

CDP Private Cloud Data Services Installation on the OpenShift Container Platform

Date published: 2023-12-16

Date modified: 2024-03-23



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

Requirements.....	4
Software Support Matrix for OpenShift.....	4
Red Hat OpenShift Container Platform hardware requirements.....	5
Cloudera Data Warehouse hardware requirements.....	6
Cloudera Machine Learning requirements (OCP).....	8
CDE hardware requirements.....	10
How to use the CDP Private Cloud Data Services sizing spreadsheet.....	11
Red Hat OpenShift Container Platform software requirements.....	14
Credentials.....	14
Security context credentials.....	15
Load balancing and ingress.....	15
Certificate management and DNS.....	15
Storage classes.....	15
Volume snapshot support.....	15
CDP Private Cloud Base requirements.....	16
Preparing CDP Private Cloud Base.....	17
CDP Private Cloud Data Services Hardware Requirements.....	18
CDP Private Cloud Data Services deployment considerations.....	18
Storage requirements.....	19
CDP Private Cloud Data Services network infrastructure considerations.....	19
CDP Private Cloud Data Services Software Requirements.....	20
External vault requirements.....	20
Docker repository access.....	21
CML software requirements for Private Cloud.....	22
 Installation on the OpenShift Container Platform (OCP).....	 23
CDP Private Cloud Data Services pre-installation checklist.....	23
CDP Private Cloud Base checklist.....	23
OpenShift Container Platform (OCP) checklist.....	25
Cloudera Data Warehouse checklist.....	26
Cloudera Machine Learning checklist.....	26
Cloudera Data Engineering checklist.....	27
Installing in internet environment.....	27
Installing in air gap environment.....	34
Uninstall CDP Private Cloud Data Services.....	43
Dedicating OCP nodes for specific workloads.....	46

Requirements

Software Support Matrix for OpenShift

This support matrix lists the supported software for the CDP Private Cloud Base cluster and the CDP Private Cloud Data Services containerized cluster when installing using the OpenShift Container Platform (OCP).

Base Cluster	Version	<ul style="list-style-type: none"> Cloudera Manager 7.11.3 CHF 4 7.1.9 CHF 3 7.1.7 SP 2 7.1.8 CHF 19
	Base OS	<ul style="list-style-type: none"> See Private Cloud Base OS requirements
	TLS	<ul style="list-style-type: none"> AutoTLS (Custom CMCA) AutoTLS (Self-signed)
	Kerberos	<ul style="list-style-type: none"> AD FreeIPA
	JDK	<ul style="list-style-type: none"> openjdk version "11.0.21" 2023-10-17 LTS OpenJDK Runtime Environment (Red_Hat-11.0.21.0.9-1) (build 11.0.21+9-LTS) OpenJDK 64-Bit Server VM (Red_Hat-11.0.21.0.9-1) (build 11.0.21+9-LTS, mixed mode)
	Custom service principals	<ul style="list-style-type: none"> Not supported
	Data Lake Storage	<ul style="list-style-type: none"> HDFS Ozone Iceberg v2 (with HDFS and Ozone)
	Base DB (HMS access from CDW Data Services*)	<ul style="list-style-type: none"> MySQL 5.7, 8.0 Maria DB 10.2, 10.3, 10.4, 10.5, 10.6 Oracle 19.19 Postgres 12, 14 <p>* CDW uses a TLS enabled connection</p>
Containerized Cluster	Kubernetes	<ul style="list-style-type: none"> OCP 4.12 (K8s 1.25.8) [Fresh install, Upgrade from 1.5.2 to 1.5.3] OCP 4.11 [Upgrade from 1.5.1 to 1.5.3, no fresh install] <p> Important: OCP upgrades from earlier versions are incremental. See OCP upgrade steps for CDP Private Cloud Data Services 1.5.3 for more information.</p>
	Control Plane Metadata DB	<ul style="list-style-type: none"> Embedded
	Vault	<ul style="list-style-type: none"> External v1.9 (OCP only) Embedded

	Docker registry type	<ul style="list-style-type: none"> Secure registry with self signed CA certs (pwd protected + self signed certs), trusted CA certs
	Storage	<ul style="list-style-type: none"> OCS (rebranded ODF) Pure Portworx
	NFS	<ul style="list-style-type: none"> Embedded External
	IdP	<ul style="list-style-type: none"> FreeIPA ActiveDirectory (LDAP) OpenLDAPs
	Network Access	<ul style="list-style-type: none"> Airgap HTTP proxy (CML)
	TLS	<ul style="list-style-type: none"> Manual - CA signed

Red Hat OpenShift Container Platform hardware requirements

Cloudera Data Platform (CDP) Private Cloud requires hardware for a dedicated OpenShift Container Platform (OCP) cluster. An OpenShift cluster consists of several master nodes for managing OpenShift and many worker nodes for running your application on CDP.

The sizing of the OpenShift cluster depends on:

- The OpenShift cluster setup on the master nodes
- Application workloads deployed on the worker nodes

The CDP Private Cloud Data Services is installed on the OpenShift worker nodes.

The following table lists the hardware requirements for each node type. You require at least 3 minimum OpenShift Master Nodes + 1 Cluster System Admin Host (CSAH) Node + 1 Bootstrap Node. You need worker nodes based on your application workload requirements.

Role	CPU cores	Memory	Storage
Master	4	16 GB	120 GB
CSAH	4	64 GB	200 GB
Bootstrap	4	16 GB	120 GB
Worker	Depends on your workloads	Depends on your workloads	Depends on your workloads

Additionally, if you plan to run Cloudera Data Warehouse (CDW) or Cloudera Machine Learning (CML) data services workloads, you need to ensure that you meet the minimum requirements for each of those Data Services.

You can install CDP Private Cloud Data Services in a low resource mode for Cloudera Data Warehouse (CDW) workloads. For more information about OpenShift low resource mode requirements for CDW, see *Get started with OpenShift low resource mode requirements* using the link in the related information section.



Important: Lowering the minimum hardware requirement reduces the up-front investment to deploy CDW on OpenShift or ECS pods, but it does impact performance. Cloudera recommends that you use the Low Resource Mode option for proof of concept (POC) purposes only. This feature is not recommended for production deployment.

Complex queries and multiple queries on HS2 may fail due to limited memory configurations for HMS and HS2 in the low resource mode.

Cloudera Data Warehouse hardware requirements

Review the requirements needed to get started with the Cloudera Data Warehouse (CDW) service on Red Hat OpenShift.

You can also use the CDP Private Cloud Data Services Spreadsheet to model the number and specification of hosts required for a deployment. See [How to use the CDP Private Cloud Data Services sizing spreadsheet](#) on page 11.

- CDP Cloudera Manager must be installed and running.
- CDP Private Cloud must be installed and running. See [Installing on OpenShift](#) and [Installing on ECS](#) for more details.
- An environment must have been registered with Management Console on the private cloud. See [CDP Private Cloud Environments](#) for more details.
- In addition to the general requirements, CDW also has the following minimum memory, storage, and hardware requirements for each worker node using the standard resource mode:

Depending on the number of executors you want to run on each physical node, the per-node requirements change proportionally. For example, if you are running 3 executor pods per physical node, you require 384 GB of memory and approximately 1.8 TB of locally attached SSD/NVMe storage.



Important:

When you add memory and storage, it is very important that you add it in the increments stated:

- increments of 128 GB of memory
- increments of 600 GB of locally attached SSD/NVMe storage

If you add memory or storage that is not in the above increments, the memory and storage that exceeds these increments is not used for executor pods. Instead, the extra memory and storage can be used by other pods that require fewer resources.

For example, if you add 200 GB of memory, only 128 GB is used by the executor pods. If you add 2 TB of locally attached storage, only 1.8 TB is used by the executor pods.

Security requirements

The CDW service requires the "cluster-admin" role on the OpenShift and ECS cluster in order to install correctly. The "cluster-admin" role enables namespace creation and the use of the OpenShift Local Storage Operator for local storage.

Low resource mode requirements

Review the memory, storage, and hardware requirements for getting started with the Cloudera Data Warehouse (CDW) service in low resource mode on Red Hat OpenShift and (ECS). This mode reduces the minimum amount of hardware needed.

To get started with the CDW service on Red Hat OpenShift or ECS low resource mode, make sure you have fulfilled the following requirements:



Important: Lowering the minimum hardware requirement reduces the up-front investment to deploy CDW on OpenShift or ECS pods, but it does impact performance. Cloudera recommends that you use the Low Resource Mode option for proof of concept (POC) purposes only. This feature is not recommended for production deployment.

Complex queries and multiple queries on HS2 may fail due to limited memory configurations for HMS and HS2 in the low resource mode.

- CDP Cloudera Manager must be installed and running.
- CDP Private Cloud must be installed and running. See [Installing on OpenShift](#) and [Installing on ECS](#) for more details.
- An environment must have been registered with Management Console on the private cloud. See [CDP Private Cloud Environments](#) for more details.

- In addition to the general requirements, CDW also has the following minimum memory, storage, and hardware requirements for each worker node using the standard resource mode:

Component	Low resource mode deployment
Nodes	4
CPU	4
Memory	48 GB
Storage	3 x 100 GB (SATA) or 2 x 200 GB (SATA)
Network Bandwidth	1 GB/s guaranteed bandwidth to every CDP Private Cloud Base node



Important: When you add memory and storage for low resource mode, it is very important that you add it in the increments stated in the above table:

- increments of 48 GB of memory
- increments of at least 100 GB or 200 GB of SATA storage

If you add memory or storage that is not in the above increments, the memory and storage that exceeds these increments is not used for executor pods. Instead, the extra memory and storage can be used by other pods that require fewer resources.

Virtual Warehouse low resource mode resource requirements

The following requirements are in addition to the low resource mode requirements listed in the previous section.

Table 1: Impala Virtual Warehouse low resource mode requirements

Component	vCPU	Memory	Local Storage	Number of pods in XSMALL Virtual Warehouse
Coordinator (2)	2 x 0.4	2 x 24 GB	2 x 100 GB	2
Executor (2)	2 x 3	2 x 24 GB	2 x 100 GB	2
Statestore	0.1	512 MB	--	1
Catalogd	0.4	16 GB	--	1
Auto-scaler	0.1	1 GB	--	1
Hue (backend)	0.5	8 GB	--	1
Hue (frontend)	--	--	--	1
Total for XSMALL Virtual Warehouse	8 (7.9)	121.5 GB	400 GB - 3 volumes	--

Impala Admission Control Configuration

- Maximum concurrent queries per executor: 4
- Maximum query memory limit: 8 GB

Table 2: Hive Virtual Warehouse low resource mode requirements

Component	vCPU	Memory	Local Storage	Number of pods in XSMALL Virtual Warehouse
Coordinator (2)	2 x 1	2 x 4 GB	2 x 100 GB	2
Executor (2)	2 x 4	2 x 48 GB (16 GB heap; 32 GB off-heap)	2 x 100 GB	2
HiveServer2	1	16 GB	--	1

Component	vCPU	Memory	Local Storage	Number of pods in XSMALL Virtual Warehouse
Hue (backend)	0.5	8 GB	--	1
Hue (frontend)	--	--	--	1
Standalone compute operator	0.1	100 MB (.1 GB)	--	--
Standalone query executor (separate)	Same as executor	Same as executor	Same as executor	--
Total for XSMALL Virtual Warehouse	21 (20.6)	237 GB (236.1)	400 GB - 4 volumes	--

Database Catalog low resource mode requirements

The HiveMetaStore (HMS) requires 2 CPUs and 8 GB of memory. Because HMS pods are in High Availability mode, they need a total of 4 CPUs and 16 GB of memory.

Data Visualization low resource requirements

Table 3: Data Visualization low resource mode requirements

vCPU	Memory	Local Storage	Number of pods in XSMALL Virtual Warehouse
0.5	8 GB	--	1

Cloudera Machine Learning requirements (OCP)

To launch the Cloudera Machine Learning service, the OpenShift Container Platform (OCP) host must meet several requirements. Review the following CML-specific software, NFS server, and storage requirements.

Requirements

If necessary, contact your Administrator to make sure the following requirements are satisfied:

1. If you are using OpenShift, the installed OpenShift Container Platform must be version 4.10 or 4.8 (when upgrading from 1.4.x to 1.5.x, not performing a fresh installation). For ECS, refer to the *Hardware and Software Requirements* section in *CDP Private Cloud Experiences Installation Hardware Requirements and Managing a Private Cloud Experience Cluster 1.5.x*.
2. CML assumes it has cluster-admin privileges on the cluster.
3. Storage:
 - a. Persistent volume block storage per ML Workspace: 600 GB minimum, 4.5 TB recommended.
 - b. 1 TB of external NFS space recommended per Workspace (depending on user files). If using embedded NFS, 1 TB per workspace in addition to the 600 GB minimum, or 4.5 TB recommended block storage space.
 - c. Access to NFS storage is routable from all pods running in the cluster.
 - d. For monitoring, recommended volume size is 60 GB.
4. On OCP, CephFS is used as the underlying storage provisioner for any new internal workspace on PVC 1.5.x. A storage class named ocs-storagecluster-cephfs with csi driver set to "openshift-storage.cephfs.csi.ceph.com" must exist in the cluster for new internal workspaces to get provisioned.
5. A block storage class must be marked as default in the cluster. This may be rook-ceph-block, Portworx, or another storage system. Confirm the storage class by listing the storage classes (run `oc get sc`) in the cluster, and check that one of them is marked default.
6. If external NFS is used, the NFS directory and assumed permissions must be those of the cdsw user. For details see Using an External NFS Server in the Related information section at the bottom of this page.

7. If CML needs access to a database on the CDP Private Cloud Base cluster, then the user must be authenticated using Kerberos and must have Ranger policies set up to allow read/write operations to the default (or other specified) database.
8. Ensure that Kerberos is enabled for all services in the cluster. Custom Kerberos principals are not currently supported. For more information, see [Enabling Kerberos for authentication](#).
9. Forward and reverse DNS must be working.
10. DNS lookups to sub-domains and the ML Workspace itself should work.
11. In DNS, wildcard subdomains (such as *.cml.yourcompany.com) must be set to resolve to the master domain (such as cml.yourcompany.com). The TLS certificate (if TLS is used) must also include the wildcard subdomains. When a session or job is started, an engine is created for it, and the engine is assigned to a random, unique subdomain.
12. The external load balancer server timeout needs to be set to 5 min. Without this, creating a project in an ML workspace with `git clone` or with the API may result in API timeout errors. For workarounds, see Known Issue DSE-11837.
13. If you intend to access a workspace over https, see [Deploy an ML Workspace with Support for TLS](#).
14. For non-TLS ML workspaces, websockets need to be allowed for port 80 on the external load balancer.
15. Only a TLS-enabled custom Docker Registry is supported. Ensure that you use a TLS certificate to secure the custom Docker Registry. The TLS certificate can be self-signed, or signed by a private or public trusted Certificate Authority (CA).
16. On OpenShift, due to a [Red Hat issue](#) with OpenShift Container Platform 4.3.x, the image registry cluster operator configuration must be set to Managed.
17. Check if storage is set up in the cluster image registry operator. See Known Issues DSE-12778 for further information.

For more information on requirements, see [CDP Private Cloud Base Installation Guide](#).

Hardware requirements

Storage

The cluster must have persistent storage classes defined for both block and filesystem volumeModes of storage. Ensure that a block storage class is set up. The exact amount of storage classified as block or filesystem storage depends on the specific workload used:

- Machine Learning workload requirements for storage largely depend on the nature of your machine learning jobs. 4 TB of persistent volume block storage is required per Machine Learning Workspace instance for storing different kinds of metadata related to workspace configuration. Additionally, Machine Learning requires access to NFS storage routable from all pods running in the cluster (see below).
- Monitoring uses a large Prometheus instance to scrape workloads. Disk usage depends on scale of workloads. Recommended volume size is 60 GB.

	Local Storage (for example, ext4)	Block PV (for example, Ceph or Portworx)	NFS (for ML user project files)
Control Plane	N/A	250 GB	N/A
CML	N/A	1.5 TB per workspace	1 TB per workspace (dependent on size of ML user files)

NFS

Cloudera Machine Learning (CML) requires NFS 4.0 for storing project files and folders. NFS storage is to be used only for storing project files and folders, and not for any other CML data, such as PostgreSQL database and LiveLog.

ECS requirements for NFS Storage

Cloudera managed ECS deploys and manages an internal NFS server based on LongHorn which can be used for CML. This is the recommended option for CML on ECS clusters. CML requires `nfs-utils` in order to mount longhorn-nfs provisioned mounts.

CML requires the `nfs-utils` package be installed in order to mount volumes provisioned by `longhorn-nfs`. The `nfs-utils` package is not available by default on every operating system. Check if `nfs-utils` is available, and ensure that it is present on all ECS cluster nodes.

Alternatively, the NFS server can be external to the cluster, such as a NetApp filer that is accessible from the private cloud cluster nodes.

OpenShift requirements for NFS storage

An internal user-space NFS server can be deployed into the cluster which serves a block storage device (persistent volume) managed by the cluster's software defined storage (SDS) system, such as Ceph or Portworx. This is the recommended option for CML on OpenShift. Alternatively, the NFS server can be external to the cluster, such as a NetApp filer that is accessible from the private cloud cluster nodes. NFS storage is to be used only for storing project files and folders, and not for any other CML data, such as PostgreSQL database and LiveLog.

CML does not support shared volumes, such as Portworx shared volumes, for storing project files. A read-write-once (RWO) persistent volume must be allocated to the internal NFS server (for example, NFS server provisioner) as the persistence layer. The NFS server uses the volume to dynamically provision read-write-many (RWX) NFS volumes for the CML clients.

CDE hardware requirements

Review the requirements needed to get started with the Cloudera Data Engineering (CDE) service on Red Hat OpenShift.

Requirements

- CDE assumes it has cluster-admin privileges on the OpenShift cluster.
- Openshift cluster should be configured with [route admission policy](#) set to namespaceOwnership: InterNamespaceAllowed. This allows Openshift cluster to run applications in multiple namespaces with the same domain name.

```
oc -n openshift-ingress-operator patch ingresscontroller/default --patch
'{"spec":{
  "routeAdmission":{
    "namespaceOwnership":"InterNamespaceAllowed"}}}' --type=merge
```


- **Table 4: The following are the CDE Service requirements:**

Component	vCPU	Memory	Block PV or NFS PV	Number of replicas
Embedded DB	4	8 GB	100 GB	1
Config Manager	500 m	1 GB	--	2
Dex Downloads	250 m	512 MB	--	1
Knox	250 m	1 GB	--	1
Management API	1	2 GB	--	1
NGINX Ingress Controller	100 m	90 MB	--	1
FluentD Forwarder	250 m	512 MB	--	1
Grafana	250 m	512 MB	10 GB	1
Data Connector	250 m	512 MB	--	1
Total	7	15 GB	110 GB	

- CDE Service requirements: Overall for a CDE service, it requires 110 GB Block PV or NFS PV, 7 CPU cores, and 15 GB memory.

Table 5: The following are the CDE Virtual Cluster requirements for Spark 3:

Component	vCPU	Memory	Block PV or NFS PV	Number of replicas
Airflow API	350 m	612 MB	100 GB	1
Airflow Scheduler	1	1 GB	100 GB	1
Airflow Web	250 m	512 MB	--	1
Runtime API	250 m	512 MB	100 GB	1
Livy	3	12 GB	100 GB	1
SHS	250 m	1 GB		1
Pipelines	250 m	512 MB	--	1
Total	5350 m	15.6 GB	400 GB	

- CDE Virtual Cluster requirements:
 - For Spark 3: Overall storage of 400 GB Block PV or Shared Storage PV, 5.35 CPU cores, and 15.6 GB per virtual cluster.
 - For Spark 2: If you are using Spark 2, you need additional 500 m CPU, 4.5 GB memory and 100 GB storage, that is, the overall storage of 500 GB Block PV or Shared Storage PV, 5.85 CPU cores, and 20.1 GB per virtual cluster.
-  **Important:** The above requirements does not include workloads. See the below workload information on the additional resources based on workload.
- Workloads: Depending upon the workload, you must configure resources.
 - The Spark Driver container uses resources based on the configured driver cores and driver memory and additional 40% memory overhead.
 - In addition to this, Spark Driver uses 110 m CPU and 232 MB for the sidecar container.
 - The Spark Executor container uses resources based on the configured executor cores and executor memory and additional 40 % memory overhead.
 - In addition to this, Spark Executor uses 10 m CPU and 32 MB for the sidecar container.
 - Minimal Airflow jobs need 100 m CPU and 200 MB memory per Airflow worker.

How to use the CDP Private Cloud Data Services sizing spreadsheet

You can use the sizing spreadsheet to model the hardware requirements for a CDP Private Cloud Data Services deployment.

Overview

The CDP Private Cloud Data Services Sizing spreadsheet is a spreadsheet that you can use to model the quantity and specifications for worker hosts required in a CDP Private Cloud Data Services deployment.

This spreadsheet is intended to use information about workloads you are planning to run and hardware specifications for worker nodes to arrive at an approximate number of worker nodes required for your deployment. Due to the complexity of estimating workloads, Cloudera recommends you review any sizing or purchasing decisions with Cloudera Professional Services before committing to those decisions.

How to access the spreadsheet

You can access the spreadsheet here: [CDP Private Cloud Data Services Sizing](#). The file is in Microsoft Excel format. You can open the file in Excel, or upload it to Google Sheets.

There are three tabs in the spreadsheet. You will make your inputs only on the Worker Node Totals tab. Do not modify the following tabs (these tabs contain data used to calculate values in the spreadsheet and should not be modified):

- Component Lookup
- K8s Resources



Important: Do not modify any cells except for the ones indicated below. Modifying the formulas in other cells will result in inaccurate calculations.

Workload inputs

The spreadsheet calculates the total amount vcores, RAM, and storage required based on information you enter about the combined workloads you intend to deploy. Then based on the hardware specifications entered, calculates the number of worker nodes required, which is displayed in cell E24.

The following sections describe values you must enter into the spreadsheet. Values are required for each Data Service you intend to deploy, and values to enter for the hardware specifications for your worker nodes.

Control plane monitoring

Label	Cell	Description
CP Monitoring	B3	Increment this number by one for each environment.

Cloudera Data Warehouse (CDW)

If you will deploy CDW, on the Worker Node Totals tab, enter the following information:

Label	Cell	Description
CDW Data Catalog (min 1 per env)	B5	Enter the number of Data Catalogs you will need in your deployment. You must have at least one Data Catalog.
CDW LLAP warehouses	B6	Enter the number of LLAP warehouses you will need for each Virtual Warehouse in your deployment.
-- LLAP Executors	B7	Enter the total number of LLAP Executors you will need in your deployment.
CDW Impala warehouses	B8	Enter the number of CDW Impala warehouses for each Virtual Warehouse you will need in your deployment.
-- Impala Coordinators (2 x for HA)	B9	Enter the number of Impala Warehouses you will need in your deployment. If you have enabled high availability, enter twice the number of Warehouses.
-- Impala Executors	B10	Enter the number of Impala Executors you will need in your deployment.
CDW Cache	B11	Enter the amount of CDW Cache space for each coordinator and executor (Default 600)
Data Viz - small instances	B12	Enter the size selected when creating a Data Visualization instance.
Data Viz - medium instances	B13	
Data Viz - large instances	B14	

For more information about sizing Cloudera Data Warehouse deployments, see:

- [Standard resource mode requirements](#)

- [Low resource mode requirements](#)

Cloudera Machine Learning (CML)

Sizing for a CML deployment depends on the number of concurrent jobs you expect to run and the number of Workspaces you provision.

Label	Cell	Description
CML Workspace (min of 1)	B16	Enter the number of workspaces you need in your deployment.
-- CML Small concurrent sessions	B17	Enter the number of concurrent small-sized sessions you intend to run.
-- CML Average concurrent sessions	B18	Enter the number of concurrent average-sized sessions you intend to run.

For more information about sizing the Cloudera Data Engineering service, see the following topics:

- [Additional resource requirements for Cloudera Machine Learning](#).
- (OCP) [Cloudera Machine Learning requirements](#)
- (ECS) [Cloudera Machine Learning requirements](#)

Cloudera Data Engineering (CDE)

Label	Cell	Description
CDE Service (min/max 1 per cluster)	B20	Enter the number of CDE clusters you will need in your deployment.
CDE Virtual Cluster	B21	Enter the number of CDE Virtual Clusters you will need in your deployment.
-- CDE Small concurrent jobs	B22	Enter the number of concurrent small-sized jobs you intend to run.
-- CDE Average concurrent jobs	B23	Enter the number of concurrent average-sized jobs you intend to run.

For more information about sizing the Cloudera Data Engineering service, see [Additional resource requirements for Cloudera Data Engineering](#).

Worker node hardware specifications

Based on the inputs you supplied for your workloads, the spreadsheet totals the number of vcores, RAM, and storage required for the cluster in cells C20-C26. Then, based on the worker node hardware specifications you enter in cells B26-B29, divides the totals for vcores, RAM and storage by each of the worker node specifications to arrive at the required number of nodes for vcores, RAM and storage shown in cells D5-D29. The final number, in cell E27 chooses the higher value of these cells.

You may notice that the calculated values in cells D26 and D27 are different. This indicates that some nodes are oversubscribed for RAM or vcores. Adjust the hardware specifications for CPU and RAM until the two cells are closer together in value. Changing these values may also change the calculated number of worker nodes.

Label	Cell	Description
CPU recommend 40+ cores (80 vcores)	B27	Enter the number of vcores for each worker node.
RAM (GB) recommend 415 GB RAM	B28	Enter the amount of RAM, in gigabytes, for each worker node.

Label	Cell	Description
Disk (GB) Block (OCP CSI block, ECS Longhorn)	B29	Enter the number of gigabytes Block required for: - OpenShift Container Platform: CSI block - Embedded Container Service: ECS Longhorn
Disk (GB) Fast Cache for CDW (nvme,ssd)	B30	Enter the number of gigabytes of Fast Cache used in Cloudera Data Warehouse.
CP Block Overhead per host (300 to 1024)	B31	Enter the Control Plane block overhead
NFS (GB) (choose 1 from below)	B33	Enter required storage in either cell B34 or cell B35
-- Embedded nfs - (subtract from Block provider) non-prod	B34	Enter the number of gigabytes storage for an embedded NFS.
-- External nfs	B35	Enter the number of gigabytes of storage for an External NFS.
ECS Master Node requires 1 for non HA - 3 for HA If you are using the Embedded Container Service, you will also need to provision a host for the ECS Master Node (a node running the ECS Server component). The values described here contain Cloudera's recommendations for specifications for the ECS Master node.	B38	Minimum: 16 vcores Recommended: 32 vcores
	B39	Minimum: 32 GB RAM Recommended: 64 GB RAM
	B40	Minimum: 300 GB HDD (This amount is adequate for a proof-of-concept cluster.) Recommended: 1 TB HDD

Red Hat OpenShift Container Platform software requirements

You must understand the various OpenShift Container Platform (OCP) requirements before you install CDP Private Cloud Data Services. CDP Private Cloud Data Services requires at least one OpenShift cluster for the control plane and the environments. The Cloudera Data Warehouse (CDW), Cloudera Machine Learning (CML), and Cloudera Data Engineering (CDE) Data Services run on these environments.

Review the [Software Support Matrix for OpenShift](#) on page 4.

Read the following topics to understand the various OpenShift integration requirements:

- Credentials
- Security context credentials
- Load balancing and ingress
- Certificate management and DNS
- Storage classes
- Docker registry access

Credentials

You must have a kubeconfig file that has the cluster access information and authentication information for a single user, who has the “cluster-admin” pre-provisioned ClusterRole assigned.

Cloudera recommends that you use a kubeconfig file that does not expire, to avoid access issues to the installed software.

Security context credentials

The Cloudera software must have privileged access at runtime. Cloudera recommends that you configure security context in your OpenShift cluster to ensure access to CDP Private Cloud Data Services.

You must install additional scc definitions into OpenShift that Cloudera provides as part of the installation software. For more information about security context credentials in OpenShift, see [Introduction to Security Contexts and SCCs](#).

Load balancing and ingress

OpenShift Route must be the default ingress controller setup on the cluster.

A non-terminating external load balancer must be configured to route ingress traffic on HTTP/HTTPS to the OpenShift cluster.

When a load balancer is used in front of the OCP external API, it must allow “Websocket traffic”, in addition to https.

Certificate management and DNS

You must be aware of the reasons why an external DNS is required for CDP Private Cloud Data Services installation along with the required setup in the cluster.

An external DNS must be available to route inbound traffic to the cluster through the load balancer. The external DNS should contain forward and reverse zones for both the OpenShift and the CDP Private Base cluster nodes.

Ensure that the canonical load balancers required for OpenShift is routable from within the OpenShift cluster and from any other location that you want to access resources in the Management Console; this is a standard requirement for on-premises load balancers communicating Kubernetes clusters.

There must also be a set of certificates set up for use by the OpenShift Route ingress controller as defined in the *OpenShift bare metal install guide* that the Cloudera services use.

Storage classes

You need to have persistent storage classes defined in your OpenShift cluster. Storage classes can be defined by OpenShift cluster administrators.

The exact amount of storage classified as block or filesystem storage depends on the specific workloads (Machine Learning or Data Warehouse) and how they are used.

See the *Red Hat OpenShift documentation* for more information about OpenShift storage classes and persistent volumes.

To use Portworx as a storage platform, you must first create a Portworx storage class on your OCP cluster and then specify it in the Storage Class field while installing CDP Private Cloud Data Services on the OCP cluster. For information on how to create the storage class, see [Step 4: StorageClass Setup](#) in the Portworx documentation. See [Installing in internet environment](#) for information on how to install CDP Private Cloud Data Services on OCP.

After you specify the storage class while installing CDP Private Cloud Data Services, all other data services can use it.

Volume snapshot support

Volume snapshot support for the storage class must be installed in your OpenShift cluster.

Run the following commands to determine whether or not volume support for the storage class is installed in your OpenShift cluster:

```
kubectl get sc
NAME                                PROVISIONER
      RECLAIMPOLICY  VOLUMEBINDINGMODE  ALLOWVOLUMEEXPANSION  AGE
cdw-scratch          Delete             WaitForFirstConsumer  false                 82d
localblock           Delete             WaitForFirstConsumer  false                 82d
```

nfs			nfs-server-provisioner	
Delete	Immediate		true	82d
ocs-storagecluster-ceph-rbd (default)			openshift-storage.rbd.csi.ceph.com	
Delete	Immediate		true	82d
ocs-storagecluster-ceph-rgw			openshift-storage.ceph.rook.io/buck	
Delete	Immediate		false	82d
ocs-storagecluster-cephfs			openshift-storage.cephfs.csi.ceph	
.com Delete	Immediate		true	82d
openshift-storage.noobaa.io			openshift-storage.noobaa.io/obc	
Delete	Immediate		false	82d

```
kubectl get volumesnapshotclasses
```

NAME	DELETIONPOLICY	AGE	DRIVER
ocs-storagecluster-cephfsplugin-snapclass	Delete	82d	openshift-storage.cephfs.csi.
ceph.com	Delete	82d	
ocs-storagecluster-rbdplugin-snapclass	Delete	82d	openshift-storage.rbd.csi.ceph.c
om	Delete	82d	

A storage class has a volume snapshot installed if there is an entry with the value in the DRIVER column returned by the second command that matches one of the values in the PROVISIONER column returned by the first command. In the example above, the following storage classes have volume snapshot support:

- ocs-storagecluster-ceph-rbd (default)
- ocs-storagecluster-cephfs

Related Information

[CSI volume snapshots](#)

CDP Private Cloud Base requirements

Your CDP Private Cloud Base cluster must have the operating system, JDK, database, CDP components, and CDP Runtime version required to install CDP Private Cloud Data Services.

Operating system, JDK, and database:

- See [CDP Private Cloud Base Requirements and Supported Versions](#)

The PostgreSQL database instance must be configured to accept inbound TLS requests to the Hive Metastore database. A TLS connection is required when initiated from CDW in OpenShift.

CDP Runtime components (services):

- Hive Metastore (HMS)
- Ranger
- Atlas
- HDFS
- Ozone
- YARN
- Kafka
- Solr

Additionally, do the following:

- Set up Kerberos on these clusters using an Active Directory.
- Enable TLS on the Cloudera Manager cluster for communication with components and services.
- Ensure that the CDP Private Cloud Base cluster is on the same network as the OpenShift cluster.
- Configure PostgreSQL database as an external database for the CDP Private Cloud Base cluster components.
- Configure the CDP Private Cloud Base cluster hostnames to be forward and reverse resolvable in DNS from the OpenShift cluster.

- Allow websocket traffic and https traffic when you use a load balancer with the OpenShift external API.
- Ensure hive user is able to create and list an Ozone bucket. For information about creating and listing ozone bucket, see *Managing buckets*.

You can use the CDP Management Console to create one or more environments. These environments can be associated with any of the Data Lake from the CDP Private Cloud Base clusters. The CDP Private Cloud Base Cloudera Manager deploys the CDP Management Console.

Cloudera currently does not support associating an environment with many CDP Private Cloud Base cluster installations.

Related Information

[Managing buckets](#)

Preparing CDP Private Cloud Base

Use Cloudera Manager to configure your CDP Private Cloud Base in preparation for the CDP Private Cloud Data Services installation.

Procedure

1. Configure the CDP Private Cloud Base cluster to use TLS.
For configuration steps, see [Configuring TLS Encryption for Cloudera Manager Using Auto-TLS](#).
2. Configure Cloudera Manager with a JKS-format (not PKCS12) TLS truststore.
For configuration steps, see [Database requirements](#).
3. Configure Cloudera Manager to include a root certificate that trusts the certificate for all Cloudera Manager server hosts expected to be used with Private Cloud.
 - a. Import the necessary certificates into the truststore configured in `Configure Administration Settings Security Cloudera Manager TLS/SSL Client Trust Store File`.



Note: This requires a Cloudera Manager restart.

4. Configure Ranger and LDAP for user authentication. Ensure that you have configured Ranger user synchronization.

For configuration steps, see [Configure Ranger authentication for LDAP](#) and [Ranger usersync](#).



Note: Upgrading to Oracle JDK 1.8.351 causes a Kerberos issue when deprecated 3DES and RC4 permitted encryption types are used.

Workaround: Remove the deprecated 3DES and RC4 encryption types in the `krb5.conf` and `kdc.conf` files.

5. Enable Kerberos for all the services in the cluster.
For configuration steps, see [Enabling Kerberos for authentication](#).
6. Configure LDAP using Cloudera Manager. Only Microsoft Active Directory (AD) and OpenLDAP are currently supported.
For configuration steps, see [Configure authentication using an LDAP-compliant identity service](#).
7. Check if all the running services in the cluster are healthy. To check this using Cloudera Manager, go to `Cloudera Manager Clusters [***CLUSTER NAME***] Health Issues`. If there are no health issues, the No Health Issues message is displayed.
8. Verify if you have the necessary CDP entitlements from Cloudera to access the Private Cloud installation. To check this using Cloudera Manager, go to `Cloudera Manager Private Cloud Select Repository [***REPOSITORY URL***]`. If you have the required entitlements, the You are about to install CDP Private Cloud version `[*VERSION*]` message with a list of prerequisites is displayed. An error message is displayed if you do not have the necessary entitlements.

Contact your Cloudera account team to get the necessary entitlements.

9. If you want to reuse data from your legacy CDH or HDP deployment in your Private Cloud, ensure that you have migrated that data into your CDP Private Cloud Base. You must be using Cloudera Runtime 7.1.7 for migrating your data from your CDH or HDP cluster.

For more information about data migration, see the [Data Migration Guide](#).

10. For installing CDP Private Cloud Base, see [Install CDP Private Cloud Base](#)

CDP Private Cloud Data Services Hardware Requirements

You must learn about the minimum and recommended hardware and network infrastructure requirements before deploying CDP Private Cloud Data Services.

Architects and infrastructure administrators must understand these requirements to install CDP Private Cloud Data Services in your data center.

You must know the minimum hardware requirements prior to:

- Installing a dedicated Red Hat OpenShift Container Platform cluster required for CDP Private Cloud
- Installing and configuring CDP Private Cloud Data Services
- Deploying and running the Cloudera Data Warehouse (CDW) and Cloudera Machine Learning (CML) Data services

Related Information

[Cloudera Data Warehouse hardware requirements](#)

CDP Private Cloud Data Services deployment considerations

You must understand the deployment requirements to sufficiently provision node counts, CPU, memory, and other hardware resources required to install CDP Private Cloud.

The CDP Private Cloud Data Services are installed on the OpenShift Cluster and run on the provisioned worker nodes. CDP Private Cloud Data Services deployment consists of a Private Cloud Management Console and one or more environments that are created for deploying the Data Services. The Management Console is a service used by CDP administrators to manage environments, users, and services.

The worker node hardware requirements are described below. The number of worker nodes needed depends on factors such as the number of virtual warehouses or machine learning workspaces required for your workloads. The recommendation here is a guideline for a basic CDP Private Cloud Data Services installation. For hardware sizing in production environments, contact Cloudera Support or your Cloudera Account Team.

Component	Minimum	Recommended
Node Count	10	20
CPU	16	32 +
Memory	128 GB	384 GB
Storage	2 TB (SATA)	4 TB (SSD/NVMe)
Network Bandwidth	1 Gbps guaranteed bandwidth (minimum) dedicated to every CDP Private Cloud Base node	10 Gbps guaranteed bandwidth (minimum) dedicated to every CDP Private Cloud Base node



Important:

- You must be a Cluster System Admin Host for OpenShift system administration.
- You need the bootstrap node for the initial installation. It can be converted into an OpenShift worker after initial deployment.

To know about architecture, design choices, and deployment guidelines to use CDP Private Cloud Data Services with Dell EMC and Intel Infrastructure, and Cisco Intelligent Data Platform, see [Dell EMC and Intel Infrastructure Guide for Cloudera Data Platform Private Cloud](#) and [Cisco Data Intelligence Platform on Cisco UCS C240 M5 with Cloudera Data Platform Private Cloud Plus Design Guide](#).

Storage requirements

Storage requirements for Data Services.

Storage Requirements

Data Services	Storage type	Storage required	Purpose
CDE	Block	500GB per Virtual Cluster in Embedded NFS	Stores all information related to virtual clusters
CDW	Local	100 GB per executor in LITE mode and 600 GB per executor in FULL mode	Used for caching
Control Plane	Block	118 GB total if using an External Database, 318 GB total if using the Embedded Database	Storage for CDP infrastructure including Fluentd logging, Prometheus monitoring, and Vault. Backing storage for an embedded DB for control plane configuration purpose, if applicable
CML	Block	600 GB per node (minimum), 4.5 TB (recommended)	Stores all CML workspace information
	External NFS or Block	1 TB per Node	Stores all user project files. VFS storage can either use Longhorn NFS-provisioner on Longhorn OR directly connect to your NFS.
MonitoringApp	Block	30 GB + (Env cnt x 100 GB)	Stores metrics collected by Prometheus.
Data Catalog	Requires Control Plane database and not a dedicated storage space	100 GB extra in Control plane database	Stores profiling metadata.

CDP Private Cloud Data Services network infrastructure considerations

Learn about the networking infrastructure consideration necessary to install CDP Private Cloud. The networking considerations for CDP Private Cloud Data Services are similar to the networking requirements for Cloudera Manager Virtual Private Clusters (CM VPC).

In CDP Private Cloud Data Services, the network bandwidth requirements are less stringent than those of the Cloudera Manager Virtual Private Cluster (VPC) because of data caching technology introduced at the compute layer, which is not available in VPCs.

While the initial load of data from the remote storage would require significant bandwidth between the compute and storage clusters, subject to the quantity of data ingested; subsequently, the network bandwidth requirements are lower.

The following list of network considerations will help you plan your network infrastructure before you install CDP Private Cloud Data Services:

- Use 1 Gbps guaranteed bandwidth between each OpenShift worker node and each CDP Private Cloud Base DataNode. Cloudera recommends 10 Gbps guaranteed bandwidth.
- Stress test the network infrastructure with all the OpenShift nodes trying to read or write from the CDP Private Cloud Data Services nodes at the same time.
- Use the Spine-Leaf network architecture with no more than a 4:1 oversubscription between the spine and leaf switches.
- Check the applicable [ports used by Cloudera Runtime components](#).

For more information about minimum network performance requirements, network sizing, and designing a network topology, see [Networking Considerations for Virtual Private Clusters](#).

CDP Private Cloud Data Services Software Requirements

You must learn about the software and configuration requirements before deploying CDP Private Cloud. Administrators and operators must understand these requirements to install CDP Private Cloud Data Services in your data center.

You must understand the following software requirements before you install CDP Private Cloud:

- OpenShift integration requirements
- CDP Private Cloud Base requirements
- External database requirements
- External vault requirements

External vault requirements

You can learn about how to configure an external HashiCorp Vault for CDP Private Cloud Data Services. Hashicorp Vault securely stores your passwords, tokens, certificates, and encryption keys.



Note: [Vault namespaces](#) are not supported.

Vault Token Policy

CDP Private Cloud Data Services can be installed using an internal or external Vault. If you are installing CDP Private Cloud Data Services with an external Vault, a Vault token with the following permissions is required.

- Create/Update/List/Read a secret engine of type kv-2 at the applicable path.
- Create/Update/List/Read auth of type kubernetes at the applicable path.
- Create/Update/List/Read policies.
- Access to List and Read the Vault token details.

Example Vault policy:

```
# Manage auth methods broadly across Vault
path "auth/*"
{
  capabilities = ["create", "read", "update", "list"]
}
# Create, update auth methods
path "sys/auth/*"
{
  capabilities = ["create", "update", "sudo"]
}

# List auth methods
path "sys/auth"
{
  capabilities = ["read"]
}

# List existing policies
path "sys/policies/acl"
{
  capabilities = ["list"]
}

# Create and manage ACL policies via API & UI
path "sys/policies/acl/*"
{
  capabilities = ["create", "read", "update", "list"]
}
```

```
# Manage secrets engines
path "sys/mounts/*"
{
  capabilities = ["create", "read", "update", "list"]
}

# List existing secrets engines.
path "sys/mounts"
{
  capabilities = ["read"]
}
```

For more information, see [HashiCorp Vault Policy Requirements](#).

Vault Token Use

The Vault token should be created using the preceding policy. It is recommended that the Vault administrator delete this token after the installation is complete.

External Vault Installation Parameters

- Vault Address – The external Vault FQDN (Fully Qualified Domain Name) with the port number.
- Token – The Vault token described above
- CA Certificate – A valid certificate for the Vault server in PEM format.

Vault Secrets Engine, Auth, and Policies

During installation, CDP enables a kv-v2 secrets engine and kubernetes authentication at unique paths in the following format:

```
cloudera-[***CONTROL PLANE NAMESPACE***]-[***SERVER-URL***]
```

It is recommended that you do not have any kv-v2 secrets and kubernetes auth enabled at the same path in your Vault server.

CDP also creates Vault policies that provide access to control plane services to write their protected data. These two policies have the following format:

```
[***NAMESPACE***]-[***SERVER URL***]
```

```
admin-[***NAMESPACE***]-[***SERVER URL***]
```

Docker repository access

You must ensure that the cluster has access to the Docker Container Repository in order to retrieve the container images for deployment.

There are several types of Docker Repositories you can use:

Cloudera Repository

Using the Cloudera Repository requires that the cluster have internet connectivity to the Cloudera public repository. Using the Cloudera Repository is the fastest option.

The Cloudera-hosted Docker Repository option may increase the time required to deploy or start the services in the cluster. Cloudera generates Docker Repository credentials that are identical to your payroll credentials. Refer to your welcome letter for the credentials or use the credential generator on cloudera.com to generate credentials from your license key.

This option is best suited for proof-of-concept, non-production deployments or deployments that do not have security requirements that disallow internet access.

Custom Repository

A Custom Repository is a repository that you manage in your environment and can be Enterprise grade and highly available.

During installation and upgrade, a custom script is generated that you use to copy the images. Copying images can take 4 - 5 hours.

Only TLS-enabled custom Docker Registry is supported. Ensure that you use a TLS certificate to secure the custom Docker Registry. The TLS certificate can be self-signed, or signed by a private or public trusted Certificate Authority (CA).



Important: When using an Embedded Container Service cluster, passwords must not contain the \$ character.

Related Information

[Installation on the OpenShift Container Platform \(OCP\)](#)

[Installation using the Embedded Container Service \(ECS\)](#)

CML software requirements for Private Cloud

To launch the Cloudera Machine Learning service, the Private Cloud host must meet several software requirements. Review the following CML-specific software requirements.

Requirements

If necessary, contact your Administrator to make sure the following requirements are satisfied:

1. If you are using OpenShift, the installed OpenShift Container Platform must be version 4.10 or 4.8 (when upgrading from 1.4.x to 1.5.x, not performing a fresh installation). For ECS, refer to the *Hardware and Software Requirements* section in *CDP Private Cloud Experiences Installation Hardware Requirements and Managing a Private Cloud Experience Cluster 1.5.x*.
2. CML assumes it has cluster-admin privileges on the cluster.
3. Storage:
 - a. Persistent volume block storage per ML Workspace: 600 GB minimum, 4.5 TB recommended.
 - b. 1 TB of external NFS space recommended per Workspace (depending on user files). If using embedded NFS, 1 TB per workspace in addition to the 600 GB minimum, or 4.5 TB recommended block storage space.
 - c. Access to NFS storage is routable from all pods running in the cluster.
 - d. For monitoring, recommended volume size is 60 GB.
4. On OCP, CephFS is used as the underlying storage provisioner for any new internal workspace on PVC 1.5.x. A storage class named ocs-storagecluster-cephfs with csi driver set to "openshift-storage.cephfs.csi.ceph.com" must exist in the cluster for new internal workspaces to get provisioned.
5. A block storage class must be marked as default in the cluster. This may be rook-ceph-block, Portworx, or another storage system. Confirm the storage class by listing the storage classes (run `oc get sc`) in the cluster, and check that one of them is marked default.
6. If external NFS is used, the NFS directory and assumed permissions must be those of the cdsu user. For details see Using an External NFS Server in the Related information section at the bottom of this page.
7. If CML needs access to a database on the CDP Private Cloud Base cluster, then the user must be authenticated using Kerberos and must have Ranger policies set up to allow read/write operations to the default (or other specified) database.
8. Ensure that Kerberos is enabled for all services in the cluster. Custom Kerberos principals are not currently supported. For more information, see [Enabling Kerberos for authentication](#).
9. Forward and reverse DNS must be working.
10. DNS lookups to sub-domains and the ML Workspace itself should work.

11. In DNS, wildcard subdomains (such as *.cml.yourcompany.com) must be set to resolve to the master domain (such as cml.yourcompany.com). The TLS certificate (if TLS is used) must also include the wildcard subdomains. When a session or job is started, an engine is created for it, and the engine is assigned to a random, unique subdomain.
12. The external load balancer server timeout needs to be set to 5 min. Without this, creating a project in an ML workspace with `git clone` or with the API may result in API timeout errors. For workarounds, see Known Issue DSE-11837.
13. If you intend to access a workspace over https, see Deploy an ML Workspace with Support for TLS.
14. For non-TLS ML workspaces, websockets need to be allowed for port 80 on the external load balancer.
15. Only a TLS-enabled custom Docker Registry is supported. Ensure that you use a TLS certificate to secure the custom Docker Registry. The TLS certificate can be self-signed, or signed by a private or public trusted Certificate Authority (CA).
16. On OpenShift, due to a [Red Hat issue](#) with OpenShift Container Platform 4.3.x, the image registry cluster operator configuration must be set to Managed.
17. Check if storage is set up in the cluster image registry operator. See Known Issues DSE-12778 for further information.

For more information on requirements, see CDP Private Cloud Base Installation Guide.

Installation on the OpenShift Container Platform (OCP)

CDP Private Cloud Data Services pre-installation checklist

Before starting the installation, you must ensure that you have configured all the required hardware and software. There are several pre-installation tasks that you must complete using Cloudera Manager and OpenShift Container Platform.

Use the following checklists to ensure that you have completed all the pre-installation tasks:

- CDP Private Cloud Base
- OpenShift Container Platform
- Cloudera Data Warehouse
- Cloudera Machine Learning
- Cloudera Data Engineering

CDP Private Cloud Base checklist

Use this checklist to ensure that your CDP Private Cloud Base is configured and ready for installing CDP Private Cloud Data Services.



Note: The Cloudera Manager mentioned in this checklist is the CDP Private Cloud Base Cloudera Manager using which you want to install CDP Private Cloud Data Services.

Table 6: CDP Private Cloud Base checklist to install CDP Private Cloud Data Services

Item	Summary	Documentation	Notes
Runtime components	Ensure that you have Ranger, Atlas, Hive, HDFS, and Ozone installed in your CDP Private Cloud Base cluster.	<ul style="list-style-type: none"> • Software Support Matrix for OpenShift on page 4 • CDP Private Cloud Base requirements 	If you do not install these components, you see an error when creating an environment in CDP Private Cloud Data Services.

Item	Summary	Documentation	Notes
Network requirement	Ensure that all the network routing hops in production. Cloudera recommends not to use more than 4:1 oversubscription between the spine-leaf switches.		
Cloudera Manager database requirement	Refer to the the CDP Private Cloud Base database requirements.	<ul style="list-style-type: none"> Database Requirements Cloudera Support Matrix 	N/A
Cloudera Manager TLS configuration	Ensure that Cloudera Manager in the CDP Private Cloud Base cluster is configured to use TLS.	Configuring TLS Encryption for Cloudera Manager Using Auto-TLS	You can also manually configure TLS to complete this task. See Manually Configuring TLS Encryption for Cloudera Manager
Cloudera Manager JKS-format TLS truststore	Ensure that the Cloudera Manager is configured with a JKS-format (not PKCS12) TLS truststore.	Obtain and Deploy Keys and Certificates for TLS/SSL	N/A
Cloudera Manager truststore and root certificate	Ensure that the Cloudera Manager truststore contains a root certificate that trusts the certificate for all Cloudera Manager server hosts used with CDP Private Cloud Data Services.	How to Add Root and Intermediate CAs to Truststore for TLS/SSL	Import the necessary certificates into the truststore configured in <code>Configure Administration > Settings > Security > Cloudera Manager TLS/SSL Client Trust Store File</code> .
LDAP configuration	Ensure that you configure LDAP using Cloudera Manager.	N/A	Only Microsoft Active Directory (AD) and OpenLDAP are currently supported.
Apache Ranger configuration for LDAP	Ensure that the CDP Private Cloud Base cluster is configured with Apache Ranger and LDAP for user authentication.	Configure Ranger authentication for LDAP	N/A
Apache Ranger usersync configuration	Ensure that you have configured Apache Ranger and Apache Ranger usersync.	Ranger usersync	Apache Ranger user synchronization is used to get users and groups from the corporate ActiveDirectory to use in policy definitions.
Kerberos configuration	Ensure that Kerberos is enabled for all services in the cluster.	Enabling Kerberos for authentication	Custom Kerberos principals are not currently supported.
Internet access or air gap installation	Ensure that CDP Private Cloud Base and the ECS hosts have access to the Internet. If you do not have access to the Internet, you must do an air gap installation.	Install CDP Private Cloud Data Services in air gap environment	You need access to the Docker registries and the Cloudera repositories during the installation process.
Services health check	Ensure that all services running in the cluster are healthy.	Cloudera Manager Health Tests	N/A
CDP Private Cloud entitlement	Ensure that you have the necessary CDP entitlement from Cloudera to access the Private Cloud installation.	N/A	
Reuse data from CDH or HDP (Optional)	To reuse data from your legacy CDH or HDP deployment in your Private Cloud, ensure that you have migrated that data into your CDP Private Cloud Base. You must be using Cloudera Runtime 7.1.7 for migrating data from your CDH or HDP cluster.	Data Migration Guide	N/A

Item	Summary	Documentation	Notes
(Recommended) Configure HDFS properties to optimize logging	CDP uses “out_webhdfs” Fluentd output plugin to write records into HDFS, in the form of log files, which are then used by different data services to generate diagnostic bundles. To optimize the size of logs that are captured and stored on HDFS, you must update a few HDFS configurations in the hdfs-site.xml file using Cloudera Manager.	Configuring HDFS properties to optimize logging	N/A

OpenShift Container Platform (OCP) checklist

Use this checklist to ensure that your OpenShift Container Platform (OCP) is configured and ready for installing CDP Private Cloud Data Services.

Table 7: OpenShift Container Platform (OCP) checklist to install CDP Private Cloud Data Services

Item	Summary	Documentation	Notes
OpenShift Platform version	Check the the installed OpenShift Container Platform version.	<ul style="list-style-type: none"> OpenShift software requirements Software Support Matrix for OpenShift on page 4 	N/A
DNS configuration	Ensure that you have set up the DNS and Reverse DNS between OpenShift Container Platform (OCP) hosts and CDP Private Cloud Base. This is required for obtaining Kerberos ticket-granting tickets.	Certificate management and DNS	A wildcard DNS entry is required for resolving the ingress route for applications. The ingress route is usually behind a load balancer.
Check if you can access the OpenShift hostnames outside the cluster	Ensure that OpenShift Container Platform (OCP) application hostnames can be accessed from outside the cluster.	A minimal Ingress resource example	Perform a DNS query on the route generated, to check if you can access the hostnames outside the cluster.
Storage classes configuration	Ensure that you have configured separate storage classes for the control plane and the compute clusters. Both the storage classes must be provisioned from Persistent Volumes.	Storage classes	N/A
OpenShift Container Platform (OCP)Kubeconfig file	Ensure that you have access to the OpenShift Container Platform (OCP) Kubeconfig file, cluster administrator privileges, and sufficient expiry time for you to complete your installation.	Download Kubernetes Configuration	The kubeconfig should have valid certificates in it for the cluster. If the kubeconfig does not have certificates, then the you must upload custom certifications during CDP installation.
Allow WebSocket traffic in addition to HTTPS	When a load balancer is used for your OpenShift Container Platform external API, you must allow WebSocket traffic in addition to HTTPS. The load balancer must allow WebSockets on port 80. Also, ensure that you set the load balancer server timeout to 5 minutes.	N/A	N/A

Item	Summary	Documentation	Notes
Clock time from NTP source	Ensure that the NTP clock in CDP Private Cloud Base is in sync with the time configured in the OpenShift Container Platform (OCP) cluster. This is an important step if your setup does not have access to the Internet.	Enable an NTP Service	Install CDP Private Cloud Data Services in air gap environment
Route admission policy	Ensure OpenShift Container Platform (OCP) cluster is configured to run applications in multiple namespaces with the same domain name.	Configuring the route admission policy	N/A

Cloudera Data Warehouse checklist

Use this checklist to ensure that you have all the requirements for Cloudera Data Warehouse in CDP Private Cloud Data Services.

Table 8: Cloudera Data Warehouse installation checklist for CDP Private Cloud Data Services

Item	Summary	Documentation	Notes
OpenShift requirements	Ensure that you have the required memory, storage, and hardware requirements for getting started with the Cloudera Data Warehouse service on Red Hat OpenShift.	OpenShift requirements	N/A
Security requirements	Ensure that you have all the security requirements needed to install and run the Cloudera Data Warehouse Private Cloud service on Red Hat OpenShift clusters.	Security requirements for Cloudera Data Warehouse Private Cloud	N/A
Database requirements	Ensure that you fulfill the requirements for the database that is used for the Hive Metastore on the base cluster (Cloudera Manager side) for Cloudera Data Warehouse (CDW) Private Cloud.	Database requirements	N/A

Cloudera Machine Learning checklist

Use this checklist to ensure that you have all the requirements for Cloudera Machine Learning in CDP Private Cloud Data Services.

Table 9: Cloudera Machine Learning installation checklist for CDP Private Cloud Data Services

Item	Summary	Documentation	Notes
Network File System (NFS) support	Ensure that you have either configured an external or embedded NFS.	CML requirements	N/A
NFS Provisioner	When OCP 4.8 is in use, NFS version 4.0 is required.		
Ranger policy configuration	Ensure that the user who is authenticated using Kerberos needs to have Ranger policies that are configured to allow read/write to the default (or other specified) databases.	CML requirements	N/A

Cloudera Data Engineering checklist

Use this checklist to ensure that you have all the requirements for Cloudera Data Engineering in CDP Private Cloud Data Services.

Table 10: Cloudera Data Engineering installation checklist for CDP Private Cloud Data Services

Item	Summary	Documentation	Notes
Ozone in Base cluster	For workloads to store logs, Ozone in Base cluster is a must. Ensure Ozone is installed on CDP Private Cloud Base cluster.	CDP Private Cloud Base Installation	N/A
Ranger policy configuration	Ensure that the user who is authenticated using Kerberos needs to have Ranger policies that are configured to allow read/write to the default (or other specified) databases.	Kerberos authentication for Apache Ranger	N/A

Installing in internet environment

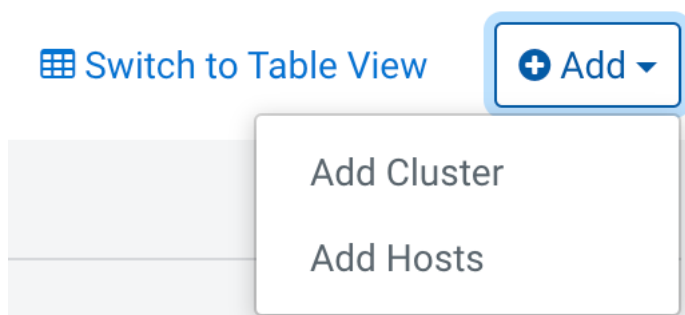
Follow the steps in this topic to install CDP Private Cloud.

Before you begin

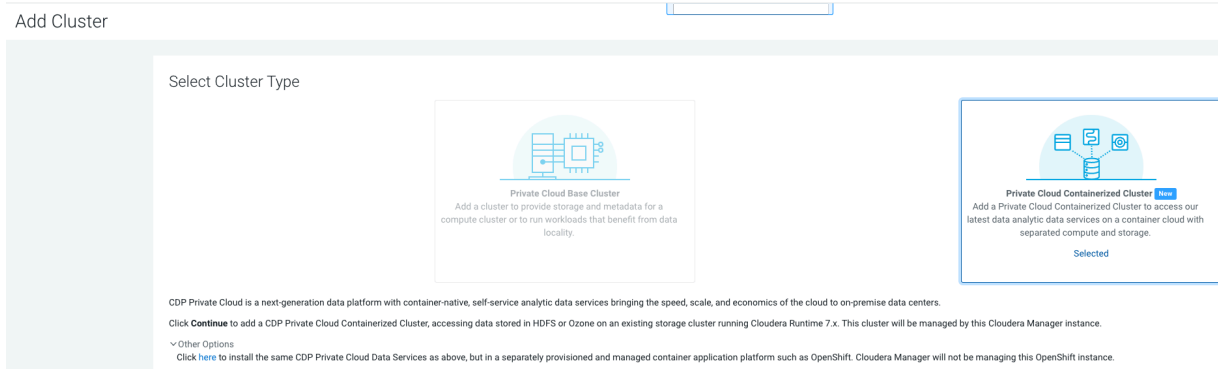
- Ensure that your Kubernetes kubeconfig has permissions to create Kubernetes namespaces.
- You require persistent storage classes defined in your OpenShift cluster. Storage classes can be defined by OpenShift cluster administrators.
- Only TLS-enabled custom Docker Registry is supported. Ensure that you use a TLS certificate to secure the custom Docker Registry. The TLS certificate can be self-signed, or signed by a private or public trusted Certificate Authority (CA).
- Only TLS 1.2 is supported for authentication with Active Directory/LDAP. You require TLS 1.2 to authenticate the CDP control plane with your LDAP directory service like Active Directory.
- OCP network configurations that restrict pod communication are not supported. For example, [multi-tenancy isolation with network policy](#) is not supported.

Procedure

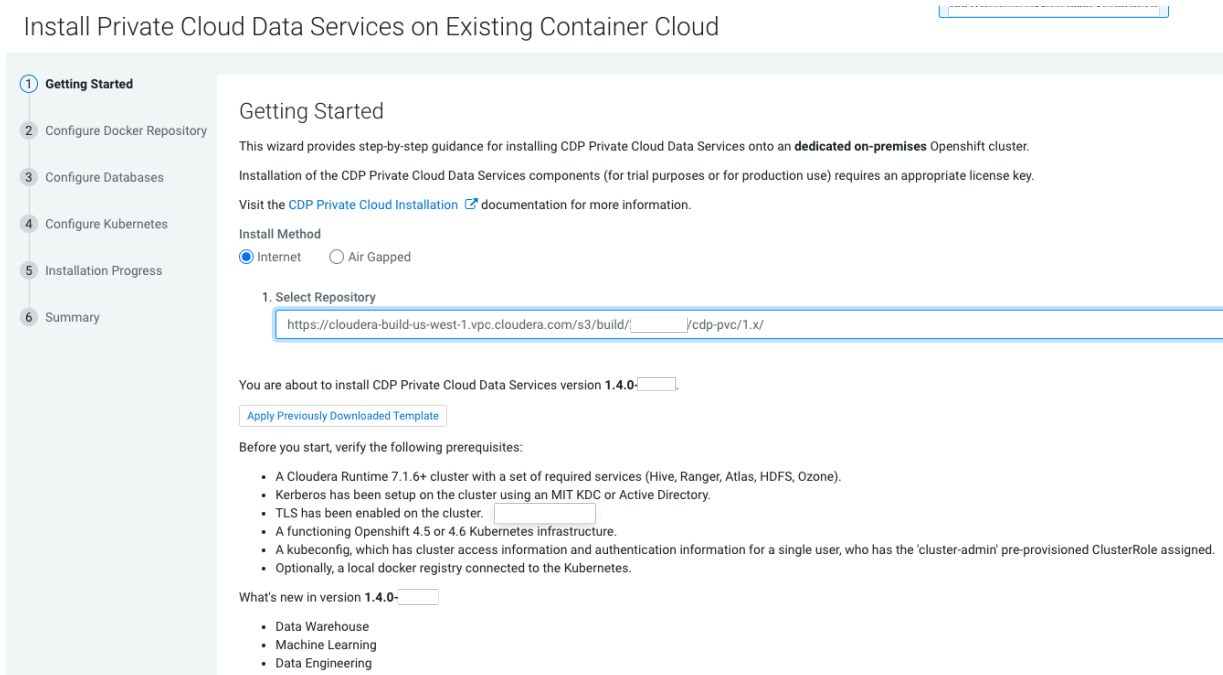
1. In Cloudera Manager, on the top right corner, click Add > Add Cluster. The Select Cluster Type page appears.



- On the Select Cluster Type page, select the cluster type as Private Cloud Containerized Cluster. Under Other Options, click here to install CDP Private Cloud Data Services, then click Continue.



- On the Getting Started page of the installation wizard, select Internet as the Install Method. To use a custom repository link provided to you by Cloudera, click Custom Repository. Click Next.



Note:

- Verify the prerequisites for the version that you're installing and then click Next.
 - You can also apply a template that you may have downloaded during a previous installation. The template contains all the installation configurations. Click Apply Previously Download Template to browse and upload a template stored on your machine.
- On the Configure Docker Repository page, you must select one of the Docker repository options. If you select Use a custom Docker Repository option, enter your local Docker Repository in the Custom Docker Repository field in

the following format: *[*DOCKER REGISTRY*]/[*REPOSITORY NAME*]*. Alternatively, you can use Cloudera's default Docker Repository if you are setting up CDP Private Cloud in non-production environments.

**Note:**

- Use a custom Docker Repository - Copies all images (Internet or Air Gapped) to the embedded registry
- Use Cloudera's default Docker Repository - Copies images from Internet to the embedded registry. This uses the default repository that is in manifest.json. Use Cloudera's default Docker Repository option can be selected only if you have selected Internet as the install method.

You can follow these steps to prepare your Docker Repository from a machine that is running Docker locally and has access to all the Docker images either directly from Cloudera or a local HTTP mirror in your network.

- a) Click Generate the copy-docker script on the wizard or download the script file.
- b) Log in to your custom Docker Registry and run the script using the following commands.

```
docker login <your_custom_registry> -u <user_with_write_access>
```

```
bash copy-docker.txt
```



Note: This command downloads 100+ Docker images and it will take some time to download.

- c) Enter your Docker user name and password.
- d) Click Choose File to upload your Docker certificate.
- e) Click Next.

Install Private Cloud Data Services on Existing Container Cloud

Install Private Cloud Data Services on Existing Container Cloud

5. On the Configure Databases page, click Next.

Install Private Cloud Data Services on Existing Container Cloud

Configure Databases

CDP Private Cloud Control Plane uses an embedded Database to store configuration and other metadata information for the cluster being managed.

Embedded Database Disk Space (GiB) ⓘ

200

Cancel ← Back Next →

6. On the Configure Kubernetes page, enter your Kubernetes, Docker, database, and vault information.
- Upload a Kubernetes configuration (kubeconfig) file from your existing environment. You can obtain this file from your OpenShift Container Platform administrator. Ensure that this kubeconfig has permissions to create Kubernetes namespaces.
 - In the Kubernetes Namespace field, enter the Kubernetes namespace that you want to use with this CDP Private Cloud deployment. Kubernetes virtual clusters are called namespaces. For more information, see [Kubernetes namespaces](#)
 - Enter your Vault information and upload a CA certificate. Cloudera recommends that you use an external Vault for production environments. Enter the Vault address and token, and upload a CA certificate.
 - Enter a Storage Class to be configured on the Kubernetes cluster. CDP Private Cloud uses Persistent Volumes to provision storage. You can leave this field empty if you have a default storage class configured on your Openshift cluster. Click Continue.
 - Under the Additional Certificates section, click Choose File and add the SSL certificate for your HMS database (MariaDB, MySQL, PostgreSQL, or Oracle). For Cloudera Data Warehouse, it is mandatory to

secure the network connection between the default Database Catalog Hive MetaStore (HMS) in CDW and the relational database hosting the base cluster's HMS.

Install Private Cloud Data Services on Existing Container Cloud

Configure Kubernetes

Kubernetes Environment

CDP Private Cloud uses the Kubernetes platform. Please provide a Kubernetes configuration file (also known as a kubeconfig file) from your existing Kubernetes environment.

Kubernetes Configuration

[Choose File](#)

Kubernetes Namespace

cdp

After the installation, CDP management console can be accessed from <https://console-cdp.apps.shared-os-qe-04.kcloud.cloudera.com>

Additional Certificates

Optional additional Certificates to be used during installation and during the runtime of CDP. Examples: Custom Ingress, Custom Kubernetes API,...

Miscellaneous Certificates

[Choose File](#)

Configure Vault

Vault is a secret management tool. You can connect to an existing customer Vault or create a new Vault with this installer. [Learn more](#) on Vault on CDP Private Cloud Data Services.

☒ Embedded vault

☐ External Vault (Recommended for production)

Embedded Vault Disk Space (GiB)

2

Storage

CDP Private Cloud Data Services uses Persistent Volumes to provision storage. This wizard requires a Storage Class to be configured on the Kubernetes cluster prior to launching installation.

Storage Class

[Tip: Before clicking Next, download the current installation configurations as a file template and apply it if you need to reinstall using the same settings.](#)

- If you want to use this installation configuration again to install CDP Private Cloud, you have the option to download this information as a template.



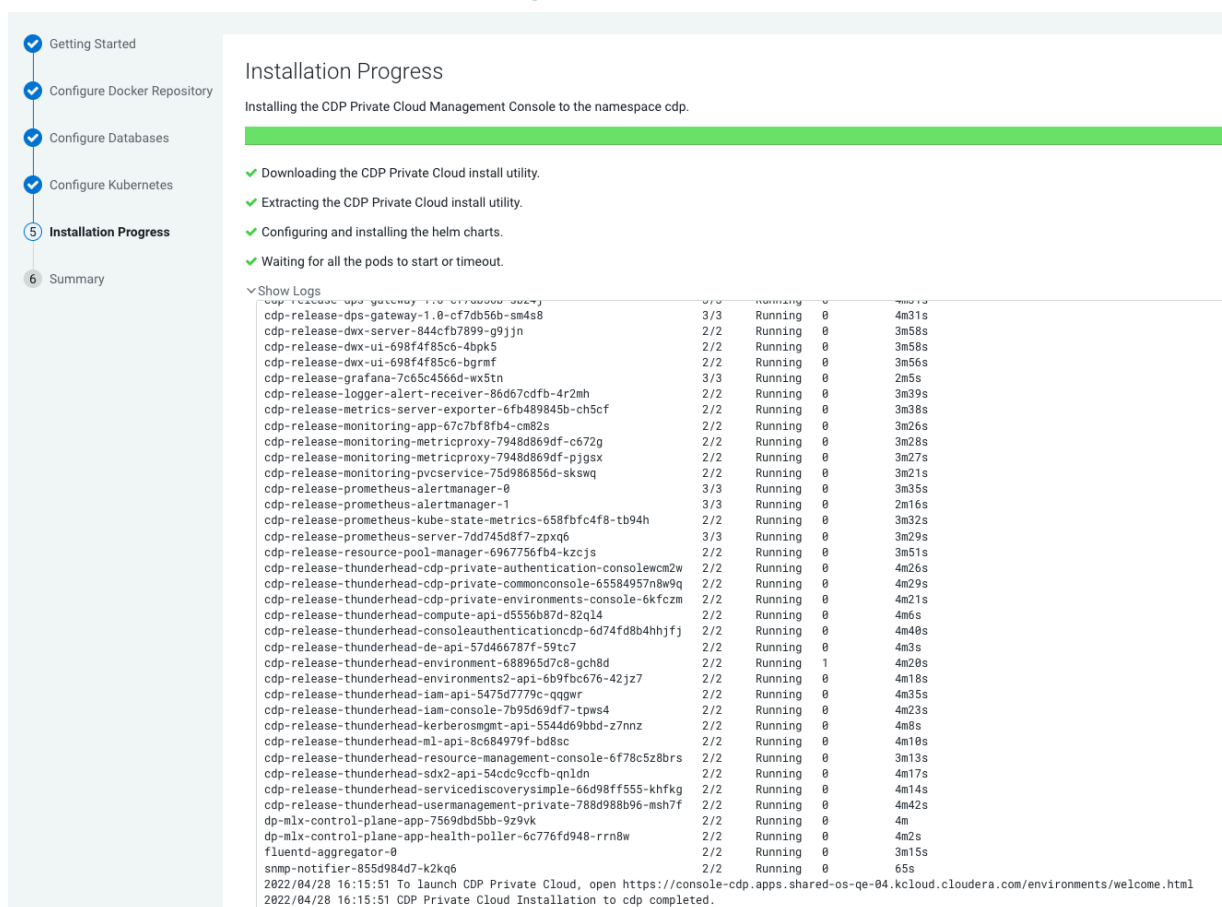
Tip: Before clicking Next, download the current installation configurations as a file template and apply it if you need to reinstall using the same settings.

[Download as Template](#)

The template file is a text file that contains the database and vault information that you entered for this installation. This template is useful if you will be installing Private Cloud again with the same databases, as the template will populate the fields here automatically. Note that the user password information is not saved in the template.

8. The Installation Progress page appears. When the installation is complete, click Next.

Install Private Cloud Data Services on Existing Container Cloud



Installation Progress

Installing the CDP Private Cloud Management Console to the namespace cdp.

- ✓ Downloading the CDP Private Cloud install utility.
- ✓ Extracting the CDP Private Cloud install utility.
- ✓ Configuring and installing the helm charts.
- ✓ Waiting for all the pods to start or timeout.

▼ Show Logs

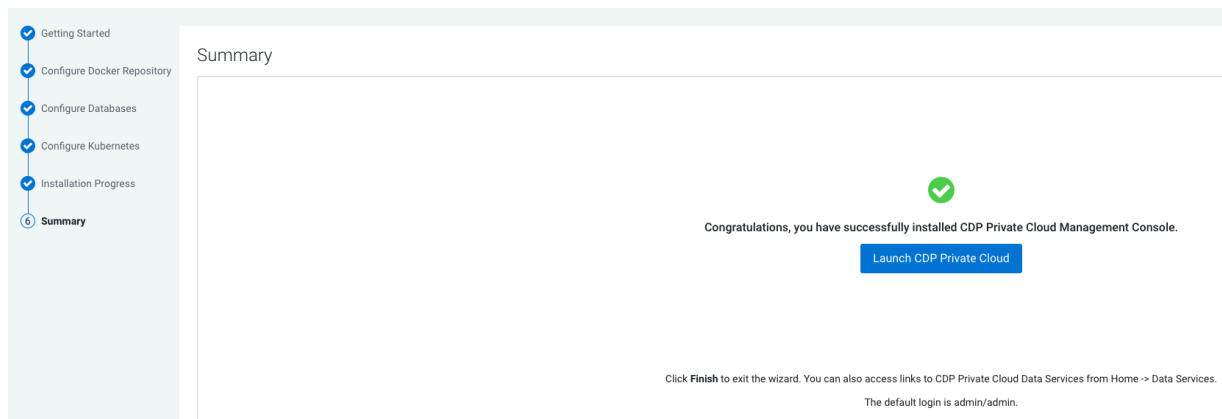
Service	Progress	Status	Age
cdp-release-dps-gateway-1.0-cf7db56b-sm48	3/3	Running	4m31s
cdp-release-dwx-server-844cfb7899-g9jjn	2/2	Running	3m58s
cdp-release-dwx-ui-698f4f85c6-4bpk5	2/2	Running	3m58s
cdp-release-dwx-ui-698f4f85c6-bgrmf	2/2	Running	3m56s
cdp-release-grafana-7c65c4566d-wx5tn	3/3	Running	2m5s
cdp-release-logger-alert-receiver-86d67cdfb-4r2mh	2/2	Running	3m39s
cdp-release-metrics-server-exporter-6fb489845b-ch5cf	2/2	Running	3m38s
cdp-release-monitoring-app-67c7bf8fb4-cm82s	2/2	Running	3m26s
cdp-release-monitoring-metricproxy-7948d869df-c672g	2/2	Running	3m28s
cdp-release-monitoring-metricproxy-7948d869df-pjgsx	2/2	Running	3m27s
cdp-release-monitoring-pvcservice-75d986856d-skswq	2/2	Running	3m21s
cdp-release-prometheus-alertmanager-0	3/3	Running	3m35s
cdp-release-prometheus-alertmanager-1	3/3	Running	2m16s
cdp-release-prometheus-kube-state-metrics-658fbfc4f8-tb94h	2/2	Running	3m32s
cdp-release-prometheus-server-7dd745d8f7-zpxq6	3/3	Running	3m29s
cdp-release-resource-pool-manager-6967756fb4-kzcjs	2/2	Running	3m51s
cdp-release-thunderhead-cdp-private-authentication-consolewcm2w	2/2	Running	4m26s
cdp-release-thunderhead-cdp-private-commonconsole-65584957n8w9q	2/2	Running	4m29s
cdp-release-thunderhead-cdp-private-environments-console-6kfczm	2/2	Running	4m21s
cdp-release-thunderhead-compute-api-d5556b87d-82q14	2/2	Running	4m6s
cdp-release-thunderhead-consoleauthenticationcdp-6d74fd8b4hhjfj	2/2	Running	4m48s
cdp-release-thunderhead-de-api-57d466787f-59tc7	2/2	Running	4m3s
cdp-release-thunderhead-environment-688965d7c8-gch8d	2/2	Running	4m28s
cdp-release-thunderhead-environments2-api-6b9fbc676-42jz7	2/2	Running	4m18s
cdp-release-thunderhead-iam-api-5475d7779c-qgqwr	2/2	Running	4m35s
cdp-release-thunderhead-iam-console-7b95d69df7-tpws4	2/2	Running	4m23s
cdp-release-thunderhead-kerberosgmt-api-5544d69bbd-z7nnz	2/2	Running	4m8s
cdp-release-thunderhead-ml-api-8c684979f-bd8sc	2/2	Running	4m10s
cdp-release-thunderhead-resource-management-console-6f78c5z8brs	2/2	Running	3m13s
cdp-release-thunderhead-sdx2-api-54cdc9ccfb-qnldn	2/2	Running	4m17s
cdp-release-thunderhead-servicediscoverysimple-66d98ff555-khfkq	2/2	Running	4m14s
cdp-release-thunderhead-usermanagement-private-788d988b96-msh7f	2/2	Running	4m42s
dp-mlx-control-plane-app-7569dbd5bb-9z9vk	2/2	Running	4m
dp-mlx-control-plane-app-health-poller-6c776fd948-rnn8w	2/2	Running	4m2s
fluentd-aggregator-0	2/2	Running	3m15s
snmp-notifier-855d984d7-k2kq6	2/2	Running	65s

2022/04/28 16:15:51 To launch CDP Private Cloud, open <https://console-cdp.apps.shared-os-qe-04.kcclouda.cloudera.com/environments/welcome.html>

2022/04/28 16:15:51 CDP Private Cloud Installation to cdp completed.

9. The summary message with a link to Launch CDP appears.

Install Private Cloud Data Services on Existing Container Cloud



Summary

✓

Congratulations, you have successfully installed CDP Private Cloud Management Console.

[Launch CDP Private Cloud](#)

Click **Finish** to exit the wizard. You can also access links to CDP Private Cloud Data Services from Home -> Data Services.

The default login is admin/admin.

What to do next

- Click Launch CDP to launch your CDP Private Cloud.
- Log in using the default user name and password admin.
- In the Welcome to CDP Private Cloud page, click Change Password to change the Local Administrator Account password.
- Set up external authentication using the URL of the LDAP server and a CA certificate of your secure LDAP. Follow the instructions on the Welcome to CDP Private Cloud page to complete this step.

- Click Test Connection to ensure that you are able to connect to the configured LDAP server.
- [Register a CDP Private Cloud environment](#)
- [Create your first Virtual Warehouse in the CDW Data Services](#)
- [Provision an ML Workspace in the CML Data Services](#)

Installing in air gap environment

You can launch the Private Cloud installation wizard from Cloudera Manager and follow the steps to install CDP Private Cloud Data Services in an air gap environment where your Cloudera Manager instance or your Kubernetes cluster does not have access to the Internet.

Before you begin

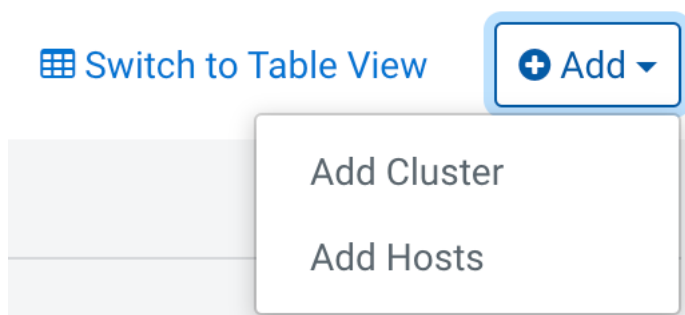
- Ensure that your Kubernetes kubeconfig has permissions to create Kubernetes namespaces.
- You require persistent storage classes defined in your OpenShift cluster. Storage classes can be defined by OpenShift cluster administrators.
- Only TLS-enabled custom Docker Registry is supported. Ensure that you use a TLS certificate to secure the custom Docker Registry. The TLS certificate can be self-signed, or signed by a private or public trusted Certificate Authority (CA).
- Only TLS 1.2 is supported for authentication with Active Directory/LDAP. You require TLS 1.2 to authenticate the CDP control plane with your LDAP directory service like Active Directory.
- OCP network configurations that restrict pod communication are not supported. For example, [multi-tenancy isolation with network policy](#) is not supported.

About this task

If this Cloudera Manager instance or your Kubernetes cluster does not have connectivity to <https://archive.cloudera.com/p/cdp-pvc-ds/>, you must mirror the Cloudera archive URL using a local HTTP server.

Procedure


1. In Cloudera Manager, on the top right corner, click Add > Add Cluster. The Select Cluster Type page appears.




2. On the Select Cluster Type page, select the cluster type as Private Cloud Containerized Cluster. Under Other Options, click here to install CDP Private Cloud Data Services. Click Continue.

Add Cluster

Select Cluster Type



Private Cloud Base Cluster
Add a cluster to provide storage and metadata for a compute cluster or to run workloads that benefit from data locality.



Private Cloud Containerized Cluster [here](#)
Add a Private Cloud Containerized Cluster to access our latest data analytic data services on a container cloud with separated compute and storage.
Selected

CDP Private Cloud is a next-generation data platform with container-native, self-service analytic data services bringing the speed, scale, and economics of the cloud to on-premise data centers.

Click **Continue** to add a CDP Private Cloud Containerized Cluster, accessing data stored in HDFS or Ozone on an existing storage cluster running Cloudera Runtime 7.x. This cluster will be managed by this Cloudera Manager instance.

< Other Options
Click [here](#) to install the same CDP Private Cloud Data Services as above, but in a separately provisioned and managed container application platform such as OpenShift. Cloudera Manager will not be managing this OpenShift instance.

- On the Getting Started page of the installation wizard, select Air Gapped as the Install Method. When you select the Air Gapped install option, extra steps are displayed. Follow these steps to download and mirror the Cloudera archive URL using a local HTTP server.

- Download everything under <https://archive.cloudera.com/p/cdp-pvc-ds/latest>

```
wget -l 0 --recursive --no-parent -e robots=off -nH --cut-dirs=2 --reject="index.html*" -t 10 https://<username>:<password>@archive.cloudera.com/p/cdp-pvc-ds/latest
```

- Edit the manifest.json file in the downloaded directory. Change "http_url": "..." to "http_url": "http://your_local_repo/cdp-pvc-ds/latest"
- Mirror the downloaded directory to your local http server, e.g. http://your_local_repo/cdp-pvc-ds/latest
- Click Custom Repository and add http://your_local_repo/cdp-pvc-ds/latest as a custom repository.
- Click the Select Repository drop-down and select http://your_local_repo/cdp-pvc-ds/latest
- Click Next.

Install Private Cloud Data Services on Existing Container Cloud

Getting Started

This wizard provides step-by-step guidance for installing CDP Private Cloud Data Services onto an **dedicated on-premises** OpenShift cluster. Installation of the CDP Private Cloud Data Services components (for trial purposes or for production use) requires an appropriate license key. Visit the [CDP Private Cloud Installation](#) documentation for more information.

Install Method
☐ Internet ☒ Air Gapped

Installing via a local mirror with an http server. You will need to setup a full mirror of Cloudera's repositories via a temporary http server within the perimeter network of all hosts.

- Download everything under <https://archive.cloudera.com/p/cdp-pvc-ds/latest>

```
$ wget -l 0 --recursive --no-parent -e robots=off -nH --cut-dirs=2 --reject="index.html*" -t 10 https://<username>:<password>@archive.cloudera.com/p/cdp-pvc-ds/latest
```

- Modify the file manifest.json inside the downloaded directory, change "http_url": "..." to "http_url": "http://your_local_repo/cdp-pvc-ds/latest"
- Mirror the downloaded directory to your local http server, e.g. http://your_local_repo/cdp-pvc-ds/latest
- Add http://your_local_repo/cdp-pvc-ds/latest to your Custom Repository settings and select it from the dropdown below.
- Select Repository

[https://cloudera-build-us-west-1.vpc.cloudera.com/s3/build/<id>/cdp-pvc/1.x/](#)

You are about to install CDP Private Cloud Data Services version 1.4.0:

[Apply Previously Downloaded Template](#)

Before you start, verify the following prerequisites:

- A Cloudera Runtime 7.1.6+ cluster with a set of required services (Hive, Ranger, Atlas, HDFS, Ozone).
- Kerberos has been setup on the cluster using an MIT KDC or Active Directory.
- TLS has been enabled on the cluster.
- A functioning OpenShift 4.5 or 4.6 Kubernetes infrastructure.
- A kubeconfig, which has cluster access information and authentication information for a single user, who has the 'cluster-admin' pre-provisioned ClusterRole assigned.
- Optionally, a local docker registry connected to the Kubernetes.

What's new in version 1.4.0:

- Data Warehouse
- Machine Learning
- Data Engineering



Note: You can also apply a template that you may have downloaded during a previous installation. The template contains all the installation configurations. Click Apply Previously Download Template to browse and upload a template stored on your machine.

- On the Configure Docker Repository page, you must select one of the Docker repository options. If you select Use a custom Docker Repository option, then enter your local Docker Repository in the Custom Docker Repository

field in the following format: *[/*DOCKER REGISTRY*/[/*REPOSITORY NAME*]].* Alternatively, you can use Cloudera's default Docker Repository if you are setting up CDP Private Cloud in non-production environments.

**Note:**

- Use a custom Docker Repository - Copies all images (Internet or Air Gapped) to the embedded registry
- Use Cloudera's default Docker Repository - Copies images from Internet to the embedded registry. This uses the default repository that is in manifest.json. Use Cloudera's default Docker Repository option can be selected only if you have selected Internet as the install method.

You can follow these steps to prepare your Docker Repository from a machine that is running Docker locally and has access to all the Docker images either directly from Cloudera or a local HTTP mirror in your network.

- a) Click Generate the copy-docker script on the wizard or download the script file.
- b) Log in to your custom Docker Registry and run the script using the following commands.

```
docker login <your_custom_registry> -u <user_with_write_access>
```

```
bash copy-docker.txt
```



Note: This command downloads 100+ Docker images and it will take some time to download.

- c) Enter your Docker user name and password.
- d) Click Choose File to upload your Docker certificate.
- e) Click Continue.

Install Private Cloud Data Services on Existing Container Cloud

If you select Use an embedded Docker Repository option, then you can download and deploy the Data Services that you need for your cluster.

- a. By selecting Default, all the data services will be downloaded and deployed.
- b. By selecting Select the optional images:
 - If you switch off the Machine Learning toggle key, then the Machine Learning runtimes will not be installed.
 - If you switch on the Machine Learning toggle key, then the Machine Learning runtimes will be installed.

Install Private Cloud Data Services on Existing Container Cloud

Install Private Cloud Data Services on Existing Container Cloud

CDP Deployment from 2022-Mar-14 07:47

Getting Started

Configure Docker Repository

Configure Databases

Configure Kubernetes

Installation Progress

Summary

Configure Docker Repository

Cloudera uses a Docker Repository to deliver CDP Private Cloud Data Services. [Learn more](#) about how to set up custom Docker Repository for CDP Private Cloud Data Services.

☒ Use a custom Docker Repository (Recommended for production)
 ☐ Use Cloudera's default Docker Repository

This release comes with 232 container images that need to be deployed to the Docker repository. Some images are optional and can be skipped by toggling them from the list below. Other images are always installed.

☐ Default
 ☒ Select the Optional Images

☐ Cloudera Machine Learning
 Docker images required to create a Cloudera Machine Learning workspace. Without these images, it will not be possible to use Cloudera Machine Learning.

You will need to deploy 203 container images, approximately 88.7 GiB, to the specified Docker repository. Run the generated script available from the link below.

Custom Docker Repository [🔗](#)

Prepare your Docker Repository from a machine that is running Docker locally and has access to all the Docker images either directly from Cloudera or from a local http mirror in your network.

1. [Generate the copy-docker script](#)
2. Optionally, review the script. The file contains usage information and lists the Docker images that it will download and push.
3. Login to your custom Docker Registry and run the script with the following commands (Note: this downloads 100+ Docker images and it will take a while):

```
docker login <your_custom_registry> -u <user_with_write_access>
bash copy-docker.txt
```

☐ I confirm that I have downloaded all the Docker images to my custom Docker Repository.

Docker Username [🔗](#)

Docker Password [🔗](#)

Docker Certificate [🔗](#)
[Choose File](#)

Install Private Cloud Data Services on Existing Container Cloud

CDP Deployment from 2022-Mar-14 07:47

Getting Started

Configure Docker Repository

Configure Databases

Configure Kubernetes

Installation Progress

Summary

Configure Docker Repository

Cloudera uses a Docker Repository to deliver CDP Private Cloud Data Services. [Learn more](#) about how to set up custom Docker Repository for CDP Private Cloud Data Services.

☒ Use a custom Docker Repository (Recommended for production)
 ☐ Use Cloudera's default Docker Repository

This release comes with 232 container images that need to be deployed to the Docker repository. Some images are optional and can be skipped by toggling them from the list below. Other images are always installed.

☐ Default
 ☒ Select the Optional Images

☒ Cloudera Machine Learning
 Docker images required to create a Cloudera Machine Learning workspace. Without these images, it will not be possible to use Cloudera Machine Learning.

You will need to deploy 204 container images, approximately 99.7 GiB, to the specified Docker repository. Run the generated script available from the link below.

Custom Docker Repository [🔗](#)

Prepare your Docker Repository from a machine that is running Docker locally and has access to all the Docker images either directly from Cloudera or from a local http mirror in your network.

1. [Generate the copy-docker script](#)
2. Optionally, review the script. The file contains usage information and lists the Docker images that it will download and push.
3. Login to your custom Docker Registry and run the script with the following commands (Note: this downloads 100+ Docker images and it will take a while):

```
docker login <your_custom_registry> -u <user_with_write_access>
bash copy-docker.txt
```

☐ I confirm that I have downloaded all the Docker images to my custom Docker Repository.

Docker Username [🔗](#)

Docker Password [🔗](#)

Docker Certificate [🔗](#)
[Choose File](#)

Click Continue.

5. On the Configure Databases page, click Next.

Install Private Cloud Data Services on Existing Container Cloud

The screenshot shows the 'Configure Databases' step of the installation wizard. On the left is a vertical sidebar with six steps: 'Getting Started' (checked), 'Configure Docker Repository' (checked), '3 Configure Databases' (active), '4 Configure Kubernetes', '5 Installation Progress', and '6 Summary'. The main content area is titled 'Configure Databases' and contains the text: 'CDP Private Cloud Control Plane uses an embedded Database to store configuration and other metadata information for the cluster being managed.' Below this is a label 'Embedded Database Disk Space (GiB)' with a help icon, followed by a text input field containing the value '200'. At the bottom of the form are three buttons: 'Cancel', '← Back', and 'Next →'.

Getting Started

Configure Docker Repository

3 Configure Databases

4 Configure Kubernetes

5 Installation Progress

6 Summary

Configure Databases

CDP Private Cloud Control Plane uses an embedded Database to store configuration and other metadata information for the cluster being managed.

Embedded Database Disk Space (GiB) ⓘ

200

Cancel

← Back

Next →

6. On the Configure Kubernetes page, enter your Kubernetes, Docker, database, and vault information.
 - a) Upload a Kubernetes configuration (kubeconfig) file from your existing environment. You can obtain this file from your OpenShift Container Platform administrator. Ensure that this kubeconfig has permissions to create Kubernetes namespaces.
 - b) In the Kubernetes Namespace field, enter the Kubernetes namespace that you want to use with this CDP Private Cloud deployment. Kubernetes virtual clusters are called namespaces. For more information, see [Kubernetes namespaces](#)
 - c) Enter your Vault information and upload a CA certificate. Cloudera recommends that you use an external Vault for production environments. Enter the Vault address and token, and upload a CA certificate.
 - d) Enter a Storage Class to be configured on the Kubernetes cluster. CDP Private Cloud uses Persistent Volumes to provision storage. You can leave this field empty if you have a default storage class configured on your Openshift cluster. Click Continue.

Install Private Cloud Data Services on Existing Container Cloud

Configure Kubernetes

Kubernetes Environment

CDP Private Cloud uses the Kubernetes platform. Please provide a Kubernetes configuration file (also known as a kubeconfig file) from your existing Kubernetes environment.

Kubernetes Configuration

[Choose File](#)

Kubernetes Namespace

After the installation, CDP management console can be accessed from <https://console-cdp.apps.shared-os-qe-04.kcloud.cloudera.com>

Additional Certificates

Optional additional Certificates to be used during installation and during the runtime of CDP. Examples: Custom Ingress, Custom Kubernetes API,...

Miscellaneous Certificates ⓘ

[Choose File](#)

Configure Vault

Vault is a secret management tool. You can connect to an existing customer Vault or create a new Vault with this installer. [Learn more](#) on Vault on CDP Private Cloud Data Services.

☒ Embedded vault

☐ External Vault (Recommended for production)

Embedded Vault Disk Space (GiB) ⓘ

Storage

CDP Private Cloud Data Services uses Persistent Volumes to provision storage. This wizard requires a Storage Class to be configured on the Kubernetes cluster prior to launching installation.

Storage Class ⓘ

ⓘ Tip: Before clicking Next, download the current installation configurations as a file template and apply it if you need to reinstall using the same settings.

7. If you want to use this installation configuration again to install CDP Private Cloud, you have the option to download this information as a template.

ⓘ Tip: Before clicking Next, download the current installation configurations as a file template and apply it if you need to reinstall using the same settings. [Download as Template](#)

The template file is a text file that contains the database and vault information that you entered for this installation. This template is useful if you will be installing Private Cloud again with the same databases, as the template will populate the fields here automatically. Note that the user password information is not saved in the template.

8. The Installation Progress page appears. Click Continue.

Install Private Cloud Data Services on Existing Container Cloud

Installation Progress

Installing the CDP Private Cloud Management Console to the namespace cdp.

- Downloading the CDP Private Cloud install utility.
- Extracting the CDP Private Cloud install utility.
- Configuring and installing the helm charts.
- Waiting for all the pods to start or timeout.

Show Logs

Service	Version	Status	Age
cdp-release-dps-gateway-1.0-cf7db56b-sm48	3/3	Running	0
cdp-release-dwx-server-844cfb7899-g9jjn	2/2	Running	0
cdp-release-dwx-ui-698f4f85c6-4bpk5	2/2	Running	0
cdp-release-dwx-ui-698f4f85c6-bgrmf	2/2	Running	0
cdp-release-grafana-7c65c4566d-wx5tn	3/3	Running	0
cdp-release-logger-alert-receiver-86d67cdfb-4r2mh	2/2	Running	0
cdp-release-metrics-server-exporter-6fb489845b-ch5cf	2/2	Running	0
cdp-release-monitoring-app-67c7bf8fb4-cm82s	2/2	Running	0
cdp-release-monitoring-metricproxy-7948d869df-c672g	2/2	Running	0
cdp-release-monitoring-metricproxy-7948d869df-pjgsx	2/2	Running	0
cdp-release-monitoring-pvcservice-75d986856d-skswq	2/2	Running	0
cdp-release-prometheus-alertmanager-0	3/3	Running	0
cdp-release-prometheus-alertmanager-1	3/3	Running	0
cdp-release-prometheus-kube-state-metrics-658fbfc4f8-tb94h	2/2	Running	0
cdp-release-prometheus-server-7dd745d8f7-zpxq6	3/3	Running	0
cdp-release-resource-pool-manager-6967756fb4-kzcjs	2/2	Running	0
cdp-release-thunderhead-cdp-private-authentication-consolewcm2w	2/2	Running	0
cdp-release-thunderhead-cdp-private-commonconsole-65584957n8w9q	2/2	Running	0
cdp-release-thunderhead-cdp-private-environments-console-6kfczm	2/2	Running	0
cdp-release-thunderhead-compute-api-d5556b87d-82q14	2/2	Running	0
cdp-release-thunderhead-consoleauthenticationcdp-6d74fd8b4hhjf	2/2	Running	0
cdp-release-thunderhead-de-api-57d466787f-59tc7	2/2	Running	0
cdp-release-thunderhead-environment-688965d7c8-gch8d	2/2	Running	1
cdp-release-thunderhead-environments2-api-6b9fbc676-42jz7	2/2	Running	0
cdp-release-thunderhead-iam-api-5475d7779c-qgqwr	2/2	Running	0
cdp-release-thunderhead-iam-console-7b95d69df7-tpws4	2/2	Running	0
cdp-release-thunderhead-kerberosgmt-api-5544d69bbd-z7nnz	2/2	Running	0
cdp-release-thunderhead-ml-api-8c684979f-bd8sc	2/2	Running	0
cdp-release-thunderhead-resource-management-console-6f78c5z8brs	2/2	Running	0
cdp-release-thunderhead-sdx2-api-54cdc9ccfb-qnlnd	2/2	Running	0
cdp-release-thunderhead-servicediscoverysimple-66d98ff555-khfkq	2/2	Running	0
cdp-release-thunderhead-usermanagement-private-788d988b96-msh7f	2/2	Running	0
dp-mlx-control-plane-app-7569dbd5bb-9z9vk	2/2	Running	0
dp-mlx-control-plane-app-health-poller-6c776fd948-rnn8w	2/2	Running	0
fluentd-aggregator-0	2/2	Running	0
snmp-notifier-855d984d7-k2kq6	2/2	Running	0

2022/04/28 16:15:51 To launch CDP Private Cloud, open <https://console-cdp.apps.shared-os-qe-04.kcclouda.cloudera.com/environments/welcome.html>

2022/04/28 16:15:51 CDP Private Cloud Installation to cdp completed.

9. The summary message with a link to Launch CDP appears.

Install Private Cloud Data Services on Existing Container Cloud

Summary

Congratulations, you have successfully installed CDP Private Cloud Management Console.

[Launch CDP Private Cloud](#)

Click **Finish** to exit the wizard. You can also access links to CDP Private Cloud Data Services from Home -> Data Services.

The default login is admin/admin.

What to do next

1. Click Launch CDP to launch your CDP Private Cloud Data Services.
2. Log in using the default user name and password admin/admin.
3. In the Welcome to CDP Private Cloud page, click Change Password to change the Local Administrator Account password.
4. Set up external authentication using the URL of the LDAP server and a CA certificate of your secure LDAP. Follow the instructions on the Welcome to CDP Private Cloud page to complete this step.

5. Click Test Connection to ensure that you can connect to the configured LDAP server.
6. [Register a CDP Private Cloud Data Services environment.](#)
7. [Create your first Virtual Warehouse in the CDW Data Services](#) and/or [Provision an ML Workspace in the CML Data Services.](#)

Uninstall CDP Private Cloud Data Services

You can uninstall CDP Private Cloud Data Services from your CDP Private Cloud Base Cloudera Manager.

Before you begin

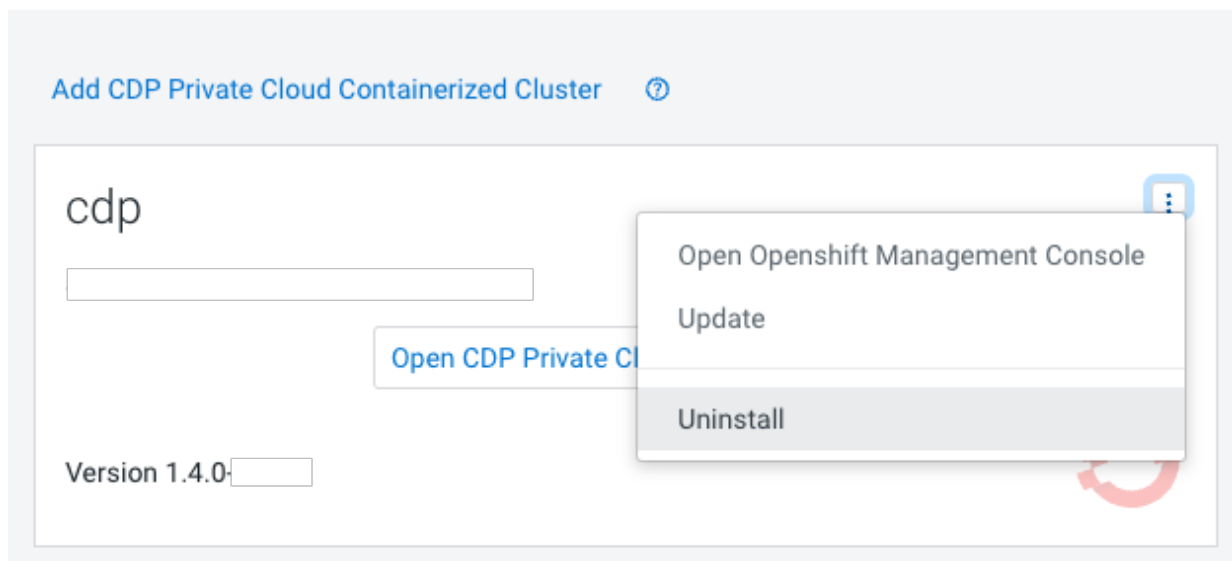
Before you uninstall CDP Private Cloud Data Services, ensure that you have deleted all the CDP Private Cloud environments registered in your CDP Private Cloud Data Services. You can delete your registered environments using Management Console.

Procedure

1.

In Cloudera Manager, navigate to CDP Private Cloud Data Services and click . Click Uninstall.

CDP Private Cloud Data Services



- The Collect Information page appears. You must select the checkbox associated with your CDP Private Cloud Environments. Click Choose File to upload your kubeconfig file associated with your Kubernetes cluster.

Uninstall Private Cloud Data Services (cdp)

CDEP De

1 Collect Information
2 Uninstallation Progress
3 Summary

Collect Information

This wizard uninstalls CDP Private Cloud Data Services.

Visit the [CDP Private Cloud Data Services Uninstallation](#) documentation for more information.

✓

Before you proceed, delete all CDP Private Cloud environments from the [Management Console](#).

☒ All CDP Private Cloud environments have been deleted. (Required)

Kubernetes Environment

Kubernetes Configuration

Choose File

Kubernetes Cluster

Delete shared Cloudera installed artifacts on this Kubernetes Cluster?

☐ Keep shared artifacts

1 Choose this option if there are other CDP Private Cloud instances running in this Kubernetes cluster or if you are not sure.

☒ Delete shared artifacts

1 Choose this option if you are uninstalling the **only** CDP Private Cloud instance in this Kubernetes cluster.

- Select Keep shared artifacts if you have other CDP Private Cloud Data Services instances running in your Kubernetes cluster, or select Delete shared artifacts to remove any cluster global security policies or objects associated with this Kubernetes namespace.

Uninstall Private Cloud Data Services (cdp)

CDEP De

1 Collect Information
2 Uninstallation Progress
3 Summary

Collect Information

This wizard uninstalls CDP Private Cloud Data Services.

Visit the [CDP Private Cloud Data Services Uninstallation](#) documentation for more information.

✓

Before you proceed, delete all CDP Private Cloud environments from the [Management Console](#).

☒ All CDP Private Cloud environments have been deleted. (Required)

Kubernetes Environment

Kubernetes Configuration

Choose File

Kubernetes Cluster

Delete shared Cloudera installed artifacts on this Kubernetes Cluster?

☐ Keep shared artifacts

1 Choose this option if there are other CDP Private Cloud instances running in this Kubernetes cluster or if you are not sure.

☒ Delete shared artifacts

1 Choose this option if you are uninstalling the **only** CDP Private Cloud instance in this Kubernetes cluster.

- Click Continue to complete the process.

Uninstall Private Cloud Data Services (cdp)

✓ Collect Information

2 Uninstallation Progress

3 Summary

Uninstallation Progress

Uninstalling the CDP Private Cloud Management Console in the namespace cdp.

- ✓ Downloading the CDP Private Cloud uninstall utility.
- ✓ Extracting the CDP Private Cloud uninstall utility.
- ✓ Uninstalling CDP Private Cloud.

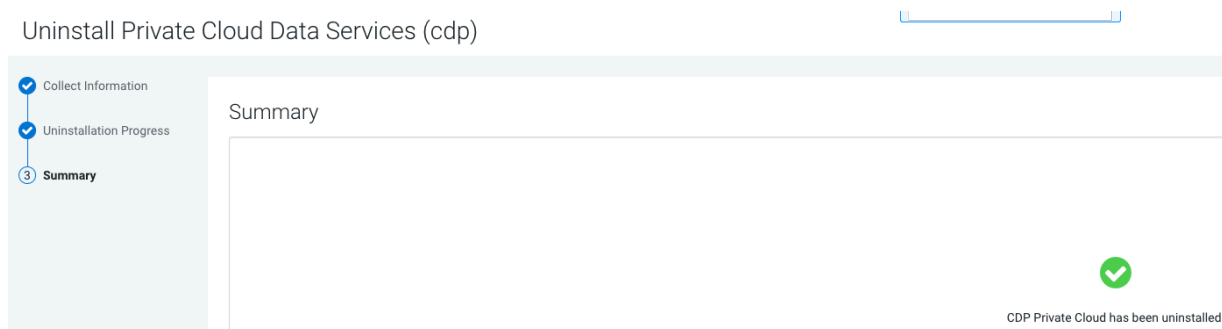
▼ Show Logs

```

2022/04/28 16:29:26 Delete entities of type deployment in namespace yunikorn.
deployment.apps "yunikorn-admission-controller" deleted
deployment.apps "yunikorn-scheduler" deleted
2022/04/28 16:29:26 Delete entities of type pod in namespace yunikorn.
pod "yunikorn-admission-controller-66bd9fdff5-6prpd" deleted
pod "yunikorn-scheduler-5774d5954d-7kc5k" deleted
2022/04/28 16:30:01 Delete entities of type rolebinding in namespace yunikorn.
rolebinding.rbac.authorization.k8s.io "system:deployers" deleted
rolebinding.rbac.authorization.k8s.io "system:image-builders" deleted
rolebinding.rbac.authorization.k8s.io "system:image-pullers" deleted
2022/04/28 16:30:02 Delete entities of type serviceaccount in namespace yunikorn.
serviceaccount "builder" deleted
serviceaccount "default" deleted
serviceaccount "deployer" deleted
serviceaccount "yunikorn-admin" deleted
2022/04/28 16:30:03 Delete entities of type role in namespace yunikorn.
No resources found
2022/04/28 16:30:03 Delete entities of type pvc in namespace yunikorn.
No resources found
2022/04/28 16:30:03 Delete entities of type configmap in namespace yunikorn.
configmap "kube-root-ca.crt" deleted
configmap "openshift-service-ca.crt" deleted
configmap "yunikorn-quotamanager-configs" deleted
configmap "yunikorn-scheduler-plugin-configs" deleted
2022/04/28 16:30:03 Delete entities of type secret in namespace yunikorn.
secret "builder-dockercfg-hcrmv" deleted
secret "builder-token-qzk6c" deleted
secret "builder-token-wgdc4" deleted
secret "cdp-private-installer-docker-cert" deleted
secret "cdp-private-installer-docker-registry" deleted
secret "default-dockercfg-wkkf9" deleted
secret "default-token-69gdb" deleted
secret "deployer-dockercfg-4rj7q" deleted
secret "deployer-token-hkfgf" deleted
secret "deployer-token-tfmc6" deleted
2022/04/28 16:30:06 Delete entities of type networkpolicy in namespace yunikorn.
No resources found
namespace "yunikorn" deleted
2022/04/28 16:30:19 Global Shared Objects Deletion completed.

```

You will now see that CDP Private Cloud has been uninstalled.



Dedicating OCP nodes for specific workloads

You can use the `kubectl taint` command to dedicate OCP cluster nodes for specific workloads. You can dedicate GPU nodes for CML workloads, and NVME nodes for CDW workloads.

About this task

Run the following command to get a list of all of the cluster nodes:

```
kubectl get nodes
```

Run the following command to list information about a specific cluster node:

```
kubectl describe node <node_name>
```

In the returned output, look for the Taints field.

Dedicate a GPU node for CML workloads

1. Run the following command to dedicate a GPU node for CML workloads:

```
kubectl taint nodes <node_name> nvidia.com/gpu=true:NoSchedule
```

No other workload pods will be allowed to run on the tainted node.

2. Run the following command to confirm that the taint has been successfully applied:

```
kubectl describe node <node_name>
```

In the returned output, look for the Taints field.

- To remove the taint, run the following command:

```
kubectl taint nodes <node_name> nvidia.com/gpu=true:NoSchedule-
```

This command returns:

```
node/<node_name> untainted
```

Dedicate a SSD node for CDW workloads

1. Run the following command to dedicate a GPU node for CDW workloads:

```
kubectl taint nodes <node_name> ssd/nvme=true:NoSchedule
```

No other workload pods will be allowed to run on the tainted node.

2. Run the following command to confirm that the taint has been successfully applied:

```
kubectl describe node <node_name>
```

In the returned output, look for the Taints field.

- To remove the taint, run the following command:

```
kubectl taint nodes <node_name> ssd/nvme=true:NoSchedule-
```

This command returns:

```
node/<node_name> untainted
```

Additional Notes



Note:

To taint the node of an existing cluster which already has CML and CDW workspaces running, you must also run the following commands:

```
kubectl drain <node_name> --ignore-daemonsets --delete-emptydir-data --  
timeout=600s  
kubectl uncordon <node_name>
```



Note:

You cannot apply a taint to a master node, or to a single-node cluster.

Related Information

[Taints and Tolerations](#)