CDP Private Cloud Data Services 1.4.1

CDP Private Cloud Data Services Installation using the OpenShift Container Platform

Date published: 2020-12-16
Date modified: 2022-11-18

CLOUDERA

https://docs.cloudera.com/

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

CDP Private Cloud Data Services overview....4

Requirements....4

Red Hat OpenShift Container Platform hardware requirements....4
- Cloudera Data Warehouse hardware requirements....5
- Cloudera Machine Learning requirements (OCP)....7
- CDE hardware requirements....9
- How to use the CDP Private Cloud Data Services sizing spreadsheet....10

Red Hat OpenShift Container Platform software requirements....13
- Credentials....13
- Security context credentials....13
- Load balancing and ingress....13
- Certificate management and DNS....14
- Storage classes....14

CDP Private Cloud Base requirements....14
- Preparing CDP Private Cloud Base....15

CDP Private Cloud Data Services Hardware Requirements....16
- CDP Private Cloud Data Services deployment considerations....16
- Storage requirements....16

CDP Private Cloud Data Services network infrastructure considerations....17

CDP Private Cloud Data Services Software Requirements....18
- External vault requirements....18
- Docker repository access....19
- CML software requirements for Private Cloud....20

Installation on the OpenShift Container Platform (OCP)....21

Set up the external databases....21

CDP Private Cloud Data Services pre-installation checklist....22
- CDP Private Cloud Base checklist....22
- OpenShift Container Platform (OCP) checklist....23
- Cloudera Data Warehouse checklist....24
- Cloudera Machine Learning checklist....25
- Cloudera Data Engineering checklist....25

Installing in internet environment....26

Installing in air gap environment....32

Uninstall CDP Private Cloud Data Services....40

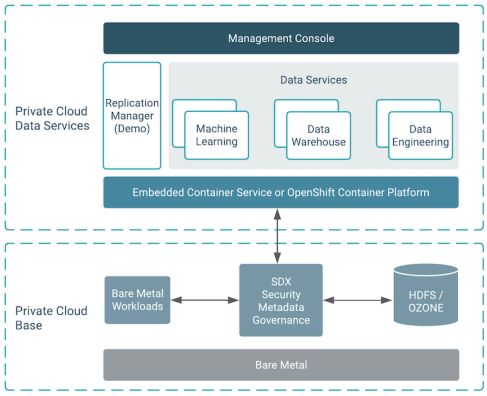
Upgrading....43

Update from 1.3.4 or 1.4.0-H1 to 1.4.1 (OCP)....43
- Preparing for CDP Private Cloud Data Services update for CDE....48
- Updating CDP Private Cloud Data Services....48
- Completing post OCP update tasks....53
- Recovering a corrupted CDE Embedded database....53
- Post Upgrade Steps for CDP PVC 1.4.1 in OCP Clusters....54

CDP Private Cloud Data Services overview

Review the installation requirements and core tasks for installing CDP Private Cloud. CDP Private Cloud Data Services works on top of CDP Private Cloud Base and is the on-premise offering of CDP that brings many of the benefits of the public cloud deployments to the on-premise CDP deployments. CDP Private Cloud Data Services lets you deploy and use the Cloudera Data Warehouse (CDW), Cloudera Machine Learning (CML), and Cloudera Data Engineering (CDE) Data Services.

You must install CDP Private Cloud Data Services on an existing deployment of CDP Private Cloud Base. To install CDP Private Cloud, you need an isolated hardware environment with dedicated infrastructure and networking. CDP Private Cloud Data Services uses containers on the Red Hat OpenShift Container Platform.



You can install CDP Private Cloud Base on virtual machines or bare-metal hardware. CDP Private Cloud Base provides the following components and services that are used by CDP Private Cloud Data Services:

- SDX Data Lake cluster for security, metadata, and governance
- HDFS or Ozone for storage
- Cloudera Runtime components such as Ranger, Atlas, and Hive Metastore (HMS)
- Networking infrastructure that supports network traffic between storage and compute environments

Requirements

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 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.

Cloudera Data Warehouse hardware requirements

Review the requirements needed to get started with the Cloudera Data Warehouse (CDW) service on Red Hat OpenShift 4.7 or 4.8 clusters.

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 10.

- CDP Cloudera Manager must be installed and running.
- CDP Private Cloud must be installed and running. See Private Cloud Installation Overview and Requirements 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 listed in the Private Cloud Installation Requirements documentation, CDW also has the following minimum memory, storage, and hardware requirements for each OpenShift 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
- increments of 100 GB (in 5 chunks of 20 GB each) of persistent volume storage per Virtual Warehouse

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.

Get started with OpenShift and 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 4.7 or 4.8 and (ECS).

To get started with the CDW service on Red Hat OpenShift 4.7 or 4.8 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 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.

- CDP Cloudera Manager must be installed and running.
- CDP Private Cloud must be installed and running. See Private Cloud Installation Overview and Requirements 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 listed in the Private Cloud Installation Requirements documentation, CDW also has the following minimum memory, storage, and hardware requirements for each OpenShift 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

Component	vCPU	Memory	Local Storage	Number of pods in XSMALL Virtual Warehouse
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
DAS	0.5	4 GB	--	1
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	--
Data Visualization (optional)	0.5	8 GB	--	1
Total for XSMALL Virtual Warehouse	21 (20.6)	237 GB (236.1)	400 GB - 4 volumes	--

Database Catalog low resource mode requirements

The metastore and the DAS event processor each use 2 CPUs and 8 GB of memory.

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 4.7 or 4.8. For ECS, refer to the Hardware and Software Requirements section in Installing and Managing a Private Cloud Experience Cluster 1.4.0.
2. CML assumes it has cluster-admin privileges on the cluster.
3. Storage:
   a. 4 TB of persistent volume block storage per ML Workspace.
   b. 1 TB of NFS space recommended per Workspace (depending on user files).
   c. Access to NFS storage is routable from all pods running in the cluster.
   d. For monitoring, recommended volume size is 60 GB.

4. 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.
5. 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.
6. 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.
7. 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.
8. Forward and reverse DNS must be working.
9. DNS lookups to sub-domains and the ML Workspace itself should work.
10. 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.
11. 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.
12. If you intend to access a workspace over https, see Deploy an ML Workspace with Support for TLS.
13. For non-TLS ML workspaces, websockets need to be allowed for port 80 on the external load balancer.
14. 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).
15. 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.
16. 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	4 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

To launch the Cloudera Data Engineering service in the CDP Private Cloud Data Services host must meet several requirements. Review the requirements to get started with the Cloudera Data Engineering (CDE) service on Red Hat OpenShift 4.7 or 4.8 clusters.

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.
- Storage Class: CDE assumes a storage class named nfs exists backed by an NFS server.
- CDE Service requirements:

Table 3: The following are the CDE Service requirements:

Component	vCPU	Memory	Block PV or NFS PV	Number of pods
Embedded DB	--	--	100 GB	1
Config Manager	--	--	--	2
Dex Downloads	--	--	--	1
Knox	250m	1 GB	--	1
Management API	1	2 GB	--	1
NGINX Ingress Controller	100m	90 MiB	--	1
FluentD Forwarder	1	1 GB	--	1
Grafana	-	-	10 GB	1

- CDE Virtual Cluster requirements: Overall storage 500 GB per Virtual Cluster and a storage class named nfs exists backed by an NFS server.

  Table 4: The following are the CDE Virtual Cluster requirements:

Component	vCPU	Memory	Block PV or NFS PV	Number of pods
Airflow API	1,250 m	2 GB, 512 MiB	100 GB	2
Airflow Scheduler	1	1 GB	100 GB	1
Airflow Web	1	2 GB	100 GB	1
Runtime API	1	2 GB	100 GB	1
Livy	3	12 GB	100 GB	1
Safari	4	6 GB	--	1

- Workloads: Depending upon the workload, you must configure resources. Minimal spark/airflow jobs need 1 executor with 1 CPU and 1GB Memory.

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 E25.

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 and monitoring

Label	Cell	Description
PvC Control Plane	B3	1 required
– Monitoring	B4	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)	B6	Enter the number of Data Catalogs you will need in your deployment. You must have at least one Data Catalog.
CDW LLAP warehouses	B7	Enter the number of LLAP warehouses you will need for each Virtual Warehouse in your deployment.
-- LLAP Executors	B8	Enter the total number of LLAP Executors you will need in your deployment.
CDW Impala warehouses	B9	Enter the number of CDW Impala warehouses for each Virtual Warehouse you will need in your deployment.
-- Impala Coordinators (2 x for HA)	B10	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	B11	Enter the number of Impala Executors you will need in your deployment.
-- CDW Data Cache	B12	Enter the amount of CDW Cache space for each coordinator and executor (Default 600)
Data Viz - small	B13	Enter the size selected when creating a Data Visualization instance.
Data Viz -medium	B14	
Data Viz -large	B15	

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

- (OCP) CDE hardware requirements.
- (ECS) Additional resource requirements for Cloudera Data Engineering

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)	B17	Enter the number of workspaces you need in your deployment.
-- CML Small session	B18	Enter the number of concurrent small-sized sessions you intend to run.
-- CML Average session	B19	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 requirementsx

Cloudera Data Engineering (CDE)

Label	Cell	Description
CDE Service (min/max 1 per cluster)	B21	Enter the number of CDE clusters you will need in your deployment.
CDE Virtual Cluster	B22	Enter the number of CDE Virtual Clusters you will need in your deployment.
-- CDE Small jobs	B23	Enter the number of concurrent small-sized jobs you intend to run.
-- CDE Avg Jobs	B24	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 32+ cores (64vcores)	B28	Enter the number of vcores for each worker node.
RAM (GB) recommend 384GB RAM	B29	Enter the amount of RAM, in gigabytes, for each worker node.
Disk (GB) Block (OCP CSI block, ECS Longhorn)	B30	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)	B31	Enter the number of gigabytes of Fast Cache used in Cloudera Data Warehouse.
NFS (GB) (choose 1 from below)	B33	Enter required storage in either cell B30 or cell B31:
-- Embedded nfs - (subtract from Block provider) non-prod	B33	Enter the number of gigabytes storage for an embedded NFS.
-- External nfs	B35	Enter the number of gigabytes of storage for an External NFS.

Label	Cell	Description
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: 8 vcores Recommended: 16 vcores
	B39	Minimum 16 GB RAM Recommended: 32 GB RAM
	B40	Minimum: 300 GB HDD (This amount is adequate for 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.

Ensure that you have OpenShift Container Platform 4.7.x or 4.8.x.

Note: If you want to update your existing CDP Private Cloud Data Services 1.3.4 or 1.4.0 to 1.4.1 and you are running OCP 4.6.x, you can do an in-place update, and then upgrade to OCP 4.7.x or 4.8.x. For more information about updating your CDP Private Cloud Data Services, see Update from 1.3.4 or 1.4.0-H1 to 1.4.1 (OCP) on page 43.

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.

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:

- CentOS 8.4, 7.x, Red Hat Enterprise Linux 8.4. 7.x, and Oracle Linux 8.4
- JDK 11 (any distribution)
- PostgreSQL 10 and 12
- For CDW, Mysql 5.7 and Maria DB 10.2

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.

Ensure that you have CDP Private Cloud Base 7.1.8 or 7.1.7 SP1 with a Data Lake cluster (CDP Private Cloud Base 7.1.7 SP1 is compatible with CDP Private Cloud Data Services 1.4.1 if you upgrade to Cloudera Manager 7.8.1). For the Private Cloud Base cluster setup, you can use the latest version of Cloudera Manager 7.8.1

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 or MIT KDC
- 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

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.

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.
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 Private Cloud OpenShift memory and storage 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	600GB per Node	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.

Note: You must not use CephFS as it is not yet supported. CephFS support is planned for a future release.

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.

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 paywall 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: The Custom Docker Registry is not supported for CML in CDP Private Cloud Data Services version 1.4.x.

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 4.7 or 4.8. For ECS, refer to the Hardware and Software Requirements section in Installing and Managing a Private Cloud Experience Cluster 1.4.0.
2. CML assumes it has cluster-admin privileges on the cluster.
3. Storage:
   a. 4 TB of persistent volume block storage per ML Workspace.
   b. 1 TB of NFS space recommended per Workspace (depending on user files).
   c. Access to NFS storage is routable from all pods running in the cluster.
   d. For monitoring, recommended volume size is 60 GB.
4. 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.
5. 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.
6. 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.
7. 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.
8. Forward and reverse DNS must be working.
9. DNS lookups to sub-domains and the ML Workspace itself should work.
10. 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.

11. 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.
12. If you intend to access a workspace over https, see Deploy an ML Workspace with Support for TLS.
13. For non-TLS ML workspaces, websockets need to be allowed for port 80 on the external load balancer.
14. 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).
15. 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.
16. 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)

Set up the external databases

You must set up the external databases to be used with CDP Private Cloud Data Services. You must enable the base cluster PostgreSQL database to use an SSL connection to encrypt client-server communication as a requirement for Cloudera Data Warehouse (CDW) data service.

To set up the external databases for CDP Private Cloud Data Services with SSL, complete the following steps:

1. Install PostgreSQL 10 or 12.
2. In the postgresql.conf file, set the following parameters:
   - parameters listen_addresses = '*'
   - max_connections = 1000
3. In the pg_hba.conf file, add the following lines:

```
host all all 0.0.0.0/0 md5
host all all ::/0 md5
```

4. Place a leaf certificate, and the certificate's private key on the Postgres server.
5. Set ssl=on in the postgresql.conf file.
6. In the postgresql.conf file, set the following parameters
   - Set the ssl_cert_file parameter to the location of the leaf certificate.
   - Set the ssl_key_file parameter to the location of the key file.
7. When you are prompted during the installation, upload the CA certificate that signed the leaf certificate.
8. Restart the Postgres server.

Related Information

Setting Parameters in PostgreSQL

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 5: CDP Private Cloud Base checklist to install CDP Private Cloud Data Services

Item	Summary	Documentation	Notes
Runtime version and components	Ensure that you have CDP Private Cloud Base 7.1.8 or 7.1.7 SP1 with a Data Lake cluster (CDP Private Cloud Base 7.1.7 SP1 is compatible with CDP Private Cloud Data Services 1.4.1. if you upgrade to Cloudera Manager 7.8.1). Ensure that you have Ranger, Atlas, Hive, and Ozone installed in your CDP Private Cloud Base cluster.	CDP Private Cloud Base requirements	If you do not install these components, you see an error when creating an environment in your CDP Private Cloud Data Services.
Network requirement	Ensure that the CDP Private Cloud Base and OpenShift Container Platform are on the same network.	Network infrastructure considerations	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.	• 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 Configure Administration > Settings > Security > Cloudera Manager TLS/SSL Client Trust Store File .

Item	Summary	Documentation	Notes
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 your CDP Private Cloud Base and your Kubernetes cluster has access to the Internet. If you do not have access to the Internet, you must do an air gap installation.	CDP Private Cloud Base requirements 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	To check this, go to Cloudera Manager > Private Cloud > Select Repository > [***REPOSITORY URL***] . If you have the required entitlements, the You are about to install the 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.
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
(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 6: OpenShift Container Platform (OCP) checklist to install CDP Private Cloud Data Services

Item	Summary	Documentation	Notes
OpenShift Platform version	Ensure that the installed OpenShift Container Platform is version 4.7.x or 4.8.x.	OpenShift software requirements	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.	Certificatef 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
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 7: 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 8: 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 9: 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

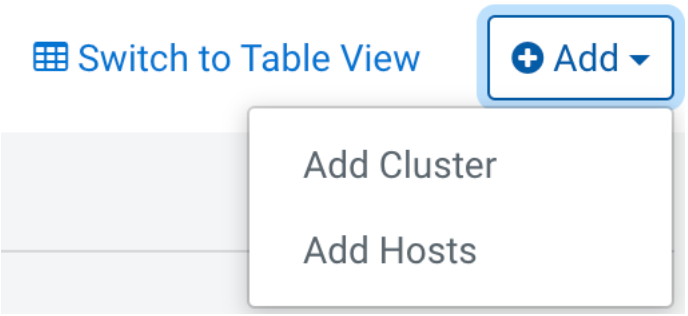
Follow the steps in this topic to install CDP Private Cloud.

Before you begin

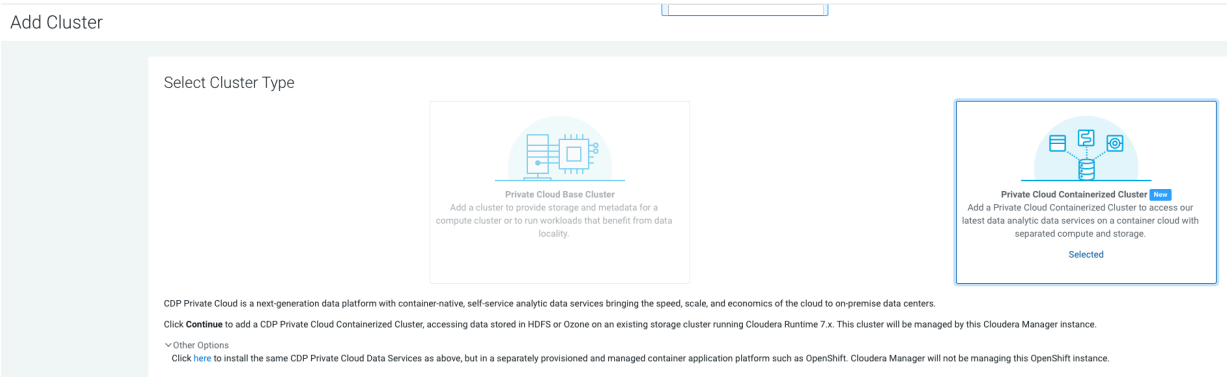
- Ensure that your Kubernetes kubeconfig has permissions to create Kubernetes namespaces.
- You require persistent storage classes defined in your OpenShift cluster. Storage classes can be defined by OpenShift cluster administrators.
- Only TLS-enabled custom Docker Registry is supported. Ensure that 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.

Procedure

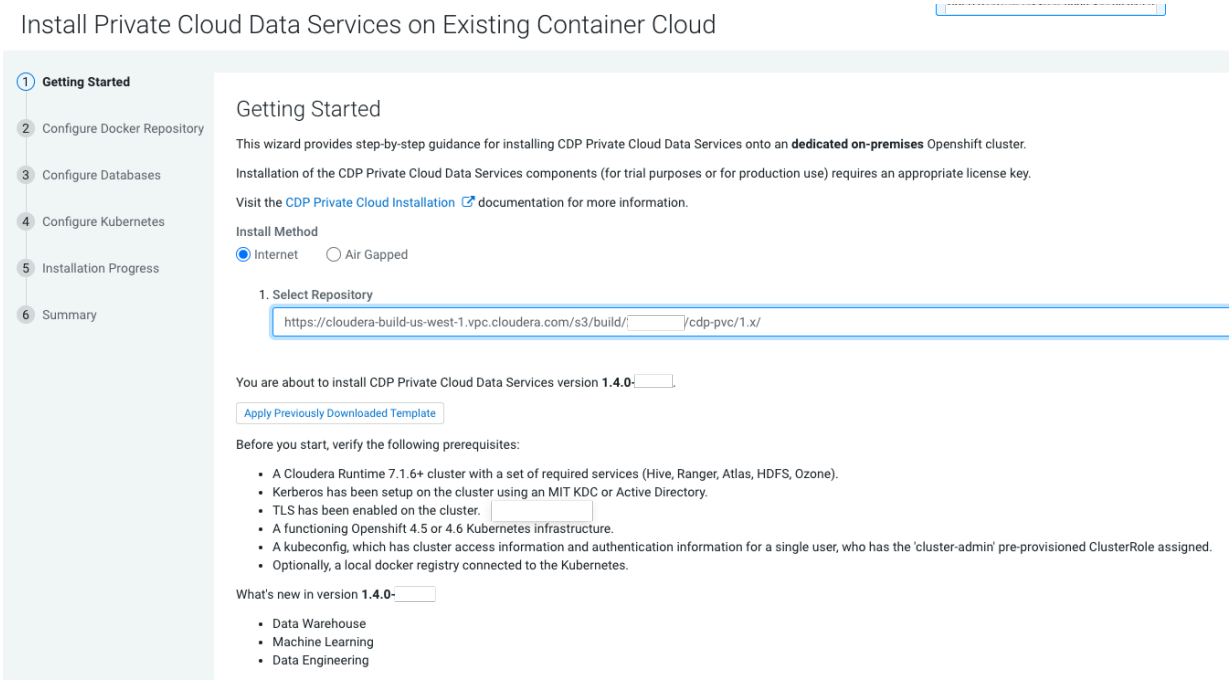
1. In Cloudera Manager, on the top right corner, click Add > Add Cluster. The Select Cluster Type page appears.



2. In the Select Cluster Type page, select the cluster type as Private Cloud Containerized Cluster and under Other Options dropdown, click here to install CDP Private Cloud Data Services. click Continue.



3. 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 Continue.



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.

4. In 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.

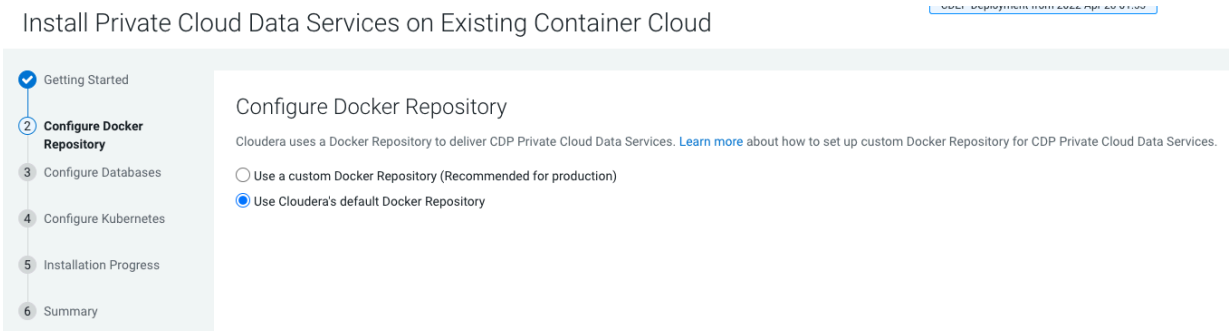
a) Click Generate the copy-docker script on the wizard or download the script file.
b) Log in to your custom Docker Registry and run the script using the following commands.

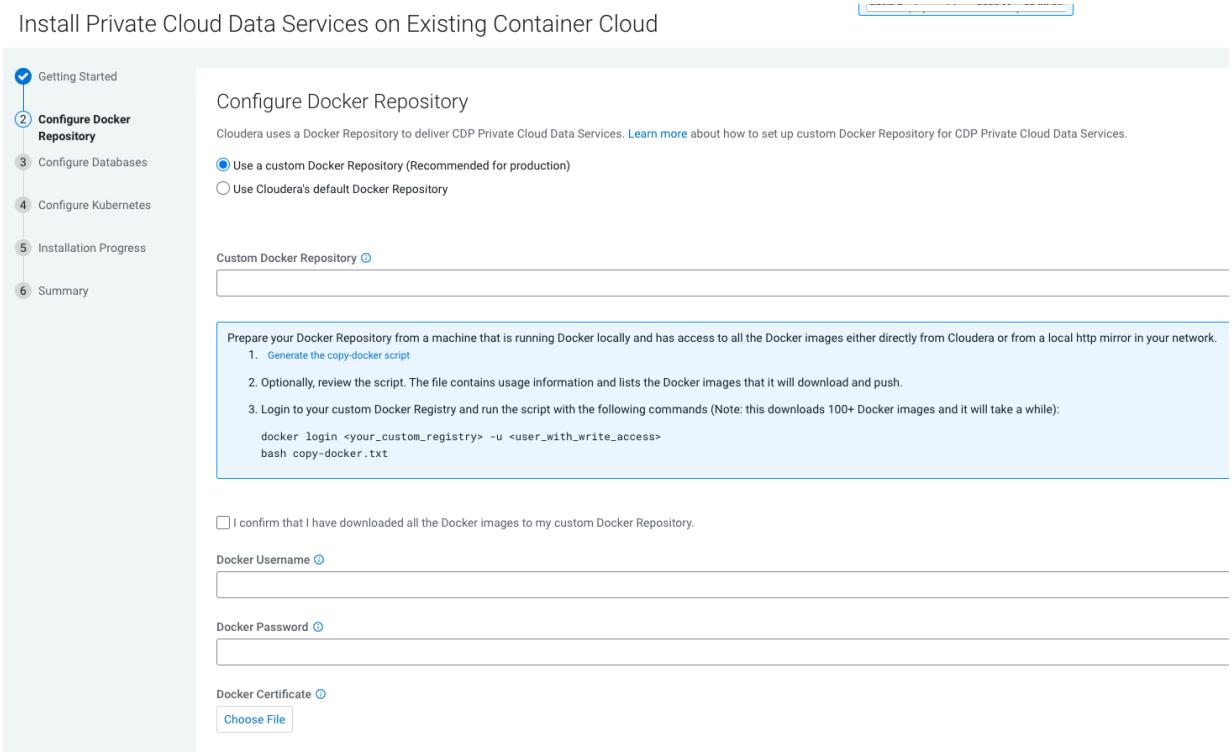
```
docker login <your_custom_registry> -u <user_with_write_access>
```

```
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.





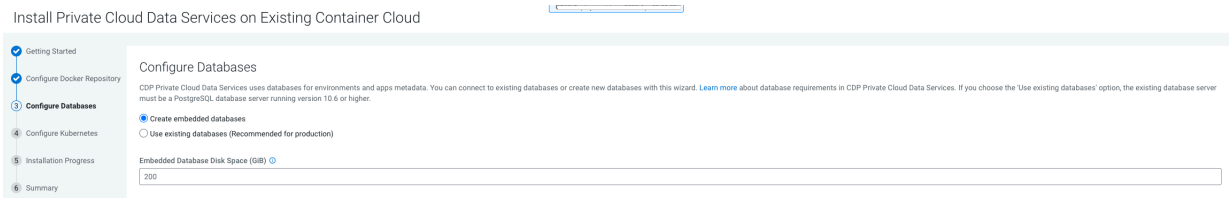
5. In the Configure Databases page, follow the instructions in the wizard to use your external existing databases with CDP Private Cloud.

   For production environments, Cloudera recommends that you use databases that you have previously created. These databases must all be on the same host and that host must be a PostgreSQL database server running version 10 or higher.

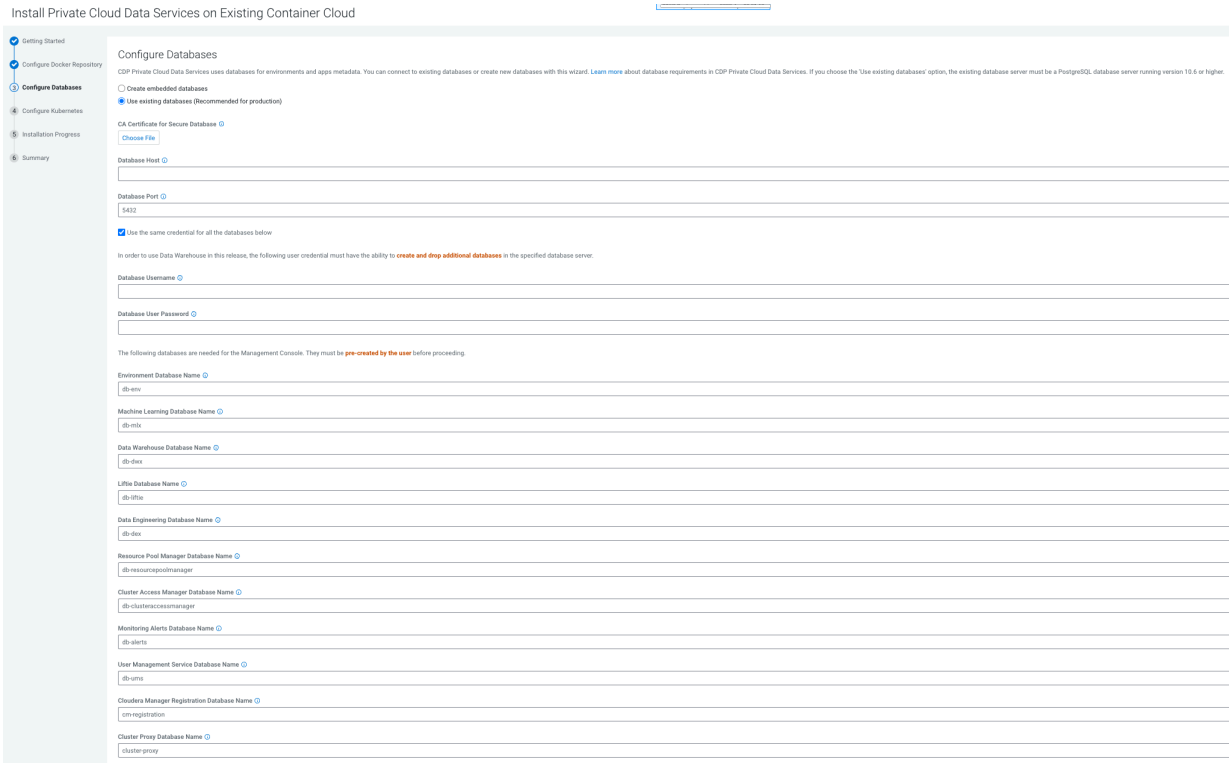
   CDW supports MariaDB, MySQL, and Oracle databases in addition to PostgreSQL. If you plan to use Cloudera Data Warehouse (CDW) Data Service, then you must add the CA certificate for your base cluster database to enable encrypted client-server communication.

   The databases must be created before you can proceed with the installation, and the database user must have permissions to create and drop databases on the server.

   To create an embedded database, select Create embedded databases option.

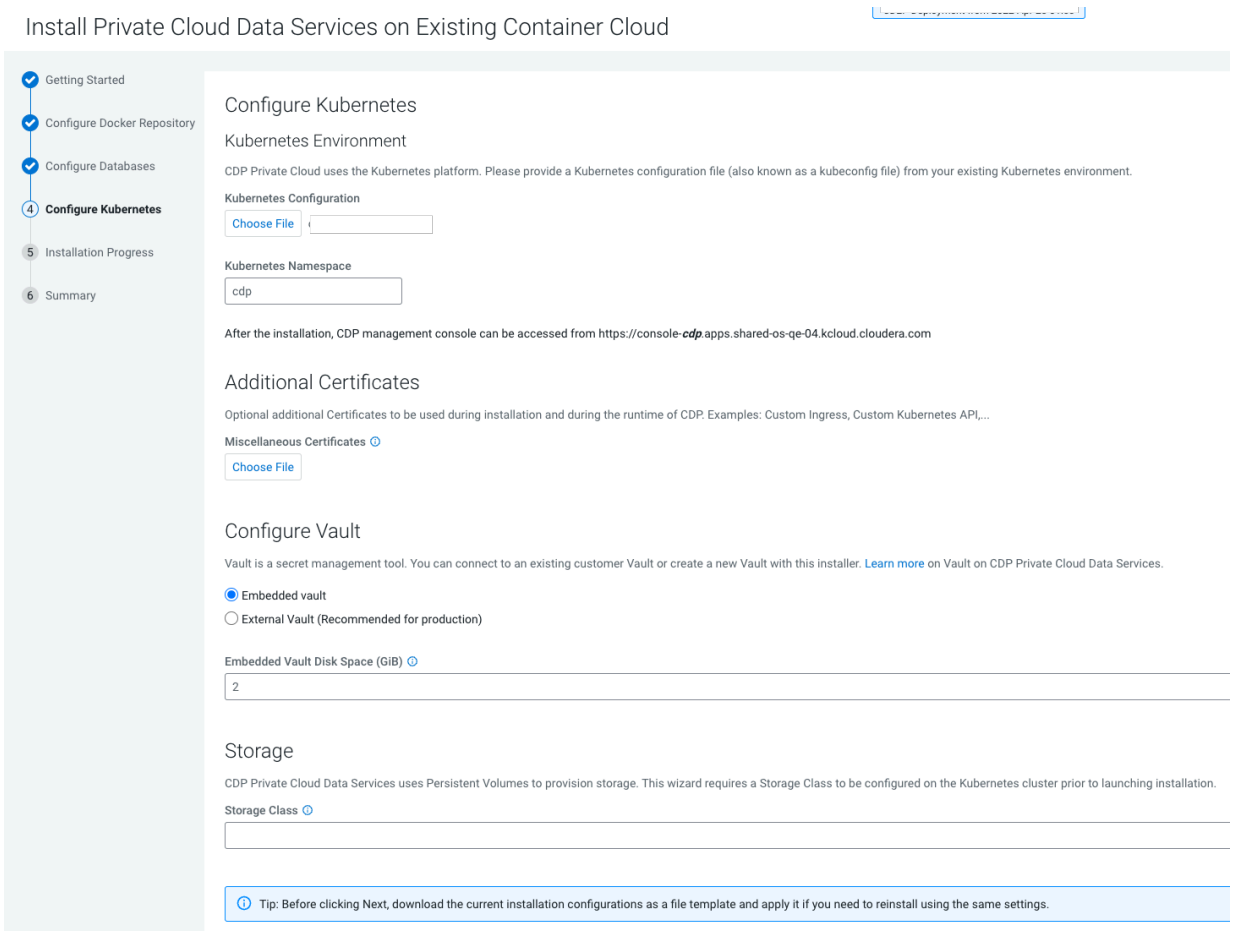


   To use an existing database, select Use existing databases (Recommended for production) option.

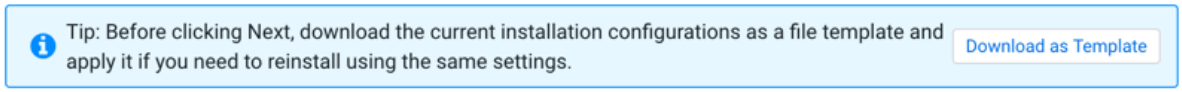


   Click Continue.

6. In 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.
   e) 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.

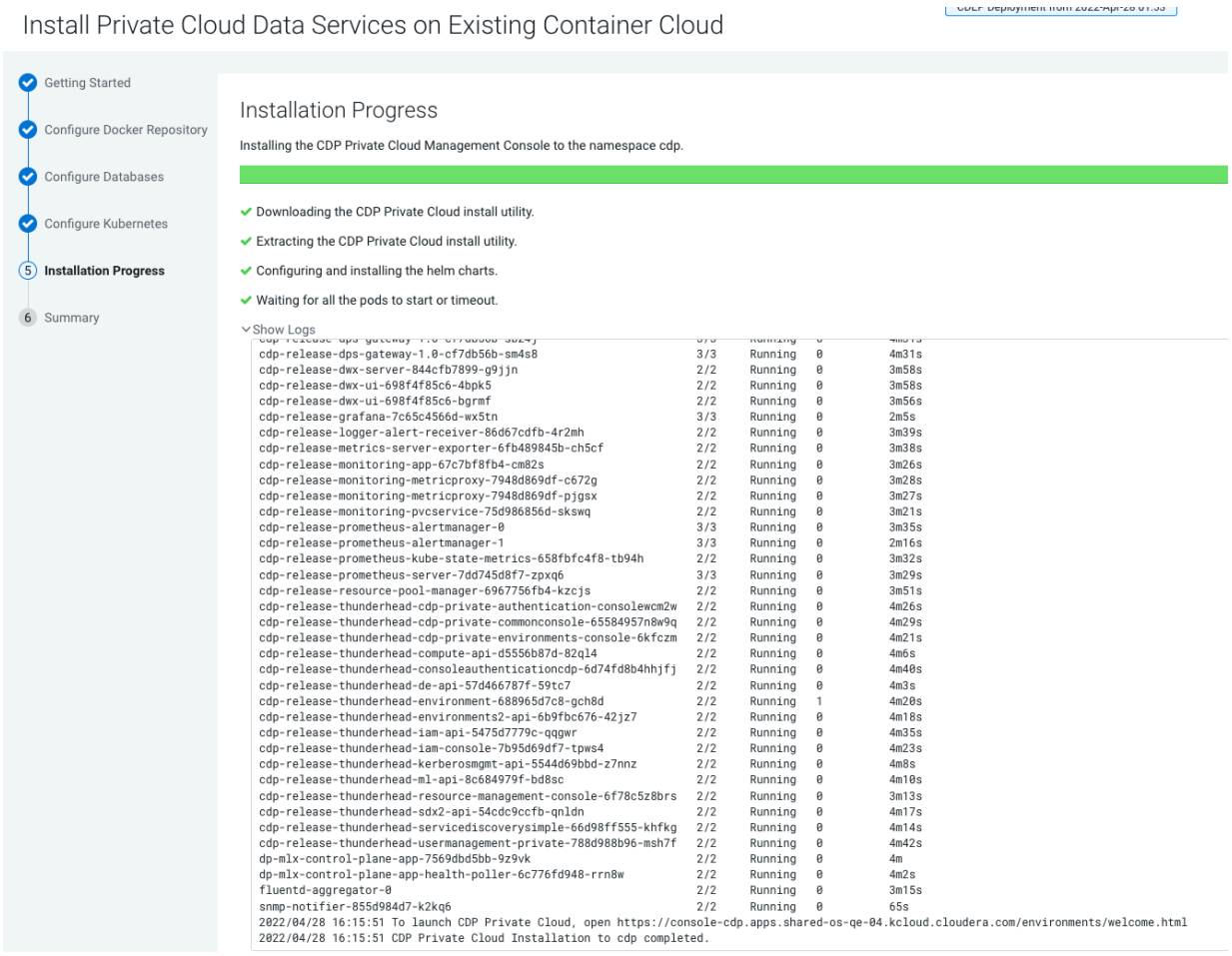


7. If you want to use this installation configuration again to install CDP Private Cloud, you have the option to download this information as a template.

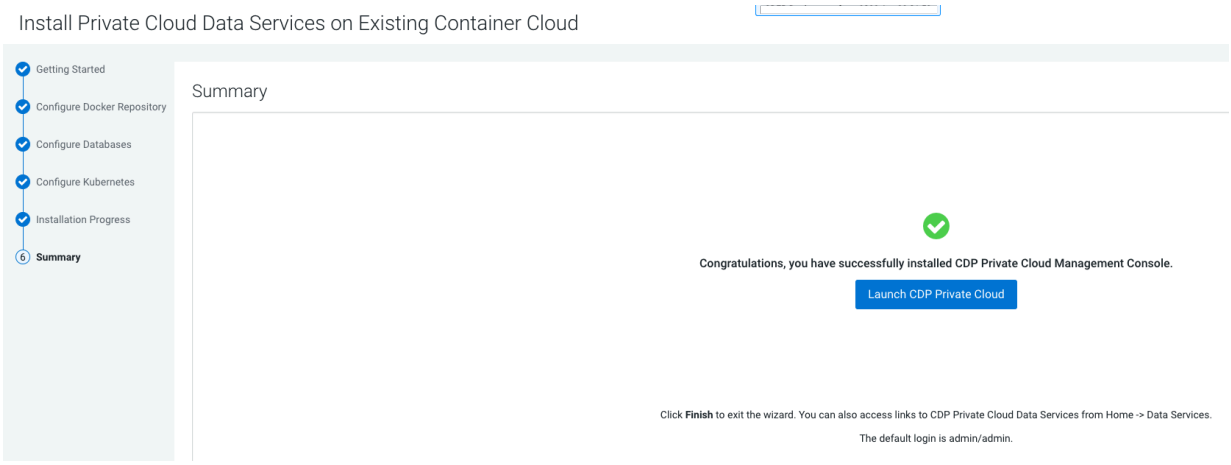


The template file is a text file that contains the database and vault information that you entered for this installation. This template is useful if you will be installing Private Cloud again with the same databases, as the template will populate the fields here automatically. Note that the user password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

8. The Installation Progress page appears. Click Continue.



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



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

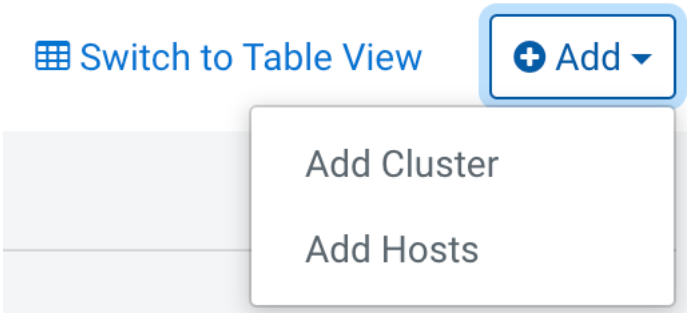
- Ensure that your Kubernetes kubeconfig has permissions to create Kubernetes namespaces.
- You require persistent storage classes defined in your OpenShift cluster. Storage classes can be defined by OpenShift cluster administrators.
- Only TLS-enabled custom Docker Registry is supported. Ensure that 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.

About this task

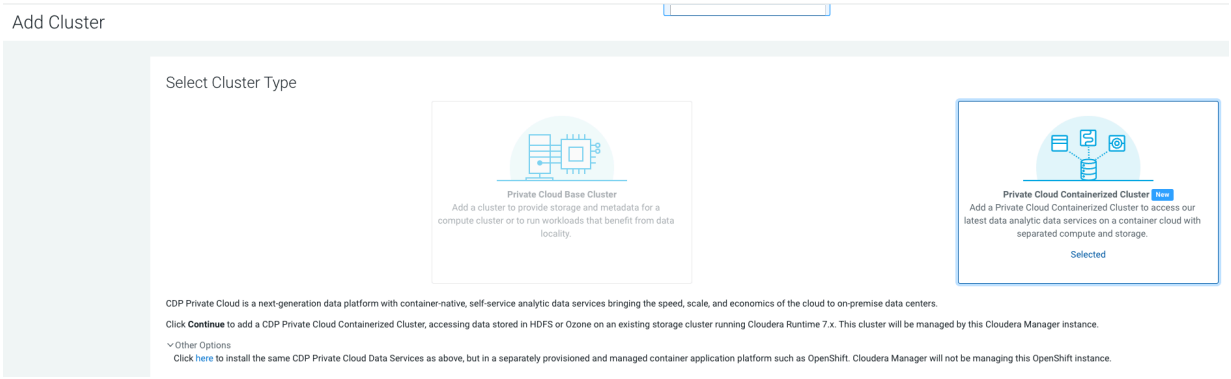
If this Cloudera Manager instance or your Kubernetes cluster does not have connectivity to https://archive.cloudera.com/p/cdp-pvc-ds/, you need to mirror the Cloudera archive URL using a local HTTP server.

Procedure

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

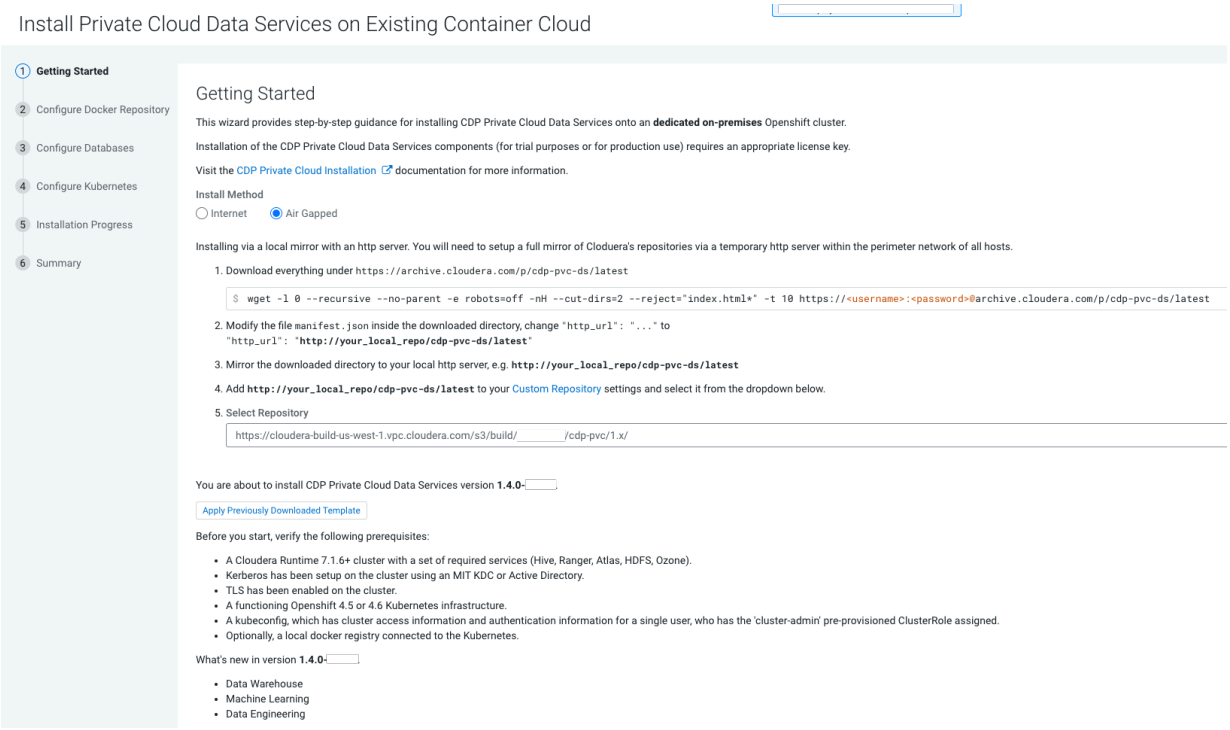


2. In the Select Cluster Type page, select the cluster type as Private Cloud Containerized Cluster and under Other Options dropdown, click here to install CDP Private Cloud Data Services. click Continue.



3. On the Getting Started page of the installation wizard, select Air Gapped as the Install Method. If you wish to use a custom repository link provided to you by Cloudera, click Custom Repository. On the Getting Started page of the installation wizard, you can select the repository that contains the installer. The Select Repository field

contains the CDP Private Cloud Data Services repository download link. To use a custom repository link provided to you by Cloudera, click Custom Repository. Click Continue.



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.

4. In 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.

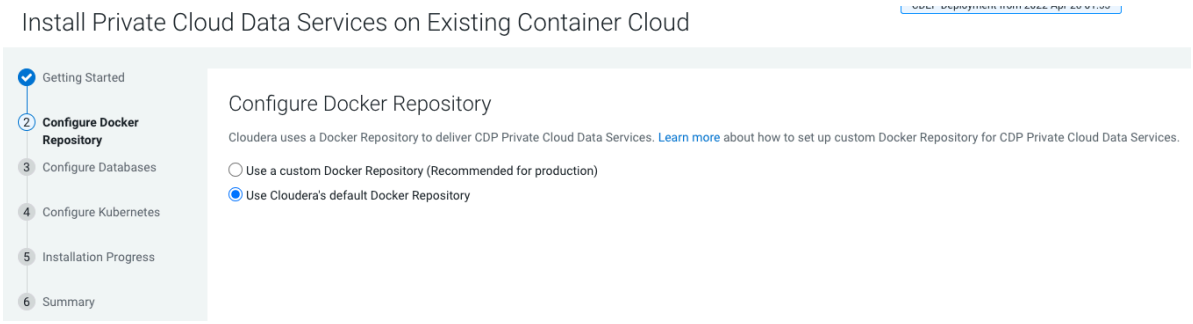
a) Click Generate the copy-docker script on the wizard or download the script file.
b) Log in to your custom Docker Registry and run the script using the following commands.

```
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.

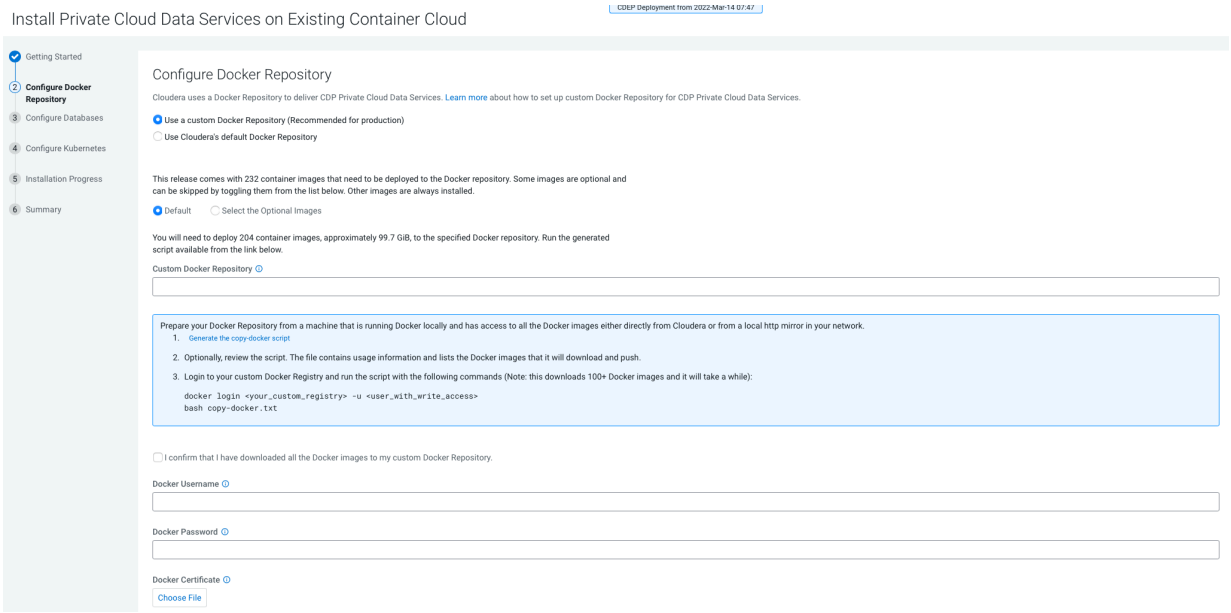
c) Enter your Docker user name and password.
d) Click Choose File to upload your Docker certificate.
e) Click Continue.

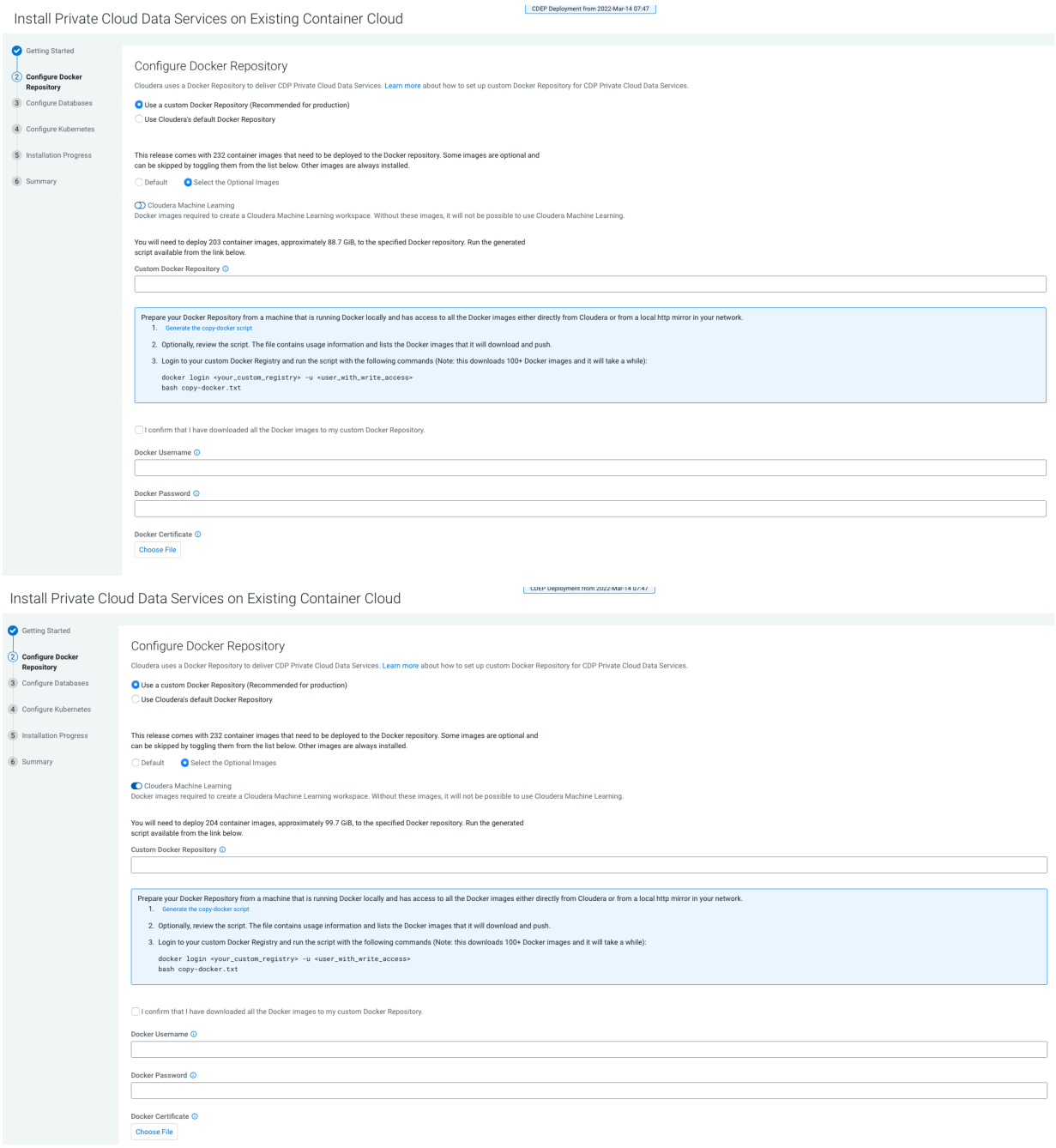


If you select Use an embedded Docker Repository option, then you can download and deploy the Data Services that you need for your cluster.

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.





Click Continue.

5. In the Configure Databases page, follow the instructions in the wizard to use your external existing databases with CDP Private Cloud.

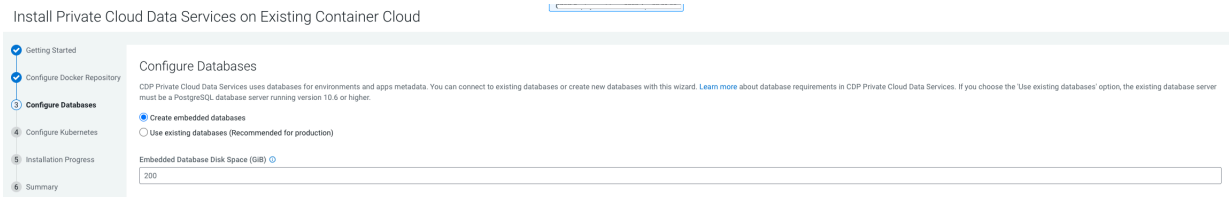
   For production environments, Cloudera recommends that you use databases that you have previously created. These databases must all be on the same host and that host must be a PostgreSQL database server running version 10 or 12.

   Ensure that you have selected the Use TLS for Connections Between the Control Plane and the Database option if you have plans to use Cloudera Data Warehouse (CDW) Private Cloud. Enabling the base cluster PostgreSQL

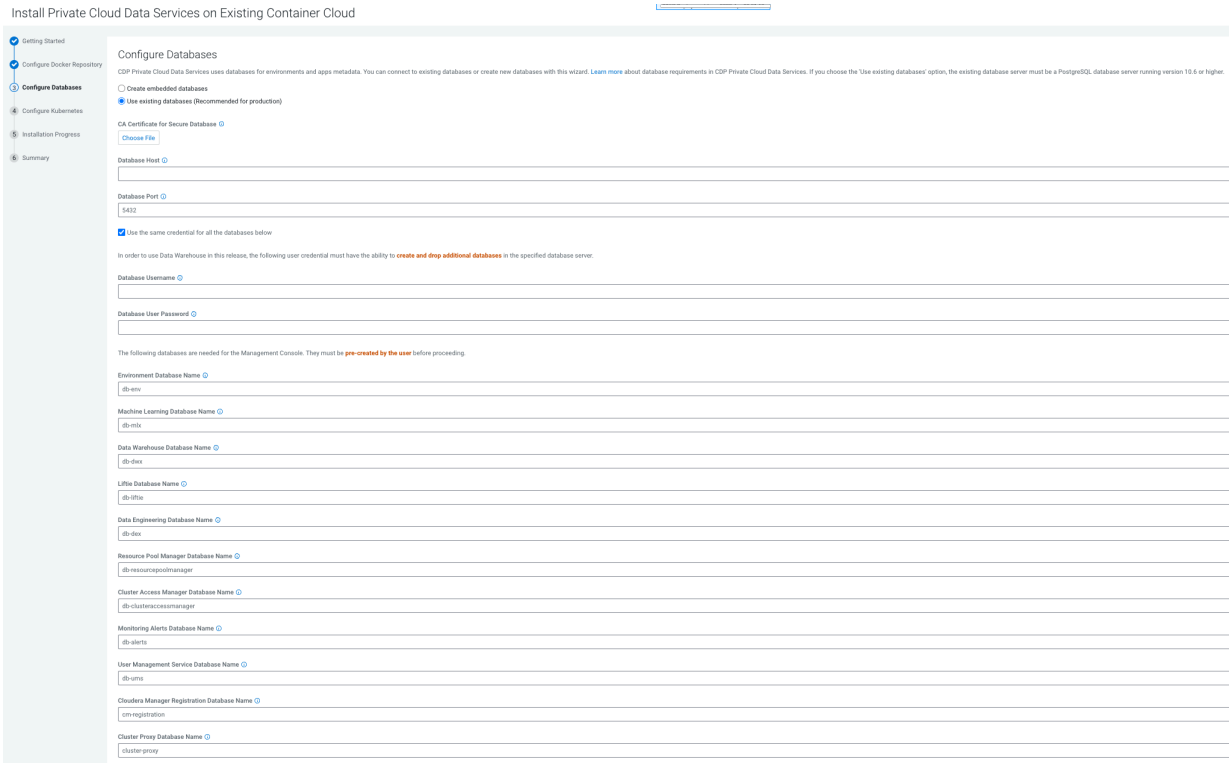
database to use an SSL connection to encrypt client-server communication is a requirement for CDW in CDP Private Cloud.

The databases must be created before you can proceed with the installation, and the database user must have permissions to create and drop databases on the server.

To create an embedded database, select Create embedded databases option.

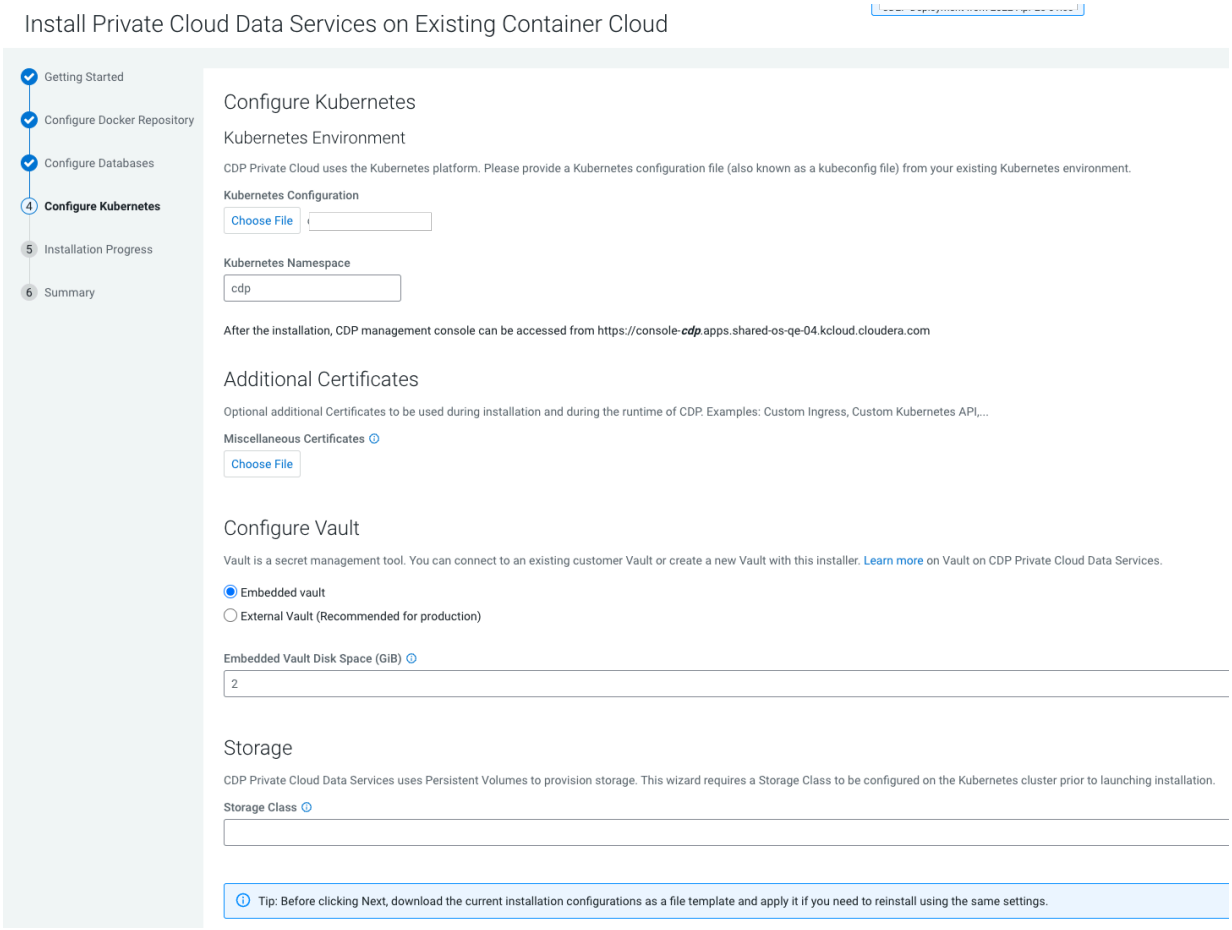


To use an existing database, select Use existing databases (Recommended for production) option.

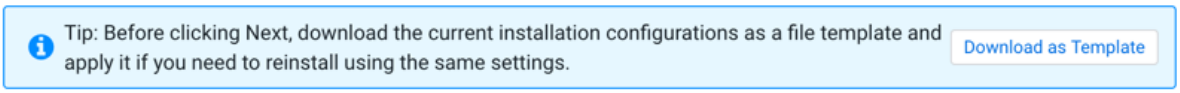


Click Next.

6. In 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.

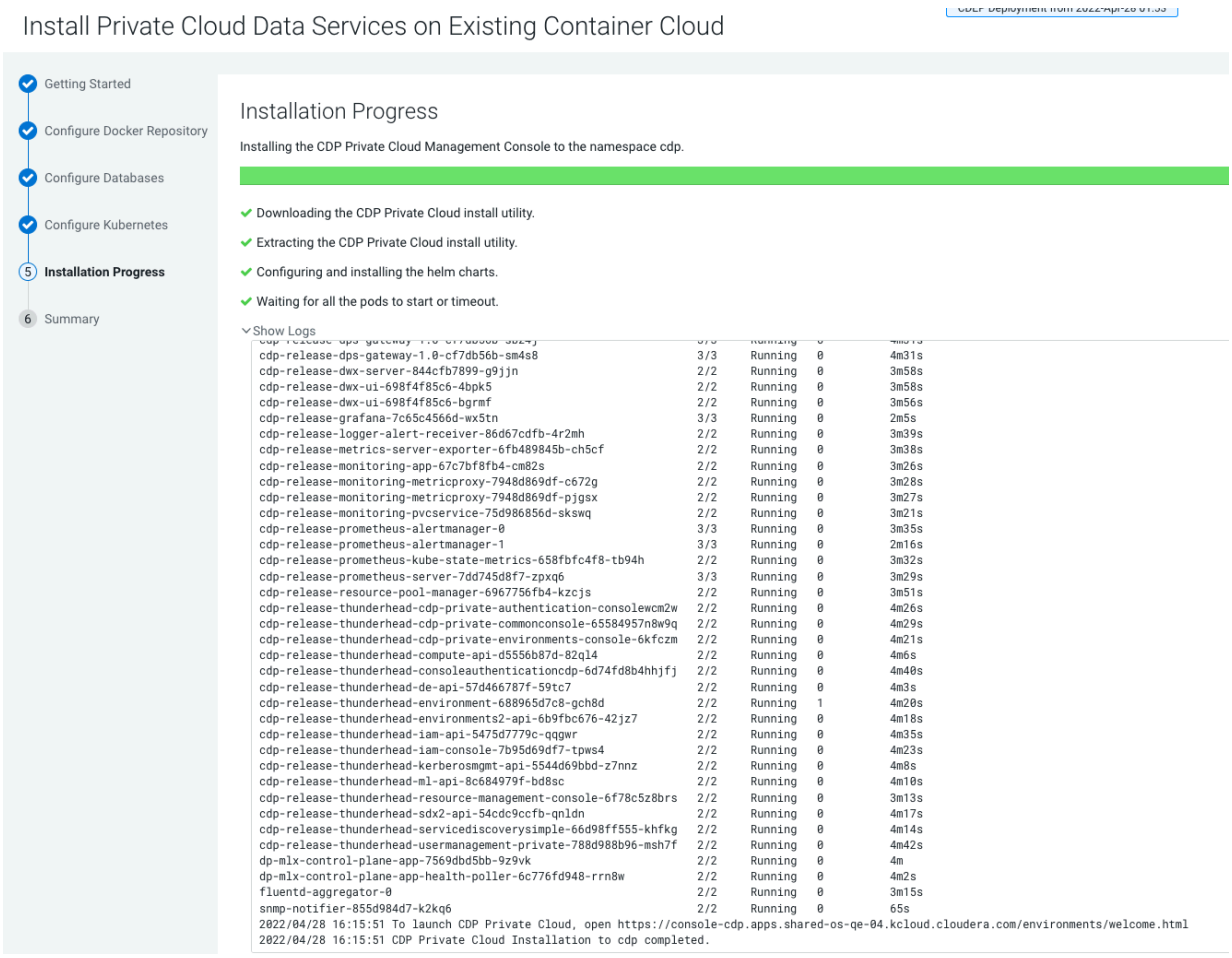


7. If you want to use this installation configuration again to install CDP Private Cloud, you have the option to download this information as a template.

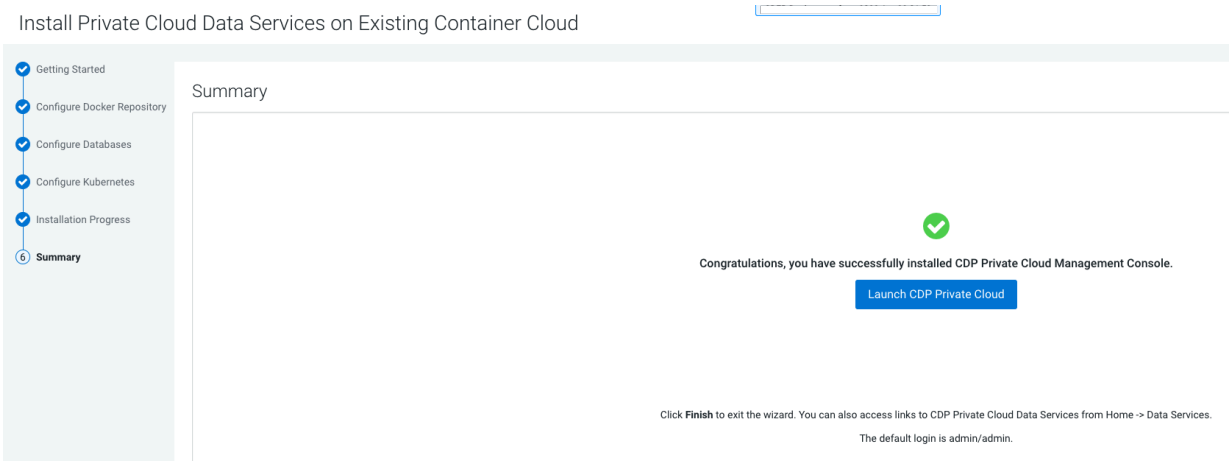


The template file is a text file that contains the database and vault information that you entered for this installation. This template is useful if you will be installing Private Cloud again with the same databases, as the template will populate the fields here automatically. Note that the user password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

8. The Installation Progress page appears. Click Continue.



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



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

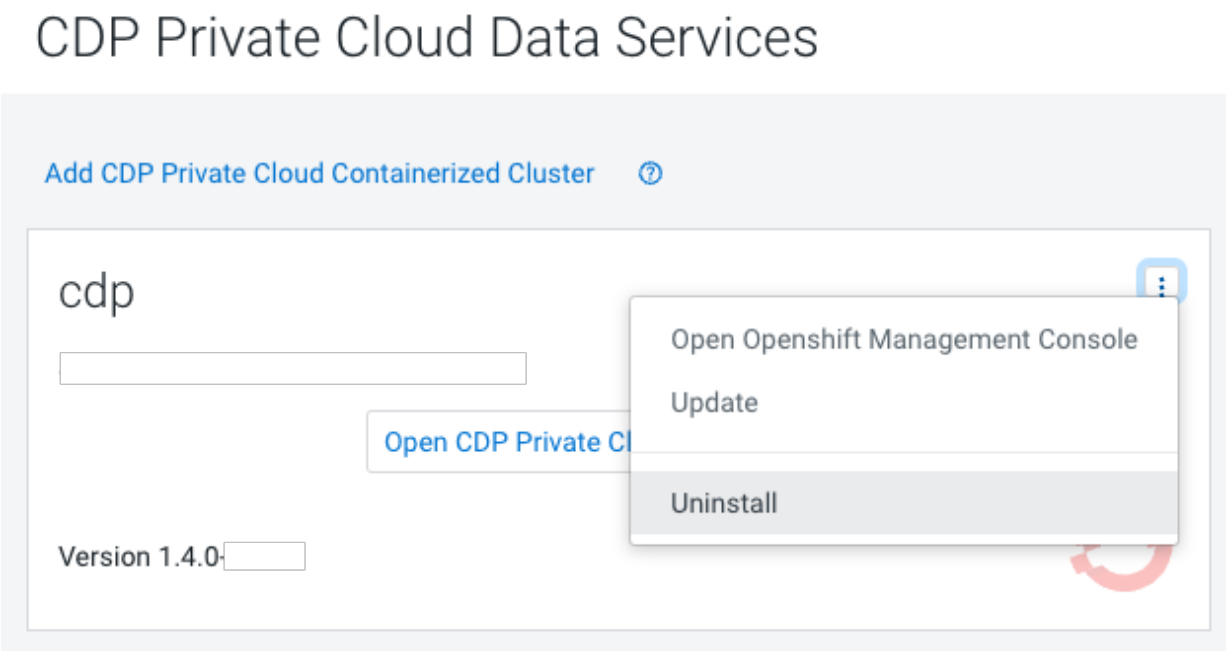
You can uninstall CDP Private Cloud Data Services from your CDP Private Cloud Base Cloudera Manager.

Before you begin

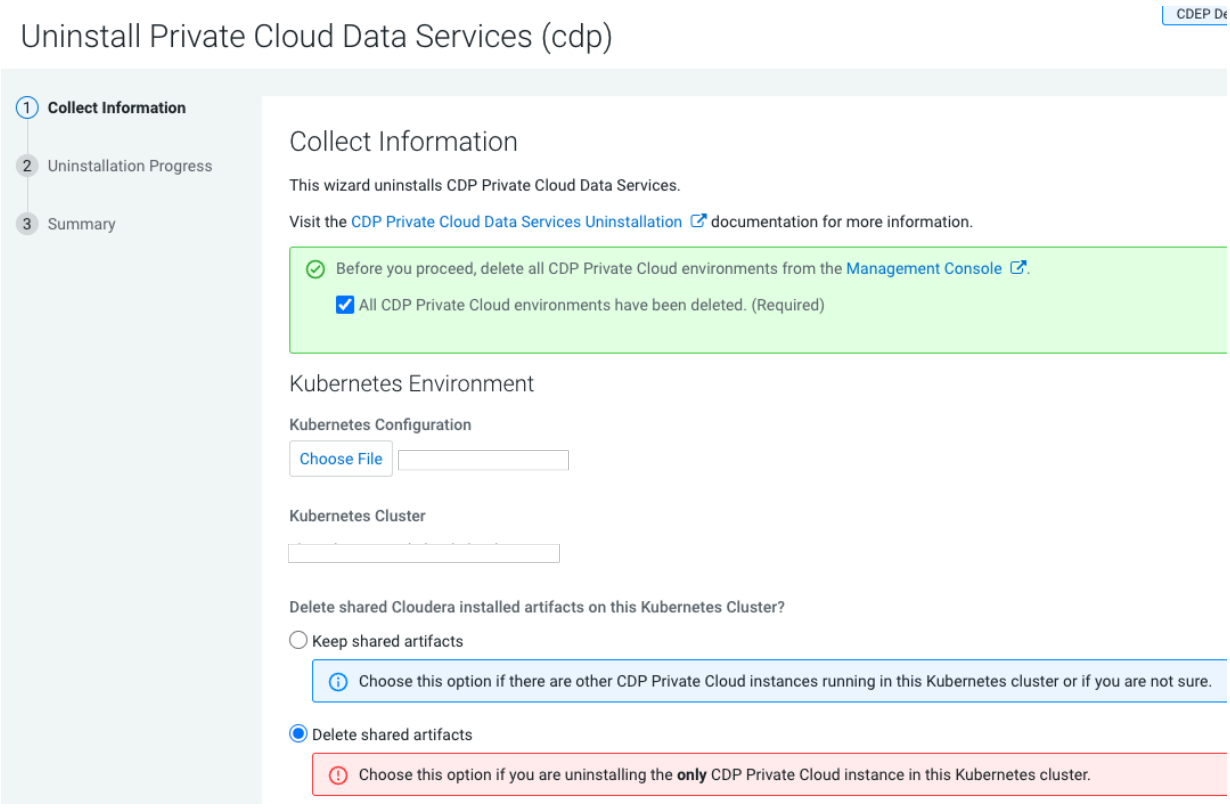
Before you uninstall CDP Private Cloud Data Services, ensure that you have deleted all the CDP Private Cloud environments registered in your CDP Private Cloud Data Services. You can delete your registered environments using Management Console.

Procedure

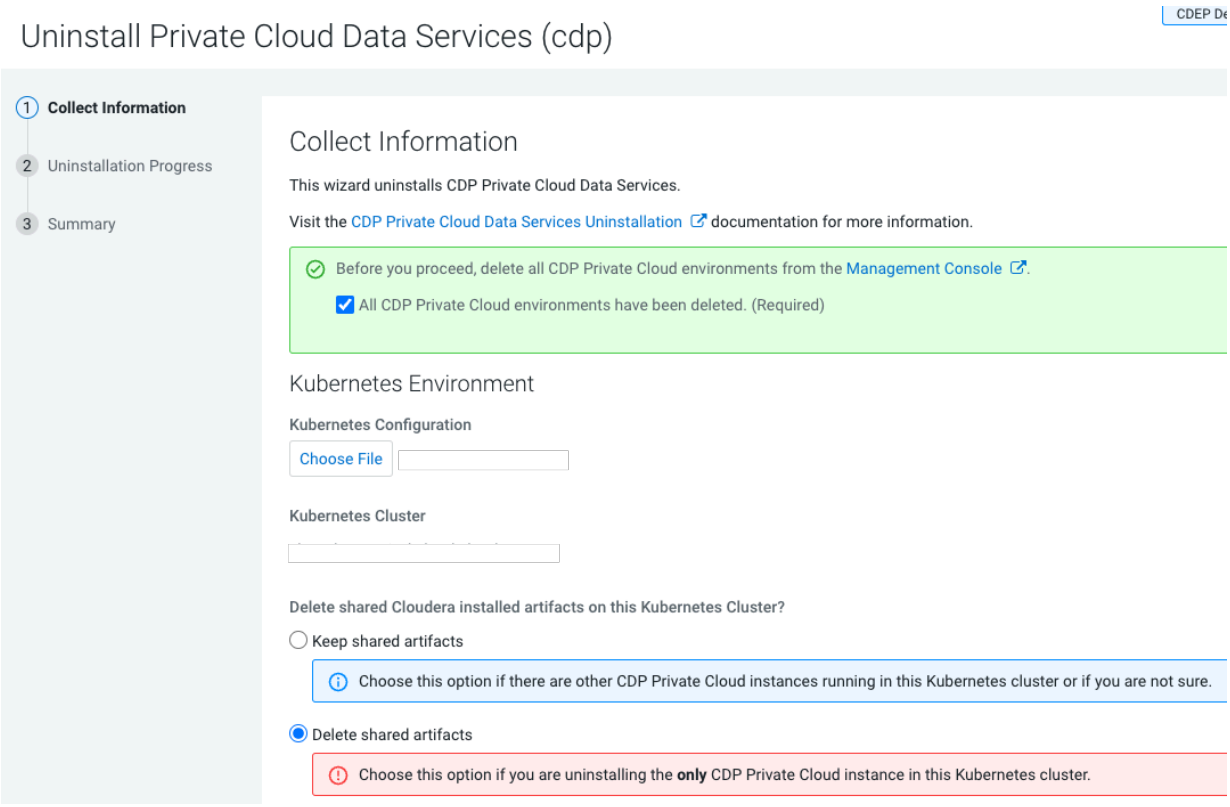
1. In Cloudera Manager, navigate to CDP Private Cloud Data Services and click . Click Uninstall.



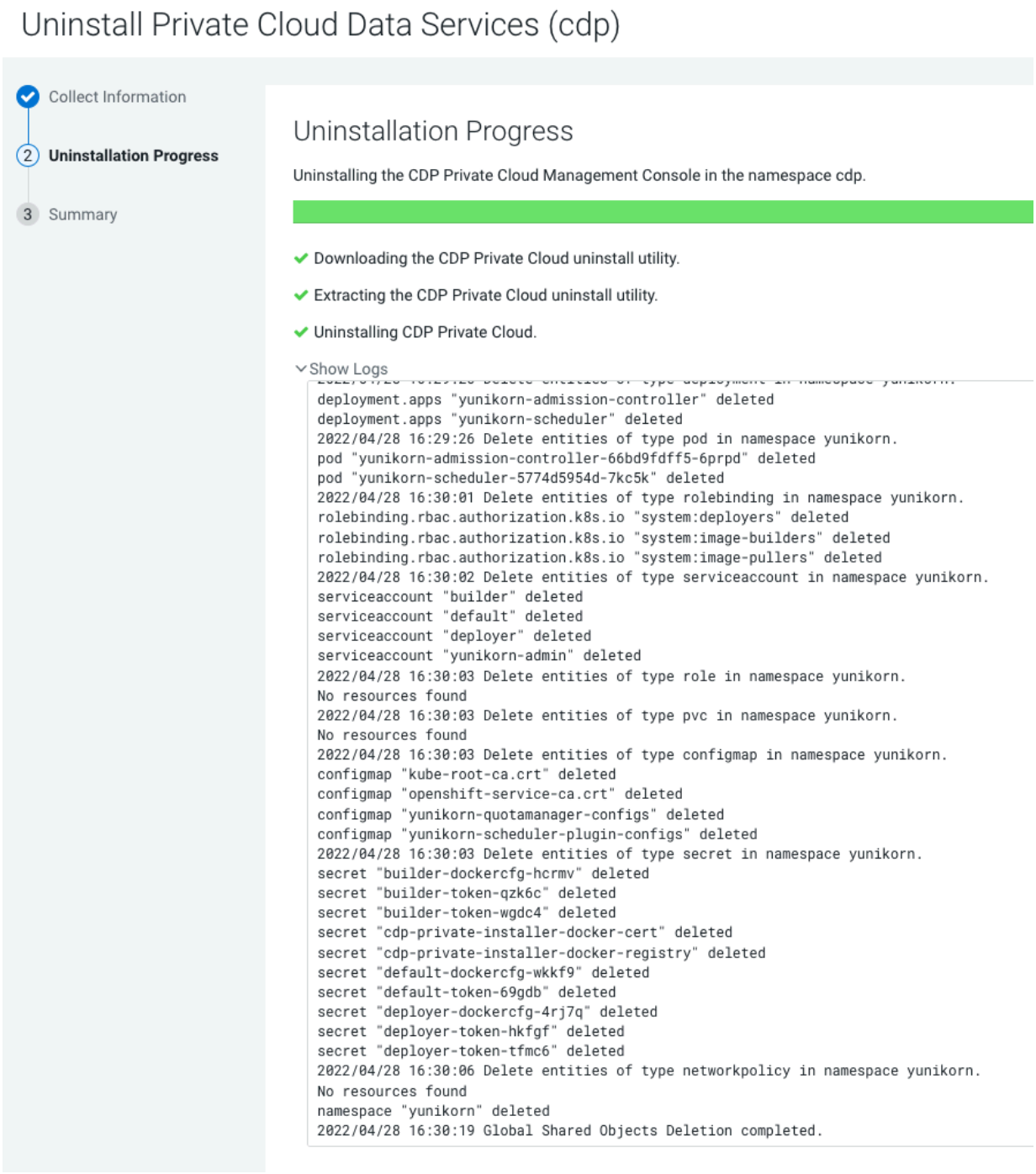
2. The Collect Information page appears. You must select the checkbox associated with your CDP Private Cloud Environments. Click Choose File to upload your kubeconfig file associated with your Kubernetes cluster.



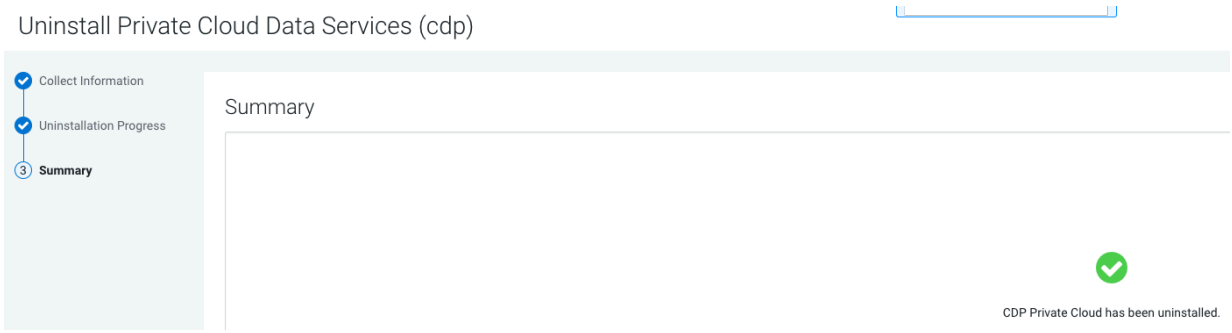
3. Select Keep shared artifacts if you have other CDP Private Cloud Data Services instances running in your Kubernetes cluster, or select Delete shared artifacts to remove any cluster global security policies or objects associated with this Kubernetes namespace.



4. Click Continue to complete the process.



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



Upgrading

Update from 1.3.4 or 1.4.0-H1 to 1.4.1 (OCP)

You can update your existing CDP Private Cloud Data Services 1.3.4 or 1.4.0-H1 to 1.4.1 without requiring an uninstall. After the update is complete, you can then upgrade the underlying OpenShift Container Platform to 4.7.x or 4.8.x.

About this task

If you are upgrading the OCP version, while the CDE service is enabled, it can cause database corruption in the embedded MySQL database used for CDE. Hence, only if you are using CDE Data Services, you must perform certain additional steps before starting the OCP version upgrade. Otherwise, you can directly *update your CDP Private Cloud Data Services*.

Before you begin

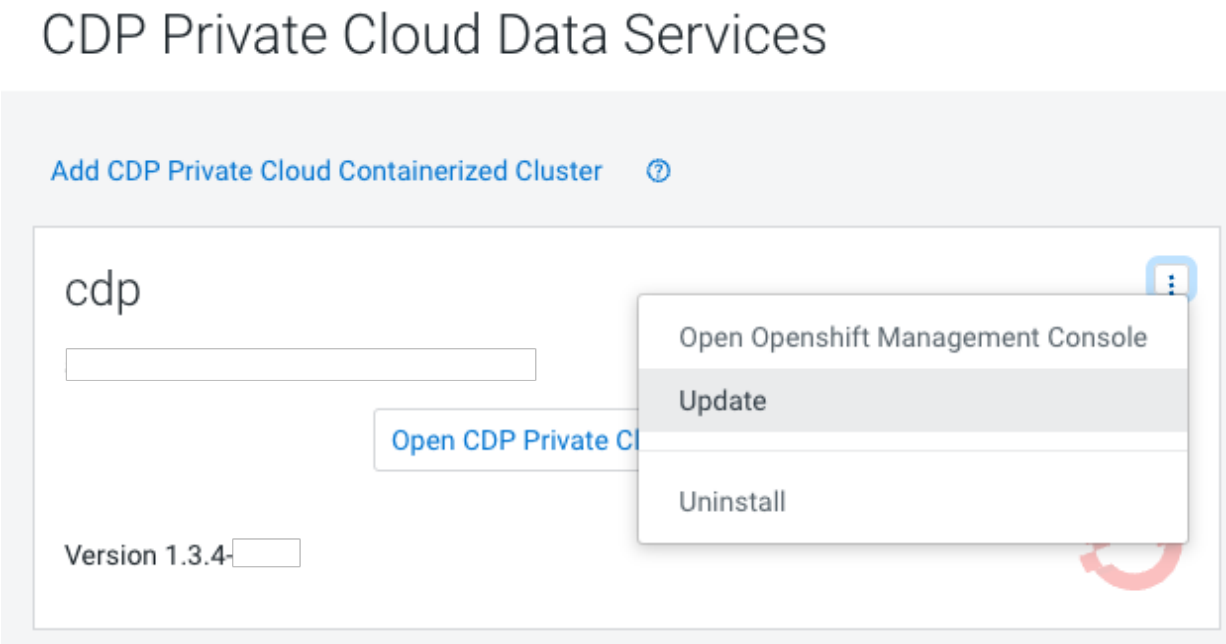
Ensure that you have the following before you update:

- Ensure that this kubeconfig has permissions to create Kubernetes namespaces.
- Back up all the external databases used by CDP Private Cloud Data Services.
- CDP Private Cloud Base 7.1.8 or 7.1.7 SP1 with a Data Lake cluster.
- Cloudera Manager 7.8.1
- Single instance of CDP Private Cloud Data Services 1.3.4 or 1.4.1 installed on your OCP cluster.
- One or more environments registered in CDP Private Cloud Data Services.
- One of the registered environment has one or more Cloudera Data Warehouse (CDW) or Cloudera Machine Learning (CML) experience workspaces.
- Kubeconfig file for the OCP 4.7 or 4.8 cluster
- Access to the Cloudera Private Cloud repositories (archive.cloudera.com)
- Administrator access to OCP 4.7 or 4.8 and Privileged access to your external Vault

Note: There is no direct upgrade from 1.0, 1.1, 1.2, 1.3.1, or 1.3.2 to 1.4.1.

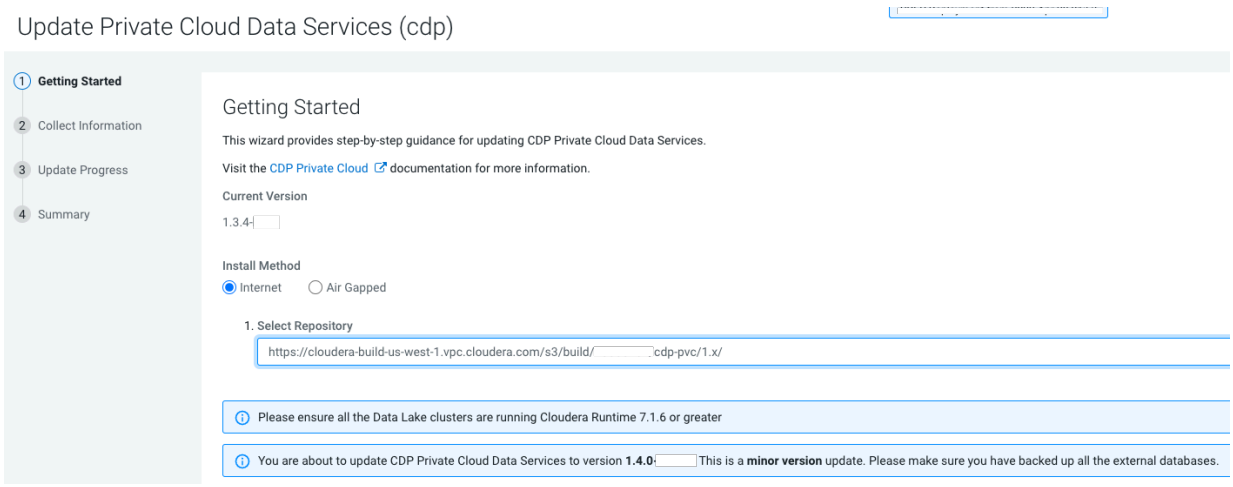
Procedure

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

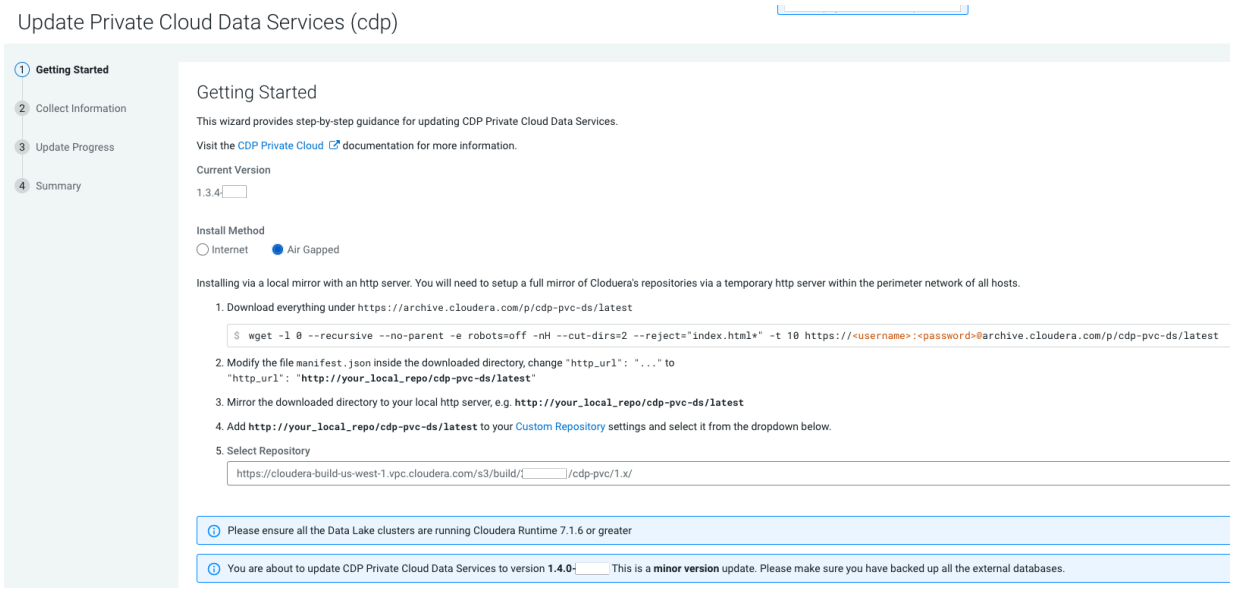


2. On the Getting Started page, you can select the Install method - Air Gapped or Internet and proceed.

   Internet install method

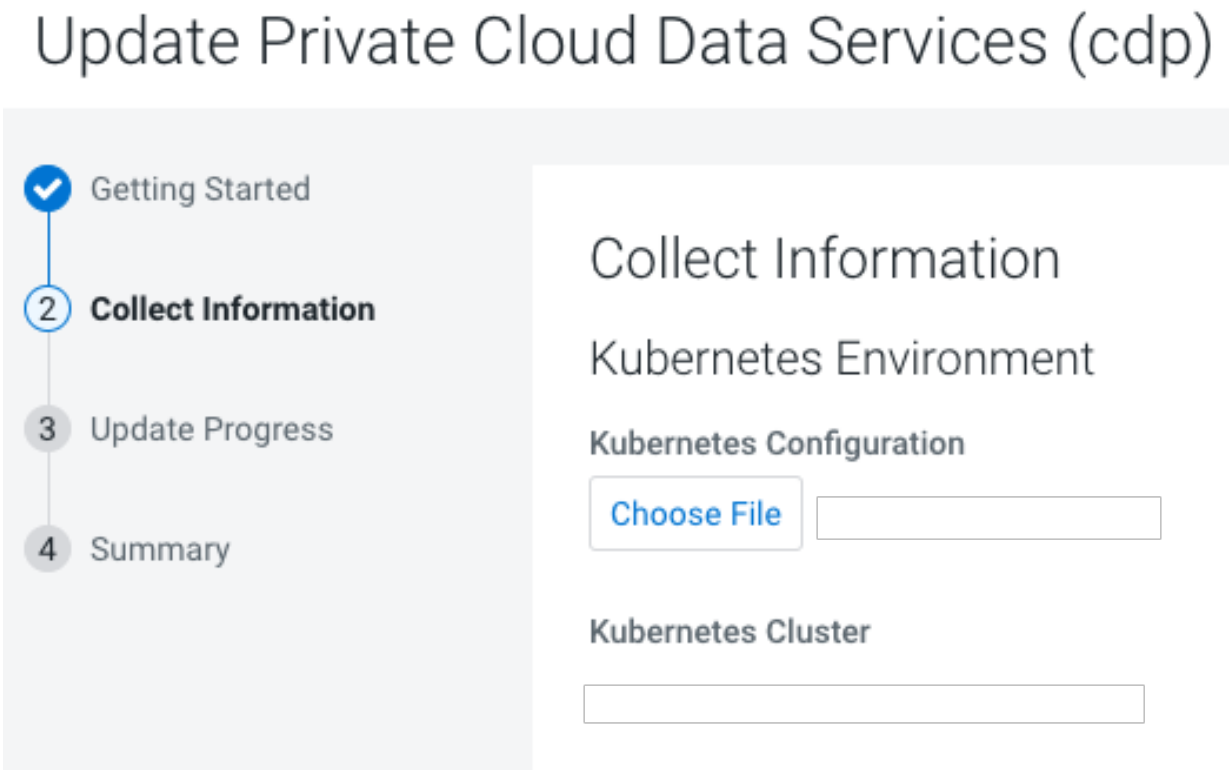


   Air Gapped install method

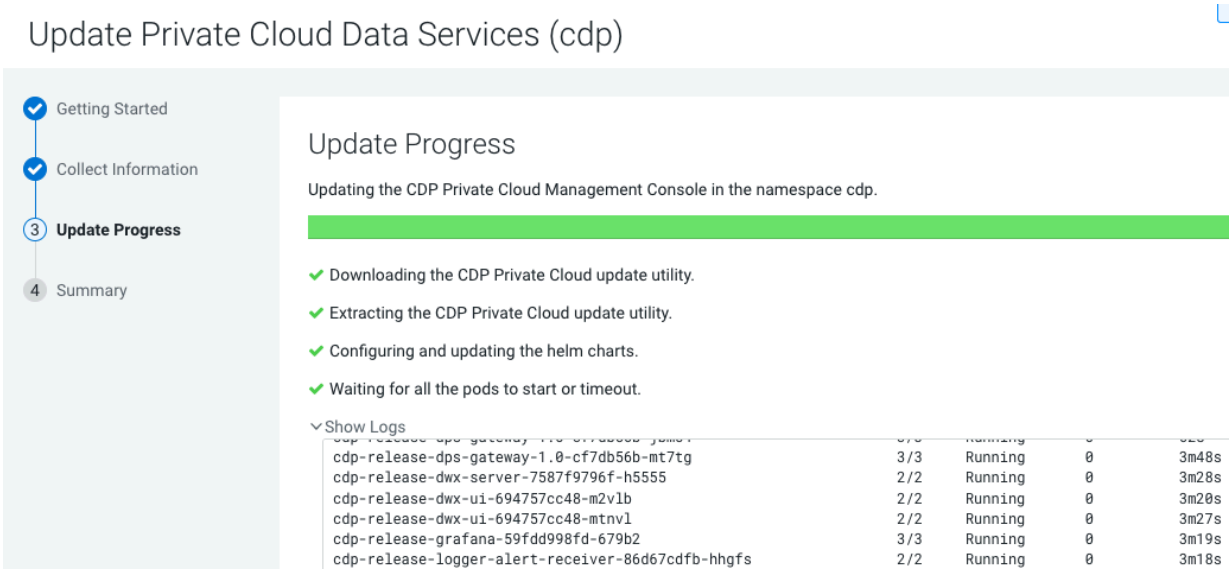


   Click Continue.

3. On the Collect Information page, upload a Kubernetes configuration (kubeconfig) file from your existing environment. You can obtain this file from your OpenShift Container Platform administrator. Click Continue.



4. On the Update Progress page, you can see the progress of your update. Click Continue.



Important:

During the "Upgrade Control Plane" step of the CDP upgrade process, the grafana pod can get stuck in the terminating sate. This usually means that all other Control Plane pods are in the running state, but for Grafana, there is one pod that is in running state and there is one pod that is stuck in terminating state. The terminating pod has the following message:

```
containers with incomplete status: [multilog-init grafana-sc-datasou
rces]
```

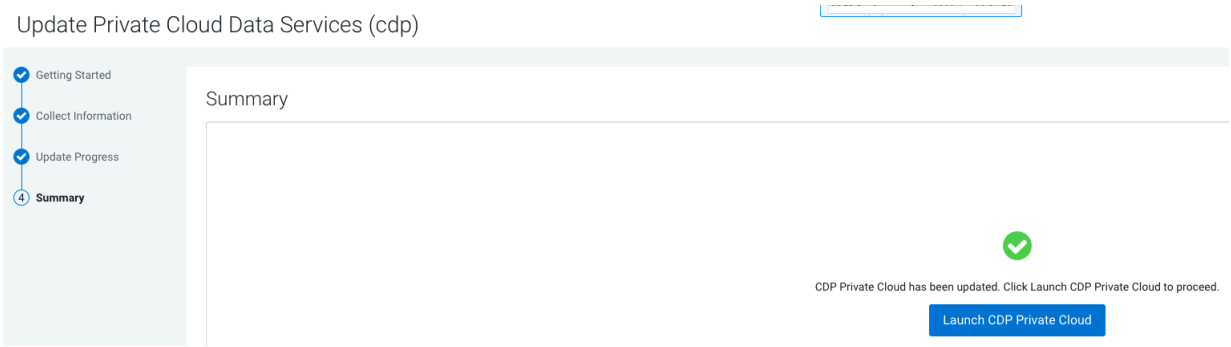
If you search for the terminating pod id in the kubelet log on the host, the following error message can be found:

```
E0531 2209 kuberuntime_sandbox.go:70] CreatePodSandbox for pod "<pod
 id>" failed: rpc error: code = Unknown desc = error reading contain
er (probably exited) json message: EOF
```

If there is a granfa pod stuck in terminating state, run the following command on the ECS Server host:

```
<grafana-pod-id> --force --grace-period=0
```

5. After the update is complete, the Summary page appears. You can now Launch CDP Private Cloud from here.



Or you can navigate to the CDP Private Cloud Data Services page and click Open CDP Private Cloud Data Services.

CDP Private Cloud Data Services opens up in a new window.

6. After the update is complete, upgrade the underlying OpenShift Container Platform from 4.5.x or 4.6.x to 4.7.x or 4.8.x.

Preparing for CDP Private Cloud Data Services update for CDE

Upgrading the OpenShift Container Platform (OCP) version while CDE service is enabled, can cause database corruption in the embedded MySQL database used for CDE. Follow the below steps before starting the OCP version upgrade.

Procedure

1. Stop running jobs and pause scheduled workloads
   a) Kill all the running Spark jobs in the CDE virtual clusters under all CDE services or wait for them to complete.
   b) Pause all Airflow jobs and scheduled Spark jobs.
2. Identifying the CDE Namespace
   a) Navigate to the Cloudera Data Engineering Overview page by clicking the Data Engineering tile in the Cloudera Data Platform (CDP) management console.
   b) In the CDE Services column, click Service Details for the CDE service.
   c) Note the Cluster ID shown in the page. For example, if the Cluster ID is *cluster-abcd1234*, then the CDE Namespace is *dex-base-abcd1234*.
   d) Use this CDE Namespace (in the above example, it is *dex-base-abcd1234*) in the following instructions to run kubernetes commands.
3. Scale down CDE embedded database

   Access the OpenShift cluster with OpenShift CLI or Kubernetes CLI, and scale down the CDE embedded database statefulset to 0 with the following command:

   OpenShift CLI

   ```
   oc scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace> --r
   eplicas 0
   ```

   Kubernetes CLI

   ```
   kubectl scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace>
    --replicas 0
   ```

Updating CDP Private Cloud Data Services

You can update your existing CDP Private Cloud Data Services 1.3.4 or 1.4.0 to 1.4.1 without requiring an uninstall. Once the update is complete, you can then upgrade the underlying OpenShift Container Platform from 4.5.x or 4.6.x to 4.7.x or 4.8.x.

Before you begin

Ensure that you have the following before you update:

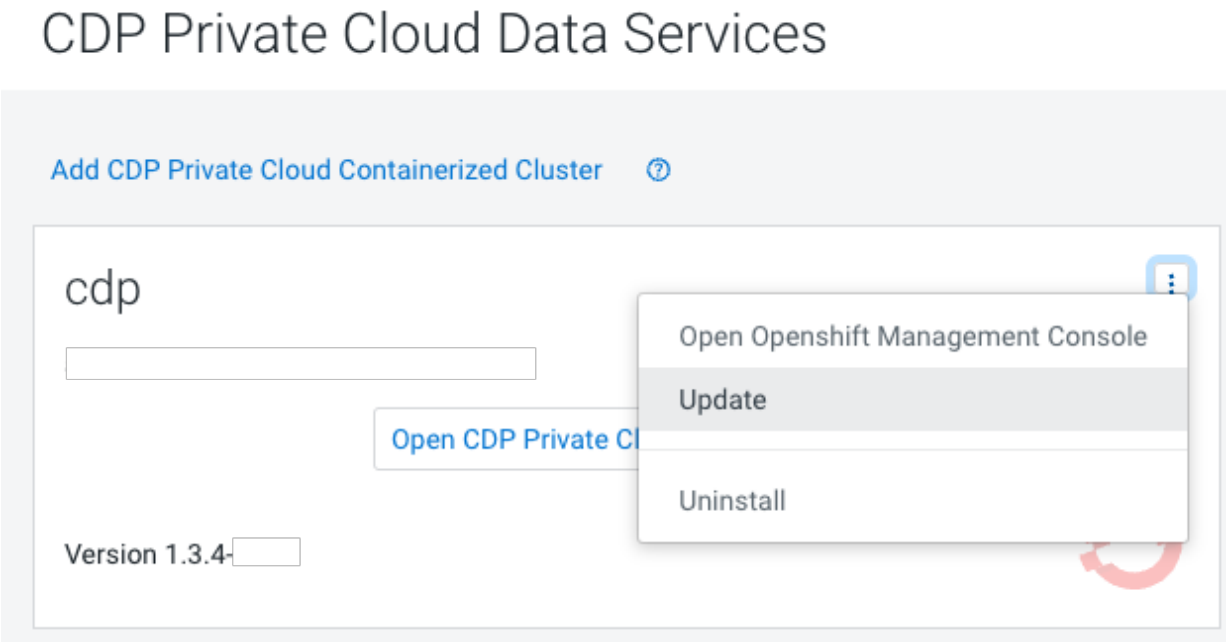
- Ensure that this kubeconfig has permissions to create Kubernetes namespaces.
- Back up all the external databases used by CDP Private Cloud Data Services.
- CDP Private Cloud Base 7.1.8, or 7.1.7 SP1 with a Data Lake cluster.
- Cloudera Manager 7.8.1
- Single instance of CDP Private Cloud Data Services installed on your OCP cluster
- One or more environments registered in CDP Private Cloud Data Services.
- One of the registered environment has one or more Cloudera Data Warehouse (CDW) or Cloudera Machine Learning (CML) experience workspaces
- Kubeconfig file for the OCP 4.7 or 4.8 cluster
- Access to the Cloudera Private Cloud repositories (archive.cloudera.com)

- Administrator access to OCP 4.7 or 4.8 and Privileged access to your external Vault

Note: There is no direct upgrade from 1.0, 1.1, 1.2, 1.3.1, or 1.3.2 to 1.4.1.

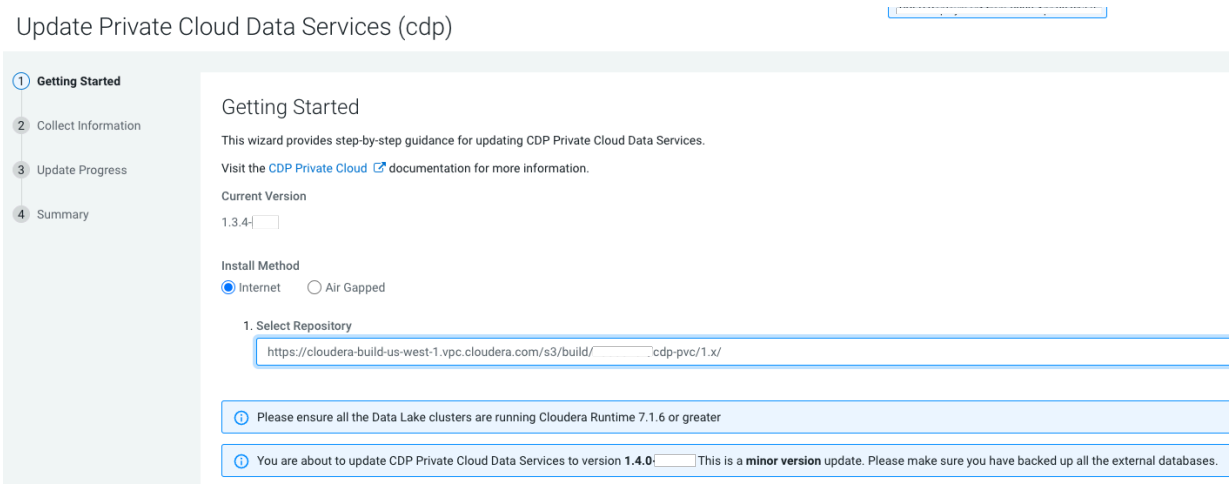
Procedure

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

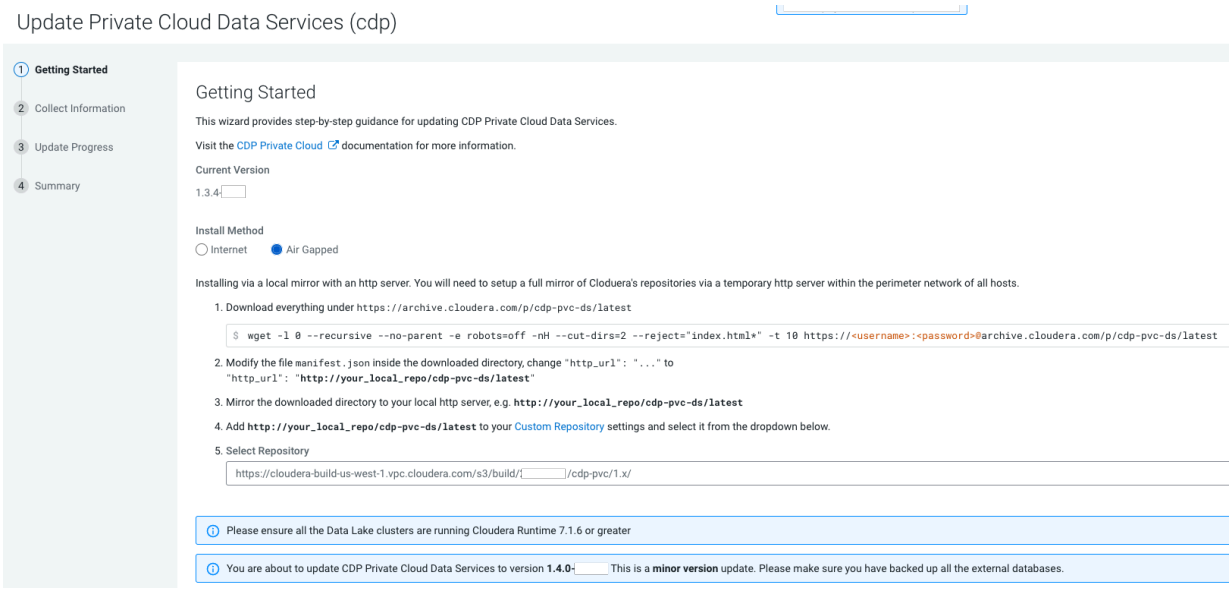


2. On the Getting Started page, you can select the Install method - Air Gapped or Internet and proceed.

Internet install method

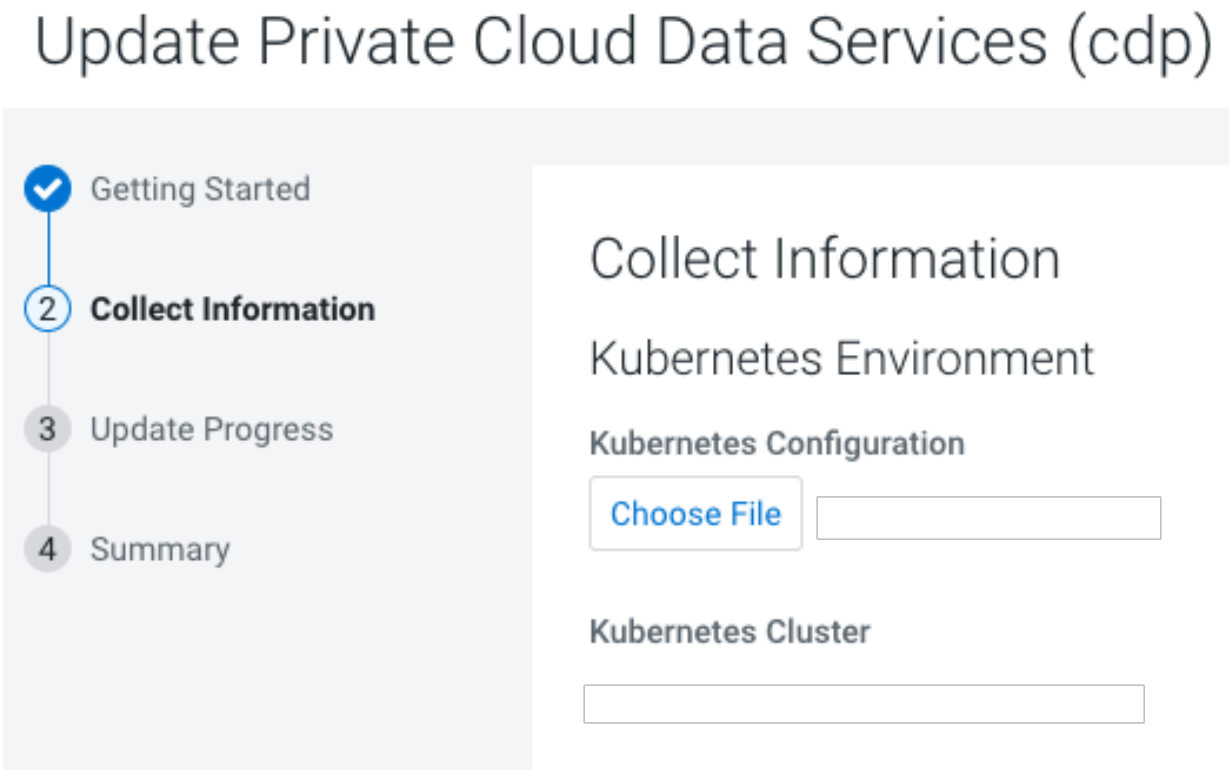


Air Gapped install method

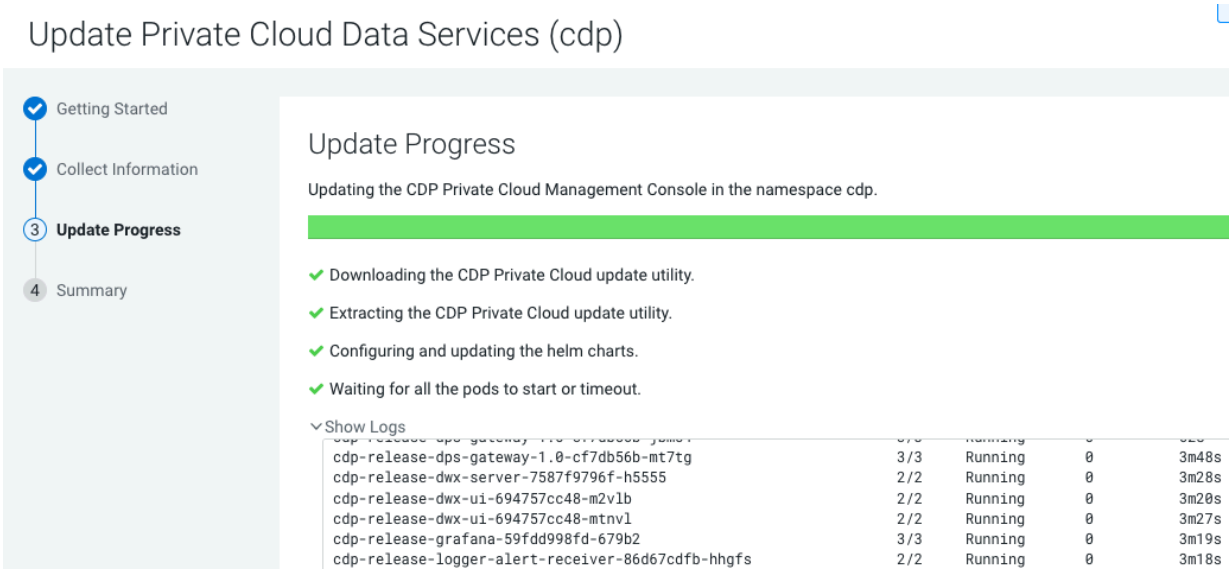


Click Continue.

3. On the Collect Information page, upload a Kubernetes configuration (kubeconfig) file from your existing environment. You can obtain this file from your OpenShift Container Platform administrator. Click Continue.



4. On the Update Progress page, you can see the progress of your update. Click Continue.



Important:

During the "Upgrade Control Plane" step of the CDP upgrade process, the grafana pod can get stuck in the terminating sate. This usually means that all other Control Plane pods are in the running state, but for grafana, there is one pod that is in running state and there is one pod that is stuck in terminating state. The terminating pod has the following message:

```
containers with incomplete status: [multilog-init grafana-sc-datasou
rces]
```

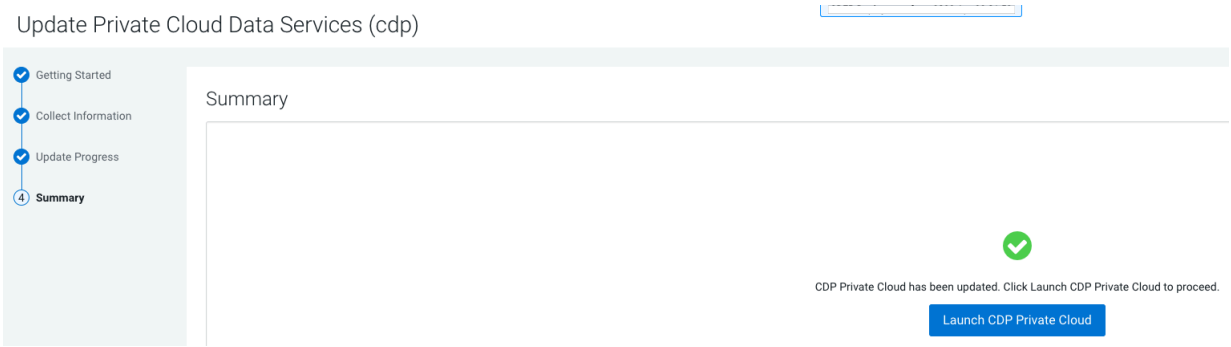
If you search for the terminating pod id in the kubelet log on the host, the following error message can be found:

```
E0531 2209 kuberuntime_sandbox.go:70] CreatePodSandbox for pod "<pod
 id>" failed: rpc error: code = Unknown desc = error reading contain
er (probably exited) json message: EOF
```

If there is a granfa pod stuck in terminating state, run the following command on the ECS Server host:

```
<grafana-pod-id> --force --grace-period=0
```

5. After the update is complete, the Summary page appears. You can now Launch CDP Private Cloud from here.



Or you can navigate to the CDP Private Cloud Data Services page and click Open CDP Private Cloud Data Services.

CDP Private Cloud Data Services opens up in a new window.

Completing post OCP update tasks

If you are using CDE, after you complete the OpenShift Container Platform (OCP) upgrade, ensure that the steps done before OCP upgrade are reverted.

Procedure

1. Unpause all Airflow jobs and scheduled Spark jobs that were paused before upgrade.
2. Scale back the CDE embedded database statefulset to 1.

   OpenShift CLI

   ```
   oc scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace> --r
   eplicas 1
   ```

   Kubernetes CLI

   ```
   kubectl scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace>
    --replicas 1
   ```

Recovering a corrupted CDE Embedded database

In case you did not stop the jobs and scale down CDE embedded databases but completed the upgrade of OpenShift Container Platform (OCP), there is a chance of the CDE embedded database getting corrupted which causes the virtual clusters to become inaccessible. Follow the below steps to recover the CDE embedded database.

Procedure

1. Identifying the CDE Namespace
   a) Navigate to the Cloudera Data Engineering Overview page by clicking the Data Engineering tile in the Cloudera Data Platform (CDP) management console.
   b) In the CDE Services column, click Service Details for the CDE service.
   c) Note the Cluster ID shown in the page. For example, if the Cluster ID is *cluster-abcd1234*, then the CDE Namespace is *dex-base-abcd1234*.
   d) Use this CDE Namespace (in the above example, it is *dex-base-abcd1234*) in the following instructions to run kubernetes commands.
2. Edit the dex-base-db-server-config configuration map and add the innodb_force_recovery=4 configuration in the [mysqld] section.

   OpenShift CLI

   ```
   oc scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace> --r
   eplicas 1
   ```

   Kubernetes CLI

   ```
   kubectl scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace>
    --replicas 1
   ```

   Example snippet:

   ```
   # Please edit the object below. Lines beginning with a '#' will be ignor
   ed,
   # and an empty file will abort the edit. If an error occurs while saving
    this file
   # will be reopened with the relevant failures.
   #
   apiVersion: v1
   data:
   ```

```
  my.cnf: |-
    [mysqld]
    port=3306
    default_authentication_plugin = mysql_native_password
    bind-address = 0.0.0.0
    innodb_force_recovery=4
```

3. Scale down and then back up the CDE embedded database statefulset to restart it.

   OpenShift CLI

   ```
   oc scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace> --r
   eplicas 0

   oc scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace> --re
   plicas 1
   ```

   Kubernetes CLI

   ```
   kubectl scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace>
    --replicas 0

   kubectl scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace>
    --replicas 1
   ```

   Wait for 10 minutes for the CDE embedded database to complete the recovery.

4. Edit the dex-base-db-server-config configuration map again by removing the previously added innodb_force_recovery=4 configuration under the [mysqld] section.

   OpenShift CLI

   ```
   oc edit configmap/dex-base-db-server-config --namespace <CDE Namespace>
   ```

   Kubernetes CLI

   ```
   kubectl edit configmap/dex-base-db-server-config --namespace <CDE Namesp
   ace>
   ```

5. Scale down and then back up the CDE embedded database statefulset to restart it again.

   OpenShift CLI

   ```
   oc scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace> --r
   eplicas 0

   oc scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace> --re
   plicas 1
   ```

   Kubernetes CLI

   ```
   kubectl scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace>
    --replicas 0

   kubectl scale statefulset/cdp-cde-embedded-db --namespace <CDE Namespace>
    --replicas 1
   ```

   Wait for all the CDE Virtual Clusters to be accessible. This usually takes about 10 minutes.

Post Upgrade Steps for CDP PVC 1.4.1 in OCP Clusters

Before you begin

Ensure that the Control Plane has successfully upgraded to version 1.4.1.

About this task

Perform the steps below after completing the upgrade. All commands in these steps should be run on the command line of your local machine where the kube config for the cluster has been exported. You can export the kube config from the CLI by running:

```
export KUBECONFIG=<absolute path to kube config for the OCP cluster>
```

Procedure

1. Ensure that the cluster master nodes have all the correct labels and taints. Run the following command:

   ```
   kubectl get nodes -o custom-columns=NAME:.metadata.name,TAINTS:.spec.tai
   nts,LABELS:.metadata.labels
   ```

2. Ensure that all the master nodes have the following label(s) (both key and value):

   ```
   node-role.kubernetes.io/master:
   ```

   Note that value for the label is empty

3. Ensure that all the master nodes have the following taint(s) (both key and effect):

   ```
   map[effect:NoSchedule key:node-role.kubernetes.io/master]
   ```

   Note that value for the taint is empty

4. If any of the labels and/or taints are wrong or completely missing, apply them using the following steps:

   a) Identify the names of all master nodes, for example master-01, master-02, master-03, and run the following commands:

   ```
   kubectl label nodes master-01 master-02 master-03 node-role.kubernetes.i
   o/master= --overwrite=true
   ```

   ```
   kubectl taint nodes master-01 master-02 master-03 node-role.kubernetes.i
   o/master=:NoSchedule --overwrite=true
   ```

   Alternatively, instead of supplying the names of every master node in the command, if all of your master nodes meet a certain filtering criteria, you can use it in the labeling and tainting commands. For example if all your master nodes have the label my-master-label-key=my-master-label-value then:

   ```
   kubectl label nodes --selector my-master-label-key=my-master-label-value
   ```

   ```
   node-role.kubernetes.io/master= --overwrite=true
   ```

   ```
   kubectl taint nodes --selector my-master-label-key=my-master-label-value
                       node-role.kubernetes.io/master=:NoSchedule
               --overwrite=true
   ```

Note:

You can also use oc instead of kubectl for all the above commands. Some commands may need minor modifications to run with oc as per the oc documentation.

5. Ensure that you have the cdp-cli command line tool setup to the latest version available at the time (0.9.71+), with a CDP Private profile that has adequate privileges. Ensure that your profile has:
   - Correct form factor: private
   - Correct CDP endpoint URL: base URL to your CDP Private dashboard
   - Correct access key and private key: generated using CDP Private console
   - See cdpcli · PyPI for more detailed information
   - Identify the name of this CDP Private profile for later use
6. Ensure that you have Python 3 installed and updated to a supported version.
7. Create a file named post_upgrade_hook.py with the following contents:

```
############################################################
# This script upgrades YuniKorn for CDP PVT OCP clusters. #
# Ensure that all prerequisites have been duly fulfilled. #
# Please read the full documentation before use.          #
############################################################

import subprocess
import json
import argparse
import sys
import time

parser = argparse.ArgumentParser()
parser.add_argument("-p", "--profile", default="", help="CDP Profile name
 as specified in ${HOME}/.cdp/credentials")
parser.add_argument("-e", "--endpoint", default="", help="CDP Private base
 endpoint URL")
args = parser.parse_args()

cdpProfileName = args.profile
controlPlanePublicEndpoint = args.endpoint

print('**************************')
print('**************************')
if cdpProfileName != "":
    print("CDP Private profile:", cdpProfileName)
if controlPlanePublicEndpoint != "":
    print("CDP Private base endpoint URL:", controlPlanePublicEndpoint)

def get_command(cmd_list_suffix):
    cmd_list = ['cdp', '--no-verify-tls',
                '--form-factor', 'private',
                '--output', 'json']
    if cdpProfileName != "":
        cmd_list = cmd_list + ['--profile', cdpProfileName]
    if controlPlanePublicEndpoint != "":
        cmd_list = cmd_list + ['--endpoint-url', controlPlanePublicEndpo
int]
    cmd_list = cmd_list + cmd_list_suffix
    return cmd_list

envNames, envCrns = [], []
process = subprocess.Popen(get_command(['environments',
                                        'list-environments']),
                           stdout=subprocess.PIPE,
                           stderr=subprocess.PIPE,
                           universal_newlines=True)
stdout, stderr = process.communicate()
try:
    data = json.loads(stdout)
```

```
except ValueError:
    print('While list environments: Something is wrong with output, Output
 JSON:', stdout)
    print('____ERROR__WHILE__CALLING__LIST__ENVIRONMENTS__COMMAND____', st
derr)
    sys.exit()
for en in data['environments']:
    if en['status'] == 'AVAILABLE':
        envNames.append(en['environmentName'])
        envCrns.append(en['crn'])

print('**************************')
print('**************************')
print('Environment names:', envNames)
print('Environment CRNs:', envCrns)

clusterIds, clusterCrns = [], []

for crn in envCrns:
    process = subprocess.Popen(get_command(['compute',
                                            'list-clusters',
                                            '--env-name-or-crn', crn]),
                               stdout=subprocess.PIPE,
                               stderr=subprocess.PIPE,
                               universal_newlines=True)
    stdout, stderr = process.communicate()
    try:
        data = json.loads(stdout)
    except ValueError:
        print('While list clusters: Something is wrong with output for en
vironment:', crn, ', Output JSON:', stdout)
        print('____ERROR__WHILE__CALLING__LIST__CLUSTER__COMMAND____', s
tderr)
        continue
    for en in data['clusters']:
        if en['status'] == 'REGISTERED':
            clusterIds.append(en['clusterId'])
            clusterCrns.append(en['clusterCrn'])

print('**************************')
print('**************************')
print('Cluster IDs:', clusterIds)
print('Cluster CRNs:', clusterCrns)
upgradeErrs = {}

for crn in clusterCrns:
    print('**************************')
    print('**************************')
    tryErr = ''
    for i in range(0, 10):
        time.sleep(60)
        print('Cluster:', crn, 'Try:', i)
        process = subprocess.Popen(get_command(['compute',
                                                'upgrade-deployment',
                                                '--cluster-crn', crn,
                                                '--namespace', 'yunikorn',
                                                '--name', 'yunikorn']),
                                   stdout=subprocess.PIPE,
                                   stderr=subprocess.PIPE,
                                   universal_newlines=True)
        stdout, stderr = process.communicate()
        try:
            data = json.loads(stdout)
        except ValueError:
```

```
            print('While upgrade deployment: Something is wrong with out
put for cluster:', crn, ', Output JSON:', stdout)
            print('____ERROR__WHILE__CALLING__UPGRADE__DEPLOYMENT__COMMAND
____', stderr)
            if i == 9:
                print('Failed upgrade deployment due to JSON error for clu
ster:', crn, ', Tries exhausted')
                tryErr = 'Error'
                break
            continue
        break
    if tryErr == 'Error':
        upgradeErrs[crn] = 'Error'
    else:
        print('Response status for cluster:', crn, 'from upgrade deployme
nt command:', data['status'])
        upgradeErrs[crn] = ''
for crn in clusterCrns:
    print('**************************')
    print('**************************')
    if upgradeErrs[crn] == 'Error':
        print('Skipping failed cluster:', crn)
        continue
    pollErr = ''
    for i in range(0, 100):
        time.sleep(5)
        print('Cluster:', crn, 'Try:', i)
        process = subprocess.Popen(get_command(['compute',
                                                'describe-deployment',
                                                '--cluster-crn', crn,
                                                '--namespace', 'yunikorn',
                                                '--name', 'yunikorn']),
                                   stdout=subprocess.PIPE,
                                   stderr=subprocess.PIPE,
                                   universal_newlines=True)
        stdout, stderr = process.communicate()
        try:
            data = json.loads(stdout)
        except ValueError:
            print('While describe deployment: Something is wrong with ou
tput for cluster:', crn, ', Output JSON:',
                  stdout)
            print('____ERROR__WHILE__CALLING__DESCRIBE__DEPLOYMENT__COM
MAND____', stderr)
            if i == 99:
                print('Failed describe deployment due to JSON error for c
luster:', crn, ', Tries exhausted')
                pollErr = 'Error'
                break
            continue
        if data['deployment']['status'] == 'DEPLOYED':
            break
        print('Response status for cluster:', crn, 'from describe deployme
nt command:', data['deployment']['status'])
        if i == 99:
            print('Failed deployment upgrade due to timeout for cluster:',
 crn, ', Tries exhausted')
            pollErr = 'Error'
    if pollErr == 'Error':
        print('Upgrade deployment failed for cluster:', crn)
    else:
        print('Upgrade deployment completed for cluster:', crn)
```

8. Run the script as follows:

```
 python3 post_upgrade_hook.py --profile <your-CDP-Private-profile> --end
point <your-CDP-Private-base-endpoint>
```

Note: If your CDP Private profile has base endpoint URL configured, then you do not need to pass --endpoint again in the python3 command

If your default CDP profile is also your CDP Private profile, then you do not need to pass --profile again in the python3 command.