

AWS Resource Panning

Date published: 2021-04-06

Date modified: 2024-06-03

CLOUDERA

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

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

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

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

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

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

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

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

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

Contents

AWS requirements for Cloudera DataFlow.....	4
Cloudera DataFlow networking in AWS.....	5
Use your own VPC.....	6
Allow Cloudera to create a VPC.....	7
Limitations on AWS.....	8
AWS restricted policies.....	8
Using Customer Managed Keys with Cloudera DataFlow.....	8
Define a new default KMS key for AWS account level EBS encryption.....	8
Define a new default KMS key for Cloudera environment level EBS encryption.....	10

AWS requirements for Cloudera DataFlow

As the administrator for your AWS environment, ensure that the environment meets the requirements for Cloudera Public Cloud and Cloudera DataFlow. Then set up your AWS cloud credential and register the environment.

Follow the steps to ensure that your AWS environment meets the Cloudera and Cloudera DataFlow requirements:

Understand your AWS account requirements for Cloudera

- Review the *Cloudera AWS account requirements*. The link is in the *Related information* section below.
- Verify that your AWS account for Cloudera has the required resources.
- Verify that you have the permissions to manage these resources.

Understand the Cloudera DataFlow requirements

- Verify that the following services are available in your environment for Cloudera DataFlow to use:
 - Network – Amazon VPC
 - Compute – Amazon Elastic Kubernetes Service (EKS)
 - Load Balancing – Amazon ELB Classic Load Balancer
 - Persistent Instance Storage – Amazon Elastic Block Store (EBS)
 - Database – Amazon Relational Database Service (RDS)
- Determine your networking option:
 - Use your own VPC
 - Allow Cloudera to create a VPC

To understand each option, see: *Cloudera DataFlow Networking*. The link is in the *Related information* section below.

- Regions:
 - Select a Cloudera Public Cloud-supported region that also includes the AWS Elastic Kubernetes Service (EKS).

For more information, see: *Cloudera Supported AWS regions* and the Region Table in *AWS Regional Services*. The links are in the *Related information* section below.

- Ports and outbound network access:
 - Review the port requirements for the Cloudera default security group. See: *Cloudera Management Console - Security groups*. The link is in the *Related information* section below.
 - Configure ports for NiFi to access your source and destination systems in the data flow.
 - If you are using a firewall or a security group setting to prevent egress from the workspace, you must ensure that the outbound destinations required by Cloudera DataFlow are reachable. For more information, see *Outbound network access destinations for AWS*. The link is in the *Related information* section below.
 - If the egress is blocked to these URLs, then autoscaling fails to pull new images and the instances will have broken pods.

Follow the recommended and minimum required security group settings by AWS. For more information, see *Amazon EKS security group considerations*. The link is in the *Related information* section below.

Set up an AWS Cloud credential

Create a role-based AWS credential that allows Cloudera Public Cloud to authenticate with your AWS account and has authorization to provision AWS resources on your behalf. Role-based authentication uses an IAM role with an attached IAM policy that has the minimum permissions required to use Cloudera.

To set up an AWS Cloud credential, see *Creating a role based provisioning credential for AWS*. The link is in the *Related information* section below.

After you have created this IAM policy, register it in Cloudera as a cloud credential. Reference this credential when you register an AWS environment in Cloudera environment as described in the next step.

Register an AWS environment in Cloudera Public Cloud

A Cloudera user must have the PowerUser role in order to register an environment. An environment determines the specific cloud provider region and virtual network in which resources can be provisioned, and includes the credential that should be used to access the cloud provider account.

To register an AWS environment in Cloudera Public Cloud, see *Cloudera AWS Environments*.

Related Concepts

[Cloudera DataFlow networking in AWS](#)

Related Information

[Cloudera Public Cloud](#)

[Cloudera Public Cloud supported AWS regions](#)

[AWS Regional Services](#)

[Cloudera Management Console](#)

[Amazon EKS security group considerations](#)

[Creating a role based provisioning credential for AWS](#)

[Cloudera Public Cloud](#)

[Outbound network access destinations for AWS](#)

Cloudera DataFlow networking in AWS

Cloudera DataFlow supports different networking options depending on how you have set up your VPC and subnets. If you want Cloudera DataFlow to use specific subnets, make sure that you specify them when registering a Cloudera environment.

If you specified a mix of public and private subnets during environment registration, Cloudera DataFlow by default will provision the Kubernetes nodes in the private subnets. For Cloudera DataFlow to work, the private subnets require outbound internet access. This can be achieved by configuring NAT gateways in separate public subnets and making sure that outbound internet traffic is routed via the NAT gateway. The VPC you are using must have an Internet Gateway set up which ultimately provides internet access to the public subnets. Following this approach allows the Cloudera DataFlow services running on Kubernetes nodes in your private network to connect to the internet while also preventing inbound connections from the internet.

You can configure Cloudera DataFlow to either use a private or public load balancer to allow users to connect to flow deployments. Using a private load balancer is possible when your Cloudera environment contains at least two private subnets. When you are using a private load balancer, you need to ensure connectivity between the client network from where your users are initiating connections and the private subnets in your VPC. This is typically done by setting up VPN access between the private subnets in AWS and the corporate network.

If you want to allow users to connect to flow deployments from the internet you can use the public load balancer option. This option will provision public load balancers in public subnets allowing your users to connect to flow deployments without the need to set up VPN connectivity between the private subnets and your corporate network.



Important: Cloudera recommends that you either use a fully private deployment in private subnets with private load balancers or a mix of private subnets with a public load balancer. Cloudera does not recommend provisioning Cloudera DataFlow in public subnets.

The image below represents a fully private deployment where Kubernetes nodes and load balancers are deployed in the private subnets.

Use your own VPC

If you choose to use your own VPC, verify that it meets the minimum requirements and review Cloudera's recommended setup.

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

Verify that your VPC meets the following requirements and recommendations:

Minimum requirements

- Cloudera DataFlow requires at least two subnets, each in a different Availability Zone (AZ). If you require a public endpoint for Cloudera DataFlow, provision at least one public subnet.
- Ensure that the CIDR block for the subnets is sized appropriately for each Cloudera DataFlow environment. You must have enough IPs to accommodate:
 - The maximum number of autoscaling compute instances.
 - A fixed overhead of 48 IP addresses for three instances for core Cloudera DataFlow services.
- You must enable DNS for the VPC.

Cloudera's recommended setup

- Provision two subnets, each in a different Availability Zone (AZ).
 - If you do not require a public endpoint, use two private subnets.
 - If you require a public endpoint, use one private subnet and one public subnet.
- Private subnets should have routable IPs over your internal VPN. If IPs are not routable, private Cloudera DataFlow endpoints must be accessed via a SOCKS. This is not recommended.
- Tag the VPC and the subnets as shared so that Kubernetes can find them. Also, for load balancers to be able to choose the subnets correctly, you must tag either the private or public subnets.

A tag in AWS consists of a key and a value.

- To tag private subnets, enter `kubernetes.io/role/internal-elb` for the key and `1` for the value.

▼ Tags - optional

Key	Value - optional	
<input type="text" value="kubernetes.io/role/internal-elb"/>	<input type="text" value="1"/>	<input type="button" value="Remove"/>
<input type="button" value="Add new tag"/>		
You can add 48 more tags.		

- To tag public subnets, enter `kubernetes.io/role/elb` for the key and `1` for the value.

▼ Tags - optional

Key	Value - optional	
<input type="text" value="kubernetes.io/role/elb"/>	<input type="text" value="1"/>	<input type="button" value="Remove"/>
<input type="button" value="Add new tag"/>		
You can add 48 more tags.		



Note: The load balancer must be on a public subnet for access to Cloudera DataFlow. By default, if they are available, Cloudera DataFlow will configure the EKS to run on private subnets.

Related Information

[VPC Console on AWS](#)

[Create and configure your VPC](#)

Allow Cloudera to create a VPC

You can choose to create a VPC through Cloudera.

If you choose to create a VPC through Cloudera, three subnets will be automatically created.

You will be asked to specify a valid CIDR in IPv4 range that will be used to define the range of private IPs for EC2 instances provisioned into these subnets.

For more information, see the AWS documentation *Amazon EKS - Cluster VPC Considerations* and *Creating a VPC for your Amazon EKS Cluster*. The links are in the *Related information* section below.

Related Information

[Amazon EKS - Cluster VPC Considerations](#)

[Creating a VPC for your Amazon EKS Cluster](#)

Limitations on AWS

Review the default AWS service limits and your current AWS account limits.

By default, AWS imposes certain default limits for AWS services for each user account. Make sure you review your account's current usage status and resource limits before you start provisioning additional resources for Cloudera and Cloudera DataFlow.

For example, depending on your AWS account, you may only be allowed to provision a certain number of EC2 instances. Be sure to review your AWS service limits before you proceed.

For more information, see the AWS documentation: *AWS Service Limits* and *Amazon EC2 Resource Limits*.

Cloudera DataFlow environments have the following resource limits on AWS:

- Certificate creation (for TLS) uses LetsEncrypt which is limited to 2000 certs/week. As such, a single tenant in Cloudera can create a maximum of 2000 flows per week.

Related Information

[AWS Service Limits](#)

[Amazon EC2 Resource Limits](#)

[ENI Max Pods](#)

AWS restricted policies

Customers with strict security policies beyond what the default Cloudera cross-account policy permits can enable Cloudera DataFlow for a Cloudera Public Cloud environment with more restricted IAM policies.

For information on setting up the required IAM roles and policies, see [Setting up Compute Cluster IAM permissions](#).

Using Customer Managed Keys with Cloudera DataFlow

By default, Cloudera DataFlow uses your account level KMS key for EBS storage encryption. You can optionally secure your data with a custom KMS key.

You have two options to implement Customer Managed Keys (CMKs):

- define a new default KMS key for EBS encryption on AWS account level
- define a key on Cloudera DataFlow environment level

Define a new default KMS key for AWS account level EBS encryption

When you define a new account level default key in AWS, you need to add policies to your key definition that allow for storage provisioning and fulfilling scaling requests.

About this task



Important: Defining a new default key affects all EBS storage encryption within your account.

Procedure

1. Create a custom encryption key on the AWS Management Console.

The key policy section of the new key must contain additional permissions. Add the three required permission blocks in the example below.

Replace `[***YOUR ACCOUNT ID***]` and `[***YOUR ACCOUNT REGION***]` with your AWS account ID and with the AWS region where you want to deploy Cloudera DataFlow, respectively.

```
{
  "Sid": "AllowAutoscalingServiceLinkedRoleForAttachmentOfPer
sistentResources",
  "Effect": "Allow",
  "Principal": {
    "AWS": "arn:aws:iam::[***YOUR ACCOUNT ID***]:role/aws-se
rvice-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
  },
  "Action": "kms:CreateGrant",
  "Resource": "*",
  "Condition": {
    "Bool": {
      "kms:GrantIsForAWSResource": "true"
    }
  }
},
{
  "Sid": "AllowAutoscalingServiceLinkedRoleUseOfTheCMK",
  "Effect": "Allow",
  "Principal": {
    "AWS": "arn:aws:iam::[***YOUR ACCOUNT ID***]:role/aws-se
rvice-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
  },
  "Action": [
    "kms:Encrypt",
    "kms:Decrypt",
    "kms:ReEncrypt*",
    "kms:GenerateDataKey*",
    "kms:DescribeKey"
  ],
  "Resource": "*"
},
{
  "Sid": "Allow EKS access to EBS.",
  "Effect": "Allow",
  "Principal": {
    "AWS": "*"
  },
  "Action": [
    "kms:CreateGrant",
    "kms:Encrypt",
    "kms:Decrypt",
    "kms:ReEncrypt*",
    "kms:GenerateDataKey*",
    "kms:DescribeKey"
  ],
  "Resource": "*",
  "Condition": {
```

```

        "StringEquals": {
            "kms:CallerAccount": "[***YOUR ACCOUNT ID***]",
            "kms:viaService": "ec2.[***YOUR ACCOUNT REGION***].amazonaws.com"
        }
    }
}

```



Important: If you fail to add these permissions, you will encounter failures when enabling Cloudera DataFlow since it will be unable to provision the necessary encrypted storage using the custom key.

2. Set the newly created key as the default KMS key for EBS encryption.
For more information, see [Default KMS key for EBS encryption](#).
3. If you are also use restricted IAM policies with Cloudera, make sure you provide the KMS CMK for volume encryption when you [Create the restricted policies and attach them to the cross-account role](#).

Define a new default KMS key for Cloudera environment level EBS encryption

When you define a custom KMS key at the Cloudera environment level, you need to add policies to your key definition that allow for storage provisioning and fulfilling scaling requests.

Procedure

1. Create a custom encryption key on the AWS Management Console.

The key policy section of the new key must contain additional permissions. Add the three required permission blocks in the example below.

Replace `[***YOUR ACCOUNT ID***]` and `[***YOUR ACCOUNT REGION***]` with your AWS account ID and with the AWS region where you want to deploy Cloudera DataFlow, respectively.

```

{
    "Sid": "AllowAutoscalingServiceLinkedRoleForAttachmentOfPersistentResources",
    "Effect": "Allow",
    "Principal": {
        "AWS": "arn:aws:iam::[***YOUR ACCOUNT ID***]:role/aws-service-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
    },
    "Action": "kms:CreateGrant",
    "Resource": "*",
    "Condition": {
        "Bool": {
            "kms:GrantIsForAWSResource": "true"
        }
    }
},
{
    "Sid": "AllowAutoscalingServiceLinkedRoleUseOfTheCMK",
    "Effect": "Allow",
    "Principal": {
        "AWS": "arn:aws:iam::[***YOUR ACCOUNT ID***]:role/aws-service-role/autoscaling.amazonaws.com/AWSServiceRoleForAutoScaling"
    },
    "Action": [
        "kms:Encrypt",
        "kms:Decrypt",
        "kms:ReEncrypt*",
        "kms:GenerateDataKey*"
    ]
}

```

```


        "kms:DescribeKey"
      ],
      "Resource": "*"
    },
    {
      "Sid": "Allow EKS access to EBS.",
      "Effect": "Allow",
      "Principal": {
        "AWS": "*"
      },
      "Action": [
        "kms:CreateGrant",
        "kms:Encrypt",
        "kms:Decrypt",
        "kms:ReEncrypt*",
        "kms:GenerateDataKey*",
        "kms:DescribeKey"
      ],
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "kms:CallerAccount": "[***YOUR ACCOUNT ID***]",
          "kms:viaService": "ec2.[***YOUR ACCOUNT REGION***].amazonaws.com"
        }
      }
    }
  ]
}

```



Important: If you fail to add these permissions, you will encounter failures when enabling Cloudera DataFlow since it will be unable to provision the necessary encrypted storage using the custom key.

2. When registering your Cloudera environment, follow these steps on the Region, Networking and Security page to assign the custom key:

 Customer Managed Encryption Keys

The provided encryption key will be used to encrypt attached volumes and external DBs. Please note that this key will not be used to encrypt cloud storage.

☒ Enable Customer-Managed Keys

Select Encryption Key*

Please select an Encryption Key

- a. Under Customer-Managed Keys, click Enable Customer-Managed Keys.
- b. Select the CMK you want to enable for this environment from the Select Encryption Key drop-down list.

For more information on registering a Cloudera environment, see [Register an AWS environment from Cloudera UI](#).

3. If you are also using restricted IAM policies with Cloudera, make sure you provide the KMS CMK for volume encryption when you [Create the restricted policies and attach them to the cross-account role](#).