Cloudera DataFlow for Data Hub 7.2.18

Setting up your Flow Management cluster

Date published: 2019-12-16
Date modified: 2024-04-03

https://docs.cloudera.com/

Legal Notice

Contents

Checking prerequisites

Before you start creating your 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 check that you meet all the requirements before you start creating the cluster. If you need more information about CDP basics, see *Getting started as a user*.

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

  Note:

  If you want to deploy your Flow Management Data Hub cluster on multiple availability zones (use multi-AZ support), make sure that multiple subnets have been selected during environment creation. In the Network section of the Region, Networking and Security page, select as many subnets as needed.

  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.

  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.

  For more information on the Data Lake service in CDP environment, see *Introduction to Data Lakes*.
- 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.

Related Information

Getting started as a user
Working with AWS environments
Working with Azure environments
Working with GCP environments
Introduction to Data Lakes

Creating your cluster

Once you have met the prerequisites, you are ready to create a managed and secured Flow Management cluster in CDP Public Cloud by using one of the prescriptive flow management cluster definitions. You can select the cluster definition that matches your cloud provider and choose from the light and heavy duty options available.

About this task

You can create your Flow Management cluster using the CDP UI or CDP CLI. The UI offers a graphical interface that simplifies the cluster creation process. You can select the desired configurations for your cluster easily.

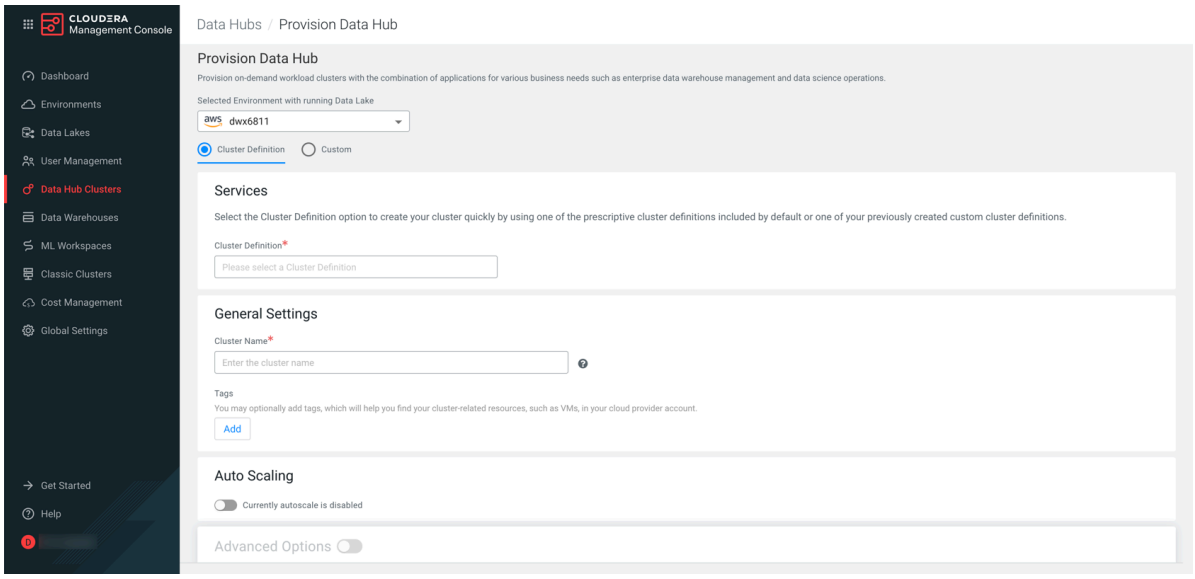
When you create a new cluster through the UI, Java 8 is used as the default JDK. The ability to deploy a new Flow Management cluster with Java 11 is available only using the CLI. The CDP command-line interface also allows you to leverage scripting and automation capabilities to create clusters programmatically or perform bulk operations.

For UI

Steps

1. Log into the CDP web interface.
2. Navigate to Management Console Environments , and select the environment where you would like to create a cluster.
3. Click Create Data Hub.

   The Provision Data Hub page is displayed:



4. Select Cluster Definition.
5. Select the appropriate Flow Management cluster definition from the Cluster Definition dropdown depending on your operational objectives.

   There are six template options available:

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

   For more information on templates, see *Flow Management cluster definitions* and *Flow Management cluster layout*.

   The list of services is automatically shown below the selected cluster definition name.
6. Give the cluster a name and add any tags you might need.

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

   Note: If you want to deploy your Flow Management Data Hub cluster across multiple availability zones (use multi-AZ support), go to the Network And Availability section and select all desired subnets from the available options. NiFi nodes will be provisioned across the selected subnets. Currently, Multi-AZ support is only available in Amazon Web Services (AWS) environments.
8. Click Provision Cluster.

Result

You will be redirected to the Data Hub cluster dashboard, and a new tile representing your cluster will appear at the top of the page.

For CLI

Before you begin

You have installed and configured CDP CLI. For more information about setting up the CLI client, see *CDP CLI*.

Steps

Use the create-<cloud_provider>-cluster command and include the --java-version parameter to create your Flow Management cluster.

Valid inputs for the --java-version parameter are 8 and 11.

Example

```
cdp datahub create-aws-cluster \
  --cluster-name nifi-cluster \
  --environment-name example-sandbox-aws \
  --cluster-template-name "7.2.17 - Flow Management Light Duty with Apache
 NiFi, Apache NiFi Registry, Schema Registry" \
  --instance-grou
ps nodeCount=3,instanceGroupName=nifi,instanceGroupType=CORE,instanceType=m5.2xlarge,
\[\{volumeSize=500,volumeCount=4,volumeType=standard
\}\],recoveryMode=MANUAL
 nodeCount=1,instanceGroupName=management,instanceGroupType=GATEWAY,instanceType=m5.2
\[\{volumeSize=100,volumeCount=1,volumeType=standard
\}\],recoveryMode=MANUAL
 nodeCount=0,instanceGroupName=nifi_scaling,instanceGroupType=CORE,instanceType=m5.2x
\[\{volumeSize=500,volumeCount=4,volumeType=standard
\}\],recoveryMode=MANUAL \
  --image id=6e038b51-0aea-43ce-873a-a0aa438b54f9,catalogName=cdp-default
 \
  --datahub-database NON_HA \
  --java-version 11 \
  --no-enable-load-balancer
```

After you finish

After you have created your Flow Managament Data Hub cluster, you are ready to build your first data flow in CDP Public Cloud.

Related Information

Flow Management cluster definitions

Flow Management cluster layout

Tags

Giving access to your cluster

As an EnvironmentAdmin, you need provide users access to your environment and to the new Flow Management cluster by assigning user roles and adding users to Ranger policies.

About this task

The cluster you have created using the Flow Management cluster definition is kerberized and secured with SSL. Users can access cluster UIs and endpoints through a secure gateway powered by Apache Knox. Before you can begin developing data flows, you must give users access to the Flow Management cluster components.

Procedure

1. Assign the EnvironmentUser role to the users to grant access to the CDP environment and the Flow Management cluster.
2. Sync the updated user permissions to the Data Lake (Ranger) using ActionsSynchronize Users to FreeIPA.

   This synchronizes the user to the FreeIPA identity management system to enable SSO.
3. Add the user to the appropriate pre-defined Ranger policies.

   To access the NiFi canvas and view or manage NiFi resources, the user must be assigned to at least:

   - Flow
   - Root Process Group

   If needed, you can also create a custom Ranger access policy and add the user to it.

   For more information, see *Flow Management security overview*.

Related Information

Security for Flow Management Clusters and Users in CDP Public Cloud