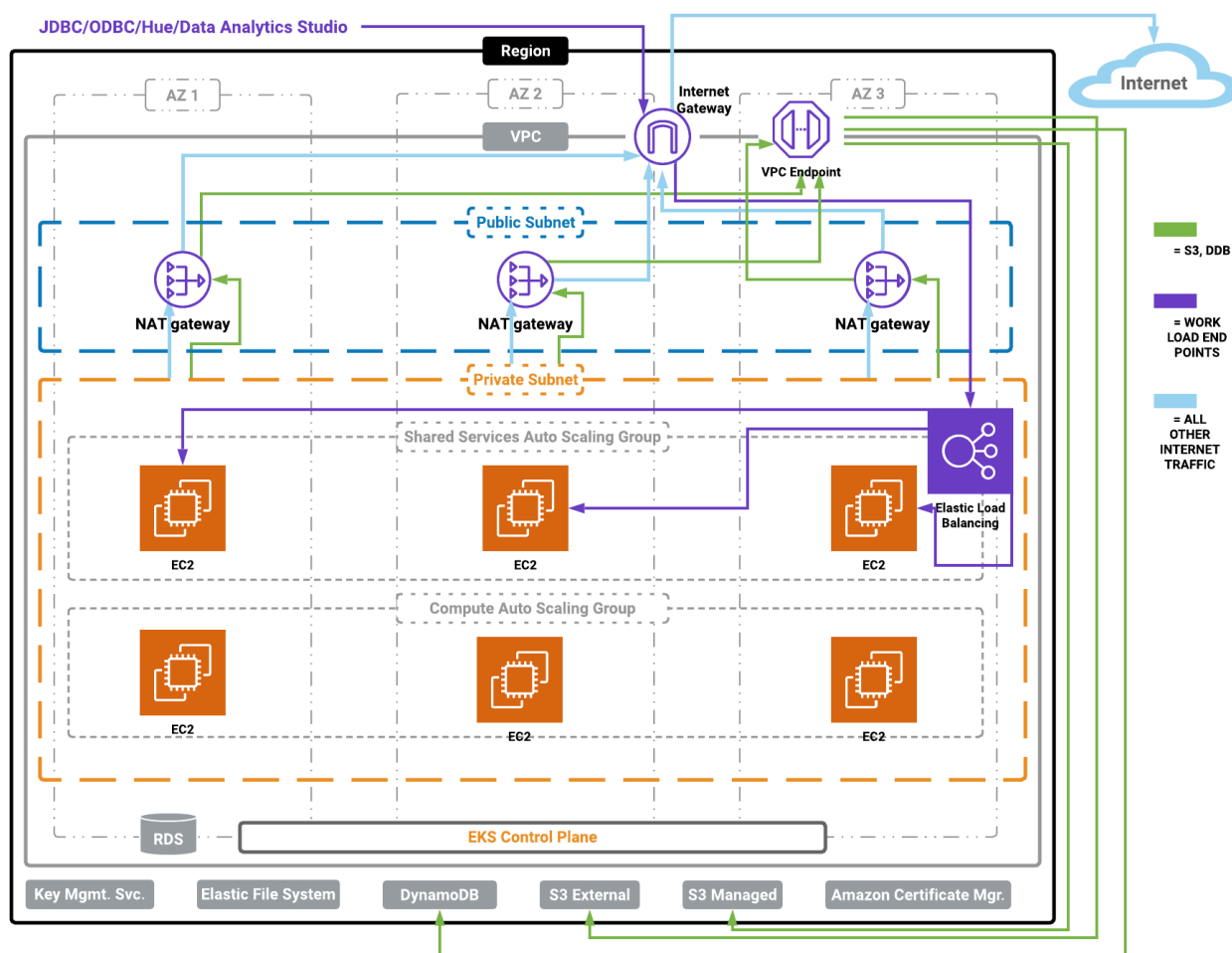
Related Information

[Supported deployment modes](#)

Architecture for Private Load Balancer, Private Worker Nodes deployment on AWS

This topic contains a network architecture diagram of Cloudera Data Warehouse configured to use private networking in an AWS environment.

The following diagram shows the architecture when you specify the Private Load Balancer, Private Worker Nodes deployment mode during environment activation in the Cloudera Data Warehouse service. In this deployment mode only the Network Address Translation (NAT) gateways reside in public subnets. Network administrators must make sure there is outbound internet access from private subnets by way of a transit gateway or by another means.



Custom tags in AWS environments

You can apply Virtual Warehouse-level tags and tenant-level custom tags to monitor, search, or filter on the key tag or key value in your AWS account.

Use the Virtual Warehouse-level tags to search and monitor the Virtual Warehouses within your environment. You can apply Virtual Warehouse-level tags even after creating the Cloudera environment. The tenant-level tags can only be applied while creating a new Cloudera environment. The Cloudera environment custom tags are propagated to the compute instances within Cloudera Data Warehouse.

You can identify and monitor the following resources within the Cloudera Data Warehouse service for AWS environments:

- Compute instances in Hive and Impala Virtual Warehouses with tenant-level and Virtual Warehouse tags
- Compute instances in Druid cluster with tenant-level tags
- AWS S3 buckets
- Amazon Elastic Block Store (EBS) volumes for EC2 instances used for compute pods with Virtual Warehouse tags. You cannot tag EBS volumes for EC2 instances used for shared service pods.

Upgrades

Upgrading Amazon Elastic Kubernetes Service (EKS) can prevent incompatibilities issues between Cloudera Data Warehouse and AWS resources.

Upgrading PostgreSQL 9.6 before EOL

You need to plan for the end-of-life (EOL) of PostgreSQL 9.6, announced by Amazon. As an AWS environment user, you need to check the version of PostgreSQL you use for Cloudera Data Warehouse environments. If your database version is still PostgreSQL 9.6, you need to upgrade to PostgreSQL 11.12.

About this task

AWS has published the following announcement regarding the imminent PostgreSQL upgrade.

Upgrade your Amazon RDS for PostgreSQL 9.6 databases before Jan 18, 2022

The RDS for PostgreSQL 9.6 end-of-life date is approaching. You have 1 database using RDS for PostgreSQL 9.6 that must be upgraded to 12 or higher as soon as possible.

We plan to automatically upgrade RDS for PostgreSQL 9.6 databases to 12 starting January 18, 2022.

For more information, see the [RDS for PostgreSQL deprecation timeline](#) in the Amazon RDS forum.

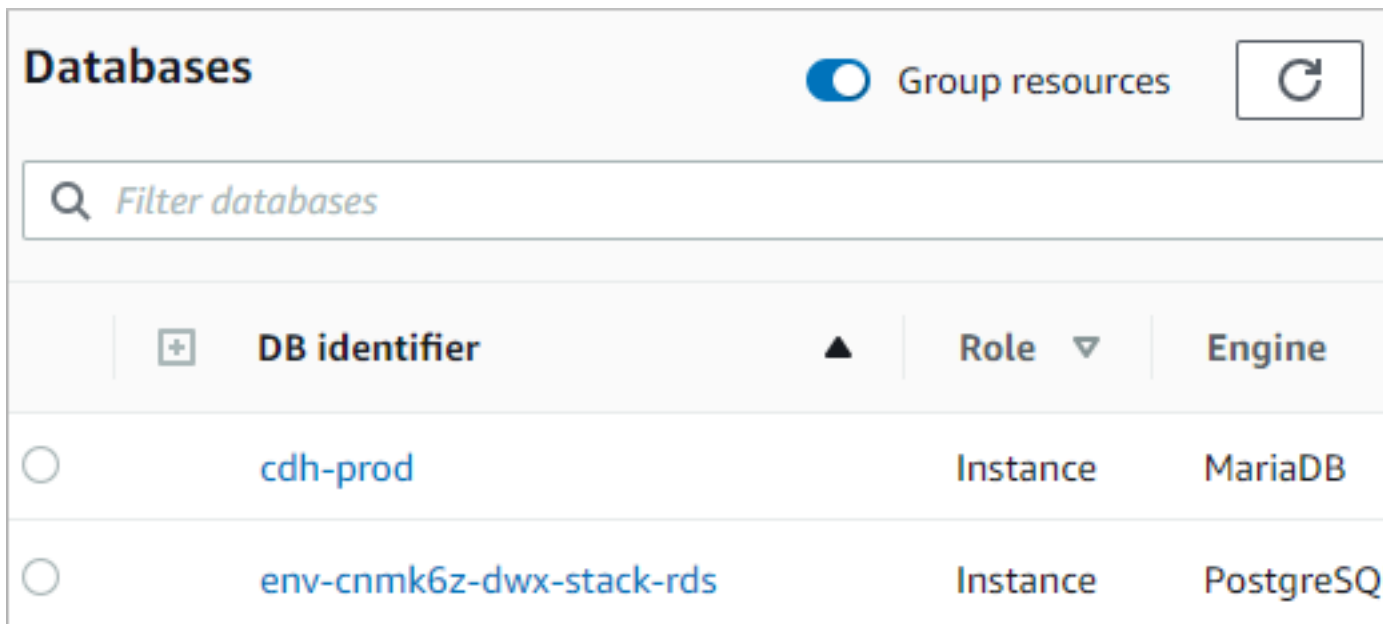
You can initiate an upgrade of your database instance — either immediately or during your next maintenance window — to the Cloudera-recommended version of PostgreSQL 11.12 using the AWS Management Console or the AWS Command Line Interface. Follow the procedure below to perform the upgrade.

The upgrade process shuts down the database instance, performs the upgrade, and restarts the database instance. The database instance may be restarted multiple times during the upgrade process. While major version upgrades typically complete within the standard maintenance window, the duration of the upgrade depends on the number of objects within the database. To estimate the time required, take a snapshot of your database and test the upgrade.

Before you begin

Procedure

1. Go to the AWS Management Console Amazon RDS Databases .



Databases			
<input checked="" type="checkbox"/> Group resources			
<input type="text"/> Filter databases			
	DB identifier	Role	Engine
<input type="radio"/>	cdh-prod	Instance	MariaDB
<input type="radio"/>	env-cnmk6z-dwx-stack-rds	Instance	PostgreSQL

2. Check the version of PostgreSQL used in your environment, and if it is 9.6.6, go to the next step to start the upgrade process.
3. In Cloudera, go to your environment, and in a Database Catalog for that environment, create a Virtual Warehouse or use an existing one.
4. Run some basic queries in Hive or Impala to see if your PostgreSQL 9.6.6 is alive and well.

```
show tables;
use default;
show tables;
create table tbl1(col1 string, col2 string);
show tables;
describe tbl1;
create table tbl2 as (select * from tbl1);
describe tbl2;
insert into table tbl1 values ("Hello", "World");
select * from tbl1;
```

5. Go to the AWS Management Console Amazon RDS Databases . Select DB Engine version 10.16 to upgrade Amazon RDS to 10.16.

Amazon prevents a direct upgrade to 11.x unless you are on PostgreSQL 9.6.20 or higher.

6. In Cloudera, rerun the basic queries in your Virtual Warehouse, and if all goes well, proceed to the next step.
 7. Look for errors, such as those shown below, in the metastore log.
- Errors in the metastore might look something like this:

```
<14>1 2021-07-31T00:52:38.223Z metastore-0.metastore-service.warehouse-1
627669911-v16x.svc.cluster.local metastore 1 0b245ec4-8419-4968-94bc-ee1
22960b1aa [mdc@18060
class="txn.TxnHandler" level="INFO" thread="pool-9-thread-200"] Non-ret
ryable error in enqueueLockWithRetry(LockRequest(component:[LockComponen
t(type:SHARED_READ, level:
DB, dbname:default, operationType:NO_TXN, isDynamicPartitionWrite:false)]
, txnid:79, user:hive, hostname:hiveserver2-0.hiveserver2-service.comput
e-1627670377-mbqx.svc.
```



```
cluster.local, agentInfo:hive_20210731005238_4aa14bf0-4e46-448d-b6c7-cdc3ca4ec863, zeroWaitReadEnabled:false)) : Batch entry 0 INSERT INTO "HIVE_LOCKS" ( "HL_LOCK_EXT_ID", "HL_LOCK_INT_ID", "HL_TXNID", "HL_DB", "HL_TABLE", "HL_PARTITION", "HL_LOCK_STATE", "HL_LOCK_TYPE", "HL_LAST_HEARTBEAT", "HL_USER", "HL_HOST", "HL_AGENT_INFO") VALUES (4561258935320160815, 1, 79, 'default', NULL, NULL, 'w', 'r', 0, 'hive', 'hiveserver2-0.hiveserver2-service.compute-1627670377-mbqx.svc.cluster.local', 'hive_20210731005238_4aa14bf0-4e46-448d-b6c7-cdc3ca4ec863') was aborted: ERROR: index "hl_txnid_index" has wrong hash version^M Hint: Please REINDEX it. Call getNextException to see other errors in the batch. (SQLState=XX002, ErrorCode=0)
```

8. Connect to the HiveServer pod or metastore pod, and using psql, connect to the RDS instance. For example,

```
psql -h env-kg.us-west-2.rds.amazonaws.com --u hive --d postgres
```

Login using the hostname from the AWS console. The postgres database password is stored in JCEKS file which is mounted using a secret volume inside the HiveServer pod or metastore pod. Note the namespace of pod and [obtain the password](#).

```
[hive@metastore-0 lib]$ psql -h env-kg.us-west-2.rds.amazonaws.com --u hive --d postgres
Password for user hive:
psql (9.2.24, server 9.6.6)
WARNING: psql version 9.2, server version 9.6.
         Some psql features might not work.
SSL connection (cipher: ECDHE-RSA-AES256-GCM-SHA384, bits: 256)
Type "help" for help.

postgres=> █
```

Related Information

[Upgrading the PostgreSQL DB engine for Amazon RDS](#)

Validate the upgrade to PostgreSQL 10.16

You need to connect to the postgres 10.16 database, and validate the upgrade.

Procedure

1. Connect to the PostgreSQL 10.16 database using the namespace of the pod and the password you obtained in the last procedure.

- On the Postgres command line, look at all the databases.

For example, type \l.

```
postgres=> \l
```

List of databases					
Name	Owner	Encoding	Collate	Ctype	Access privileges
metastore	hive	UTF8	en_US.UTF-8	en_US.UTF-8	
postgres	hive	UTF8	en_US.UTF-8	en_US.UTF-8	
rdsadmin	rdsadmin	UTF8	en_US.UTF-8	en_US.UTF-8	rdsadmin=CTc/rdsadmin
template0	rdsadmin	UTF8	en_US.UTF-8	en_US.UTF-8	=c/rdsadmin
template1	hive	UTF8	en_US.UTF-8	en_US.UTF-8	=c/hive
viz-1629222151-5xqq_vizdb	hive	UTF8	en_US.UTF-8	en_US.UTF-8	hive=CTc/hive
warehouse-1629193710-rmjh-das	hive	UTF8	en_US.UTF-8	en_US.UTF-8	
warehouse-1629193710-rmjh-metastore	hive	UTF8	en_US.UTF-8	en_US.UTF-8	

- Fix any problematic indexes. Go to a database in your metastore named something like
For example, go to a database in your metastore named warehouse-1629221321-pf2h-metastore.

```
\c warehouse-1629221321-pf2h-metastore
```

```
postgres=> \c warehouse-1629221321-pf2h-metastore
psql (9.2.24, server 9.6.6)
WARNING: psql version 9.2, server version 9.6.
        Some psql features might not work.
SSL connection (cipher: ECDHE-RSA-AES256-GCM-SHA384, bits: 256)
You are now connected to database "warehouse-1629221321-pf2h-metastore" as user "hive".
warehouse-1629221321-pf2h-metastore=> █
```

- Run the REINDEX command.

```
REINDEX INDEX hl_txnidx_index;
```

- In Cloudera, rerun the basic queries in your Virtual Warehouse, and if all goes well, proceed to the next step.
- Upgrade the Amazon RDS version of PostgreSQL to 11.12, and rerun the basic queries.

Upgrade to PostgreSQL 11.12

Procedure

- Go to the AWS Management Console Amazon RDS Databases , and upgrade the Amazon RDS version to PostgreSQL 11.12.
- Rerun the basic queries in your Virtual Warehouse to validate the upgrade.

Upgrading Amazon Kubernetes Service (EKS)

Amazon Elastic Kubernetes Service (EKS) cluster requires regular updates to the Kubernetes versions. Using the latest EKS version supported by Cloudera avoids compatibility issues between Cloudera Data Warehouse and AWS resources.

Cloudera Data Warehouse automatically provisions the latest supported EKS version when you activate an environment in Cloudera Data Warehouse. For more information on the EKS version, refer [Version mapping for supported release](#).

To upgrade to the latest EKS, you must deactivate and reactivate your Cloudera Data Warehouse environment. To restore the state of your Cloudera Data Warehouse cluster after reactivation, Cloudera recommends using the backup-restore functionality. For more information, see [Backing up and restoring Cloudera Data Warehouse](#).

Upgrading using your own AMI or reduced permissions

You learn the prerequisite for using your own AMI, the risks of using your own AMI, and the steps involved in upgrading.

About this task

You can [upgrade the Amazon Elastic Kubernetes Service \(EKS\)](#) using your own Amazon Machine Image (AMI) if you meet either one of the following conditions:

- You are in reduced permissions mode.
- You have obtained the CDW_CUSTOM_AMI entitlement.

The option to enter your own AMI name does not appear when you perform the upgrade steps in this document unless you meet one of these conditions.

Using your own AMI to upgrade worker executors is error-prone due to EKS version matching requirements, and therefore not recommended. Cloudera recommends that you upgrade to the recommended EKS instead of using your own AMI. If you decide to use your own AMI, you must choose the AMI version that matches the EKS version used by Cloudera Data Warehouse. Using your own AMI, you can upgrade only to the next minor EKS version, so more than one upgrade might be required to match the EKS version used by Cloudera Data Warehouse.

If you do not use a custom AMI in step 4 below, but use reduced permissions, you must upgrade three times from EKS v1.17 to get to v.1.20.

Before you begin


- Check that you are using the [reduced permissions mode](#), or that you have obtained the CDW_CUSTOM_AMI entitlement.
- Perform this workaround described in the known issue [DWX-8573: EKS upgrade from DWX UI to K8s v1.20 fails in reduced permissions mode](#). In AWS, add the following permissions to your [Reduced permissions mode JSON IAM permissions](#) policy right after "iam:PutRolePolicy":

```
{
  ...
  "cloudformation:GetTemplate",
  "cloudformation:GetTemplateSummary",
  "iam:GetRole",
  "eks:ListUpdates",
  "ec2:CreateLaunchTemplateVersion",
  "ec2:DescribeLaunchTemplateVersions",
  "ec2:DescribeLaunchTemplates",
  "autoscaling:UpdateAutoScalingGroup",
  "autoscaling>DeleteAutoScalingGroup",
  "ec2:RunInstances",
  "autoscaling:DescribeScalingActivities",
  "autoscaling:TerminateInstanceInAutoScalingGroup",
  "autoscaling:DescribeScheduledActions",
  "autoscaling:SetDesiredCapacity",
  "iam:PassRole",
  "rds:DescribeDBInstances",
  "ec2:DescribeInstances"
  ...
}
```

Procedure

1. In the Cloudera Data Warehouse service, go to the Environments tab.
2. Locate the environment on AWS for upgrading the Kubernetes version.
3. Click Upgrade.

In reduced permissions mode, the option to upgrade using a custom AMI appears:



Upgrade ✕

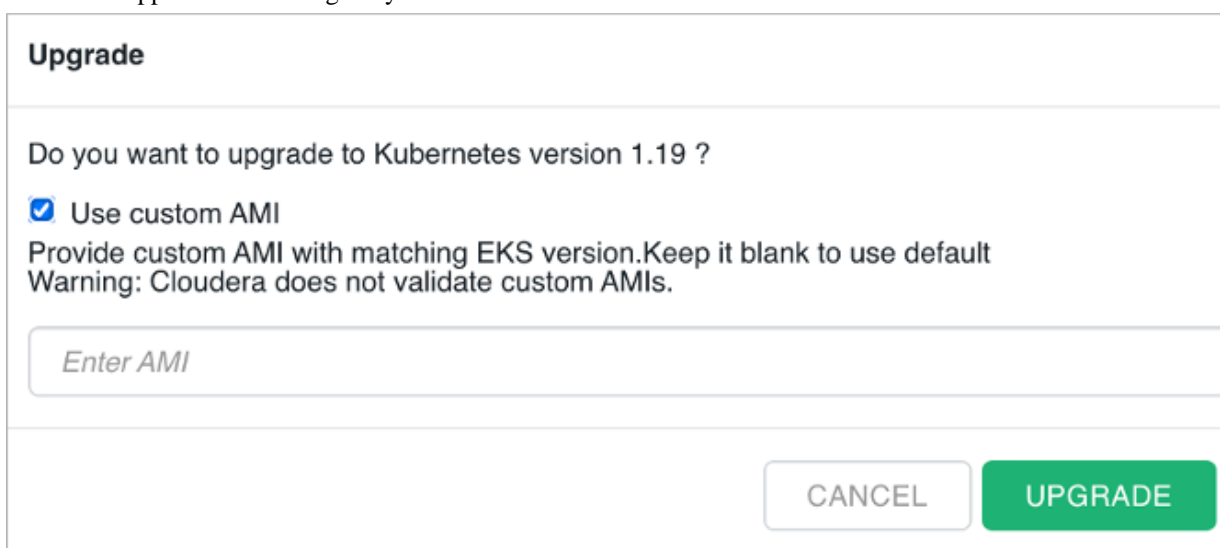
Do you want to upgrade to Kubernetes version 1.19 ?

☐ Use custom AMI

CANCEL UPGRADE

4. Select Use custom AMI.

A text box appears for entering the your AMI version.



Upgrade

Do you want to upgrade to Kubernetes version 1.19 ?

☒ Use custom AMI

Provide custom AMI with matching EKS version.Keep it blank to use default
Warning: Cloudera does not validate custom AMIs.

Enter AMI

CANCEL UPGRADE

You can upgrade only to the next minor version for each environment upgrade, for example from 1.18 to 1.19. For example, to upgrade to EKS 1.20, you must upgrade three times from EKS 1.17.

5. Enter an AMI version that is consistent with the Cloudera Data Warehouse upgrade to the EKS version and region of the activated cluster.
6. Click Upgrade to start the upgrade process.

Dynamically updating the Amazon Machine Image

You need the capability to update the Amazon Machine Image (AMI) to prevent potential problems running workloads on an old AMI. You can update the AMI of the Cloudformation stack while keeping the current Elastic Kubernetes Service (EKS) version.

Before you begin

You must meet the following prerequisites to get dynamic updates:

- You are running release 1.7.1 (released August, 2023) or later, which supports this feature.

- You must add the following IAM roles to your cross account role:

```
{
  "Sid": "AllowSsmParams",
  "Effect": "Allow",
  "Action": [
    "ssm:DescribeParameters",
    "ssm:GetParameter",
    "ssm:GetParameters",
    "ssm:GetParameterHistory",
    "ssm:GetParametersByPath"
  ],
  "Resource": [
    "arn:aws:ssm:*:*:parameter/aws/service/eks/optimized-ami/*"
  ]
}
```

This action creates and upgrades DWX stacks with read permissions on AWS Systems Manager (SSM). The SSM parameter in the Cloudformation dwx-stack template obtains the latest AMI to use for your EKS version.

- If you use reduced permissions mode, you must update the following fields of the stack template while performing a Kubernetes version upgrade:
 - EksBaseVersion - The value of this parameter must be the next upgrade version of kubernetes.
 - EksImageSSMParam - The value of this parameter must be formatted as follows: `/aws/service/eks/optimized-ami/<K8s-version>/amazon-linux-2/recommended/image_id`

Update only the K8s-version of EksImageSSMParam. For example:

- Old value
`/aws/service/eks/optimized-ami/1.21/amazon-linux-2/recommended/image_id`
- New value
`/aws/service/eks/optimized-ami/1.22/amazon-linux-2/recommended/image_id`

Procedure

- Log into the AWS Management Console.
- From the list of stacks, select the running CloudFormation stack.
- In the stack details pane, click Update.
A new AMI, if available, will be applied in the stack update.

Managed storage access

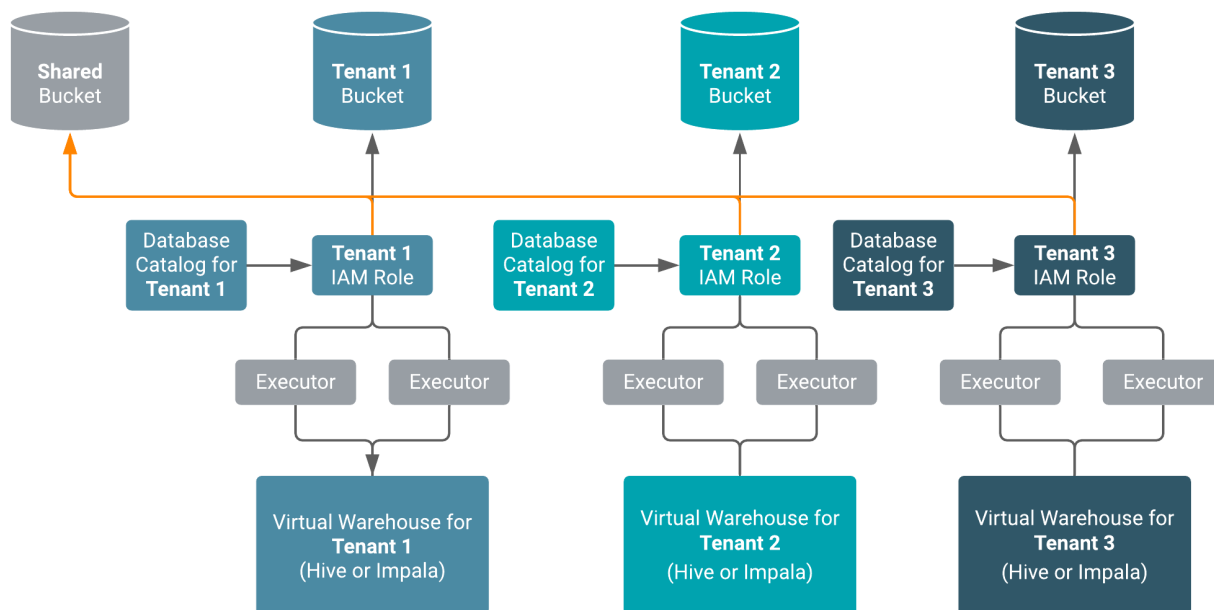
Understanding how Cloudera Data Warehouse stores data for multiple tenants and a high-level overview of the configuration tasks prepares you as DWAdmin to set up a RAZ-controlled warehouse.

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

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

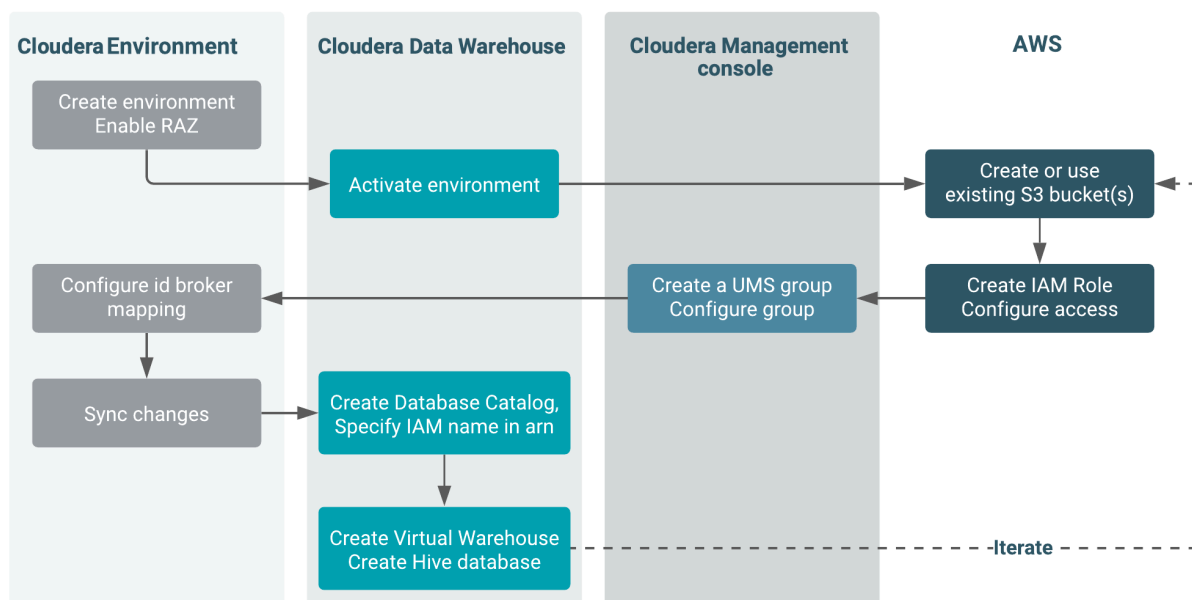
- The tenant-specific bucket(s)
- Potentially shared bucket(s)
- The SDX Data Lake bucket

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



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

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



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

- In Cloudera, register an environment that enables RAZ and uses the SDX Data Lake.
- Activate the environment in Cloudera Data Warehouse.

- Manually create S3 bucket(s), or use existing ones, for access by the IAM tenant role.
- In AWS, create an IAM role for the tenant.
- Give the IAM role certain permissions.
- Configure a trust relationship with the [IDBROKER_ROLE](#).
- Give the Instance Profile ([idbroker assume role](#)) the permission to assume the IAM role of the tenant.
- In Cloudera Management Console, create an UMS group.
- Add the UMS machine user (created when you activated the Cloudera Data Warehouse environment) to the UMS group.
- Add the id broker mapping to the environment.
- Sync user group changes for the RAZ-enabled environment with FreeIPA.
- Create a separate Database Catalog with a unique IAM role.
- Create a tenant-specific Virtual Warehouse based on the Database Catalog, SDX Data Lake, and RAZ-enabled environment.
- Create a tenant-specific Hive or Impala database that points to tenant-specific buckets.

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

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

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

The following topics describe step-by-step how to set up your environment, IAM roles, Database Catalog, and Virtual Warehouse for storing RAZ-enabled data.

Setting up managed storage access

You learn the consequences of setting up RAZ control, which cannot be removed, and requirements you must meet to enable managed storage access.

About this task

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

Before you begin

- Request activation of the following entitlements:
 - CDW_ALLOW_MULTI_DEFAULT_DBC
 - CDW_STORAGE_ROLES
- You meet the requirements described in the [AWS requirements documentation](#).

Required role: PowerUser

Creating the Cloudera environment and IAM roles

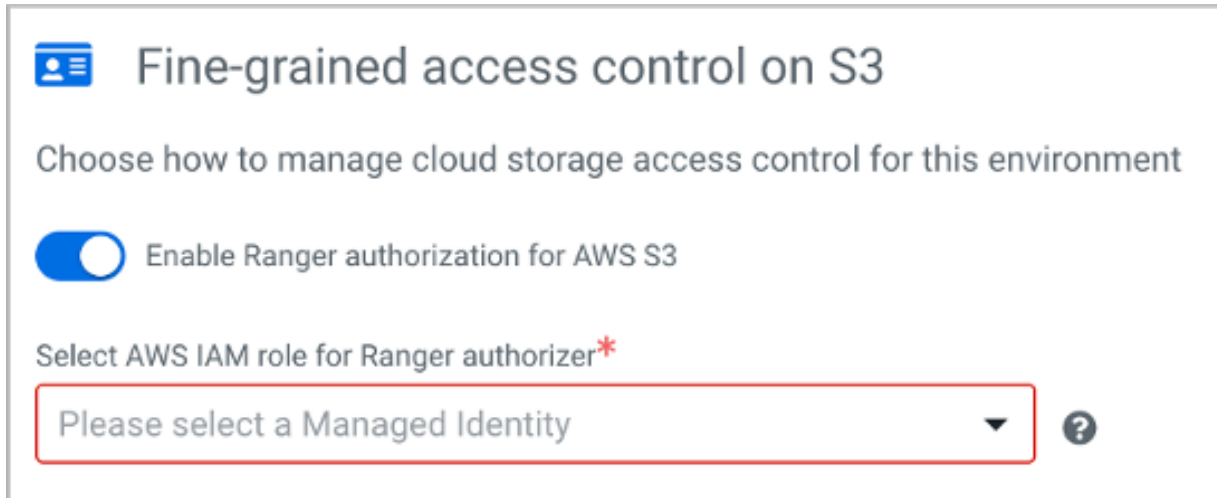
You follows steps to register and activate a new environment to enable RAZ in Cloudera Data Warehouse.

About this task

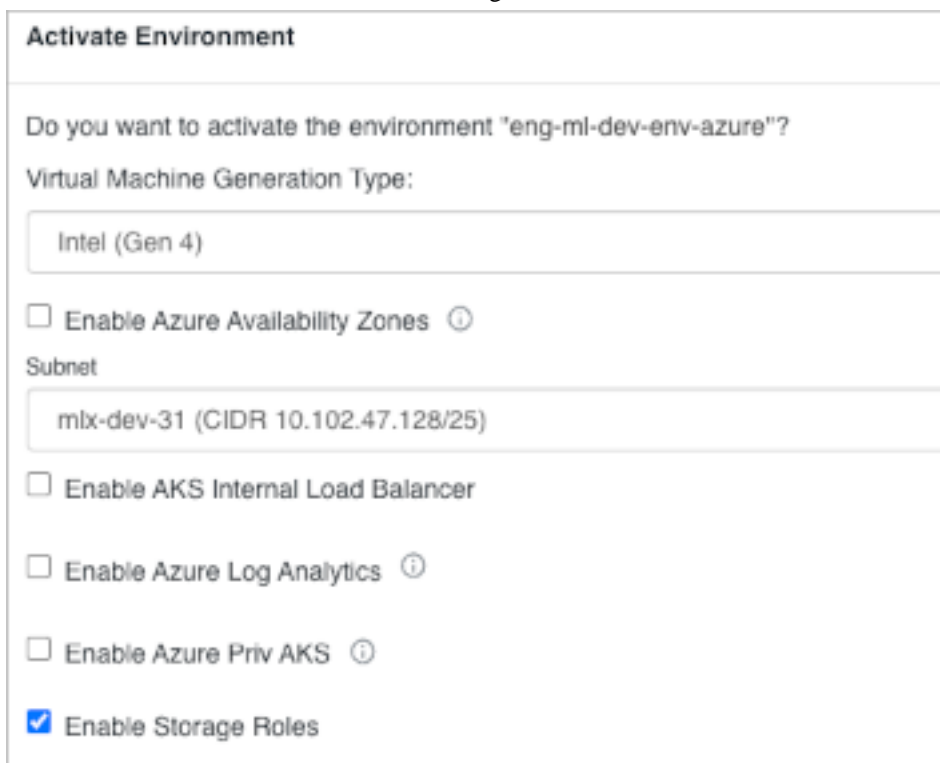
You need to register and activate a new environment to enable RAZ in Cloudera Data Warehouse.

Procedure

1. Register an environment with RAZ using the Cloudera web interface.
In the web interface, in Fine-grained access control on S3, select Enable Ranger Authorization for AWS S3.



2. In Cloudera Data Warehouse Overview, locate and select the RAZ-enabled environment, and click Start to activate the environment for Cloudera Data Warehouse.
This action disables the standard default Database Catalog that is automatically created after activation. The UMS machine user is created and attached to the environment when you activate the Cloudera Data Warehouse environment. Later, you see how to add this same UMS machine user to a different UMS group for each tenant.
3. In Activate Environment, select Enable Storage Roles.



Unchecking Enable Storage Roles disables the metadata proxy associated with managed storage access.

Repeat the following steps for each tenant:

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

5. In AWS, [create an IAM role](#) that has significant privileges to read/write to a tenant-specific buckets as well as across all tenants' shared SDX Data Lake bucket.
For example, create a role named role-tenant-1 and a bucket called s3-tenant-1.
6. In the IAM role, configure a trust relationship with the [IDBROKER_ROLE](#).
For example:

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

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

7. Make a note of the IAM role ARN.
8. Attach the following AWS policies to the IAM role: AmazonEKSWorkerNodePolicy, AmazonEKS_CNI_Policy, AmazonEC2ContainerRegistryReadOnly.

The AmazonEKSWorkerNodePolicy is shown in the next topic.

9. Give the AWS Instance Profile ([idbroker assume role](#)), the permission to assume role of role-tenant-1.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "VisualEditor0",
            "Effect": "Allow",
            "Action": "sts:AssumeRole",
            "Resource": [
                "arn:aws:iam::<AWS_ACCOUNT_ID>:role/idbroker-data-access-role",
                "arn:aws:iam::<AWS_ACCOUNT_ID>:role/mock-idbroker-admin-role",
                "arn:aws:iam::<AWS_ACCOUNT_ID>:role/role-tenant-1"
            ]
        }
    ]
}
```

Creating a UMS group and machine users

This procedure ensures that the [Cloudera machine user](#) gets permission to access the tenant bucket.

Procedure

Repeat the following steps for each tenant:

1. In Cloudera Management Console, User Management Groups , click CREATE GROUP and [create a User Management Service \(UMS\) group](#), for example group-tenant-1.

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

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

Creating a new Database Catalog

You repeat a step-by-step procedure to set up a Database Catalog for each tenant, taking care to correctly enter your IAM Role ARN.

Procedure

Repeat the following steps for each tenant.

1. Click **Data Warehouse Database Catalog New Database Catalog**.
2. In **Tenant Storage Role**, enter the IAM Role ARN you copied earlier.
Take care to enter the ARN correctly because Cloudera Data Warehouse does not validate your ARN.
3. In **Tenant Storage Location**, enter the tenant bucket name, for example `s3-tenant-1`, and click **CREATE**.
4. Select the RAZ-enabled environment.

In Data Lake, SDX is the required value. The backend Data Lake and Database Catalog database must be the same.

Creating a tenant-specific Virtual Warehouse

You follow a step-by-step procedure to create a Virtual Warehouse based on the Database Catalog and SDX Data Lake you created in the RAZ-enabled environment.

About this task

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

Procedure

Repeat the following steps for each tenant.

1. In the Cloudera Data Warehouse service, click **Virtual Warehouses New Virtual Warehouse**.
2. Specify a name, select either the Hive or Impala type, and select the Database Catalog you created for the tenant.
3. In **Overview**, find your Virtual Warehouse, click **Hue**.
4. Create a tenant-specific Hive or Impala database where the location for external and managed tables are pointing to the tenant-specific buckets.

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

Do not set **LOCATION** and **MANAGEDLOCATION** to the same path. For more information, see [Create a default directory for managed tables](#).

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

Tenant IAM role policy

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

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

```

        "Sid": "VisualEditor0",
        "Effect": "Allow",
        "Action": [
            "kms:Decrypt",
            "s3:ListAllMyBuckets",
            "kms:Encrypt",
            "kms:ListAliases",
            "kms:GenerateDataKey",
            "kms:DescribeKey"
        ],
        "Resource": "*"
    }
}
s3-read-write
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Action": [
                "s3:Get*",
                "s3:Delete*",
                "s3:Put*",
                "s3:ListBucket",
                "s3:ListBucketMultipartUploads",
                "s3:AbortMultipartUpload",
                "s3:GetBucketLocation"
            ],
            "Resource": [
                "arn:aws:s3:::<SDX_BUCKET>",
                "arn:aws:s3:::<SDX_BUCKET>/*",
                "arn:aws:s3:::tenant-1-bucket-1",
                "arn:aws:s3:::tenant-1-bucket-1/*"
            ],
            "Effect": "Allow"
        }
    ]
}

```

In Cloudera Data Warehouse, you need full access to the SDX bucket, even if you used only folder level while creating the CB environment.

Using AWS Vault to manage credentials

Instead of using a jceks file to store credentials, or exposing your user name and password through some other means of credential management, you learn how to use AWS Secrets Service for storing credentials in the AWS Vault.

Procedure

1. Log into AWS.
2. From the IAM console Roles on AWS select your role.
3. Click your IAM role to navigate to its summary, and copy the Role ARN.
4. Create your AWS secret or look for it in your .aws/credentials directory.
5. In the AWS Secrets manager, modify secrets permissions, substituting your role arn and secret resource arn for the variables:

```

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

```

```

    "Statement" : [ {
      "Effect" : "Allow",
      "Principal" : {
        "AWS" : "arn:aws:iam::NNNN:role/env-AAAAAA-dwx-stack-NodeInstanceRol
e-XXXXXX"
      },
      "Action" : [ "secretsmanager:GetSecretValue", "secretsmanager:Descr
ibeSecret" ],
      "Resource" : "arn:aws:secretsmanager:us-west-2:NNNN:secret:<secret-nam
e>-dsffd"
    } ]
  }
}

```

6. Create a table that specifies the JDBC-URL, your role arn, and secrets arn in the table properties.
For example:

```

create external table customer_address (
  column definitions ...
)
STORED BY 'org.apache.hive.storage.jdbc.JdbcStorageHandler'
TBLPROPERTIES (
  "hive.sql.database.type" = 'POSTGRES',
  "hive.sql.jdbc.driver" = 'com.mysql.jdbc.Driver',
  "hive.sql.jdbc.url" = '<jdbc-url>',
  "hive.sql.dbcp.username" = '<role-name>',
  "hive.sql.dbcp.password.uri" = 'aws-sm:///<secret-arn>',
  "hive.sql.table" = "<table-name>"
);

```

7. Create a Hive table that specifies the JDBC-URL, your database user name, and secrets arn in the table properties.

```

create external table customer_address (
  column definitions ...
)
STORED BY 'org.apache.hive.storage.jdbc.JdbcStorageHandler'
TBLPROPERTIES (
  "hive.sql.database.type" = 'POSTGRES',
  "hive.sql.jdbc.driver" = 'org.postgresql.Driver',
  "hive.sql.jdbc.url" = '<jdbc-url>',
  "hive.sql.dbcp.username" = '<database-user-name>',
  "hive.sql.dbcp.password.uri" = 'aws-sm:///<secret-arn>',
  "hive.sql.table" = "<table-name>"
);

```

Only the password field is extracted from the secret.

Identifying the spill location for Impala temporary data

Impala writes temporary data to S3 to prevent a memory overflow. The result is successful completion of a query instead of an out-of-memory error. When you create an Impala Virtual Warehouse, a path to write temporary data to the Data Lake S3 bucket is configured automatically. You need to know the name of the bucket and then allow Impala to access it.


About this task

The Impala Virtual Warehouse attempts to write temporary data to the local Non-Volatile Memory Express (NVMe) SSD on compute instances before spilling to Data Lake S3 bucket. The attempt succeeds only if you configure Impala access to the default scratch location on DataLake S3 bucket.

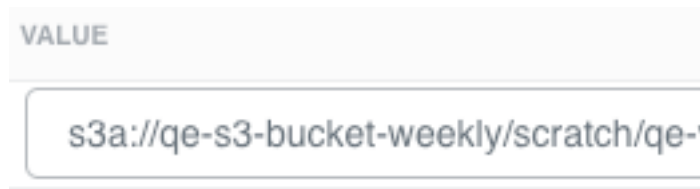
Before you begin

- Obtain the DWAdmin role.
- Your Impala Virtual Warehouse spill-to-Data-Lake-bucket was configured automatically when you created a new Virtual Warehouse; you are not using an existing Virtual Warehouse that required a manual spill-to-S3 configuration.

Procedure

1. From the Cloudera Management Console or Cloudera landing page, navigate to Data Warehouses.
2. Go to the Impala Virtual Warehouse  Edit Configurations Impala coordinator and select flagfile from the Configuration files drop-down list.
3. Note the value of the scratch_dirs property.

The first path segment of the value is the default scratch location in the Data Lake bucket. For example, qe-s3-bucket-weekly is the S3 bucket name of the default scratch location.



4. Go to the **Impala executor** tab and select flagfile from the Configuration files drop-down list.
5. Note the value of the scratch_dirs property.
6. Configure access to the default scratch location using the bucket name, for example qe-s3-bucket-weekly, as described in [Accessing S3 buckets](#).

Related Information

[Tuning Impala Virtual Warehouses on on cloud](#)

Managing temporary scratch files in S3

Learn how to clean up temporary scratch files left in S3 when an Impala daemon stops unexpectedly during query execution involving S3 spilling.

About this task

When queries involve data spilling to S3 and an Impala daemon crashes or is terminated outside the graceful shutdown period, temporary scratch files may remain in S3 storage. This task outlines steps to identify and safely delete these files.

To determine when the action is required, if no active queries are running, but the directory still contains files, action may be necessary. Additionally, if the directory has files that do not belong to any running queries and these files have been present for a significant period, such as several hours, this may indicate the need for cleanup.

Before you begin

- Ensure that there are no ongoing queries accessing the S3 directories.
- Determine if action is necessary by verifying no active queries or checking for files not linked to active queries.
- Verify the scratch_dirs configuration path to locate S3 scratch directories.

Procedure

1. Get the path from the `scratch_dirs` configuration. For example: `s3a://dw-bucket/scratch/impala-test,/opt/impala/scratch/remote-buffer:150G,/opt/impala/scratch/local-scratch:150G`

```
s3a://dw-bucket/scratch/impala-test,/opt/impala/scratch/remote-buffer:150G,/opt/impala/scratch/local-scratch:150G
```

2. If no active queries, run the following command to delete the directory:

```
aws s3 rm s3://dw-bucket/scratch/impala-test --recursive
```

3. If active queries are running, filter directories associated with running queries before cleanup.
 - a) Get a list of active query IDs from the **admission** page on the coordinator's WebUI under the "Running queries" section. If using multiple active coordinators then get active query IDs from each coordinator's **admission** page.
 - b) Use the following script to delete unneeded directories while excluding active query directories.



Important: Avoid running the below command during high query workload. It removes all directories except those explicitly excluded, which may disrupt running queries.

This example excludes two queries, you can add more as needed:

```
aws s3 rm s3://dw-bucket/scratch/impala-test/impala-scratch \ --exclude '*query_id_1*' --exclude '*query_id_2*' --recursive
```



Note: The query ID in file names differs from the actual query ID. Colons (:) are replaced with underscores (_). For example, `xxxx:yyyy` appears as `xxxx_yyyy`.

- c) `aws s3 rm s3://dw-bucket/scratch/impala-test/impala-scratch/[***HOSTNAME***] --recursive`

If an Impala daemon shuts down unexpectedly and does not restart on the original host, it may leave behind scratch files in a directory named after the hostname. These files remain in remote storage under the configured remote scratch path and are not automatically removed. If the host is no longer in use by the Impala daemon and the directory exists in remote storage, Cloudera recommends you remove it using the above command.

Results

- Temporary scratch files in S3 are removed, freeing up storage.
- Active query data remains unaffected by the cleanup process.

Configuring an existing Impala Virtual Warehouse to spill to S3

A new Impala Virtual Warehouse requires no configuration to spill to S3. However, if you have an existing Impala Virtual Warehouse that you did not configure to spill to S3 when you created the Virtual Warehouse, configuration is required.

About this task

The Impala Virtual Warehouse on an AWS environment writes temporary data to S3 when you configure a spill to S3. This capability in an existing, but not a new, Virtual Warehouse, requires an entitlement.

If you have an existing Impala Virtual Warehouse, you need to take the following actions:

- Edit your existing Virtual Warehouse to specify an S3 URI to spill to S3.
 - After editing, you cannot change the S3 URI.
 - After editing, you cannot select additional storage in Scratch Space Limit per node. The default 300 instance storage is used.
- Ensure that Impala has read/write access to the configured scratch location on the Data Lake bucket using steps in [Configuring a policy to spill Impala temporary data to S3](#).


Alternatively, instead of using the automatic default scratch location or to configuring the location, you can run the following Cloudera CLI command `create-vw` to configure a custom scratch location. Specify the spill location using the `impala-options` option using the `spillToS3Uri` field.

After you have created the Virtual Warehouse configured to spill to a specific S3 location, you cannot change the S3 URI. The field becomes uneditable.

Before you begin

- To use an external S3 bucket for spilled data, [add an external S3 bucket to Cloudera Data Warehouse](#).
- Note the URI of the external S3 bucket you added. For example, `S3://mybucket/scratch/path`.

Procedure

1. From the Cloudera Management Console or Cloudera landing page, navigate to Data Warehouses.
2. Go to the Virtual Warehouses tab.
3. Click  Edit .
4. Set the spill to S3 location.
5. Click Save.
6. Configure read/write access to the configured scratch location on the Data Lake bucket using steps in [Identifying the spill location for Impala temporary data](#).

Setting the scratch space limit for spilling Impala queries

Running out of space to store query data on the SSD causes query failure. To prevent these failures, you need to know how to configure scratch space limits for Impala Virtual Warehouse (Cloudera Data Warehouse) service.

About this task

Configuring scratch space on EBS volumes incurs additional cost and requires an entitlement specified below. The compute nodes that are used for coordinator and executor Kubernetes pods in Impala Virtual Warehouses are equipped as follows:

- Attached solid state drive (SSD) storage, using NVMe (non-volatile memory express) protocol.
- Instance storage 2x300 GB allocated as follows:
 - Data cache (200 GiB default value)
 - Scratch space (200 GiB default value)

Scratch space requires a base 2 storage increment, such as GibiBytes.

If you have spilling queries that require more scratch space than your compute nodes have, you need to configure additional scratch space for Impala Virtual Warehouses on AWS Elastic Block Store (EBS). You configure scratch space limits when you create an Impala Virtual Warehouse. After the Virtual Warehouse is created, you cannot add scratch space limits by editing the configuration. Configuring scratch space on EBS volumes incurs additional cost. See [Amazon EBS Pricing](#) for details.

Before you begin

You must obtain the CDW_IMPALA_EBS_SCRATCH_SPACE entitlement to use this feature.

Procedure

1. Create a Virtual Warehouse as described in [Adding a new Virtual Warehouse](#), selecting the Impala engine type.



The screenshot shows a web form titled "New Virtual Warehouse". It has two main sections. The first section is labeled "Name" with a red asterisk, indicating it is required. Below it is a text input field containing the word "impala". The second section is labeled "Type" with a red asterisk. Below it are two radio button options: "HIVE" and "IMPALA". The "IMPALA" option is selected, indicated by a dark grey background and white text.

2. Select a database catalog, or accept the default.
3. Set User Groups that can access endpoints, keys and values for Tagging the Virtual Warehouse.
4. Select a size for the Virtual Warehouse.

5. In Scratch Space Limit per node (in GiBs), choose to only use the 300 GiB of instance storage for scratch space, or choose additional EBS volumes for spilling queries.

☒ Enable Impala HA

☐ Disable AutoSuspend

Scratch Space Limit per node (in GiBs) :

- ✓ 300 instance storage only
- 634 instance storage + EBS
- 968 instance storage + EBS
- 1302 instance storage + EBS
- 1636 instance storage + EBS
- 1970 instance storage + EBS
- 2304 instance storage + EBS
- 2638 instance storage + EBS
- 2972 instance storage + EBS
- 6550 instance storage + EBS
- 7000 instance storage + EBS
- 8000 instance storage + EBS
- 9000 instance storage + EBS
- 10000 instance storage + EBS
- 11000 instance storage + EBS
- 12000 instance storage + EBS
- 13000 instance storage + EBS

6. Click Create Virtual Warehouse.