Cloudera Data Warehouse Private Cloud 1.5.1

Planning and setting up CDW on Private Cloud

Date published: 2020-08-17
Date modified: 2023-06-13

https://docs.cloudera.com/

Legal Notice

Contents

Planning and setting up CDW Private Cloud

As a Cloudera Data Warehouse (CDW) Administrator on CDP Private Cloud, learn what the CDW hardware requirements are, how to deploy CDW, and understand the various interfaces and clients that you can use to access CDW.

- Review the hardware, security, and database requirements for deploying CDW.
- Create the required CDP resource roles such as DWAdmin and DWUser.
- Activate your environment in CDW.
- Create your first Virtual Warehouse.

Requirements for deploying Cloudera Data Warehouse on Private Cloud

Review the hardware requiremetns for deploying Cloudera Data Warehouse (CDW) in low and starndard resource modes, security and database requirements, user roles required to access and administer CDW.

Low resource mode requirements

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

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

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

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

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

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

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

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

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

Virtual Warehouse low resource mode resource requirements

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

Table 1: Impala Virtual Warehouse low resource mode requirements

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

Impala Admission Control Configuration

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

Table 2: Hive Virtual Warehouse low resource mode requirements

Component	vCPU	Memory	Local Storage	Number of pods in XSMALL Virtual Warehouse
Coordinator (2)	2 x 1	2 x 4 GB	2 x 100 GB	2
Executor (2)	2 x 4	2 x 48 GB (16 GB heap; 32 GB off-heap)	2 x 100 GB	2
HiveServer2	1	16 GB	--	1
Hue (backend)	0.5	8 GB	--	1
Hue (frontend)	--	--	--	1
Standalone compute operator	0.1	100 MB (.1 GB)	--	--
Standalone query executor (separate)	Same as executor	Same as executor	Same as executor	--
Total for XSMALL Virtual Warehouse	21 (20.6)	237 GB (236.1)	400 GB - 4 volumes	--

Database Catalog low resource mode requirements

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

Data Visualization low resource requirements

Table 3: Data Visualization low resource mode requirements

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

Standard resource mode requirements

Review the memory, storage, and hardware requirements for getting started with the Cloudera Data Warehouse (CDW) service in standard resource mode on Red Hat OpenShift and Embedded Container Service.

To get started with the CDW service on standard resource mode, make sure you have fulfilled the following requirements:

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

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

The following table lists the minimum and recommended compute (processor), memory, storage, and network bandwidth required for each OpenShift or ECS worker node using the Standard Resource Mode for production use case. Note that the actual node still needs some extra resources to run the operating system, Kubernetes engine, and Cloudera Manager agent on ECS.

Component	Minimum	Recommended
Node Count	4	10
CPU per worker	16 cores [or 8 cores or 16 threads that have Simultaneous Multithreading (SMT) enabled]	32+ cores (can also be achieved by enabling SMT)
Memory per worker	128 GB per node	384 GB* per node
FAST (Fully Automated Storage Tiering) Cache - Locally attached SCSI device(s) on every worker. Preferred: NVMe and SSD. OCP uses Local Storage Operator. ECS uses Local Path Provisioner.	1.2 TB* SATA, SSD per host	1.2 TB* NVMe/SSD per host
Network Bandwidth	1 GB/s guaranteed bandwidth to every CDP Private Cloud Base node	10 GB/s guaranteed bandwidth to every CDP Private Cloud Base node

* 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 1.8TB (600GB per executor) of locally attached SSD/NVMe storage for FAST Cache.

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

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

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

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

Related Information

Hyper-Threading

Security requirements for Cloudera Data Warehouse Private Cloud

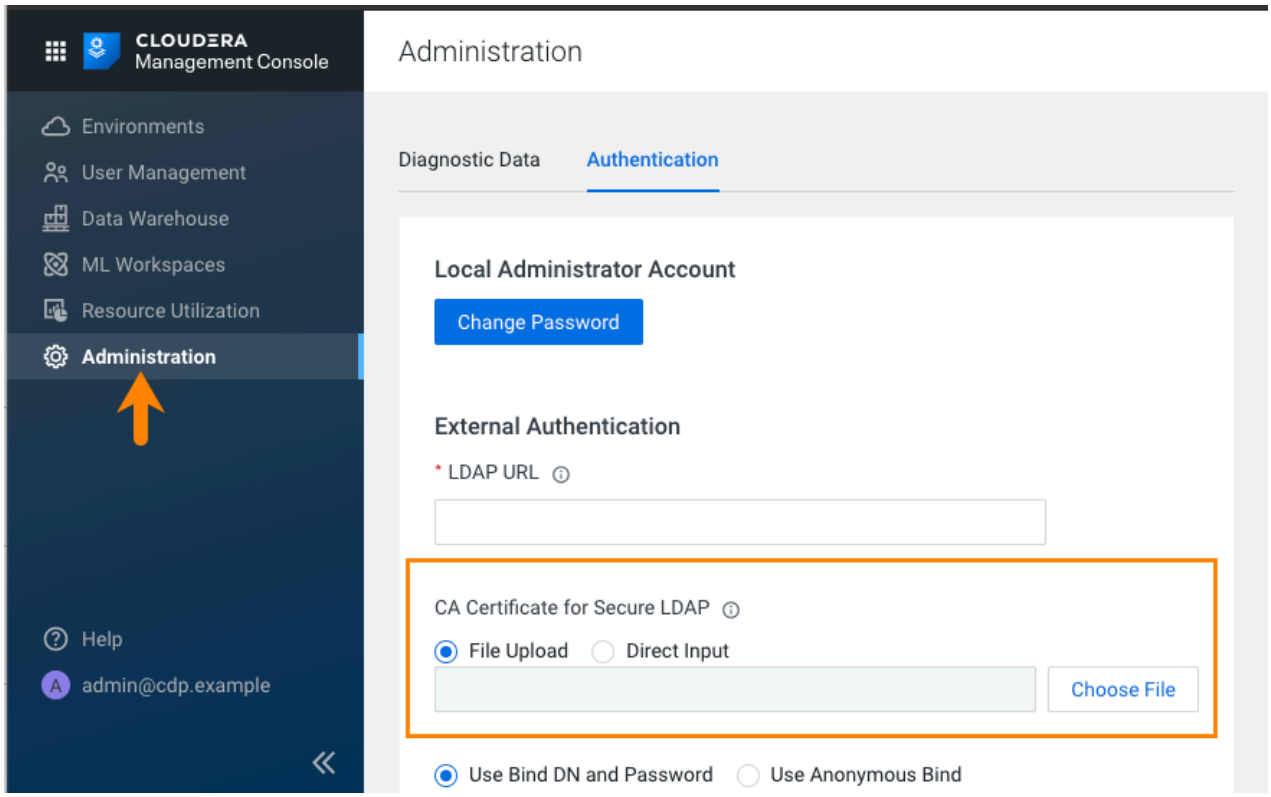
This topic describes security requirements needed to install and run Cloudera Data Warehouse (CDW) Private Cloud service on Red Hat OpenShift and Embedded Container Service (ECS) clusters.

Required OpenShift/ECS cluster permissions

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.

CDP Private Cloud LDAP certificate requirement

A certificate authority (CA) certificate for secure LDAP must be uploaded to the Administration page of Management Console to run CDW Private Cloud service:



Base cluster database requirements for Cloudera Data Warehouse Private Cloud

You must be aware of 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.

CDW supports MariaDB, MySQL, PostgreSQL, and Oracle databases for the Hive Metastore (HMS) on the base CDP cluster (Cloudera Manager side). On a default Database Catalog, Hue and HMS use an embedded PostgreSQL database that is defined when you install CDP Private Cloud.

Note: Cloudera recommends that you use an embedded database for the HMS and the Control Plane service. You can use the Data Recovery Service for backing up and restoring Kubernetes namespaces behind CDW entities (Database Catalogs and Virtual Warehouses).

If you are using PostgreSQL, MySQL, MariaDB, or Oracle database for the Hive Metastore on the base cluster, then it must meet the following requirements:

- SSL-enabled.
- Uses the same keystore containing an embedded certificate as Ranger and Atlas.

If your HMS database is not SSL/TLS-enabled and you want to continue using CDW, then you must disable the SSL requirement in CDW, so that the Database Catalog does not fail to start in an attempt to establish a secure connection with an unsecured database. For instructions, see *Disable the SSL or TLS requirement for HMS database*.

To use the same keystore with an embedded certificate for Ranger and Atlas:

- If you are using Auto-TLS:

  In the Management Console **Administration** page, go to the **CA Certificates** tab and select External Database from the CA Certificate Type drop-down menu. Upload the CA certificates either by uploading a file or by direct input.
- If you are not using Auto-TLS:

  Ensure that the public certificate of the certificate authority (CA) that signed the Hive metastore database's certificate is present in Cloudera Manager's JKS truststore. If the certificate is self-signed, import that certificate into Cloudera Manager's JKS truststore: In the Management Console Administration page, find the path to Cloudera Manager's JKS truststore by navigating to Administration Settings Security Cloudera Manager TLS/ SSL Client Trust Store File . Import the CA's certificate into that JKS file.

  To add the certificate name to an existing or a new JKS file, use the following keytool command, which uses the same example certificate name:

  keytool -import -alias postgres -file /path/to/postgres.pem -storetype JKS -keystore /path/to/cm.jks

  Where /path/to/cm.jks is the JKS file that is configured by Cloudera Manager.

  This ensures that the file specified for Cloudera Manager TLS/SSL Client Trust Store File is passed to Management Console and workloads.

  Note: If you have a JRE11 keystore you must convert it to a JRE8 keystore using the following keytool command:

```
keytool -importkeystore -srckeystore
            <path-to-my-pfx-file.pfx> -srcstoretype pkcs12 -srcstore
pass
               <***password***>  -destkeystore
               <path-to-client-certificate.jks> -deststoretype JKS
               -deststorepass <***password***>
```

Related Information

Disable the SSL or TLS requirement for HMS database

CDP resource roles and other prerequisites

To get started in Cloudera Data Warehouse (CDW), your data must conform to supported compression codecs, and you must obtain CDP resource roles to grant users access to a private cloud environment. Users can then get started on CDW tasks, such as activating the environment from CDW.

Unsupported compression

CDW does not support LZO compression due to licensing of the LZO library. You cannot query tables having LZO compression in Virtual Warehouses, which use CDW Impala or Hive LLAP engines.

CDP resource roles

Required role: PowerUser

The following CDP resource roles are associated with the CDW service. A CDP PowerUser must assign these roles to users who require access to the Database Catalogs and Virtual Warehouses that are associated with specific environments. After granting these roles to users and groups, they then have access to the Data Catalogs and Virtual Warehouses that are associated with the environment.

- **DWAdmin:** This role enables users or groups to grant a CDP user or group the ability to activate, terminate, launch, stop, or update services in Database Catalogs and Virtual Warehouses.
- **DWUser:** This role enables users or groups to view and use CDW clusters (Virtual Warehouses) that are associated with specific environments.

Requirements for Hue

Hue in CDW requires WebHDFS to be enabled on the CDP Private Cloud Base cluster. Worker nodes for both, Embedded Container Service (ECS) and OpenShift Container Platform (OCP), must have access to the WebHDFS (HTTPFS) port 14000.

Recommended HAProxy timeout for HA deployments

If you have enabled High Availability (HA) for CDP Private Cloud Data Services on ECS or OCP, then set the HAProxy timeout values to 10 minutes or more, depending on how long your queries run. Setting a higher timeout value is needed to support long-running queries and prevent timeouts.

Related Information

Understanding roles in CDP Private Cloud Data Services

Activating OpenShift environments

This topic describes how to activate an environment to use for Cloudera Data Warehouse (CDW) Private Cloud on Red Hat OpenShift Container Platform (OCP).

About this task

Before you can create a Database Catalog to use with a Virtual Warehouse, you must activate a CDP environment. Activating an environment causes CDP to connect to the Kubernetes cluster, which provides the computing resources for the Database Catalog. In addition, activating an environment enables the Cloudera Data Warehouse (CDW) service to use the existing data lake that was set up for the environment, including all data, metadata, and security.

Before you begin

- Determine which environment that uses a particular data lake is the environment you want to activate for use with a Database Catalog and Virtual Warehouse.

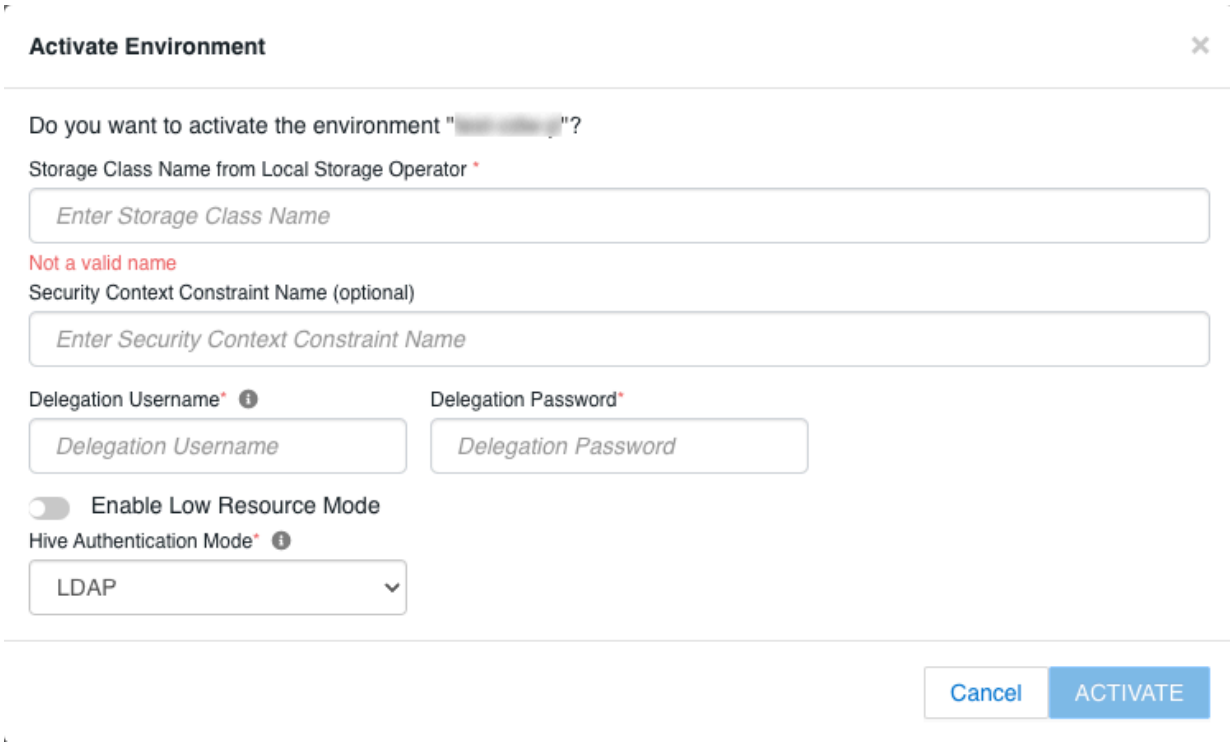
- For local caching, ensure that an administrator uses the Local Storage Operator to create a local file system on an SSD/NVMe for each OpenShift worker node and then mounts it to a known location on the worker node. Make sure that this local caching location allows temporary data to be stored in a way that supports performance. You need to specify the Storage Class Name from the Local Storage Operator when you activate the environment for the CDW service in Step 4 below. For more information about creating a local file system on OpenShift worker nodes using the Local Storage Operator, see Persistent storage using local volumes in the OpenShift documentation.
- (Optional) Go to Advanced Configuration Advanced Settings and enable the Use deterministic namespace names option to use deterministic namespaces for Kerberos principals and keytabs. You cannot enable this option after activating an environment.
- (Optional) Go to Advanced Configuration Advanced Settings and enable the Create databases for Virtual Warehouses option if you are upgrading the CDP Private Cloud Data Services platform from an older release to the latest release, and you want to continue using external database for Hue and HMS. You cannot enable this option after activating an environment.

Note: If you have more than one OCP clusters managed using different instances of Cloudera Manager, but using the same AD server and using the same environment name, then go to the Advanced Configuration Advanced Settings page and ensure that the Use deterministic namespace names option is disabled before activating the environment in CDW.

Procedure

1. Log in to Data Warehouse service as DWAdmin.
2. Expand the Environments column by clicking More… and locate the Environment that you want to activate.
3. Click the activation icon.

   The **Activate Environment** dialog box is displayed.



4. Specify the Storage Class Name from Local Storage Operator:

   This is the Storage Class Name you specified when you created the local file system for caching as described in the above section Before you begin. It is the location where temporary data is stored.

   Important: Be sure to specify the correct Storage Class Name when activating an environment. If an incorrect Storage Class Name is specified, the environment might activate successfully, but Virtual Warehouses that use the environment do not start.

   Optionally, you can specify the Security Context Constraint Name.
5. Specify Delegation Username and Delegation Password to impersonate authorization requests from Hue to the Impala engine.

   Note:

   - The delegation user and the LDAP Bind user configured on the **Administration** page of the Management Console are not necessarily the same user.
   - The special characters used in the LDAP Bind user password are not exactly the same as the ones that can be used in the delegation user password, because only the following characters are supported to be used in the LDAP Bind user password: ! # $ % ( ) * + , - . / : ; = ? @ [ ] ^ _ ` { | } ~.
   - The following special characters are not supported to be used in the name of the delegation user or in the Distinguished Name of the LDAP Bind user: < > & ' ".
6. Enable low resource mode to deploy CDW on minimum hardware.

   Note: Cloudera recommends that you use the Low Resource Mode option for proof of concept (POC) purposes only. This feature is not recommended for production deployment.

   Complex queries and multiple queries on HS2 may fail due to limited memory configurations for HMS and HS2 in the low resource mode.
7. If you are using an external database on your base cluster and want to use a default Database Catalog, then you must specify custom database name for Hue in the Pre-created database names for default database catalog field.
8. Select KERBEROS from the Hive Authentication Mode drop-down menu to use Kerberos as the authentication mode for Hive Virtual Warehouses.

   The default authentication mode is LDAP.
9. Click ACTIVATE.

Related Information

Advanced Configuration in CDW Private Cloud

How predefined Kerberos principals are used in CDW Private Cloud

Activating Embedded Container Service environments

This topic describes how to activate an environment to use for Cloudera Data Warehouse (CDW) Private Cloud on Embedded Container Service (ECS).

About this task

Before you can create a Database Catalog to use with a Virtual Warehouse, you must activate a CDP environment. Activating an environment causes CDP to connect to the Kubernetes cluster, which provides the computing resources for the Database Catalog. In addition, activating an environment enables the Cloudera Data Warehouse (CDW) service to use the existing data lake that was set up for the environment, including all data, metadata, and security.

Before you begin

- Determine which environment that uses a particular data lake is the environment you want to activate for use with a Database Catalog and Virtual Warehouse.

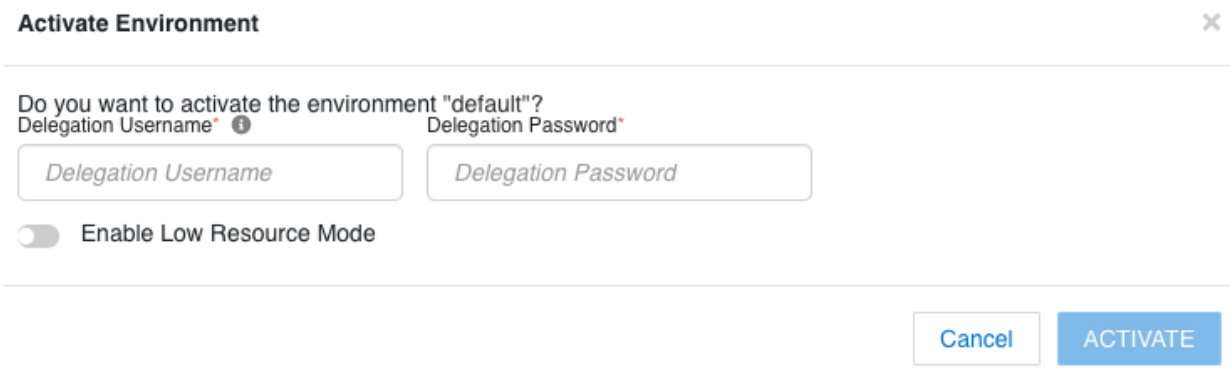
- In ECS environments, the Storage Class Name is automatically obtained from Cloudera Manager.
- (Optional) Go to Advanced Configuration Advanced Settings and enable the Use deterministic namespace names option to use deterministic namespaces for Kerberos principals and keytabs. You cannot enable this option after activating an environment.
- (Optional) Go to Advanced Configuration Advanced Settings and enable the Create databases for Virtual Warehouses option if you are upgrading the CDP Private Cloud Data Services platform from an older release to the latest release, and you want to continue using external database for Hue and HMS. You cannot enable this option after activating an environment.

Note: A “default” environment is created by the Control Plane when you add a Private Cloud cluster. If you have more than one ECS clusters managed using different instances of Cloudera manager, but using the same Active Directory (AD) server and using the same “default” environment, then go to the Advanced Configuration Advanced Settings page and ensure that the Use deterministic namespace names option is disabled before activating the environment in CDW.

Procedure

1. Log in to Data Warehouse service as DWAdmin.
2. Expand the Environments column by clicking More… and locate the Environment that you want to activate.
3. Click the activation icon.

   The **Activate Environment** dialog box is displayed.



4. Specify Delegation Username and Delegation Password to impersonate authorization requests from Hue to the Impala engine.

   Note:

   - The delegation user and the LDAP Bind user configured on the **Administration** page of the Management Console are not necessarily the same user.
   - The special characters used in the LDAP Bind user password are not exactly the same as the ones that can be used in the delegation user password, because only the following characters are supported to be used in the LDAP Bind user password: ! # $ % ( ) * + , - . / : ; = ? @ [ ] ^ _ ` { | } ~.
   - The following special characters are not supported to be used in the name of the delegation user or in the Distinguished Name of the LDAP Bind user: < > & ' ".

5. Enable low resource mode to deploy CDW on minimum hardware.

   Note: Cloudera recommends that you use the Low Resource Mode option for proof of concept (POC) purposes only. This feature is not recommended for production deployment.

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

6. Click ACTIVATE.

Related Information

Advanced Configuration in CDW Private Cloud

How predefined Kerberos principals are used in CDW Private Cloud

Creating your first Virtual Warehouse

After you activate an environment in Cloudera Data Warehouse (CDW), a default Database Catalog is automatically created. After the Database Catalog is in the running state, you can create Virtual Warehouses.

About this task

You can create Hive, Impala, or Impala Virtual Warehouses with Unified Analytics mode enabled.

Before you begin

Important: (On OpenShift environments) To activate an environment for the CDW service, someone with adequate permissions must use the Red Hat OpenShift Local Storage Operator to create a local file system on an SSD/NVMe for each OpenShift worker node and then mount it to a known location on the worker node. This creates space for local caching. The process is documented in Activating OpenShift environments.

On ECS clusters, CDW automatically creates the local file system. No additional steps are needed.

Procedure

1. Log in to the data Warehouse service as DWAdmin.
2. On the **Overview** page, click Create Virtual Warehouse.

   The **New Virtual Warehouse** modal screen is displayed.
3. Specify a name for your Virtual Warehouse, select the type, specify a size and click Create.

   To create a first test Virtual Warehouse, you can proceed with the default values. For fine-tuning, sizing, configuring, creating, and upgrading Virtual Warehouses, see Managing Virtual Warehouses.

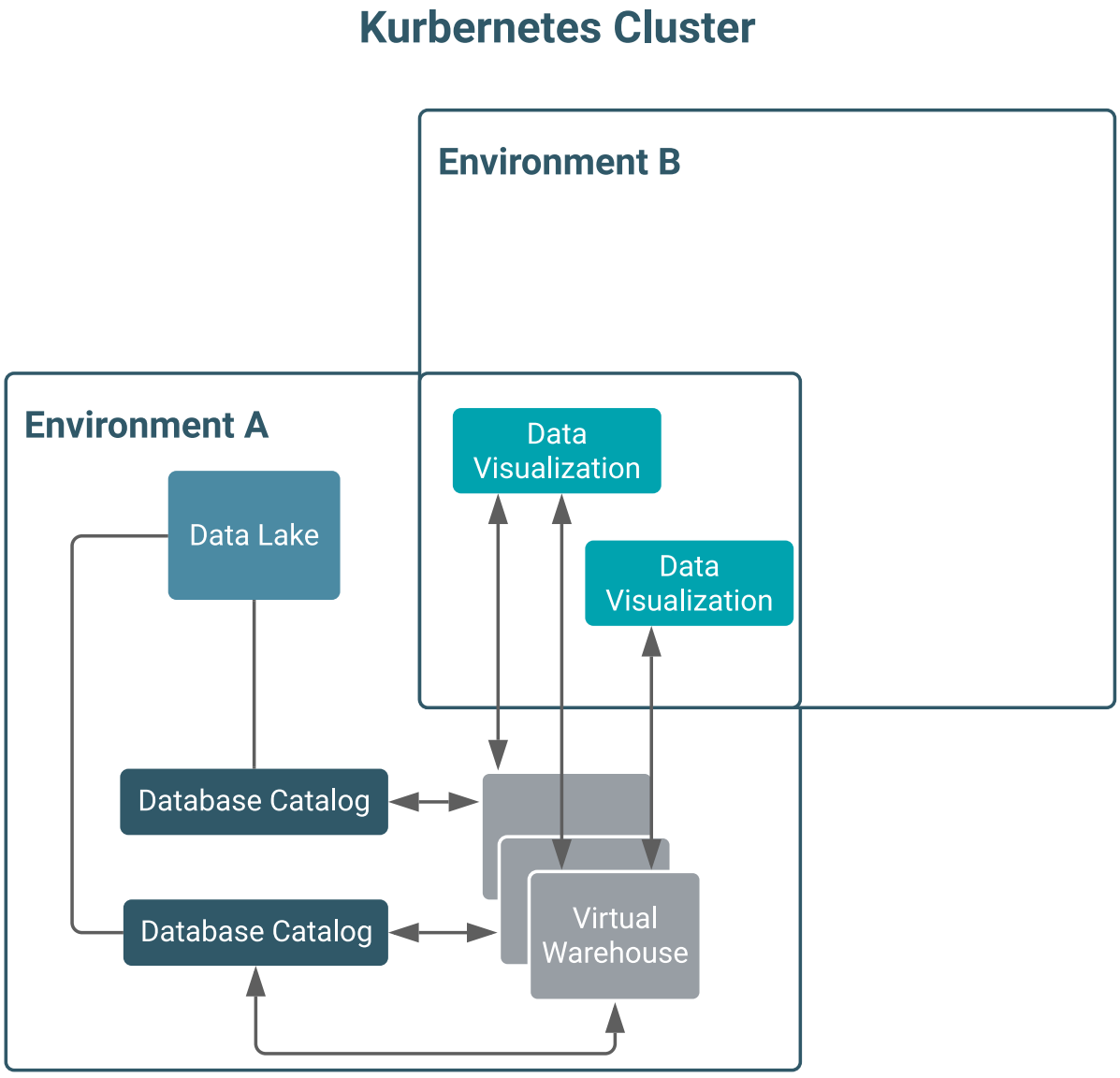
Results

You can submit workloads and run queries using Hue. You can also use SQL clients such as beeline, impala-shell, and so on to submit workloads after you connect them to your Virtual Warehouses.

Data Visualization in Cloudera Data Warehouse

Cloudera Data Warehouse (CDW) integrates Data Visualization for building graphic representations of data, dashboards, and visual applications based on CDW data, or other data sources you connect to. You, and authorized users, can explore data across the entire CDP data lifecycle using graphics, such as pie charts and histograms. You arrange visuals on a dashboard for collaborative analysis.

You connect Data Visualization to a Virtual Warehouse as described in Starting Data Visualization integrated in CDW. Similar to using a BI client, you can configure and connect to Virtual Warehouses from different clusters. You configure the connection in a familiar way, providing an IP address or host name. Data Visualization is not tied to a particular Virtual Warehouse (VW). You can access data for your visualization from multiple Data Catalogs using multiple Hive or Impala Virtual Warehouses and multiple environments.



Having multiple Data Visualization instances attached to an environment, you can create dashboards for different groups. For example, Marketing and Sales can have their own private dashboards. When you delete a Virtual Warehouse, your visuals remain intact.