Cloudera Data Engineering 1.20.3

Enabling and disabling Cloudera Data Engineering

Date published: 2020-07-30
Date modified: 2024-02-26

https://docs.cloudera.com/

Legal Notice

Contents

Enabling a Cloudera Data Engineering service

Before you can use the Cloudera Data Engineering (CDE) service, you must enable it on each environment that you want to use CDE on.

Before you begin

Make sure that you have a working environment for which you want to enable the CDE service. For more information about environments, see Environments.

Note: CDE on Microsoft Azure does not currently support SSD or Spot instances.

Procedure

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile. The CDE Home page displays.
2. Click Administration in the left navigation menu. The Administration page displays.
3. In the Services column, click the plus icon at the top to enable CDE for an environment.
4. Type the name of the service that you want to enable CDE for.
5. Select the Environment from the drop-down menu.
6. Select the Workload Type.

   The workload type corresponds to the instance size that will be deployed to run your submitted Spark jobs. When you select a type, the corresponding cloud provider instance size is displayed in the Summary section to the right.

   a) If you want to use SSD storage, check the box labeled Use SSD instances. In this configuration, SSD storage is used for the workload filesystem, such as the Spark local directory. If your workload requires more space than is available in the instance storage, select a larger instance type with sufficient local storage or select an instance type without SSD, and then configure the EBS volume size.
7. Set the Auto-Scale Range under the Capacity & Costs section:

   The range you set here creates an auto scaling group with the specified minimum and maximum number of instances that can be used. The CDE service launches and shuts down instances as needed within this range. The instance size is determined by the Workload Type you selected.

   For Azure: Set the On-demand Instances range. This option displays for Core (Tier 1) cluster types during service creation. After the service and cluster is created, you can edit the service and set the All purpose On-demand Instances range under the Capacity & Costs section.

   For AWS: After the service and clusters are created, you can set the All purpose On-demand Instances range and the All purpose Spot Instances range under the Capacity & Costs section. This option displays for the All-purpose (Tier 2) cluster types.
8. If you want to use spot instances, check the box labeled Use Spot instances and select a range of spot instances to request. This creates another auto scaling group of spot instances. Spot instances are requested with similar CPU and memory profiles as the instances selected for the Workload Type. For more information, see Cloudera Data Engineering Spot Instances.

   Note: Duplicate auto-scaling groups of the same size are created for supporting tiered pricing. These can be edited at a later time if needed.
9. Optional: Enable a private network (preview feature). This feature ensures that the CDE service is deployed with a secure network setup based on the cloud provider such as Microsoft Azure or Amazon Web Services (AWS). For more information see the links below on how to configure this feature as there are prerequisites needed before you see the Enable Private Network option in the CDE user interface under Network & Storage.
10. If you create the service in an AWS environment using a non-transparent proxy, you find a Proxy CIDR Source Ranges field. You are only required to enter the proxy CIDR ranges for the proxy instances if you registered

your proxies using host names, as the Cloudera Control Plane has no way to resolve those to IPs. If your proxy instances were registered using IP addresses, you can leave this field blank.
For example, if you have a load balanced proxy with static IPs running on 10.80.199.105 and 10.80.200.45, add "10.80.199.105/32" and "10.80.200.45/32". If your proxy instances are dynamic (behind a load balancer or virtual IP) then you would enter a wider range, e.g.: "10.80.200.0/24".

Note: Currently non-transparent support is only available on Amazon Web Services.

11. If you want to create a load balancing endpoint in a public subnet, check the box labeled Enable Public Loadbalancer. If you leave this unchecked, the load balancing endpoint will be created in a private subnet, and you will need to configure access manually in your cloud account.

Important: When you enable the public loadbalancer, you need at least one public subnet configured in the environment.

12. Specify a comma-separated list of CIDRs in API server Authorized IP Ranges that can access the Kubernetes master API server.

    You may specify a comma-separated list of CIDRs that can access the Kubernetes master API server.

Attention: Leaving this field empty renders the Kubernetes API server open to all traffic.

    Make sure that the provided IP addresses do not overlap with the following ranges:

    - 0.0.0.0 - 0.255.255.255
    - 10.0.0.0 - 10.255.255.255
    - 100.64.0.0 - 100.127.255.255
    - 127.0.0.0 - 127.255.255.255
    - 169.254.0.0 - 169.254.255.255s
    - 172.16.0.0 - 172.31.255.255
    - 192.0.0.0 - 192.0.0.255
    - 192.0.2.0 - 192.0.2.255
    - 192.88.99.0 - 192.88.99.255
    - 192.168.0.0 - 192.168.255.255
    - 198.18.0.0 - 198.19.255.255
    - 198.51.100.0 - 198.51.100.255
    - 203.0.113.0 - 203.0.113.255
    - 224.0.0.0 - 239.255.255.255
    - 240.0.0.0 - 255.255.255.254
    - 255.255.255.255

13. Specify a comma-separated list of client IP ranges in Load Balancer Source Ranges that should be allowed to access the load balancer.
14. Specify which subnets to use for the Kubernetes worker nodes. Select from available Subnets in the drop-down list.
15. Specify which subnets to use for the Load balancer. Select from available Subnets in the drop-down list.
16. Optional: Check the box labeled Enable Observability Analytics if you want diagnostic information about jobs and query execution sent to Cloudera Observability. This helps optimize troubleshooting.
17. Check the box labeled Enable Workload Analytics to automatically send diagnostic information from job execution to Cloudera Workload Manager.

18. Optionally add Tags as needed. Tags are applied to the cloud provider resources associated with the CDE service (including virtual clusters created in that service). For more information about tags, see the cloud provider documentation:

    Amazon AWS

    Tagging AWS resources

    Microsoft Azure

    Use tags to organize your Azure resources and management hierarchy

    Note: The following tags are added automatically by CDE along with the custom Tags: "cde-cluster-id", "cde-provisioner-id", "cde-owner-email", "Cloudera-Resource-Name", "owner". After CDE 1.20.3, the "owner" tag is no longer added. If you ever rely on the "owner" tag, please use "cde-owner-email" instead.
19. Default Virtual Cluster selection is enabled by default to create a default virtual cluster after enabling a CDE service. This will help you get a jump start to create your jobs easily, without having to wait to create a CDE virtual cluster as mentioned in Creating virtual clusters, making the onboarding smoother. You can turn this toggle off if you do not wish to use a default virtual cluster.
20. Click Enable.

Results

The CDE Administration page displays the status of the environment initialization. You can view logs for the environment by clicking on the environment vertical ellipsis menu, and then clicking View Logs.

Related Information

Cluster Connectivity Manager (CCM)

Enabling Cluster Connectivity Manager (CCM) in the Management Console

Enabling a fully private network for a CDE service for Azure (Tech Preview)

Learn how to enable a fully private network setup for a Cloudera Data Engineering (CDE) service for Azure services in Cloudera Data Platform (CDP). Additionally, you can learn how to add User Defined Routing (UDR) in the UI or CLI. The UDR helps from exposing public IP addresses in your service.

This feature ensures that all Azure services used by CDE are provisioned as private (private Azure Kubernetes Service (AKS), MySQL, and Storage Accounts). The Azure cluster is deployed as a fully private network cluster when you enable a CDE service and enables VNet access through private endpoints and private links. Lastly, CDE on Microsoft Azure does not currently support SSD or Spot instances.

For CDE UI

Before you begin

- Ensure that you have created and enabled a CDE service. Additionally, the Cloudera Data Platform (CDP) must communicate with the CDE service on a private network in order to manage the CDE service lifecycle. This communication occurs using the Cluster Connectivity Manager (CCM) v2; therefore, to enable this feature, the CDP environment must be enabled with the CCMv2. Once the CCMv2 is enabled at the CDP environment level, the Enable Private Network option displays in the CDE user interface when you enable a service. For more information on how to enable a CDE service and set up CCMv2, refer to the links below.

  Note: To enable UDR, you must enable a private network flag and you must provide a subnet.

UI steps for enabling a private network and enabling UDR

1. While enabling a CDE service for an Azure environment, select Enable Private Network. Optionally, once you've enabled a private network on Microsoft Azure, you can select the User Defined Routing checkbox. Use

this to avoid exposing public IP addresses in your service by using a user defined routing (UDR) table. After, you'll need to specify a Subnet.

2. Click Enable.

For CDP CLI

You can enable the user defined routing (UDR) with the CDP CLI using the –network-outbound-type CLI switch with a value of “UDR”. See the example command:

```
./clients/cdpcli/cdp.sh de enable-service --name "test-service-cdpcli" -
-env "dex-priv-env" --instance-type "Standard_D8s_v4" --minimum-instance
s 0 --maximum-instances 10 --enable-private-network --subnets dex-dev.in
ternal.19.westus2 --network-outbound-type UDR
```

Related Information

Enabling a CDE service

Cluster Connectivity Manager (CCM)

Enabling Cluster Connectivity Manager (CCM) in the Management Console

Enabling a semi-private network for a CDE service with AWS (Tech Preview)

Learn how to enable a semi-private network setup for a Cloudera Data Engineering (CDE) service with Amazon Web Services (AWS) services in Cloudera Data Platform (CDP). When you enable a CDE service with this feature, the Amazon Kubernetes Service (EKS) cluster is deployed as a private cluster but some services used by CDE such as MySQL and S3 are not provisioned as private.

Note: You need to contact Cloudera to have this feature enabled.

For CDE UI

Before you begin

- Ensure that you have created and enabled a CDE service. Additionally, the Cloudera Data Platform (CDP) must communicate with the CDE service on a private network in order to manage the CDE service lifecycle. This communication occurs using the Cluster Connectivity Manager (CCM) v2; therefore, to enable this feature, the CDP environment must be enabled with the CCMv2. Once the CCMv2 is enabled at the CDP environment level, the Enable Private Network option displays in the CDE user interface when you enable a service. For more information on how to enable a CDE service and set up CCMv2, refer to the links below.

UI steps for enabling a private network

1. While enabling a CDE service for an AWS environment, under Network & Storage, select Enable Private Network.
2. Click Enable.

For CDP CLI

You can enable a private network with the CDP CLI with the following commands:

```
cdp de enable-service --name dsp-private-eks-ntp-try1 --env dsp-aws-ntp-
priv --instance-type m5.2xlarge --minimum-instances 0 --maximum-instance
s 4 --initial-instances 0 --root-volume-size 50 --no-skip-validation --e
nable-private-network
```

Related Information

Enabling a CDE service

Cluster Connectivity Manager (CCM)
Enabling Cluster Connectivity Manager (CCM) in the Management Console

Managing a CDE Service

You can view configuration, metrics, and logs of existing CDE services. You can use the Edit option to make the configuration changes dynamically.

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile. The CDE Home page displays.
2. Click Administration in the left navigation menu. The Administration page displays.
3. In the Services column, click the Service Details icon for the CDE service you want to manage.
4. On the Configuration tab, you can view details about the service, including the service name and CDP environment.
5. Optional: Click the Edit option to make the configuration changes dynamically which may take a few minutes to update.

You can switch between the following tabs to view additional information:

Configuration

The Configuration tab lists details about the service name, CDP environment, and the CPU and memory capacity. You can modify the CPU and memory capacity dynamically.

Charts

The Charts tab displays the charts related to CPU Requests, Memory Requests, Jobs, and Nodes.

Logs

The Logs tab displays the latest log entries for the CDE service.

Access

The Access tab displays the option to add Amazon Web Services (AWS) user's Amazon Resource Names (ARNs) for the CDE Service. The Access tab also displays the users you have added to the CDE Service which will provide them with the following:

- Access to all pods
- Access to all secrets including TGTs
- Bypass Istio security
- Access to tgtgen host keytab with ability to create, delete, and modify users in FreeIPA

Diagnostics

The Diagnostics tab provides option to generate and download diagnostics bundle.

Removing a Cloudera Data Engineering service

Disabling an existing Cloudera Data Engineering (CDE) service stops all jobs, and deletes all associated virtual clusters and virtual cluster metadata. Do not do this unless you are certain that you no longer need any of these. Disabling CDE does not delete your CDP data. If enabling the CDE service on an environment for the first time fails, you must disable the service before you can try again. In this scenario, there are no clusters or jobs, and you can safely perform this procedure.

Before you begin

Important: The user interface for CDE 1.17 and above has been updated. The left-hand menu was updated to provide easy access to commonly used pages. The steps below will vary slightly, for example, the Overview page has been replaced with the Home page. You can remove a CDE service by clicking Administration on the left-hand menu, then proceed to step 2 listed below. The new home page still displays Virtual Clusters, but now includes quick-access links located at the top for the following categories: Jobs, Resources, and Download & Docs.

Procedure

1. In the Cloudera Data Platform (CDP) console, click the Data Engineering tile and click Overview.
2. In the CDE Services column, click the menu icon for the environment for which you want to disable the CDE service, and then click Disable CDE

   Warning: Disabling an existing Cloudera Data Engineering (CDE) service stops all jobs, deletes all associated virtual clusters and virtual cluster metadata. Do not do this unless you are certain that you no longer need any of these. Additionally, if you're prompted to perform a Force Disable, in the event that a Disable is not successful, you must perform a manual cleanup of cloud infrastructures such as Security Group, EBS Volume, and S3 Bucket. A manual cleanup is not required for a standard Disable.
3. Confirm that you want to disable CDE by typing the environment name and then clicking Disable.

Results

The CDE Administration page displays the status of the environment that is being disabled.

What to do next

If you disabled CDE as a result of a failure to enable CDE for the first time on an environment, resolve any reported issues, and then try again.

Limiting Incoming Endpoint Traffic for CDE Services For AWS

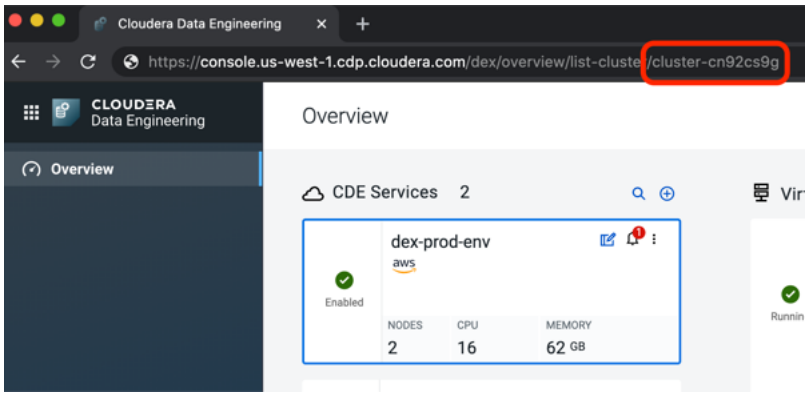
You can limit incoming endpoint traffic for a CDE service.

Before you begin

Important: The user interface for CDE 1.17 and above has been updated. The left-hand menu was updated to provide easy access to commonly used pages. The steps below will vary slightly, for example, the Overview page has been replaced with the Home page. The new home page still displays Virtual Clusters, but now includes quick-access links located at the top for the following categories: Jobs, Resources, and Download & Docs.

Procedure

1. Note the CDE service ID, which you can obtain from the URL in the CDE Management Console when the service is highlighted:



2. Go to the AWS console for the account where the CDE service is enabled.
3. Navigate to EC2 -> Load Balancers in the AWS console and enter the following filter:

```
 tag:cde-cluster-id : <id_from_step_1>
, e.g. tag:cde-cluster-id : cluster-cn92cs9g
```

4. Select the Load Balancer instance and then under the Description tab in the Detail window, click on the link to the Source Security Group.
5. In the subsequent view, select the correct security group ID.
6. In the subsequent window, click Edit inbound rules.
7. Modify the "0.0.0.0/0" CIDR ranges for the HTTPS rule to your desired CIDR ranges.

   Add additional ranges and rules as required but note that HTTPS traffic must be enabled for each range. The HTTP (port 80) and ICMP rules can be removed.