

Managing CDW on Private Cloud

Date published: 2020-08-17

Date modified: 2023-11-02



Legal Notice

© Cloudera Inc. 2025. 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

Upgrade CDW runtime components.....	4
Refresh CDW.....	5
Advanced configurations.....	6
Object storage services.....	7
Enable S3 storage.....	8
Enable ADLS storage.....	9
Using Ozone.....	9
Set up Ozone on base.....	10
Configure Database Catalog to access Ozone.....	12
Configure Virtual Warehouses to create tables on Ozone.....	13
Configuring Hive/Impala logging on Ozone for Cloudera Data Warehouse Private Cloud.....	13
Specify or create an Ozone bucket for Cloudera Data Warehouse Private Cloud logs.....	14
Update Cloudera Data Warehouse Private Cloud log configuration to point to Ozone.....	14
Monitor Cloudera Data Warehouse Private Cloud logs on Ozone storage.....	16
Analyze Cloudera Data Warehouse Private Cloud logs stored on Ozone.....	17
Enable custom Database Catalog creation.....	17
Enable group access control.....	18
List of base cluster configurations.....	19
Disable configuration copy from base.....	21
Enable workload-aware autoscaling.....	21

Upgrading Database Catalogs and Virtual Warehouses in CDW Private Cloud

After you upgrade the CDP Private Cloud Data Services platform, you must upgrade the Database Catalog and Virtual Warehouses in Cloudera Data Warehouse (CDW). Upgrading to the latest release brings you new features from Hive, Impala, Hue, and other related runtime services. This is known as an in-place upgrade.

What gets upgraded

Database Catalog in CDW uses a Hive MetaStore (HMS) instance. The Virtual Warehouses use Apache Hive, Apache Impala, and Hue runtime images that are used in CDW. These runtime images are different than those used on CDP Private Cloud Base. With every new CDP Private Cloud Data Services release, you get a new version of Apache Hive, Apache Impala, and Hue runtimes with CDW, which includes new features and fixes.

Supported upgrade path for an in-place upgrade

In-place upgrade option is available only for upgrades from CDP Private Cloud Data Services 1.5.0 to a newer release.

What you should know before you upgrade

Review the [Release Notes](#) to learn about the new features, fixes, and known issues in this release, and more importantly, the [upgrade-related known issues](#).

In-place upgrade steps

To perform an in-place upgrade:

1. Upgrade the CDP Private Cloud Data Services platform.
2. Log in to the Data Warehouse service as DWAdmin.
3. Upgrade the Database Catalog by clicking  Upgrade .
4. Upgrade individual Virtual Warehouses by clicking  Upgrade .

To verify a successful upgrade, check the version information on the Database Catalog or Virtual Warehouse details page by clicking  Edit on the Database Catalog or Virtual Warehouse tile.



Note: In CDW Private Cloud, you can upgrade Database Catalogs and Virtual Warehouses only to the latest available version.

What changes after the upgrade

- If you had enabled the setting to copy the base cluster configurations to CDW on the CDP Private Cloud Data Services 1.5.1 cluster, then the base cluster configurations will continue to get copied to CDW after the upgrade. However, you can disable this setting from the **Advanced Settings** page.
- Starting with CDP Private Cloud Data Services 1.5.1, Data Analytics Studio (DAS) has been deprecated and completely removed from CDW. After you upgrade the platform, any running DAS instances will be removed from the cluster. Cloudera recommends that you use Hue for querying and exploring data in CDW.
- If you upgrade the CDP Platform from 1.5.0 to the latest release, then the configuration of an existing environments stays the same as before. Configurations such as default file format, compression type, and transactional type are not copied from the base cluster. To copy configurations from the base cluster, you must reactivate the environment.

- Starting with CDP Private Cloud Data Services 1.5.0, Hue in CDW requires WebHDFS to be enabled on the CDP Private Cloud Base cluster. Ensure that worker nodes for both, OpenShift Container Platform (OCP) and Embedded Container Service (ECS), have access to the WebHDFS (HTTPFS) port 14000.

Related Information

[Runtime component versions for Cloudera Data Warehouse Private Cloud](#)

[Upgrading CDP Private Cloud Data Services on Embedded Container Service](#)

[Upgrading CDP Private Cloud Data Services on OpenShift Container Platform](#)

[Activating OpenShift environments on CDW](#)

[Activating Embedded Container Service environments in CDW](#)

Refreshing environments, Database Catalog, and Virtual Warehouses in CDW Private Cloud

Learn when to refresh environments, Database Catalog, and Virtual Warehouses in Cloudera Data Warehouse (CDW) on Private Cloud and understand the difference between the refresh and rebuild operations.

Where is the refresh option in CDW?

The Refresh option is available in the more options () menu at the Environment, Database Catalog, and Virtual Warehouse levels.

When to use the refresh option?

You must refresh the environment, Database Catalog, and Virtual Warehouses in this order after completing the following actions:

- Adding or updating CA certificates in the Management Console
- Modifying LDAP server configurations in the Management Console
- Adding, updating, or deleting LDAP users
- Adding, updating, or deleting user groups and admin groups in the Management Console
- Updating database settings such as host, port, database name, username, and password in the Management Console
- Changing the configurations for Ozone, Hadoop, Hive, Impala, Ranger, and Atlas on the CDP Private Cloud Base cluster. This is true only if you have allowed CDW to receive configurations from the base cluster.



Note: If you change any Database Catalog or Virtual Warehouse configuration on the CDW web interface, then these configurations are not overwritten with the configurations from the base cluster even after you refresh the Virtual Warehouse.

- Enabling or disabling the following options from the CDW **Advanced Settings** page:
 - Enable ADLS as a storage provider
 - Enable S3 and S3-compatible object store providers
 - Store logs on HDFS
 - Enable warehouse-level access control for Impala
 - Copy configurations from base cluster to CDW
 - Enable workload-aware autoscaling for Impala

Difference between refresh and rebuild

The Refresh option can be used to apply configuration changes listed in this topic. Refreshing an environment, a Database Catalog, or Virtual Warehouses does not change the runtime version. The Rebuild option is displayed only on the Database Catalog and Virtual Warehouse tiles. When you rebuild a Database Catalog or a Virtual Warehouse,

CDW upgrades the Helm charts and the runtime version (if available), and also applies any configurations that may have changed on the base cluster or in the Control Plane service. The “Rebuild” operation is a superset of the “Refresh” operation.

Related Information

[Rebuilding a Database Catalog](#)

[Rebuilding Virtual Warehouses](#)

Advanced configurations in CDW Private Cloud

You can access advanced configurations in Cloudera Data Warehouse (CDW) Private Cloud from the left navigation pane on the CDW web interface. Some of these configurations must be enabled before you activate an environment and some can be applied by refreshing the environment, Database Catalog, and Virtual Warehouses.

Configuration	Description	Enabled by default?	Condition for applying the configuration
Create databases for Virtual Warehouses	Enables the automatic creation of databases for Hue and HMS. For external tables, CDW requires permission for CREATE and DROP DATABASE operations. If disabled, you need to create the databases before creating Database Catalogs.**	No	Enable before activating an environment in CDW.
Use deterministic namespace names	Makes the namespace names deterministic, that is, given the same input, a client can get to the same name every time. Enable this option if you need to create Kerberos principals and keytabs.	Yes	Enable before activating an environment in CDW.
Enable ADLS as a storage provider	Enables you to use Azure Data Lake Storage (Gen1 and Gen2) for storing tables.	Yes	Refresh
Copy configurations from base cluster to CDW	Configurations such as default file format, compression type, and transactional type are copied from the base cluster to CDW to aid cluster setup and workload migration.	Yes	Refresh*
Customize pod sizes	Enables you to customize the number of resources allocated to the Impala coordinators, executors, and catalog daemons.	Yes	No special action needed.
Store logs on HDFS	Enables you to store CDW logs to HDFS on the base cluster.	Yes	Refresh*
Create multiple Database Catalogs	Enables you to create more than one custom Database Catalogs for your environment. Not recommended for production deployments.	No	No special action needed.

** This option is available only if you are upgrading CDP Private Cloud Data Services from a previous release.

If you enable this option, then you must specify the external database name on the **Activate Environment** page.

* Refresh the environment, Database Catalog, and Virtual Warehouses, in this order.

Configuration	Description	Enabled by default?	Condition for applying the configuration
Enable quota management	Enables you to assign quota-managed resource pools to environments, Database Catalogs, Virtual Warehouses, and Data Visualization instances.	No	Enable before activating an environment in CDW.
Enable S3 and S3-compatible object store providers	Enables you to use AWS S3 and other similar, compatible, on-premises object stores that support the S3 protocol for storing tables.	Yes	Refresh*
Enable warehouse-level access control for Impala	Enables you to allow access to an Impala Virtual Warehouse for selected user groups. Select the Enable warehouse-level access control for Hive and Unified Analytics option to create an Impala warehouse in the Unified Analytics mode.	Yes	Refresh*
Enable warehouse-level access control for Hive and Unified Analytics	Enables you to allow access to an Hive Virtual Warehouse for selected user groups in regular and Unified Analytics mode.	No	Refresh*
Enable workload-aware autoscaling for Impala	Enables you to create multiple executor group sets of different sizes that can scale independently based on the load. This is a preview feature. Not recommended for production deployments.	No	Refresh*
Skip cluster validation during environment activation	Skips the cluster validation step during environment activation. Select this option if you want to proceed with the environment activation even after seeing false positive errors in the logs.	No	No special action needed.

Supported object storage services for Cloudera Data Warehouse Private Cloud

HDFS is the default storage system for Cloudera Data Warehouse (CDW). However, you can enable CDW to access object storage such as AWS S3 and Azure Data Lake Storage (ADLS Gen1 and Gen2) if the CDP Private Cloud base cluster is configured to access it. You can query Hive and Impala tables stored on object stores using Hue.



Important: S3, S3-compatible, and ADLS object storage support is in technical preview and is not recommended for production deployments. Cloudera recommends that you try this feature in test and development environments.

When you activate an environment in CDW, all the hadoop configurations variables (fs.s3a.*/fs.azure.*) are copied from the core-site.xml file present on the base cluster to the hadoop-core-site.xml file of the Hive and Impala metastore pods, enabling CDW to establish a connection to S3/ADLS.

Following are the key configurations that must be present in the base cluster core-site.xml file for connecting to S3 or S3-compatible storage providers:

- fs.s3a.access.key
- fs.s3a.secret.key
- fs.s3a.endpoint

- `fs.s3a.connection.ssl.enabled`

Following are the key configurations that must be present in the base cluster `core-site.xml` file for connecting to ADLS storage provider:

- `fs.azure.account.oauth.provider.type`
- `fs.azure.account.oauth2.client.id`
- `fs.azure.account.oauth2.client.secret`
- `fs.azure.account.oauth2.client.endpoint`

**Important:**

Because CDW uses all the base cluster configurations, it is important that you fine-tune and debug these configurations on the base cluster before creating the CDW environment.

If you have installed the Private Cloud Data Services, including CDW, before fine-tuning the base cluster configurations, then you must upload the Amazon/Azure server certificates referenced in the `fs.s3a/fs.azure` endpoint configuration on the Management Console Administration CA Certificates tab. Select Miscellaneous as the certificate type from the CA Certificate Type drop-down menu.

The `fs.s3a.*`/`fs.azure` configurations are read-only. You can view these configurations from the CONFIGURATION tab on the Database Catalog and Virtual Warehouse details page by selecting the `hadoop-core-site.xml` option from the Configuration files drop-down menu.



Note: Disabling CDW's access to the third-party S3 providers from the **Advanced Settings** page does not affect the previously created Database Catalogs and Virtual Warehouses. To disable access, you must delete and recreate the Database Catalog and Virtual Warehouse.

Enabling S3 and S3-compatible storage providers in CDW

You can enable Cloudera Data Warehouse (CDW) data service on CDP Private Cloud to access S3 and S3-compatible object storage if the CDP Private Cloud base cluster is configured to access it.

About this task



Note: S3 and S3-compatible object storage support is in technical preview and is not recommended for production deployments. Cloudera recommends that you try this feature in test and development environments.

Before you begin

If you have installed the Private Cloud Data Services, including CDW, before fine-tuning the base cluster configurations, then you must upload the Amazon server certificates referenced in the `fs.s3a` endpoint configuration on the Management Console Administration CA Certificates tab. Select Miscellaneous as the certificate type from the CA Certificate Type drop-down menu.

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to **Advanced Configurations** **Advanced Settings** page.
3. Select the **Enable S3 and S3-compatible object store providers** option.
4. Click **Update**.



Important: If you have upgraded from an earlier release and you have enabled the option to use S3, then you need to recreate the environment in CDW.

Enabling ADLS storage providers in CDW

You can enable Cloudera Data Warehouse (CDW) data service on CDP Private Cloud to access Azure Data Lake Storage (ADLS Gen1 and Gen2) object storage if the CDP Private Cloud base cluster is configured to access it.

About this task



Important: ADLS object storage support is in technical preview and is not recommended for production deployments. Cloudera recommends that you try this feature in test and development environments.

Before you begin

If you have installed the Private Cloud Data Services, including CDW, before fine-tuning the base cluster configurations, then you must upload the Azure server certificates referenced in the `fs.azure` endpoint configuration on the **Management Console Administration CA Certificates** tab. Select **Miscellaneous** as the certificate type from the **CA Certificate Type** drop-down menu.

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to the **Advanced Configurations Advanced Settings** page.
3. Select the **Enable ADLS as a storage provider** option.
4. Click **Update**.
5. Refresh the Database Catalog and Virtual Warehouses by clicking  **Refresh** on the Database Catalog and Virtual Warehouse tile.

Using Ozone storage with Cloudera Data Warehouse Private Cloud

Apache Ozone is an object store available on the CDP Private Cloud Base cluster which enables you to optimize storage for big data workloads. You can query data residing on Ozone using Hive or Impala from Cloudera Data Warehouse (CDW) Data Service on Private Cloud.

Apache Ozone DataNodes support storage density up to 400 TB, unlike HDFS DataNodes which support storage density only up to 100 TB. Apart from the ability to scale to billions of objects or files of varying sizes, applications that use frameworks like Apache Spark, Impala, Apache YARN, and Apache Hive work natively on Ozone without any modifications.

Supported use cases

Ozone filesystem (OFS) is best suited for Hive and Impala in the following use cases:

- To retain HDFS IO performance and other characteristics critical for big data use cases.
- Recommended in an environment with dense nodes using up to 400 TB per node.
- To scale linearly and handle a large number of files and data.
- Recommended with Hadoop and S3 workloads.
- Recommended with native API, fast IO scans, streaming reads, and writes.
- Object-level rename in a bucket.

Advantages

OFS offers the following operational advantages:

- Ability to share physical storage and nodes with HDFS.
- Designed for easy Node-addition, deletion, and decommission for repair.
- Has a security model similar to HDFS.
- Supports Kerberos authentication.
- Supports Data encryption at rest and in flight.
- Supports Ranger Authorization.

Related Information

[Blog: Apache Ozone and Dense Data Nodes](#)

Setting up Ozone on the CDP Base cluster

To access and use Ozone from Cloudera Data Warehouse (CDW) data service on Private Cloud, you must add and configure the Ozone service on the base cluster.

Before you begin

Provision an Ozone cluster based on your desired storage capacity.

Procedure

1. Log in to Cloudera Manager as an Administrator.
2. Add and configure the Ozone service on the base cluster.
3. Enable Kerberos on the base cluster before you install the CDW data service.
Enabling Kerberos on the base cluster automatically enables the Ozone service to use Kerberos. To verify this, go to Ozone service Configuration . The `ozone.security.enabled` parameter should be set to true and the `hadoop.security.authentication` parameter should be set to kerberos.
4. SSH into the Ozone host on the base cluster as an Administrator.
5. Obtain the tickets for the Hive or Impala user by using the Kerberos CLI kinit command.
6. Verify the Ozone Service ID for your cluster from the Configuration tab of the Ozone service in Cloudera Manager.
7. Verify that at least one volume and a bucket is available in Ozone by using the service ID you just verified. If a volume and a bucket does not exist, then run the following commands to create a volume in Ozone using the service ID:

```
ozone sh volume create --quota=[***VOLUME-CAPACITY***] --user=[***USERNAME***] URI
```

where,

- `-q, --quota`: Used to specify the maximum size that a volume can occupy in the cluster. This is an optional parameter.
- `-u, --user`: Used to specify the name of the user who can use the volume. The designated user can create buckets and keys inside the particular volume. This is a mandatory parameter.
- `URI`: Used to specify the name of the volume to be created. Specify the URI in the following format:

```
[***PREFIX***]://[***SERVICE-ID]/[***VOLUME-NAME***]
```

```
ozone sh volume create --quota=100GB --user=hrt_1 o3://vvs1ab/testvol
```

8. Create an encrypted or a non-encrypted bucket using the service ID that you just verified by running the following commands:

To create encrypted buckets:

```
ozone sh bucket create -k [***ENCRYPTION-KEY***] [***PREFIX***]://[***SERVICE-ID]/[***VOLUME-NAME***]/[***BUCKET-NAME***]
```

```
ozone sh bucket create -k key1 o3://vvs1ab/testvol/testbucketencrypted
```



Important: You must have the GET_METADATA and GENERATE_EEK permissions on the encryption key to create encrypted buckets on Ozone. The user who needs to read from the encrypted bucket must have the DECRYPT_EEK permission. These permissions are defined in the Ranger KMS policies on the base cluster.

To create non-encrypted buckets:

```
ozone sh bucket create [***PREFIX***]://[***SERVICE-ID]/[***VOLUME-NAME***]/[***BUCKET-NAME***]
```

```
ozone sh bucket create o3://vvs1ab/testvol/testbucket
```

22/08/10 10:25:10 INFO rpc.RpcClient: Creating Bucket: testvol/testbucket, with Versioning false and Storage Type set to DISK and Encryption set to false

9. Verify that the bucket is created by listing the bucket as follows:

```
ozone sh bucket list [***PREFIX***]://[***SERVICE-ID]/[***VOLUME-NAME***] --length=[***NUMBER-OF-BUCKETS] --prefix=[***BUCKET-PREFIX] --start=[***STARTING-BUCKET***]
```

where,

- -l, --length: Used to specify the maximum number of results to return. The default is 100.
- -p, --prefix: Used to list the bucket names that match the specified prefix.
- -s, --start: Used to return results starting with the bucket after the specified value.



Note: All the existing buckets in Ozone are automatically available to query from Hive and Impala Virtual Warehouses in CDW.

To set Ozone as the default file system, you must configure OFS and add specific properties for the Ozone bucket you created.

What to do next

After setting up Ozone storage on the base cluster, configure CDW to use Hive or Impala to query data residing on the Apache Ozone object store.

Related Information

[Enabling Kerberos Authentication for CDP](#)

[Kerberos configuration for Ozone](#)

[Commands for managing buckets](#)

[Managing storage elements by using the command-line interface](#)

[Setting up ofs](#)

Configuring the Database Catalog to access the Ozone filesystem

After adding and configuring the Ozone service on the base cluster, creating buckets, and granting Ranger KMS policies to the users, you must configure the Hive MetaStore warehouse directories in the Database Catalog to point to the Ozone filesystem.

Before you begin



Note: If you have activated an environment in CDW before installing the Ozone service on the base cluster, then you must recreate the CDW environment so that Ozone configurations can be imported into CDW.

By default, the Hive MetaStore (HMS) for Database Catalogs on CDW Private Cloud points to HDFS.

- If you plan to make Ozone as the default FS, you must configure the Database Catalog to point to the Ozone storage system, as described in this topic.
- Alternatively, you can create a database with an Ozone bucket as the base directory so that all tables are created in that directory. Following is a sample command:

```
CREATE DATABASE ozone_db
[LOCATION ofs://ozone1/bucket1/ozone_db/external]
[MANAGEDLOCATION ofs://ozone1/bucket1/ozone_db/managed]
[WITH DBPROPERTIES (property_name=property_value, ...)];
```

Before you re-configure the Database Catalog settings, make sure there are no running Virtual Warehouses associated with it. Either the Database Catalog has no associated Virtual Warehouses or you have suspended all the Virtual Warehouses associated with it.

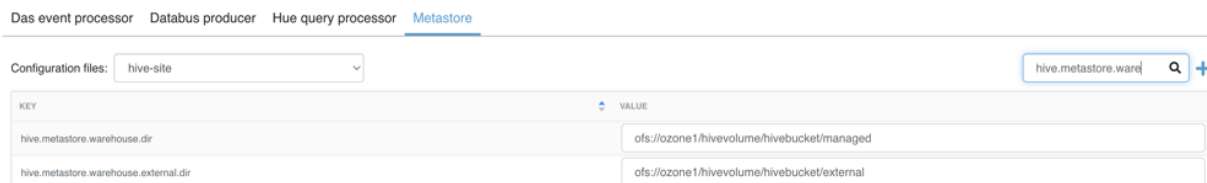
Procedure

1. Log in to the Data Warehouse service as a DWAdmin.
2. Activate an environment in CDW.
3. Go to the **Database Catalog** tab, locate your Database Catalog, and click  Edit CONFIGURATIONS Metastore and select hive-site from the Configuration files drop-down menu.
4. Search for the following configuration properties and update them to Ozone filesystem paths, which start with ofs:
 - hive.metastore.warehouse.dir
 - hive.metastore.warehouse.external.dir



Note: For the Hive Table creation, the warehouse directory must be set at bucket level or directory level under the hive.metastore.warehouse.dir or hive.metastore.warehouse.external.dir parameters. For more information, see [Changing the Hive warehouse location](#).

Following is an example of these properties set for a Database Catalog:



The screenshot shows the 'Metastore' configuration page. At the top, there are tabs for 'Das event processor', 'Databus producer', 'Hue query processor', and 'Metastore'. Below the tabs, there is a 'Configuration files' dropdown menu set to 'hive-site'. To the right, there is a search bar with 'hive.metastore.ware' and a magnifying glass icon. Below this, there is a table with two columns: 'KEY' and 'VALUE'.

KEY	VALUE
hive.metastore.warehouse.dir	ofs://ozone1/hivevolume/hivebucket/managed
hive.metastore.warehouse.external.dir	ofs://ozone1/hivevolume/hivebucket/external



Note: The example values in the screenshot show the Hive warehouse locations in Ozone (set at a directory level) where Hive stores the tables. hivevolume represents the Ozone volume, hivebucket represents the Ozone bucket, and managed and external are directories where Hive stores the managed and external tables.

5. Click Apply Changes and wait for the Database Catalog to finish applying changes.

Results

After configuring the Database Catalog's Hive metastore to point to Ozone, create a Hive or an Impala Virtual Warehouse, or restart an existing Virtual Warehouse. You can then create databases and table using Hue or other SQL clients with your Virtual Warehouse.

Creating a Virtual Warehouse and creating tables on Ozone

After you configure the Database Catalog to point to the Ozone filesystem, verify that the Hive and Impala Virtual Warehouses in Cloudera Data Warehouse (CDW) carry the right configurations, and then you can managing databases and tables residing in Ozone using Hue or other SQL clients.

Before you begin

Ensure that the Hive MetaStore warehouse directories in the Database Catalog point to the Ozone filesystem on the CDP Private Cloud Base cluster.

Procedure

1. Log in to the Data Warehouse service as a DWAdmin.
2. Create an Impala or Hive Virtual Warehouse.
Since you have already added Ozone in your base cluster, the required configuration are made available in CDW when you create a Virtual Warehouse.
3. Verify that the Ozone configurations are present in CDW. From your Virtual Warehouse tile, click  Edit CONFIGURATIONS Impala catalogd and select ozone-site from the Configuration files drop-down menu.

For Hive, click  Edit CONFIGURATIONS Hiveserver2 and select ozone-site from the Configuration files drop-down menu.

4. Use Hue or any other SQL clients to start managing databases, managed and external tables.
Following is a sample command to create an external table:

```
create external table
[***TABLE-NAME***] (id int, name string)
location 'ofs://ozone1/s3v/cdw-logs/compute-schal-pvc111-env-1-hive5/ware
house/tablespace/[***TABLE-NAME***]';
```

5. Verify that the required keys are created in the bucket by running the following command:

```
ozone sh bucket ls [***VOLUME***] -p warehouses/tablespace
```

Related Information

[Adding a new Virtual Warehouse](#)

Configuring Hive/Impala logging on Ozone for Cloudera Data Warehouse Private Cloud

This section describes how to configure Cloudera Data Warehouse (CDW) on Private Cloud to store Hive and Impala logs on Ozone storage.

You can configure CDW to store Hive and Impala logs on CDP Private Cloud storage components, such as Ozone. Ozone is a good choice to store these logs because:

- Ozone efficiently handles files regardless of their size.

- In addition to Ozone's built-in CLI interface, Ozone also supports the HDFS CLI and CLIs that are compatible with AWS clients.
- CDP Private Cloud uses [fluentd](#) to push application logs to the storage layer. Ozone is a supported logging "back-end" component and has a fluentd-compatible endpoint for collecting the logs.



Note: Ozone support is in technical preview in CDW 1.4.1. Cloudera recommends that you use Ozone with CDW in test and development environments. It is not recommended for production deployments.

Specify or create an Ozone bucket for Cloudera Data Warehouse Private Cloud logs

This topic describes how to specify an Ozone bucket to store Cloudera Data Warehouse (CDW) Private Cloud Hive and Impala logs.

About this task

You can either re-use the Ozone bucket that is automatically configured for storing Cloudera Machine Learning (CML) Private Cloud logs or create a new bucket to store CDW logs separately. The Ozone bucket used to store CML logs usually has a `cdplogs-` prefix.

Procedure

Use one of the following two methods depending on whether you want to use the existing CML log bucket or create a new one for CDW:

- To select an existing Ozone bucket, use the `ozone sh bucket list` command from the Ozone shell on your Private Cloud Base cluster. The following example shows how you can list buckets by the `cdplogs-` prefix:

```
ozone sh bucket list o3://ozone1/s3v --prefix=cdplogs
{
  "metadata" : { },
  "volumeName" : "s3v",
  "name" : "cdplogs-av-dwx-env-96c47aa9",
  "storageType" : "DISK",
  "versioning" : false,
  "creationTime" : "2020-08-01T18:29:08.686z",
  "modificationTime" : "2020-08-03T18:29:08.686z",
  "encryptionKeyName" : null,
  "sourceVolume" : null,
  "sourceBucket" : null
}
```

- To create a new bucket on Ozone, use the `ozone sh bucket create` command from the Ozone shell on your Private Cloud Base cluster. The following example shows how to create a new Ozone bucket named `cdw-logs-bucket`:

```
ozone sh bucket create o3://ozone1/s3v/cdw-logs-bucket
```



Important: Cloudera recommends that you use the `hive` user because this user automatically has create/read/write permissions on buckets that you create.

Update Cloudera Data Warehouse Private Cloud log configuration to point to Ozone

This topic describes how to configure Cloudera Data Warehouse (CDW) Private Cloud to store logs on Ozone.

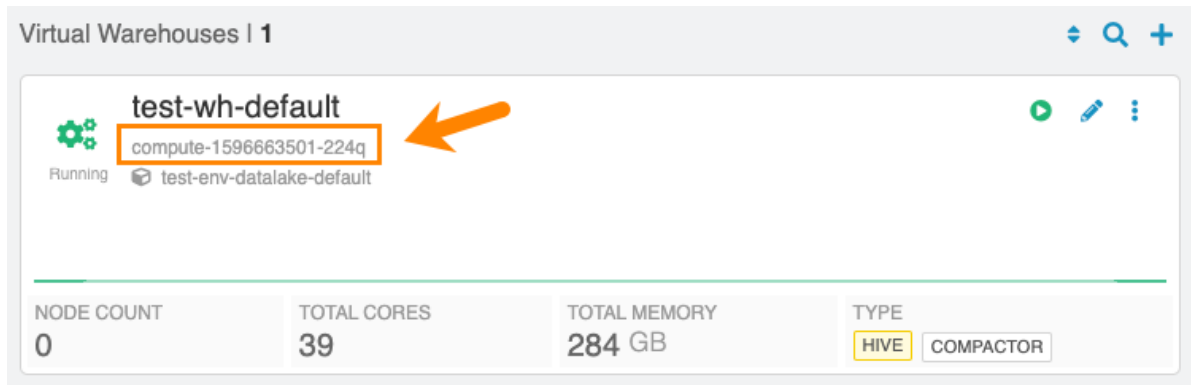
About this task

To configure CDW Private Cloud and the underlying OpenShift cluster to store Hive and Impala logs on Ozone, you must gather some information and prepare a block of code that you will insert into the Virtual Warehouse ConfigMap on the OpenShift pod. These preliminary steps are described in the following section.

Before you begin

Get the following information and prepare the block of code for the Virtual Warehouse ConfigMap before you start the steps of updating the configuration:

- Get the CDW namespace for your Virtual Warehouse:
 1. From the Management Console home page left menu, click Data Warehouse in the left menu. You are taken to the Overview page of CDW Private Cloud service.
 2. Locate the Virtual Warehouse you want to configure log storage for in the right-most column of the page, and locate the CDW namespace, which starts with compute- as shown below:



- Prepare the code block that must be pasted into the OpenShift ConfigMap:

Here is an example:

```
<match **>
  @type s3
  @log_level debug
  aws_key_id <ACCESS-ID>
  aws_sec_key <SEC-KEY>
  s3_bucket <BUCKET-NAME>
  s3_endpoint <OZONE-S3-GATEWAY-ENDPOINT>
  ssl_verify_peer false
  s3_object_key_format
    "<WAREHOUSE_PREFIX>/warehouse/tablespace/external/hive/sys.db/logs
    /dt=%Y-%m-%d/${path_tag}/${time_slice}_${unique_file_key}.log.${file_ext
    ension}"
  time_slice_format %Y-%m-%d-%H-%M
  store_as gzip
  auto_create_bucket false
  check_apikey_on_start false
  force_path_style true
  check_bucket false
  check_object false
  <buffer path_tag, unique_file_key, time, warehouse>
  @type file
  path /tmp/fluentd-buffers/${unique_file_key}-s3.buffer
  timekey 900 # minute precision for time_slice_format to have minu
te in file name
  timekey_use_utc true
  chunk_limit_size 265m
  flush_mode interval
  flush_interval "900s"
```

```

      flush_thread_count 8
      flush_at_shutdown true
    </buffer>
    <format>
      @type single_value
      message_key log
      add_newline true
    </format>
  </match>

```

In the above code block example:

- `<BUCKET-NAME>` indicates the name of the Ozone bucket used for storing the CDW Private Cloud logs.
- `<OZONE-S3-GATEWAY-ENDPOINT>` indicates the endpoint of the Ozone S3 Gateway. Get this value from the Ozone S3 Gateway Web UI page of Cloudera Manager.
- `<ACCESS_ID>` and `<SEC_KEY>` are the AWS access credentials for the Ozone S3 Gateway. Get these values by using the `kinit -kt` and the `ozone s3 getsecre` commands on the Private Cloud Base OpenShift cluster.

Procedure

1. Using OpenShift commands, view the OpenShift project for the pod where the CDW Private Cloud instance is running by specifying the CDW namespace for the Virtual Warehouse that you noted in the [Before you begin](#) section above.

For example, if the CDW namespace is `compute-1596663501-224q`, you can view the OpenShift project with the following command:

```
oc project compute-1596663501-224q
```

2. Open the ConfigMap for the Virtual Warehouse that is associated with the CDW namespace. For example:

```
oc edit configmap warehouse-fluentd-config
```

This command opens the ConfigMap in a separate editor that is similar to `vi`.

3. Replace the match section of the ConfigMap with the code block you prepared in the [Before you begin](#) section above, and then save your changes
4. Verify that the new configuration is correctly updated by running the following command:

```
oc get namespace -o yaml | grep fluentd-status
```

If the configuration is successfully updated, the value of the `fluentd-status` returns an empty string as shown in the following example:

```

com.cloudera/fluentd-status: ""
com.cloudera/fluentd-status: ""
com.cloudera/fluentd-status: ""
com.cloudera/fluentd-status: ""

```

Monitor Cloudera Data Warehouse Private Cloud logs on Ozone storage

This topic describes how to monitor Cloudera Data Warehouse (CDW) Private Cloud logs that are stored on Ozone.

About this task

You can use either the Ozone S3 Gateway Web UI in Cloudera Manager or run commands in a terminal window to monitor CDW logs.



Note: Because fluentd buffers the logs and then pushes them to the configured endpoint, Ozone might take up to 15 minutes to display the CDW logs.

Procedure

Use one of the following methods to monitor CDW logs in Ozone:

- Ozone S3 Gateway Web UI in Cloudera Manager:

Navigate to the following URL:

`https://<S3-GATEWAY-ENDPOINT>/<BUCKET-NAME>?browser=true`

Where:

- `<S3-GATEWAY-ENDPOINT>` indicates the endpoint of the Ozone S3 Gateway, which you can get from the Ozone S3 Gateway Web UI
- `<BUCKET-NAME>` indicates the Ozone bucket where you are storing the CDW logs.
- Run the following command from the Ozone shell: `ozone sh key list o3://<OZONE.SERVICE.ID>/s3v/<BUCKET-NAME>/ --prefix=<WAREHOUSE-PREFIX>`

Where:

- `<OZONE.SERVICE.ID>` indicates the identifier used for your implementation of Ozone.
- `<BUCKET-NAME>` indicates the name of the Ozone bucket where the CDW logs are stored.
- `<WAREHOUSE-PREFIX>` indicates the Virtual Warehouse identifier.

Analyze Cloudera Data Warehouse Private Cloud logs stored on Ozone

This topic describes how to analyze Cloudera Data Warehouse (CDW) Private Cloud logs that are stored on Ozone using Hue.

About this task

You can use Hue to analyze Impala logs or Hive logs.



Note: You must use the Hue instance that corresponds to the Virtual Warehouse whose logs are saved on Ozone.

Procedure

1. Using Hue, create an external table that points to the log data on Ozone:

```
CREATE EXTERNAL TABLE <TABLE-NAME> LIKE sys.logs LOCATION 'ofs://<OZONE.SERVICE.ID>/s3v/<BUCKET-NAME>/<WAREHOUSE-PREFIX>/warehouse/tablespace/external/hive/sys.db/logs';
```

2. Run the MSCK REPAIR TABLE command on the table you created in Step 1:

```
MSCK REPAIR TABLE <TABLE-NAME>;
```

Results

After completing the above steps, you can use SQL queries to analyze the log data.

Enabling the option to create additional Database Catalogs in CDW Private Cloud

Starting with the Private Cloud 1.5.0 release, you cannot create additional non-default Database Catalogs without enabling the “Create multiple Database Catalogs” option on the Advanced Settings page.

About this task

In Cloudera Data Warehouse (CDW) Private Cloud, a default Database Catalog is created when you activate an environment and is specific to that environment. The default Database Catalog uses a CDW-managed database. If you create additional Database Catalogs, then custom Database Catalogs are created which require a separate, external database that CDW does not manage.



Note: Custom Database Catalogs have been deprecated and will be completely removed in future releases. Cloudera does not support custom Database Catalogs. Cloudera recommends that you use the default Database Catalog in your production deployments.

To enable creating custom Database Catalogs:

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to the [Advanced Configuration Advanced Settings](#) page.
3. Select the Create multiple Database Catalogs option.
4. Click Update.

Enabling warehouse-level access control for Hive and Impala in CDW Private Cloud

Cloudera Data Warehouse (CDW) enables you to specify one or more user groups to access a Virtual Warehouse while creating it. As a result, only those users can connect to that Virtual Warehouse, from all supported connection channels such as Hue, JDBC, Beeline, Impala-shell, Impyla, or other Business Intelligence tools. You can enable and disable warehouse-level access control from the Advanced Settings page in the CDW web UI.

About this task

If you do not specify a user group while creating a Virtual Warehouse, then the access is not restricted. Any logged-in user can access the Virtual Warehouse.



Attention: When you enable warehouse-level access control for Hive warehouses or Impala warehouses in the Unified Analytics mode and associate a user group with that Virtual Warehouse, Kerberos authentication is disabled. Only LDAP is used for authentication. This is because of a current limitation in Hive on using LDAP to filter users and groups when Kerberos is used for authentication.



Note: Warehouse-level access control feature for Hive is in technical preview and not recommended for production deployments. Cloudera recommends that you try this feature in test or development environments.

Before you begin

You must have the user groups created in the Management Console. If you are using Kerberos for authentication, then ensure that the users for whom you are enabling access are present in LDAP as well.

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to [Advanced Configuration Advanced Settings](#) page.
3. To enable the access control feature, select the Enable warehouse-level access control for Hive and Unified Analytics option.



Note: The warehouse-level access control is enabled by default for Impala Virtual Warehouses. To enable warehouse-level access control for an Impala Virtual Warehouse in Unified Analytics mode, you must enable the Enable warehouse-level access control for Hive and Unified Analytics option.

4. Click Update.
5. Refresh the Virtual Warehouses by going to Overview Virtual Warehouses  Refresh .

Results

The User Groups drop-down menu will no longer be available on the **New Virtual Warehouse creation** tile.

Related Information

[Creating a group in Management Console](#)

[Authenticating users in CDW Private Cloud](#)

List of configurations copied from the base cluster to CDW on Private Cloud

The Cloudera Data Warehouse (CDW) data service on Private Cloud has different configurations than the base cluster. When you activate an environment in CDW, configurations such as default file format, compression type, and transactional type are copied from the base cluster to CDW by default. This enables workload migration from base clusters to CDW data service.

Understanding the scenarios in which the configurations are copied from base to CDW

If you upgrade the platform from 1.5.0 to 1.5.1, for example, then the configuration of an existing environments stays the same as before. The configurations are not copied from the base cluster. To copy configurations from the base cluster, you must reactivate the environment.

On CDW environments that have received the base cluster configurations: If you change the configurations on the base cluster, refresh the Virtual Warehouse to obtain the updates base-cluster configurations by clicking  Refresh on the Virtual Warehouse tile.



Important: If you change any Database Catalog or Virtual Warehouse configuration on the CDW web interface, then these configurations are not overwritten with the configurations from base cluster even after refreshing the Virtual Warehouse.

The CDW web interface displays all the current configurations. If the Impala or the Hive on Tez service does not exist on the base cluster or, if the specific configuration is empty on the base cluster, then the default values from the Virtual Warehouse are used.

If you do not want to use the base cluster configuration, then you can disable the Copy configurations from base cluster to CDW option from the Advanced Configurations Advanced Settings page before activating the environment.

The following table provides the list of base cluster Impala configurations that are be copied to CDW upon activating the environment:

Table 1: Base cluster configuration for Impala

Configuration category	Base cluster configuration	Description
Default query option (default_query_options)	default_file_format	The default file format for the CREATE TABLE statement, for example Parquet. The default value is Parquet.
	default_transactional_type	The default transactional type, for example insert_only or none. Creates insert-only ACID tables by default. Does not apply to external tables. Default value is insert_only.

Configuration category	Base cluster configuration	Description
	timezone	Defines the timezone used for conversions between UTC and the local time. If not set, Impala uses the system time zone where the coordinator Impalad runs. As query options are not sent to the Coordinator immediately, the timezones are validated only when the query runs.
	parquet_array_resolution	Controls the behavior of the indexed-based resolution for nested arrays in Parquet.
	parquet_fallback_schema_resolution	Allows Impala to look up columns within Parquet files by column name, rather than column order, when necessary. The allowed values are: POSITION (0) and NAME (1).
	allow_erasure_coded_files	Enables or disables the support for erasure coded files in Impala. The default value is false. When set to false, Impala returns an error when a query requires scanning an erasure coded file.
	max_row_size	Ensures that Impala can process rows of at least the specified size. Applies when constructing intermediate or final rows in the result set. Used to prevent out-of-control memory use when accessing columns containing huge strings.
	compression_codec	The underlying compression for Parquet data files when Impala writes them using the INSERT statement.
Timeout options	idle_query_timeout	Sets the idle query timeout value for the session, in seconds. It is copied from the base cluster if it is greater than 0. If this option is not set on the base cluster, then the default value is 600.
	idle_session_timeout	The time in seconds after which an idle session is cancelled. It is copied from the base cluster if it is greater than 0. If this option is not set on the base cluster, then the default value is 1200.
TLS/SSL version and ciphers	ssl_minimum_version	Controls the allowed versions of TLS/SSL used by Impala. Starting with Impala 4.0, the default value is tlsv1.2.
	ssl_cipher_list	Used to specify the allowed set of TLS ciphers that are used by Impala.

The following table provides the list of base cluster Hive on Tez configurations that are copied to CDW upon activating the environment:

Table 2: Base cluster configuration for Hive on Tez

Base cluster configuration	Description
hive.create.as.insert.only	Used to specify whether the eligible tables should be created as ACID insert-only tables by default. Does not apply to external tables that use storage handlers. If this property is not set on the base cluster, then the default value is true.
hive.create.as.acid	Used to specify whether the eligible tables should be created as full ACID tables by default. Does not apply to external tables that use storage handlers. If this property is not set on the base cluster, then the default value is true.

Base cluster configuration	Description
hive.default.fileformat	The default file format for the CREATE TABLE statement. The default value is TextFile.
hive.default.fileformat.managed	The default file format for the CREATE TABLE statement applied to the managed tables only. External tables are created with default file format. The default value is ORC.
hive.local.time.zone	Sets the timezone for displaying and interpreting time stamps. If the value of this property is either set to LOCAL, is not specified, or is an incorrect timezone, then the system default timezone is used.
hive.external.table.purge.default	If set to true, it sets external.table.purge=true on the newly created external tables, which indicates that the table data should be deleted when the table is dropped. If set to false, it maintains the existing behavior in which the external tables do not delete data when the table is dropped.

Disabling copy configuration from base cluster to CDW option on Private Cloud

When you activate an environment, configurations such as default file format, compression type, and transactional type are copied from the base cluster to Cloudera Data Warehouse (CDW) by default; this can ease workload migrations. You must disable this feature from the Advanced Settings page before activating an environment in CDW.

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to [Advanced Configurations Advanced Settings](#) page.
3. Deselect the Copy configurations from base cluster to CDW option.
4. Click Update.

Enabling workload-aware autoscaling for Impala in CDW on Private Cloud

Using workload-aware autoscaling, you can configure multiple executor groups within a single Virtual Warehouse that can independently autoscale to allow handling of different workloads in the same Virtual Warehouse. To use workload-aware autoscaling, you must enable it from the Cloudera Data Warehouse (CDW) UI before creating a Virtual Warehouse.

About this task



Note: This feature is in technical preview and not recommended for production deployments. Cloudera recommends that you try this feature in development or test environments.

Procedure

1. Log in to CDW as DWAdmin.
2. Click [Advanced Configurations](#).
3. Select the [Enable workload-aware autoscaling for Impala](#) option.

4. Click Update.

The use workload-aware autoscaling option is available in the Size drop-down menu when you create a new Impala Virtual Warehouse.