Cloudera Data Warehouse on premises 1.5.5

Managing Cloudera Data Warehouse on premises

Date published: 2020-08-17
Date modified: 2025-11-08

https://docs.cloudera.com/

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

Upgrading Database Catalogs and Virtual Warehouses in Cloudera Data Warehouse on premises

After you upgrade the Cloudera Data Services on premises platform, you must upgrade the Database Catalog and Virtual Warehouses in Cloudera Data Warehouse. 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 Cloudera Data Warehouse uses a Hive MetaStore (HMS) instance. The Virtual Warehouses use Apache Hive, Apache Impala, and Hue runtime images that are used in Cloudera Data Warehouse. These runtime images are different than those used on Cloudera on premises. With every new Cloudera Data Services on premises release, you get a new version of Apache Hive, Apache Impala, and Hue runtimes with Cloudera Data Warehouse, which includes new features and fixes.

Supported upgrade path for an in-place upgrade

In-place upgrade option is available only for upgrades from Cloudera Data Services on premises 1.5.1 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 Cloudera Data Services on premises platform.
2. Log in to the Data Warehouse service as PowerUser.
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 Cloudera Data Warehouse on premises, you can upgrade Database Catalogs and Virtual Warehouses only to the latest available version.

What changes after the upgrade

- The ability to create custom Database Catalogs has been removed. After you upgrade to Cloudera Data Services on premises 1.5.4, you can no longer create new custom Database Catalogs. The existing custom Database Catalogs remain until you deactivate the environment. You can continue to upgrade, refresh, and rebuild the existing Database Catalogs.
- Custom pod configurations that you have created before upgrading from the Cloudera Data Services on premises 1.5.3 release to a newer release are migrated to the new resource templates as read-only settings after the upgrade. You can view the pod configurations from the **Resource Templates** page. You can use these as is while creating a Virtual Warehouse or modify them by creating a copy.
- The Keep current image version option has been removed from the Cloudera Data Warehouse web interface. When you rebuild the Database Catalog or Virtual Warehouse, they always retain the image version.

- In existing Data Visualization connections, you must rename the proxy user (delegation user) to "impala" user. You can rename it manually or refresh, upgrade, or rebuild the Virtual Warehouse or the Data Visualization instance after upgrading to Cloudera Data Services on premises 1.5.4.
- If you had enabled the setting to copy the base cluster configurations to Cloudera Data Warehouse on the Cloudera Data Services on premises 1.5.1 cluster, then the base cluster configurations will continue to get copied to Cloudera Data Warehouse after the upgrade. However, you can disable this setting from the **Advanced Settings** page.
- Starting with Cloudera Private Cloud Data Services 1.5.1, Data Analytics Studio (DAS) has been deprecated and completely removed from Cloudera Data Warehouse. 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 Cloudera Data Warehouse.
- Starting with Cloudera Private Cloud Data Services 1.5.0, Hue in Cloudera Data Warehouse requires WebHDFS to be enabled on the Cloudera Base on premises 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 on premises
Upgrading Cloudera Data Services on premises on Embedded Container Service
Upgrading Cloudera Data Services on premises on OpenShift Container Platform
Activating OpenShift environments on Cloudera Data Warehouse
Activating Embedded Container Service environments in Cloudera Data Warehouse

Refreshing environments, Database Catalog, and Virtual Warehouses in Cloudera Data Warehouse on premises

Learn when to refresh environments, Database Catalog, and Virtual Warehouses in Cloudera Data Warehouse on premises and understand the difference between the refresh and rebuild operations.

Where is the refresh option in Cloudera Data Warehouse?

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

Note: If you change any Database Catalog or Virtual Warehouse configuration on the Cloudera Data Warehouse 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 Cloudera Data Warehouse **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 Cloudera Data Warehouse
  - 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, Cloudera Data Warehouse 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 Cloudera Data Warehouse on premises

You can access advanced configurations in Cloudera Data Warehouse on premises from the left navigation pane on the Cloudera Data Warehouse 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
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 Cloudera Data Warehouse.
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 Cloudera Data Warehouse	Configurations such as default file format, compression type, and transactional type are copied from the base cluster to Cloudera Data Warehouse to aid cluster setup and workload migration.	Yes	Refresh*
Back up Virtual Warehouse namespaces before an upgrade	Cloudera Data Warehouse backs up namespace-related data using the Data Recovery Service before upgrading the Virtual Warehouse.	Yes	Refresh*
Store logs on HDFS	Enables you to store Cloudera Data Warehouse logs to HDFS on the base cluster.	Yes	Refresh*

* 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	You can enable or disable quota management at any time during the lifetime of your Cloudera Data Warehouse environment.
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	Enable before activating an environment in Cloudera Data Warehouse.

Supported object storage services for Cloudera Data Warehouse on premises

HDFS is the default storage system for Cloudera Data Warehouse. However, you can enable Cloudera Data Warehouse to access object storage such as AWS S3 and Azure Data Lake Storage (ADLS Gen1 and Gen2) if the Cloudera Base on premises 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 Cloudera Data Warehouse, 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 Cloudera Data Warehouse to establish a connection to S3/ADLS.

[?] Cluster validation includes port validation, and the Kerberos keytab configuration validation, and Root CA certificate validation.

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

If you have installed the Cloudera Data Services on premises, including Cloudera Data Warehouse, 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 Cloudera Data Warehouse'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 Cloudera Data Warehouse

You can enable Cloudera Data Warehouse data service on Cloudera on premises to access S3 and S3-compatible object storage if the Cloudera Base on premises 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 Cloudera Data Services on premises, including Cloudera Data Warehouse, 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 Cloudera Data Warehouse.

Enabling ADLS storage providers in Cloudera Data Warehouse

You can enable Cloudera Data Warehouse data service on Cloudera on premises to access Azure Data Lake Storage (ADLS Gen1 and Gen2) object storage if the Cloudera Base on premises 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 Cloudera Data Services on premises, including Cloudera Data Warehouse, 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 on premises

Apache Ozone is an object store available on the Cloudera Base on premises 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 on premises.

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 Cloudera Base on premises cluster

To access and use Ozone from Cloudera Data Warehouse on premises, 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 Cloudera Data Warehouse 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=[***USERNA
   ME***] 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://vvslab/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***]://[***SE
   RVICE-ID]/[***VOLUME-NAME***]/[***BUCKET-NAME***]
   ```

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

   Important: You must have the GET_METADATA and GENERATE_EEK permissions on the encryption key to create encrupted 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://vvslab/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 Cloudera Data Warehouse.

   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 Cloudera Data Warehouse to use Hive or Impala to query data residing on the Apache Ozone object store.

Related Information

Enabling Kerberos Authentication for Cloudera

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 Cloudera Data Warehouse before installing the Ozone service on the base cluster, then you must recreate the Cloudera Data Warehouse environment so that Ozone configurations can be imported into Cloudera Data Warehouse.

By default, the Hive MetaStore (HMS) for Database Catalogs on Cloudera Data Warehouse on premises 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, ...)];
```

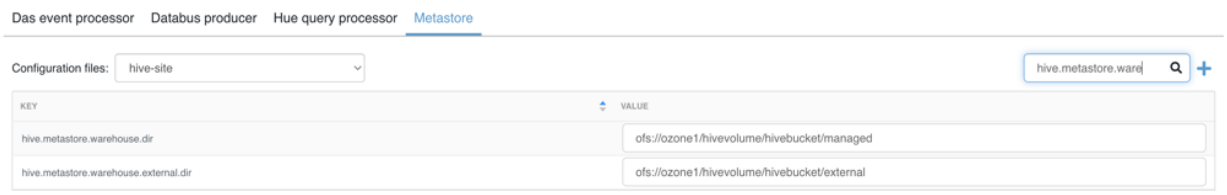
Before you re-configure the Database Catalog settings, make sure there are no running Virtual Warehouses associated with it. Either the Database Catalog has no associated Virtual Warehouses or you have suspended all the Virtual Warehouses associated with it.

Procedure

1. Log in to the Data Warehouse service as a DWAdmin.
2. Activate an environment in Cloudera Data Warehouse.
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:



   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 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 Cloudera Base on premises 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 Cloudera Data Warehouse when you create a Virtual Warehouse.
3. Verify that the Ozone configurations are present in Cloudera Data Warehouse. 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 on premises

This section describes how to configure Cloudera Data Warehouse on premises to store Hive and Impala logs on Ozone storage.

You can configure Cloudera Data Warehouse to store Hive and Impala logs on Cloudera on premises 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.
- Cloudera on premises 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 Cloudera Data Warehouse 1.4.1. Cloudera recommends that you use Ozone with Cloudera Data Warehouse in test and development environments. It is not recommended for production deployments.

Specify or create an Ozone bucket for Cloudera Data Warehouse on premises logs

This topic describes how to specify an Ozone bucket to store Cloudera Data Warehouse on premises Hive and Impala logs.

About this task

You can either re-use the Ozone bucket that is automatically configured for storing Cloudera AI Private Cloud logs or create a new bucket to store Cloudera Data Warehouse logs separately. The Ozone bucket used to store Cloudera AI logs usually has a cdplogs- prefix.

Procedure

Use one of the following two methods depending on whether you want to use the existing Cloudera AI log bucket or create a new one for Cloudera Data Warehouse:

- To select an existing Ozone bucket, use the ozone sh bucket list command from the Ozone shell on your Cloudera Base on premises 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 Cloudera Base on premises 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 on premises log configuration to point to Ozone

This topic describes how to configure Cloudera Data Warehouse on premises to store logs on Ozone.

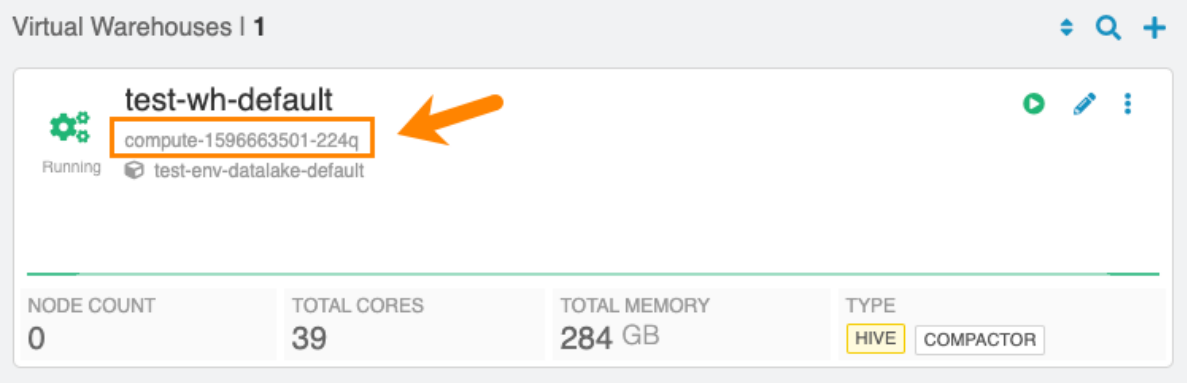
About this task

To configure Cloudera Data Warehouse on premises 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 Cloudera Data Warehouse 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 Cloudera Data Warehouse on premises service.
  2. Locate the Virtual Warehouse you want to configure log storage for in the right-most column of the page, and locate the Cloudera Data Warehouse 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 Cloudera Data Warehouse on premises 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 Cloudera Base on premises OpenShift cluster.

Procedure

1. Using OpenShift commands, view the OpenShift project for the pod where the Cloudera Data Warehouse on premises instance is running by specifying the Cloudera Data Warehouse namespace for the Virtual Warehouse that you noted in the Before you begin section above.

   For example, if the Cloudera Data Warehouse 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 Cloudera Data Warehouse 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 on premises logs on Ozone storage

This topic describes how to monitor Cloudera Data Warehouse on premises 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 Cloudera Data Warehouse 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 Cloudera Data Warehouse logs.

Procedure

Use one of the following methods to monitor Cloudera Data Warehouse 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 Cloudera Data Warehouse 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 Cloudera Data Warehouse logs are stored.
  - <*WAREHOUSE-PREFIX*> indicates the Virtual Warehouse identifier.

Analyze Cloudera Data Warehouse on premises logs stored on Ozone

This topic describes how to analyze Cloudera Data Warehouse on premises 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 warehouse-level access control for Hive and Impala in Cloudera Data Warehouse on premises

Cloudera Data Warehouse 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 Cloudera Data Warehouse 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: You cannot use the warehouse-level access control feature if you have enabled SAML authentication in the Cloudera Control Plane.

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 Cloudera 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 Cloudera Management Console

Authenticating users in Cloudera Data Warehouse on premises

List of configurations copied from the base cluster to Cloudera Data Warehouse on premises

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

Understanding the scenarios in which the configurations are copied from base to Cloudera Data Warehouse

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 Cloudera Data Warehouse 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 Cloudera Data Warehouse web interface, then these configurations are not overwritten with the configurations from base cluster even after refreshing the Virtual Warehouse.

The Cloudera Data Warehouse 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 Cloudera Data Warehouse 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 Cloudera Data Warehouse 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.
	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.

Configuration category	Base cluster configuration	Description
	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 Cloudera Data Warehouse 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.
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.

Base cluster configuration	Description
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 Cloudera Data Warehouse option

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 by default; this can ease workload migrations. You must disable this feature from the Advanced Settings page before activating an environment in Cloudera Data Warehouse.

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 Cloudera Data Warehouse option.
4. Click Update.

Enabling workload-aware autoscaling for Impala in Cloudera Data Warehouse on premises

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 UI before creating a Virtual Warehouse.

About this task

Procedure

1. Log in to Cloudera Data Warehouse 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 Imapla Virtual Warehouse.

Quota management in Cloudera Data Warehouse on premises

Review how to enable quota management in Cloudera Data Warehouse on premises and assign quota-managed resource pools to Database Catalogs, Virtual Warehouses, and Cloudera Data Visualization instances.

How quota management works in Cloudera Data Warehouse

A namespace is created when you create a Database Catalog, a Cloudera Data Visualization instance, or a Virtual Warehouse. At this stage, a resource pool is created with the necessary resources under the selected parent resource pool. You can view the resource pools on the **Resource Utilization** page by going to the Cloudera Management Console Resource Utilization Quotas tab. Virtual Warehouse namespaces allocate a minimum resource quota, known as the Guaranteed Quota in the UI, to ensure the operation of at least one executor group.

If you have configured auto-scaling of the executor groups for a Hive or an Impala Virtual Warehouse and the Virtual Warehouse needs to schedule more executors for the current load, Cloudera Data Warehouse tries to increase its resource pool's quota. Auto-scaling fails if the requested resource pool quota exceeds the parent resource pool's quota.

If you have configured auto-scaling of the executor groups for a Hive or an Impala Virtual Warehouse, the requested quota is the resources that are used by the maximum number of executors groups. During a scale-out event, new executor groups are created, and Yunikorn schedules them based on the availability of resources at that moment.

On activating an environment, a root.<environment-name>.cdw resource pool is created by default. The created resource pool is a top-level resource pool that can be assigned to any of the Cloudera Data Warehouse entities that are created under the associated environment. The newly created root.<environment-name>.cdw resource pool has infinite quotas set that can be modified by an administrator from Cloudera Management Console Resource Utilzation Quotas .

During environment activation, a monitoring namespace and a separate resource pool are also created under the root.<environment-name>.cdw resource pool for the Diagnostic Data Generation jobs regardless of whether Quota Management is enabled or not. This resource pool is a leaf level resource pool associated with the monitoring namespace.

The Database Catalog inherits the root.<environment-name>.cdw resource pool from the environment upon activation.

Note: When creating a Cloudera Data Warehouse component, such as Virtual Warehouse, Database Catalog, Cloudera Data Visualization instance, or Log Router, the Quota Management service checks whether sufficient resources are available. The component is created only if the required resources exist. For more information about available quota, see *Managing cluster resources using Quota Management*.

If the Store Logs on HDFC advanced configuration is enabled, the created log router namespace is enrolled into Quota Management based on the state of Quota Management in Advanced Configurations.

What happens when the quota is insufficient

Suppose you are creating a new Cloudera Data Warehouse entity, such as a Virtual Warehouse with a specific resource pool. The Virtual Warehouse creation process fails if there are insufficient resources when a namespace is being created.

Cloudera CLI support

By using the --resource-pool Cloudera CLI option of the type string, you can specify the resource pool for the Cloudera Data Warehouse entities, namely the Virtual Warehouse and the Cloudera Data Visualization instance, when using Cloudera CLI commands.

Related Information

Managing cluster resources using Quota Management

Adding resource pools after Cloudera Data Warehouse upgrade

You can manually add resource pools for Cloudera Data Warehouse entities that are not yet enabled for quota management, including Virtual Warehouse, Database Catalog, and Cloudera Data Visualization, if the entities did not have resource pools enabled following an upgrade from Cloudera Data Warehouse 1.5.5 to 1.5.5 SP1.

About this task

Important: You cannot change the resource pools of Cloudera Data Warehouse entities that are already enabled for quota management.

Procedure

1. Log in to the Cloudera Data Warehouse service as DWAdmin.
2. Go to the Cloudera Data Warehouse entitys details page by clicking ⋮ Upgrade .
3. Select a resource pool from the Resource Pool drop-down menu.
4. Click Apply Changes.

Note: The operation succeeds only if there is sufficient CPU and memory for the common pods, the executors for a workload, and the Cloudera Data Warehouse entity fits into the resource pool.

Resource templates for Cloudera Data Warehouse on premises pods

Cloudera Data Warehouse on premises provides Kubernetes resource templates for Hive, Impala, Data Visualization, and Database Catalogs. You can use an available resource template or create a custom one.

When you create any Cloudera Data Warehouse entity such as a Virtual Warehouse, Data Visualization instance, or Database Catalog, Cloudera Data Warehouse allocates standard resources to these instances suitable for most workloads. By using custom resource templates, you can also change the resources used by the critical subcomponents of a service, such as the coordinators, executors, catalog daemons, usage monitor, and so on, to pack a particular number of pods into a Kubernetes node or to create extra-large daemons to handle specific workloads.

Post-upgrade behavior

Custom pod configurations you created before upgrading from the Cloudera Data Services on premises 1.5.3 release to a newer release are migrated to the new resource templates as read-only settings after the upgrade. You can view the pod configurations from the **Resource Templates** page. You can use these as is while creating a Virtual Warehouse or modify them after creating a copy.

Note: After upgrading to the latest Cloudera Data Services on premises release, for the existing Virtual Warehouses that were associated with a pod configuration, you see that they are now associated with the default Hive or Impala resource template on the **Virtual Warehouses Details** page. However, in the backend, they continue to use the resources that were defined in the pod configuration as it was before the upgrade. In other words, the Cloudera Data Warehouse web interface shows a switch to the default resource template after the upgrade, but the Virtual Warehouses are still functioning with their pre-upgrade resource allocation.

List of predefined resource templates

Cloudera Data Warehouse provides predefined resource templates for Hive, Impala, Database Catalog, and Data Visualization pods. The Default Resources template is used whenever you create a new Virtual Warehouse and Data Visualization instance, or edit an existing Database Catalog. Medium and Large Resources are double and triple the default preset respectively. For initial exploration and proof of concept use cases, you can use the Reduced Resources template. The predefined resource templates are read-only.

Cloudera Data Visualization

Table 3: Default resources for Cloudera Data Visualization subcomponents

Resource type	Resource limit for webapp
CPUs	2
Memory	8192 MB

Database Catalog

Table 4: Default resources for Database Catalog subcomponents

Resource type	Resource limit	
	Hive Query Processor	HMS
CPUs	2	4
Memory	8192 MB	24576 MB
Xmx	-	6432 M

Hive

Table 5: Default resources for Hive subcomponents

Resource type	Resource limit				
	HS2	Hue backend	Query coordinator	Query executor	Sta op
CPUs	4	1	1	12	0.2
Memory	16384 MB	8192 MB	4096 MB	116736 MB	204
Scratch	-	-	-	280 GiB	-
Cache	-	-	-	280 G	-
Overhead size	-	-	-	40 GiB	-
Xms	8 G	-	2 G	24 G	-
Xmx	11468 M	-	2457 M	48 G	-
Xss	-	-	-	512 k	-
Max Direct Memory Size	-	-	-	64 G	-
Wait Queue size	-	-	-	10	-

Impala

Table 6: Default resources for Impala subcomponents

Resource type	Resource limit							
	Huebackend	autoscaler	catalogd	Usage monitor	Impala coordinator	Impala executor	Impala proxy	statestored
CPUs	1	1	1	0.1	14	14	1	1
Memory	8192 MB	1024 MB	8192 MB	256 MB	112640 MB	116736 MB	1024 MB	1024 MB
Scratch	-	-	-	-	300 GiB	300 GiB	-	-
Cache	-	-	-	-	200 GiB	200 GiB	-	-

Resource type	Resource limit							
	Huebackend	autoscaler	catalogd	Usage monitor	Impala coordinator	Impala executor	Impala proxy	statestored
Overhead size	-	-	-	-	58 GiB	58 GiB	-	-
Xms	-	-	2 G	-	2 G	2 G	-	-
Xmx	-	-	4 G	-	25 G	4 G	-	-
Xss	-	-	-	-	-	-	-	-
Max Direct Memory Size	-	-	-	-	-	-	-	-
Wait Queue size	-	-	-	-	-	-	-	-

Trino

Table 7: Default resources for Trino subcomponents

Resource Type	Hue backend	Trino Coordinator	Trino Worker
CPUs	1	14	14
Memory	8192 MB	116736 MB	116736 MB
Xms	-	64 G	64 G
Xmx	-	100 G	100 G

Note: Trino is in Technical Preview and is not ready for production deployments. Cloudera recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

Unified Analytics

Table 8: Default resources for Hive and Impala subcomponents required to run Unified Analytics

Resource type	Resource limit			
	HS2	Query coordinator	Query executor	Standalone query executor
CPUs	4	1	12	14
Memory	16384 MB	4096 MB	116736 MB	120832 MB
Scratch	-	-	280 GiB	280 GiB
Cache	-	-	280 GiB	280 GiB
Overhead size	-	-	40 GiB	40 GiB
Xms	8 G	2 G	24 G	49 G
Xmx	11468 M	2457 M	48 G	98 G
Xss	-	-	256 k	256 k
Max Direct Memory Size	-	-	64 G	20 G
Wait Queue size	-	-	10	10

Glossary

Memory

Physical Random Access Memory (RAM) available on a node (also called a worker machine) in a cluster. This memory resource is crucial for running pods in Kubernetes.

CPU

CPU (Central Processing Unit) refers to the processing power available on a node in the cluster.

Xmx

Maximum memory allocation pool for a Java Virtual Machine (VM).

Xms

Initial memory allocation pool for a Java VM.

Xss

Java VM configuration for -Xss (thread stack size).

Cache size

Size of the data cache.

Scratch size

Limit of Impala scratch space.

Overhead size

Size for resources used by tools run by the containers.

Max Direct Memory Size

Java VM configuration for -XX:MaxDirectMemorySize (limit for Direct Byte Buffers).

Wait queue size

Overhead buffer for hive.query.isolation.slots.per.node. The total number of concurrent tasks the Hive query executor can process is the sum of the available CPU cores and the value you specify in this field.

Creating custom resource templates in Cloudera Data Warehouse on premises

Suppose you have customized the node sizes in your cluster to suit a particular workload. In that case, you can customize the number of hardware resources allocated to the pods to suit your custom-sized node or workload to optimize performance or to control resource usage in the environment.

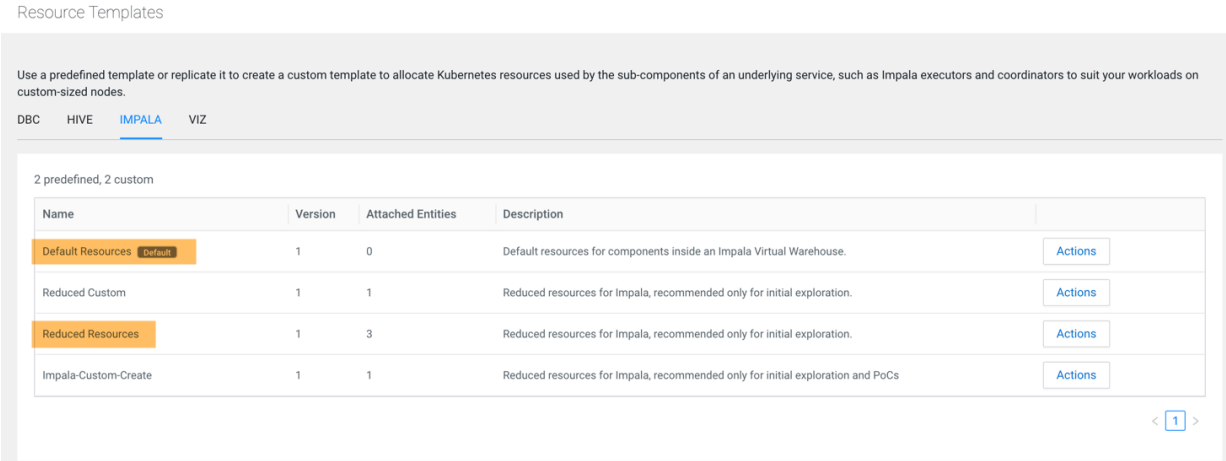
Before you begin

You must have already activated an environment in Cloudera Data Warehouse.

Procedure

1. Log in to the Cloudera Data Warehouse service as a DWAdmin.
2. Click on the Resource Templates menu on the left navigation pane.

3. Go to the respective tab for which you want to create a custom resource template. For example, IMPALA.
   Cloudera Data Warehouse displays the predefined resource templates.



4. Click Actions corresponding to the resource template you want to customize and click Make A Copy.
5. Specify a name for your template in the Name field. Optionally, specify a description and click Make A Copy.
6. Select the Set As Default option to make this resource template the default template for all new Cloudera Data Warehouse entities; in this case all new Impala Virtual Warehouses.
7. Specify or change the values of the subcomponents as needed and click Apply Changes.
   The Config creation initiated message is displayed.
8. Refresh the page on your browser to view the newly created resource template.

Results

The new pod configuration becomes available in the Resource Template drop-down menu on the Cloudera Data Warehouse entity's details page. To go to the details page, click ⋮ Edit .

Modifying a custom resource template in Cloudera Data Warehouse on premises

You can modify a custom resource template from the Cloudera Data Warehouse UI. When you modify a template that is associated with an instance, a new resource template is created while retaining the older template. You must manually apply the modified template to the respective Virtual Warehouse, Data Visualization, or Database Catalog instance.

About this task

If you modify a resource template associated with an instance by changing the resource values, a new version of the resource template is created. A new version is not created if you only change the name or the description. The new version of the resource template is automatically assigned to that Cloudera Data Warehouse entity.

If a resource template is not associated with a Cloudera Data Warehouse entity, then its older version is not displayed on the **Resource Templates** page.

Procedure

1. Log in to the Cloudera Data Warehouse service as a DWAdmin.
2. Click on the Resource Templates menu on the left navigation pane.
3. Go to the respective component tab for which you want to modify a custom resource template. For example, IMPALA.

4. Click Actions corresponding to the resource template you want to modify.
5. Modify the required values and click Apply Changes.
   The resource template is displayed with a new version number on the **Resource Templates** page.

Deleting a custom resource template in Cloudera Data Warehouse on premises

You can delete a custom resource template when it is no longer needed. However, you cannot delete a template associated with a Cloudera Data Warehouse entity such as a Virtual Warehouse, Data Visualization, or Database Catalog instance.

Before you begin

- Dissociate the custom resource template from the Virtual Warehouse, Data Visualization, or Database Catalog instance by selecting a predefined or any other custom resource template on the corresponding details page. To go to the details page, in the Cloudera Data Warehouse service, click ⋮ Edit .
- Ensure that the template you want to delete is not set as a default template.

Procedure

1. Log in to the Cloudera Data Warehouse service as a DWAdmin.
2. Click on the Resource Templates menu on the left navigation pane.
3. Go to the respective tab for which you want to create a custom resource template. For example, IMPALA.
4. Click Actions corresponding to the resource template you want to delete and click Delete.

Trino federation connectors

Trino is a distributed SQL query engine designed from the bottom up to be built around the concept of connectors and federation. Trino connectors help you connect to and access data from a variety of remote data sources, expose metadata (exposed within Trino as catalogs), and handle sending or receiving data from the remote source.

Note: This feature is in Technical Preview and is not ready for production deployments. Cloudera recommends trying this feature in test or development environments and encourages you to provide feedback on your experiences.

You can use Cloudera Data Warehouse to configure a connector for a data source, enabling a Trino Virtual Warehouse to access the data source. Cloudera enables you to configure connectors for the following data sources:

- PostgreSQL
- MySQL
- Snowflake
- AWSRedshift
- Hive
- Iceberg
- Oracle
- MariaDB

When you create a connector, a template specific to the selected data source type is provided for you to specify the connector configuration details, such as connector URL and secrets that are required to access the data source. You can also choose a "default template", which enables you to configure data source or connector types that are supported by the open-source Trino offering, however, these connectors are not supported by Cloudera.

Related Information
Trino Connectors

Creating a federation connector

Learn how you can create a federation connector that can be used by a Trino Virtual Warehouse to query and access data from different data sources.

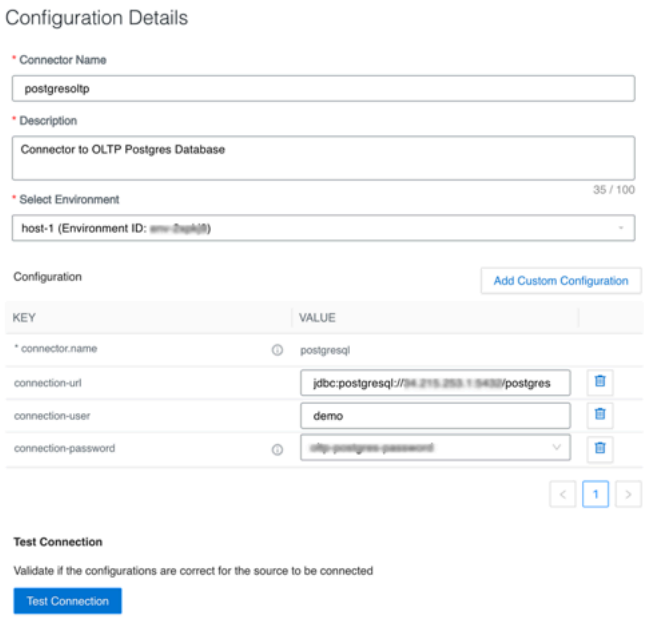
Before you begin

- You must create a Trino Virtual Warehouse.
- You must register a secret for your Environment.

Procedure

1. Log in to the Cloudera Data Warehouse service and click Federation Connectors.
   The **Federation Connectors** page is displayed that lists all the currently configured connectors.
2. Click Create Data Source to create a new connector.
3. In the **Data Source Type** page, select the type of data source, such as PostreSQL, MySQL, Snowflake, AWSRedshift, or Default template that you want to configure and then click Next.
4. In the **Configuration Details** page, enter the following information:
   a. Provide a name and description for the connector.
   b. Select the appropriate environment.
   c. Enter the appropriate URL to connect to the required data source. For example, jdbc:postgresql://example.net:3306/postgres
   d. Enter the connection username and select a registered secret for the password.
   e. Click Test Connection to verify if the configurations are correct to establish a successful connection to the data source.

      A message is displayed indicating if the connection is successful or not. If the connection fails, modify the configuration details and try again.



5. If the connection is successful, click Next to proceed to the **Review And Connect** page.
   This page provides a summary of the connector and its configuration.

6. Click Connect To Datasource to establish a connection to the data source

Results
The connector is created successfully and is listed in the **Federation Connectors** page.

Related Information
Adding a new Virtual Warehouse
Registering secrets

Modifying a federation connector

Learn how you can modify the configuration details of an existing connector.

Procedure

1. Log in to the Cloudera Data Warehouse service and click Federation Connectors.
   The **Federation Connectors** page is displayed that lists all the currently configured connectors.
2. For the connector that you want to modify, click ⋮ Edit Configuration .
3. In the **General Details** tab, modify the required fields, such as name, description, connector configuration details, or associated secret.
4. Click Test Connection to verify if the modified configurations are correct to establish a successful connection to the data source.
5. If the connection is successful, click Apply Change.

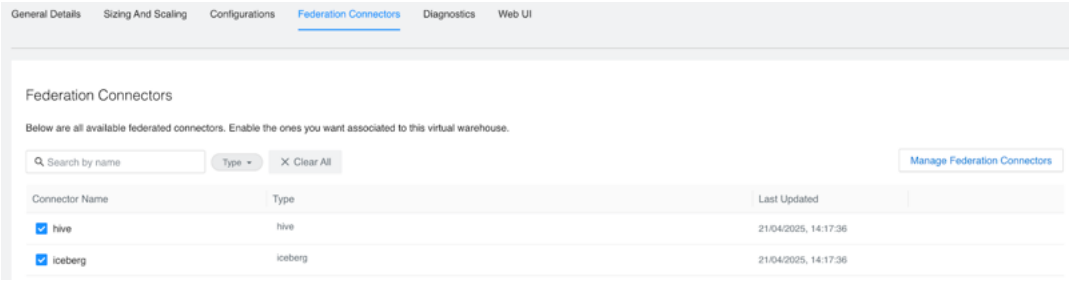
Associating connectors to a Virtual Warehouse

You can associate multiple connectors to a Trino Virtual Warehouse that allows you to access data from multiple data sources. Learn how you can edit a Trino Virtual Warehouse to associate or disassociate federation connectors.

Before you begin
You must have created one or more federation connectors.

Procedure

1. Log in to the Cloudera Data Warehouse service.
2. From the **Overview** page, click the **Virtual Warehouses** tab and click ⋮ Edit against the required Trino Virtual Warehouse.
3. In the **Virtual Warehouse Details** page, click the **Federation Connectors** tab.
   A list of available connectors are displayed. Connectors that are already associated with the Virtual Warehouse have the checkbox selected.



4. Select a connector that is not associated with the Virtual Warehouse.
   A message is displayed indicating that you have chosen to add the connector to the Virtual Warehouse.

   Tip: You can clear the selection against a connector to disassociate it with the Virtual Warehouse, however, you can no longer access the datasource using this connector.

5. Click Apply Changes.

Results
The connector is associated with the Virtual Warehouse and you can now query data from the respective data source.

Related Information
Creating a federation connector

Security management for federation connectors

The connector configuration details including credentials like user name and password are managed in a properties file by Trino. Cloudera Data Warehouse enables you to manage these credentials in a secure manner using secrets that can be stored within the native secret service.

Secrets allow you to access the data source using the configured connector in a secure manner and prevent unauthorized user access.

Registering secrets

Learn how you can register secrets for your environment that allows you to securely access your data sources using connectors.

Before you begin
You must have registered and activated your environment.

Procedure

1. Log in to the Cloudera Data Warehouse service.
2. From the **Overview** page, click the **Environments** tab and identify the Environment against which you want to register the secret, and then click ⋮ Edit .
3. In the **Environment Details** page, click the **SECRETS** tab.
   A list of secrets that are registered for this environment are displayed.
4. Click Create Secret and enter a name and value for the secret.
   A message is displayed indicating that you have chosen to register a secret.
5. Click Apply Changes.

Results
The registered secret is listed in the **Secrets** page and can be used while creating a connector.

Deregistering secrets

Learn how you can deregister secrets that are no longer required to be associated with connectors.

Procedure

1. Log in to the Cloudera Data Warehouse service.
2. From the **Overview** page, click the **Environments** tab and then click ⋮ Edit against the required Environment.

3. In the **Environment Details** page, click the **SECRETS** tab.

   A list of secrets that are registered for this environment are displayed.
4. For the secret that you want to deregister, click Deregister Secret.

   A confirmation dialog box is displayed with the list of connectors that are currently associated with the selected secret.
5. Enter the name of the secret in the text box to confirm deregisteration and then click Yes, Deregister.

   A message is displayed indicating that you have chosen to deregister the secret.
6. Click Apply Changes.

Results

The secret is deregistered from the Environment and is no longer associated with any connector.