Cloudera DataFlow for Data Hub 7.2.17

Setting up your Edge Management cluster

Date published: 2019-12-16
Date modified: 2023-06-27

https://docs.cloudera.com/

Legal Notice

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

Checking prerequisites

Before you start creating your Edge Flow Management Data Hub Cluster, you need to ensure that you have set up the environment properly and have all the necessary accesses to use CDP Public Cloud. Use this checklist to verify that you meet all the requirements before you start creating the cluster.

- You have CDP login credentials.
- You have an available CDP environment.

  When you register your environment, make sure that the correct security access settings are configured. You need to enable SSH access and specify SSH key so that you can generate certificates for the agents. For more information on creating a CDP environment, see:

  - Working with AWS environments
  - Working with Azure environments
  - Working with GCP environments
- You have a running Data Lake. For more information on the Data Lake service in CDP environment, see Introduction to Data Lakes.

  > ⚠ **Important:** Make sure that the Runtime version of the Data Lake cluster matches the Runtime version of the Data Hub cluster that you are about to create. If these versions do not match, you may encounter warnings and/or errors.
- You have a CDP username and the predefined resource role of this user is EnvironmentAdmin.
- Your CDP user is synchronized to the CDP Public Cloud environment.

If you need more information about CDP basics, see Getting started as a user.

Creating your cluster

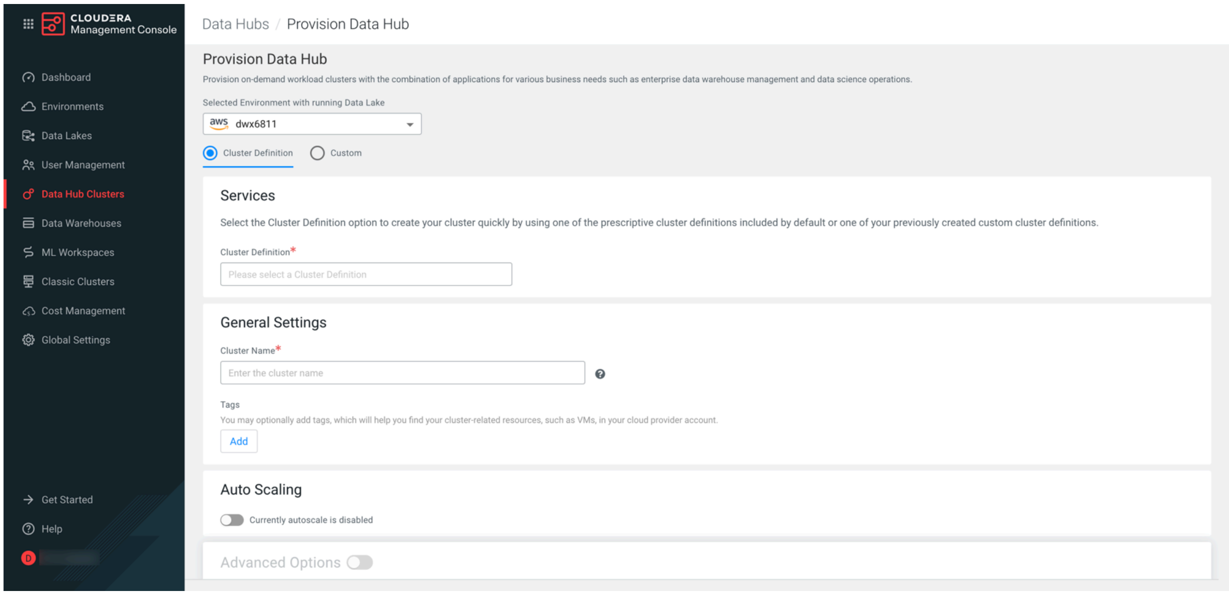
If you meet all prerequisites, you are ready to create a managed and secured Edge Flow Management cluster in CDP Public Cloud by using the prescriptive cluster definition available in Technical Preview.

Procedure

1. Log into the CDP web interface.
2. Navigate to Management Console Environments and select the environment where you want to create a cluster.

3. Click Create Data Hub.

   The Provision Data Hub page is displayed:



Provision Data Hub

Provision on-demand workload clusters with the combination of applications for various business needs such as enterprise data warehouse management and data science operations.

Selected Environment with running Data Lake: aws dwx6811

● Cluster Definition ○ Custom

Services

Select the Cluster Definition option to create your cluster quickly by using one of the prescriptive cluster definitions included by default or one of your previously created custom cluster definitions.

Cluster Definition*: Please select a Cluster Definition

General Settings

Cluster Name*: Enter the cluster name

Tags

You may optionally add tags, which will help you find your cluster-related resources, such as VMs, in your cloud provider account.

Auto Scaling

Currently autoscale is disabled

Advanced Options

4. Select Cluster Definition.
5. Select the appropriate Edge Flow Management cluster definition from the Cluster Definition dropdown depending on which cloud privider you are using.

   There are three template options available:

   - Edge Flow Management Light Duty for AWS
   - Edge Flow Management Light Duty for Azure
   - Edge Flow Management Light Duty for GCP

   The cluster template referenced in the selected cluster definition determines which services are included in the cluster. The list of services is automatically displayed below the selected cluster definition name. It shows that the cluster definition contains the Edge Flow Manager.
6. Provide a cluster name and add tags you might need.

   Note:

   The name must be between 5 and 40 characters, it must start with a letter, and should only include lowercase letters, numbers, and hyphens.

   You can define tags that will be applied to your cluster-related resources on your cloud provider account. For more information, see *Tags*.
7. Use the Configure Advanced Options section to customize the infrastructure settings.

   For more information on these options, see the *Advanced cluster options* for your cloud environment.
8. Click Provision Cluster.

Results

The new Data Hub cluster appears on the Data Hubs tab of the Clusters page. You can follow the status of the provisioning process in the Status column. When your cluster is ready, its status changes to Running.

Related Information

Tags

Advanced cluster options for AWS

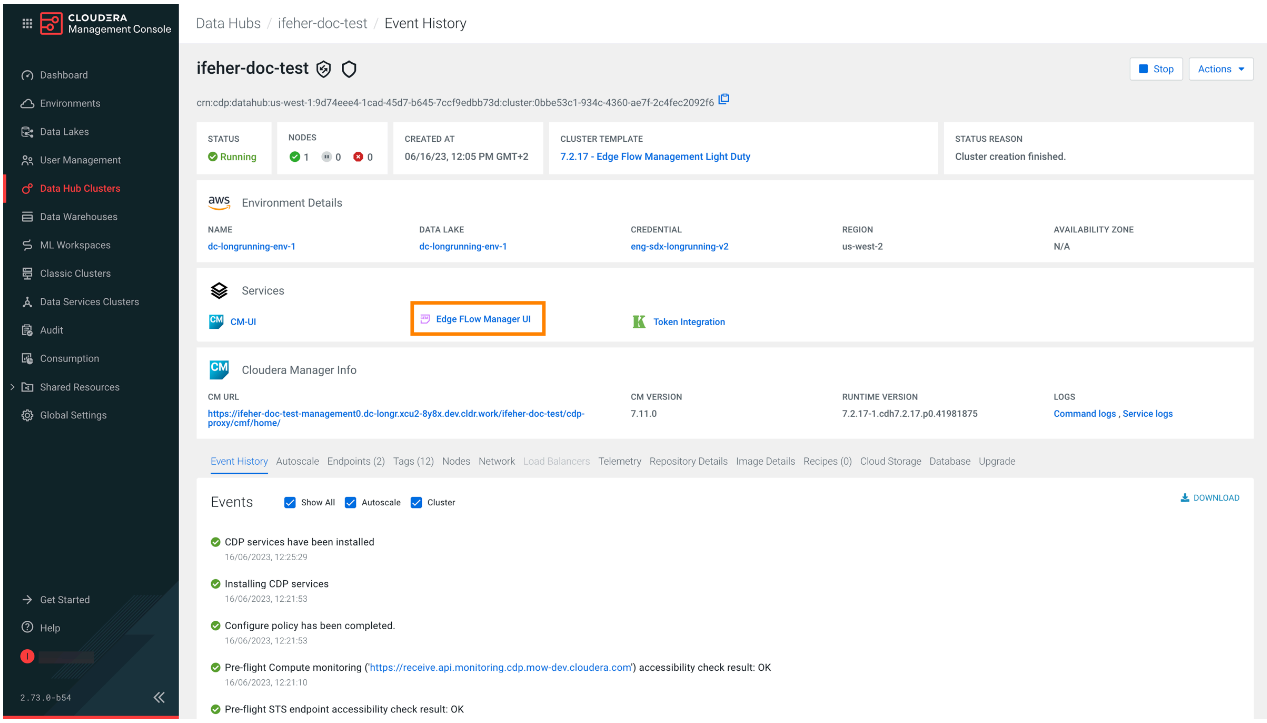
Advanced cluster options for Azure

[Advanced cluster options for GCP](#)

After creating your cluster

The cluster you have created using the Edge Flow Management cluster definition is secured by default, and it is integrated with Knox SSO.

You can access the EFM UI from the Services section of the Data Hub cluster page. Click the CEM icon or the Edge Flow Manager UI link and you are redirected to the EFM page.



Data Hubs / ifeher-doc-test / Event History

ifeher-doc-test

crn:cdp:datahub:us-west-1:9d74eee4-1cad-45d7-b645-7ccf9edbb73d:cluster:0bbe53c1-934c-4360-ae7f-2c4fec2092f6

STATUS	NODES	CREATED AT	CLUSTER TEMPLATE	STATUS REASON
Running	✓ 1 ⏸ 0 ✗ 0	06/16/23, 12:05 PM GMT+2	7.2.17 - Edge Flow Management Light Duty	Cluster creation finished.

aws Environment Details

NAME	DATA LAKE	CREDENTIAL	REGION	AVAILABILITY ZONE
dc-longrunning-env-1	dc-longrunning-env-1	eng-sdx-longrunning-v2	us-west-2	N/A

Services

CM-UI   Edge FLow Manager UI   Token Integration

Cloudera Manager Info

CM URL	CM VERSION	RUNTIME VERSION	LOGS
https://ifeher-doc-test-management0.dc-longr.xcu2-8y8x.dev.cldr.work/ifeher-doc-test/cdp-proxy/cmf/home/	7.11.0	7.2.17-1.cdh7.2.17.p0.41981875	Command logs , Service logs

Events ☑ Show All ☑ Autoscale ☑ Cluster

- CDP services have been installed
  16/06/2023, 12:25:29
- Installing CDP services
  16/06/2023, 12:21:53
- Configure policy has been completed.
  16/06/2023, 12:21:53
- Pre-flight Compute monitoring ('https://receive.api.monitoring.cdp.mow-dev.cloudera.com') accessibility check result: OK
  16/06/2023, 12:21:10
- Pre-flight STS endpoint accessibility check result: OK

The user that creates the Data Hub cluster is added as an administrator in EFM and can access the UI automatically. Other users can log in, but they must be granted access by the administrator before they can access data in EFM. To secure the communication between agents and EFM, you need to generate and utilize proper certificates. You also need to add the agents that you want to manage with EFM.

Giving access to your cluster

When your cluster has been created successfully, EFM is running as a Data Hub and the token provided by Knox is translated to an EFM token internally.

The administrator must grant access to all other users on the EFM Administration page before they can access data in EFM. For more information about user management and access control in CEM, see *Access control policies*.

Related Information

[Access control policies](#)

Enabling remote agent deployment in Edge Flow Manager

Edge Flow Manager (EFM) supports the deployment and automatic configuration of MiNiFi agents (including security settings), enabling streamlined agent provisioning.

About this task

Remote agent deployment simplifies the agent deployment process by generating a one-liner command that you can run on the target host. The selected agent binary is downloaded, configured, and started without requiring manual intervention, allowing you to start to work on your data flows immediately.

You have to perform the following steps to enable and configure the Remote Agent Deployer:

Procedure

1. Open port for direct EFM access.

   While user traffic accessing the EFM UI is routed through Knox, the agents running outside the CDP deployment need to access EFM directly. To enable this, open a specific port for the agents on the host where EFM is deployed. The default port is 10090, utilized by CEM components for the C2 Protocol.
2. Deploy MiNiFi agent binaries.

   Place the agent binaries in the agent deployer directory of EFM. The default directory is configured to /hadoopfs/fs{1-4}/agent-deployer/binaries on the EFM host.

Results

Once the port is open and the agent binaries are in place, you can access this functionality from the EFM UI. For more information on using the Remote Agent Deployer, see *Deploying agents in CEM*.

Related Information

Deploying agents in CEM

Generating certificates for MiNiFi agents

To secure the communication between agents and EFM, you need to generate and use proper certificates.

About this task

Edge Flow Manager (EFM) is a secured application, which has to be bootstrapped with the initial admin identity. The initial admin is the person who is able to assign roles and manage permissions in EFM. In the Technical Preview, the initial admin is the workload user of the person who deploys the Data Hub. For more information about authentication and authorization, see *Access control bootstrapping*.

While the user traffic accessing the UI utilizes Knox, the agents running outside of the CDP deployment need to access EFM directly. To enable this, you have to open a port for the agents on the host where EFM is deployed. By default, this port is 10090, used by CEM components for C2 Protocol.

You do not have to generate the certificates from the agent host. You can generate them on any host that has access to the management node. When created, you can copy the certificates to the appropriate agent host.

In test environments it is not necessary to create different certificates for all agents. The same certificate can be configured for all agents. However, in production environments it is highly recommended to create a certificate for each agent.

Generating certificates with this approach is similar to adding a node to the cluster using Cloudera Manager.

Note: Agents using these certificates are considered to be the members of the cluster managed by Cloudera Manager. Use your certificates with care and protect them from illegal access.

MiNiFi agents need to set up mTLS (mutual TLS) for C2 communication to be able to communicate with EFM. For information on MiNiFi Java agent authentication, see *Securing MiNiFi Java Agent*. For information on MiNiFi C++ agent authentication, see *Securing MiNiFi C++ Agent*.

In CDP Public Cloud, certificates are managed by Cloudera Manager, acting as a certificate authority. All certificates are generated by Cloudera Manager, there is no option to use custom certificates.

Note:

In the Technical Preview version of CEM for CDP Public Cloud, you have to set up agent security manually. In later versions there will be an option to set up agent security using EFM.

Before you begin

- You have a running CEM Public Cloud cluster
- SSH access is configured to the management node of the cluster
- You have an SSH user with keypair that has sudo privileges
- You have the host name of the Edge Management cluster’s management node
- An external node is available from which you are able to SSH into the Edge Management cluster’s management node

Procedure

1. Create a working directory on your external node that has SSH access to your Edge Flow Management cluster.
2. Save the following script to the previously created working directory, and name it create_certs.sh.

```
#!/bin/bash
set -eo pipefail

# input parameters
SSH_USER=$1
SSH_KEY=$2
CM_HOST=$3
AGENT_FQDN=$4

EXAMPLE_USAGE="Example usage: ./create_certs.sh sshUserName ~/.ssh/userKe
y.pem host0.company.site agent-x.company.site"

[[ -z "$SSH_USER" ]] && echo "SSH User parameter is missing. $EXAMPLE_US
AGE" && exit 1
[[ -z "$SSH_KEY" ]] && echo "SSH Key parameter is missing. $EXAMPLE_USAGE"
 && exit 1
[[ -z "$CM_HOST" ]] && echo "Cloudera Manager parameter is missing. $EXA
MPLE_USAGE" && exit 1
[[ -z "$AGENT_FQDN" ]] && echo "Agent FQDN parameter is missing. $EXAMPLE_
USAGE" && exit 1

KEYSTORE_PASSWORD=$(hexdump -vn16 -e'4/4 "%08X" 1 "\n"' /dev/urandom | tr
'[:upper:]' '[:lower:]')

# constants
GENERATED_CREDENTIALS_ARCHIVE=credentials.tar
GENERATED_CREDENTIALS_REMOTE_PATH="/tmp/$GENERATED_CREDENTIALS_ARCHIVE"
CM_SITE_PACKAGES="/opt/cloudera/cm-agent/lib/python3.8/site-packages"
ORIGINAL_CERTMANAGER_BASE_DIR="/etc/cloudera-scm-server/certs"
CUSTOM_CERTMANAGER_BASE_DIR="/root/certs"
CERT_PASSWORDS_DIR="$CUSTOM_CERTMANAGER_BASE_DIR/private"
GLOBAL_KEY_PASSWORD_FILE="$CERT_PASSWORDS_DIR/.global_key_password"
GLOBAL_TRUSTSTORE_PASSWORD_FILE="$CERT_PASSWORDS_DIR/.global_truststore
_password"

rm -rf "$AGENT_FQDN"
mkdir "$AGENT_FQDN"

remote_ssh_command=$(cat << EOF
sudo \cp -n -R $ORIGINAL_CERTMANAGER_BASE_DIR $CUSTOM_CERTMANAGER_BASE_DI
R;
sudo /opt/rh/rh-python38/root/bin/python -c "import site; site.addsitedir
('$CM_SITE_PACKAGES'); import cmf.tools.cert; passwd = cmf.tools.cert.re
ad_obfuscated_password('$GLOBAL_TRUSTSTORE_PASSWORD_FILE'); print(passwd
);"
sudo rm -f $GLOBAL_KEY_PASSWORD_FILE;
sudo /opt/rh/rh-python38/root/bin/python -c "import site; site.addsitedir(
'$CM_SITE_PACKAGES'); import cmf.tools.cert; cmf.tools.cert.write_obfusc
ated_password('$GLOBAL_KEY_PASSWORD_FILE', '$KEYSTORE_PASSWORD');";
sudo /opt/cloudera/cm-agent/bin/certmanager --location "$CUSTOM_CERTMANA
GER_BASE_DIR" gen_node_cert --output "$GENERATED_CREDENTIALS_REMOTE_PATH"
 --rotate "$AGENT_FQDN";
sudo chmod 666 "$GENERATED_CREDENTIALS_REMOTE_PATH";
EOF
)

ssh -i "$SSH_KEY" -o StrictHostKeyChecking=no "$SSH_USER"@"$CM_HOST" "$rem
ote_ssh_command" > "$AGENT_FQDN/cm-auto-in_cluster_trust.pw" 2> /dev/null
scp -r -i "$SSH_KEY" -o "StrictHostKeyChecking=no" "$SSH_USER"@"$CM_HOST":
"$GENERATED_CREDENTIALS_REMOTE_PATH" "$AGENT_FQDN/" 2> /dev/null
tar -xf "$AGENT_FQDN/$GENERATED_CREDENTIALS_ARCHIVE" -C "$AGENT_FQDN"
echo "MiNiFi-Java KeyStore File":
ls -alh "$AGENT_FQDN/cm-auto-host_keystore.jks"
echo "MiNiFi-Java TrustStore File:"
ls -alh "$AGENT_FQDN/cm-auto-in_cluster_truststore.jks"
echo "MiNiFi-CPP Client certificate":
ls -alh "$AGENT_FQDN/cm-auto-host_key_cert_chain.pem"
echo "MiNiFi-CPP Client private key":
ls -alh "$AGENT_FQDN/cm-auto-host_key.pem"
echo "MiNiFi-CPP CA certificate"
ls -alh "$AGENT_FQDN/cm-auto-in_cluster_ca_cert.pem"
echo "KeyStore / HostKey Password: sensitive data, please check for it in
 $AGENT_FQDN/cm-auto-host_key.pw"
echo "TrustStore Password: sensitive data, please check for it in $AGENT_
FQDN/cm-auto-in_cluster_trust.pw"

rm -f "$AGENT_FQDN/cm-auto-global_cacerts.pem" "$AGENT_FQDN/cm-auto-globa
l_truststore.jks" "$AGENT_FQDN/$GENERATED_CREDENTIALS_ARCHIVE" "$AGENT_F
QDN/cm-auto-host_cert_chain.pem"
```

3. Make the script executable.

```
chmod +x create_certs.sh
```

4. Run the script with the following parameters:

```
./create_certs.s
h **[ssh_user}** **[ssh_private_key]** **[management_node_host_name]** **[agent_fqdn]
```

For example:

```
./create_certs.sh adminuser ~/.ssh/adminuser.pem management-node.company
.site.com agent-1.company.site.com
```

The script should print a similar output:

```
credentials.tar

                                                     100%  420KB 222.0KB/s
   00:01
MiNiFi-Java KeyStore File:
-rw-------@ 1 user  group   5.2K Apr 24 13:33 agent-1.company.site.com/cm-
auto-host_keystore.jks
MiNiFi-Java TrustStore File:
```

```
-rw-r-----@ 1 user  group   2.3K Apr 24 13:19 agent-1.company.site.com/cm-auto-in_cluster_truststore.jks
MiNiFi-CPP Client certificate:
-rw-------@ 1 user  group   7.1K Apr 24 13:33 agent-1.company.site.com/cm-auto-host_key_cert_chain.pem
MiNiFi-CPP Client private key:
-rw-------@ 1 user  group   2.5K Apr 24 13:33 agent-1.company.site.com/cm-auto-host_key.pem
MiNiFi-CPP CA certificate
-rw-r-----@ 1 user  group   3.0K Apr 24 13:19 agent-1.company.site.com/cm-auto-in_cluster_ca_cert.pem
KeyStore / HostKey Password: sensitive data, please check for it in agent-1.company.site.com/cm-auto-host_key.pw
TrustStore Password: sensitive data, please check for it in agent-1.company.site.com/cm-auto-in_cluster_trust.pw
```

A directory is created with the same name as the agent’s FQDN, provided as a parameter for the script. The directory contains all the necessary keystores and certificates for configuring mTLS authentication.

The keystore and truststore passwords are not printed as they are sensitive information. You can find them in the directory that was created with the following names:

- cm-auto-host_key.pw
- cm-auto-in_cluster_trust.pw

5. Set the agent parameters.

- For the MiNiFi Java agent:

```
c2.security.truststore.location=/path/to/cm-auto-in_cluster_truststore.jks
c2.security.truststore.password=<password_from_cm-auto-in_cluster_trust.pw>
c2.security.truststore.type=JKS
c2.security.keystore.location=/path/to/cm-auto-host_keystore.jks
c2.security.keystore.password=<password_from_cm-auto-host_key.pw>
c2.security.keystore.type=JKS
```

- For the MiNiFi C++ agent:

```
nifi.security.client.certificate=/path/to/cm-auto-host_key_cert_chain.pem
nifi.security.client.private.key=/path/to/cm-auto-host_key.pem
nifi.security.client.pass.phrase=/path/to/cm-auto-host_key.pw
nifi.security.client.ca.certificate=/path/to/cm-auto-in_cluster_ca_cert.pem
```

Note:

Although the parameter is called Agent FQDN, it is not mandatory to use the agent’s domain name. You can use any other string. Keep in mind that the string you provide will be the common name (CN) in the generated certificate.

Related Information

Access control bootstrapping

Securing MiNiFi Java Agent

Securing MiNiFi C++ Agent

Adding agents to your cluster

When your cluster has been created successfully, you can add agents that you want to manage with EFM. Agents are deployed outside of CDP Public Cloud, so follow the standard agent deployment instructions:

Java agents

Installing the MiNiFi Java agent

C++ agents

Installing the MiNiFi C++ agent

Note:

Make sure that you point the agents to heartbeat to the Data Hub EFM deployment.