Cloudera Data Warehouse Private Cloud 1.5.4

Cloudera Data Warehouse Private Cloud Environments

Date published: 2020-08-17
Date modified: 2024-05-30

CLOUDERA

Legal Notice

Contents

About Private Cloud environments

An environment is a logical entity that represents the association of your Private Cloud user account with compute resources. Using the compute resources of the environment, you can provision and manage Cloudera Data Warehouse (CDW), Cloudera Data Engineering (CDE), or Cloudera Machine Learning (CML) workloads.

For a CDP Private Cloud deployment, you can configure the environment on an OpenShift Container Platform or an Embedded Container Service (ECS) cluster. Deploying CDP Private Cloud Data Services (CDW, CDE, and CML) on OpenShift requires you to deploy and manage the Kubernetes infrastructure. ECS service creates and manages an embedded Kubernetes infrastructure for use with the CDP Private Cloud Experiences. To deploy CDP Experiences on ECS clusters, you only need to provide hosts on which you want to install the service and Cloudera Manager sets up an ECS cluster using an embedded Kubernetes service and provides a framework to manage and monitor the cluster.

Much of the installation and deployment configuration for private cloud is performed before you register environments using the Management Console. However, in addition to the steps described in the installation guide, you must perform additional configurations to activate an environment so you can use it with the CDW service.

The following diagram shows the components of CDP Private Cloud:
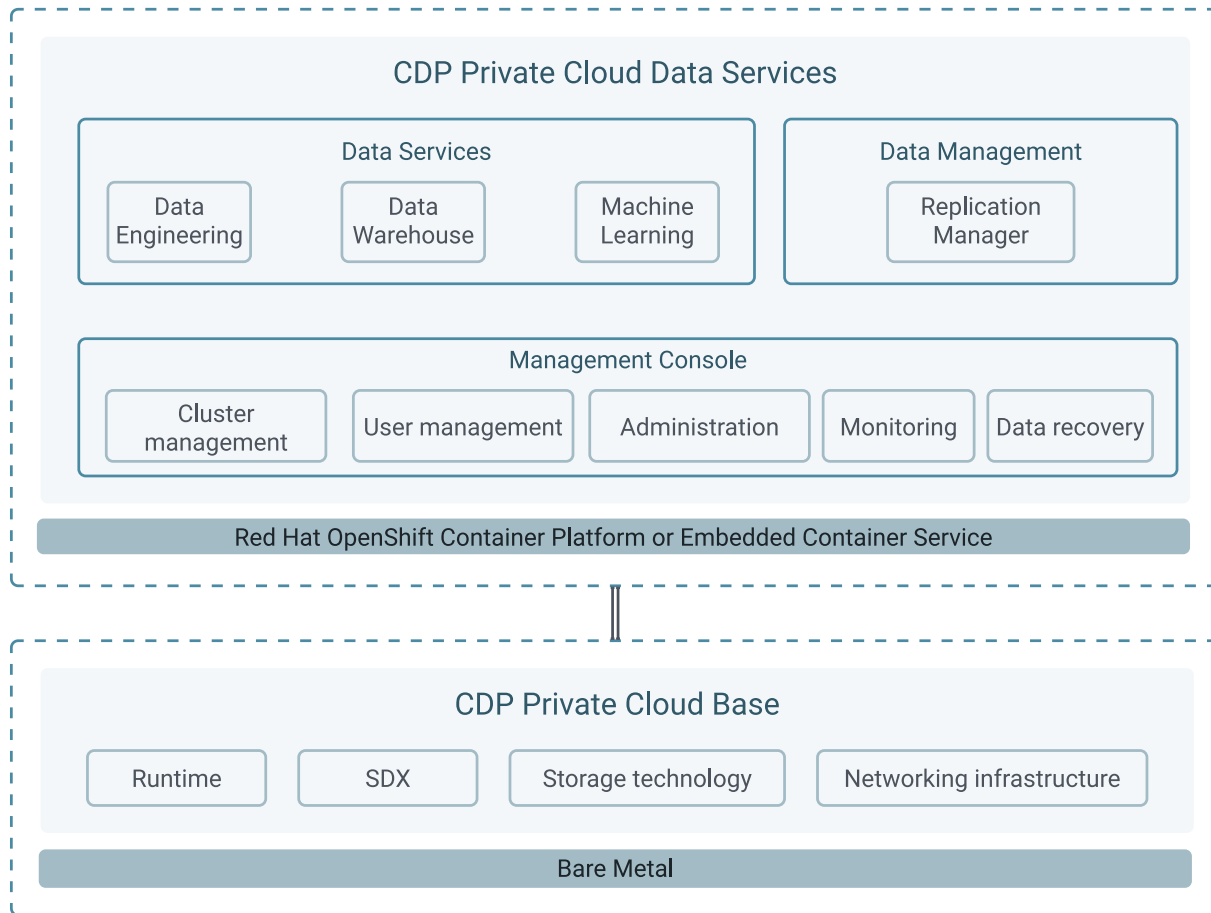
For more details about environment requirements and how to register an environment so CDP can access Kubernetes clusters on your OpenShift or ECS deployment, see CDP Private Cloud Environments.

Related Information
Activating OpenShift environments
Activating Embedded Container Service environments

Viewing environment details

You can view information about an activated environment, such as when it was created and last updated, or how many Database Catalogs and Virtual Warehouses use the environment from the Cloudera Data Warehouse (CDW) web interface.

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to the Environments tab on the **Overview** page.
3. Locate the environment you want to view and then click the ⋮ Edit .

   Environment details for the selected environment are displayed.

Scheduling executor pods on dedicated worker nodes in CDW Private Cloud

To optimize resource utilization on your cluster and improve query performance, Cloudera Data Warehouse (CDW) allows you to schedule Hive and Impala executor and coordinator pods on dedicated executor nodes that are tainted for CDW and equipped with more local storage, either using SSDs or NVMEs.

About this task

To block CDW pods, other than executors and coordinators to be scheduled on a dedicated worker node with higher local storage capacity, you must enable the Use dedicated nodes for executors option while activating an environment in CDW or by editing an existing environment. This allows nodes with available local storage to be reserved for CDW query executor pods over all other CDW or other data services pods marking them exclusively for CDW executor pods.

By default, the Use dedicated nodes for executors option is disabled. When this option is disabled, all CDW pods, including Hive MetaStore (HMS), Data Visualization, Hive and Impala executors and coordinators can be scheduled on the tainted worker nodes. When you enable this option, only Hive and Impala executor and coordinator pods can be scheduled on the tainted worker nodes.

To enable the Use dedicated nodes for executors option while activating the environment, see *Activating OpenShift environments* or *Activating Embedded Container Service environments*. This task describes how to enable the Use dedicated nodes for executors option on an existing environment.

Before you begin

On Embedded Container Service (ECS), you must dedicate the Embedded Container Service cluster nodes for specific workloads using Cloudera Manager, as described in the *Dedicating ECS nodes for specific workloads* section.

On OpenShift Container Platform (OCP), you must dedicate the OCP cluster nodes for specific workloads using the `kubectl taint` command, as described in the *Dedicate an SSD node for CDW workloads* section.

Procedure

1. Log in to the Data Warehouse service as DWAdmin.
2. Go to the **Environment(s)** tab.
3. Locate the environment on which you want to enable this feature and click ⋮ Edit .
4. Go to the **CONFIGURATIONS** tab and enable the Use dedicated nodes for executors option.

5. Click Apply Changes.

6. Refresh the associated Database Catalogs and Virtual Warehouses.

You can also upgrade or update the associated Database Catalogs and Virtual Warehouses depending on your needs.

Related Information

Dedicating ECS nodes for specific workloads

Dedicating OCP nodes for specific workloads

Dedicate a SSD node for CDW workloads

Activating OpenShift environments

Activating Embedded Container Service environments

Backup and restore in CDW Private Cloud

The backup/restore feature saves your environment parameters, making it possible to recreate your environment with the same settings, URL, and connection strings you used in your previous environment. Learn about the different backup and restore methods, objects and configurations that are included in the backup, and the backup method to use in different scenarios.

For cluster maintenance and integrity, you can back up the Cloudera Data Warehouse (CDW) cluster including its Kubernetes objects, persistent volumes, autoscaling configuration, and so on for various CDW entities such as the environment, Database Catalog, Virtual Warehouses, and Data Visualization instance. You can recreate the deployment by restoring the backed-up configurations and settings, as part of the planned maintenance or in a disaster recovery scenario.

Available backup and restore methods

There are two ways to backup and restore the Data Warehouse service:

• Using Data Recovery Service (DRS)
• Using the CDW's CDP CLI cluster management commands

Both these methods use CDP CLI commands. The CDP Management Console also provides a graphical user interface to perform backup and restore operations. See DRS automatic backups and Using DRS with CDW.

Deciding which backup method to use

Use DRS to back up all Private Cloud Data Services environment configurations for an ECS or Private Cloud Data Services upgrade. DRS takes a snapshot of the namespace on the Kubernetes cluster. You can also choose DRS if some Control Plane service fails and you need to reinstall it with a previously preserved configuration.

If you just want to back up your Database Catalog or Virtual Warehouse configurations to recreate CDW from an earlier configuration, use CDW's backup and restore commands.

Understanding what is backed up

There are differences in the cluster configurations and objects that get backed up when you create backups using DRS or Cloudera Data Warehouse (CDW)'s cluster management CDP CLI commands.

Use DRS to create on-demand backups of the Data Warehouse namespace, including Kubernetes objects, persistent volumes, and so on. See *Using DRS with CDW*. The output is backupCRN; where CRN is Customer Resource Number, a Cloudera-specific identifier for an environment.

Use the CDW's CDP CLI cluster management commands to create a backup from the configuration and settings, including all the connected Database Catalog, Virtual Warehouses, and Data Visualization instances. The following settings are included when you back up using the CDW's CDP CLI cluster management commands:

- All environment activation settings:

    - Storage Class (Openshift Container Platform)
    - Security Context Constraint (SCC) (Openshift Container Platform)
    - Low Resource Mode
    - Quota Management
    - Resource Pool
    - Dedicated Executor Nodes
    - mTLS client credentials
- All Database Catalog configurations that are associated with the environment. The Hue database is also backed up if you have created one.
- All configurations associated with the Hive and Impala Virtual Warehouses.
- All configurations associated with the Cloudera Data Visualization instances. The database that is associated with the CDV instance is also backed up so that the visualizations, dashboards, data connections, and so on are preserved.

Related Information

Using DRS with CDW

Understanding what is restored

The restore operation restores the state of the Data Warehouse service depending on the backup method.

If you have backed up Cloudera Data Warehouse (CDW) using the Data Recovery Service (DRS) CDP CLI command `dw create-backup`, then the `dw restore-backup` command restores the state of the Data Warehouse service from the backup.

If you have backed up CDW using CDW's CDP CLI cluster management commands (`dw backup-cluster`), then based on the existing state of the cluster, the restore process creates a workflow plan that decides whether to create, or skip the restore of a component. This workflow plan is returned as the response of the restore command, so you can see which components will be created, updated, or skipped during the restore.

About the restore command

CDW's restore command is as follows:

```
cdp dw restore-cluster --cluster-id <value> --data <value> [--cli-input-json
  <value>] [--generate-cli-skeleton]
```

The "data" field in the output of the `dw backup-cluster` command contains a base64-encoded zip file containing the backup data of the cluster which includes the environment activation settings, Virtual Warehouse and Cloudera Data Visualization (CDV) settings and configuration, and locations of CDV and Hue database backups on HDFS.

You can use the CDW's `dw restore-cluster` command in one of the following ways:

- By passing the environment's Cloudera resource name (crn) to activate the cluster from the backup file and restore all the entities and database contents.
- By passing an activated environment identifier to restore all the entities and database contents to the running environment. This method is useful when you need to change activation parameters, but it requires manual reactivation.

When you run the `dw restore-cluster` command, CDW:

1. Activates the environment using the settings from the backup and waits for the infrastructure to be created
2. Creates a default Database Catalog
3. Updates the Database Catalog configuration to apply custom configurations
4. Starts the Hue database restore job in the database catalog namespace asynchronously

5. Deploys the Virtual Warehouse instances
6. Deploys the Data Visualization instances
7. Starts the Cloudera Data Visualization (CDV) restore job in the individual namespaces asynchronously. It also restores the database associated with the CDV instance.

> **Note:** The CDV database restore fails for the first time as it tries to drop tables that do not exist. The CDV database restore succeeds the second time.

The restore process is designed to be an idempotent process. You can run it multiple times, if needed. If the environment is activated and healthy, you can run the restore operation multiple times to restore the Virtual Warehouse and Data Visualization objects. For every restore operation, the Hue database restore will run. This operation overwrites the Hue database contents. If a Virtual Warehouse or a Data Visualization object is not present on the cluster, but the backup file contains it, it is restored to the cluster. If such an entity is already deployed, no changes or configuration updates occur.

Prerequisites for CDW backup and restore in Private Cloud

Learn about the prerequisites that are mandatory for a successful backup and restore event of Cloudera Data Warehouse (CDW) Data Service.

• You must have an active CDW cluster with one Database Catalog and one or more Hive or Impala Virtual Warehouses.
• Only the default Database Catalog is backed up. Ensure that you do not have custom (non-default) Database Catalogs.
• You have installed and configured CDP CLI version 0.9.99 or higher on a host in your cluster from which you can access CDP Private Cloud Data Services.
• You have Cluster Administrator privileges and can access the CDW web interface.
• You must use the same CDW version to restore the files you used during the backup process.

Before you start the backup process, you must deactivate the environment in CDW to ensure successful cluster restoration. During downtime, you must prevent end-users from accessing the cluster.

> **Caution:** You lose any manual modification of the Kubernetes objects or configurations when you bring down the cluster. For example, third-party integrations with Protegrity. Modifications applied using the CDW web interface and settings defined during creation are preserved.

To deactivate the environment in CDW:

1. Log in to the Data Warehouse service as a DWAdmin.
2. Go to the Environments tab on the **Overview** page, and click Deactivate corresponding to the environment you want to deactivate.

Backing up CDW using the backup-cluster command

Use the backup-cluster command to back up the configuration and settings of all the Database Catalog, Virtual Warehouses, and Data Visualization instances within your Cloudera Data Warehouse (CDW) environment.

Procedure

1. SSH into a host on your cluster from which you can access the CDP Private Cloud Data Services cluster.

2. Run the following command to back up the cluster:

```
cdp dw backup-cluster --cluster-id [***CDW-CLUSTER-ID***] [--cli-input-js
on <value>] [--generate-cli-skeleton]
```

Replace *[***CDW-CLUSTER-ID***]* with the actual cluster ID of your environment. The cluster ID is a unique CDW environment identifier.

[--cli-input-json <value>] and [--generate-cli-skeleton] are optional parameters.

To specify the –cli-input-json parameter, you must obtain the skeleton of the JSON file by running the following command:

```
cdp dw backup-cluster --generate-cli-skeleton
```

The output of this command is a JSON object as follows:

```
{
    "clusterId": ""
}
```

You can now use the JSON string as a parameter for the --cli-input-json command option as follows:

```
cdp dw backup-cluster --cli-input-json '{"clusterId":"[***CDW-CLUSTER-
ID***]"}'
```

The output contains the following information:

- clusterId: The ID of the cluster, a unique identifier of the CDW environment.
- operationId: The ID of the backup operation. You can use the operation ID to query the backup execution details using the CLI.
- timestamp: The date of the creation.
- data: The backup data and configuration.
- md5: The md5 hash of the encoded data. In case the data and its hash are lost, the cluster objects cannot be restored automatically.

3. Save the output in a file.

You need this information during the restoration process.

Results

The Hue backup is stored in the following location:

```
hdfs://cdw-backups/[***TIMESTAMP***]_[***JOB-ID***]/[***ENVIRONMENT-
NAME***]/hue-backup
```

The CDV backup is stored in the following location:

```
hdfs://cdw-backups/[***TIMESTAMP***]_[***JOB-ID***]/[***DATAVIZ-INSTANCE-
NAME]/viz-backup
```

What to do next

Monitor the database backup jobs. The backup process automatically starts the Hue and Data Visualization database backup jobs that you can monitor. Make sure that the database backup jobs complete before destroying the cluster. If you delete the cluster before the jobs are completed, you cannot recover the application contents.

Related Information

CDP CLI documentation: backup-cluster

Monitoring Hue database backup

To monitor the Hue database backup, log into the cluster and monitor the job status under the Database Catalog namespace.

About this task

The backup-cluster command starts a job to load the database dump file, but does not wait for the job to complete. If you have a large database, the job can take up to an hour to complete. Ensure you allow enough time for the job to succeed.

Note: The job that backs up the Hue database fails on OpenShift Container Platform. This is a known issue in CDP Private Cloud Data Services 1.5.4. You must manually backup the Hue database.

Procedure

1. SSH in to a cluster host as an Administrator.
2. Run the following command to fetch the details of the backup job:

```
kubectl get jobs -n [***DATABASE-CATALOG-ID***]
```

Following is a sample output:

```
$ kubectl get jobs -n warehouse-1692037411-96hk
   NAME                                              COMPLETIONS   DURATION
    AGE
   hue-backup-ede2b8bd-1d53-4d23-a0f9-87d8ec658f74   1/1           11s
    113s
   hue-query-processor-db-create-job                 1/1           8s
    42h
```

Results

The Hue backup is stored in the following location:

```
hdfs://cdw-backups/[***TIMESTAMP***]_[***JOB-ID***]/[***ENVIRONMENT-
NAME***]/hue-backup
```

Monitoring Data Visualization database backup

To monitor Cloudera Data Visualization (CDV) database backup, log into the cluster and see the job status under the Data Visualization namespace.

About this task

The backup-cluster command starts a job to create the database dump file, but it does not wait for it to complete. In case your database size is large, it can take up to 20 minutes for the job to complete. Make sure you leave enough time for the job to succeed.

Procedure

1. SSH in to a cluster host as an Administrator.
2. Run the following command to fetch the details of the backup job:

```
kubectl get jobs -n [***DATA-VISUALIZATION-ID***]
```

Following is a sample output:

```
$ kubectl get jobs -n viz-1692216942-fc2g
```

```
NAME                                                 COMPLETIONS   DURATION
  AGE
viz-backup-d874515a-be7e-4902-ac75-269c14f9580c      1/1           3m3s
  10m
viz-webapp-vizdb-create-job                          1/1           57s
  99m
```

Results

The CDV backup is stored in the following location:

```
hdfs://cdw-backups/[***TIMESTAMP***]_[***JOB-ID***]/[***DATAVIZ-INSTANCE-
NAME]/viz-backup
```

Restoring CDW using the restore-cluster command

You can reactivate the entire Cloudera Data Warehouse (CDW) environment, which includes your cluster with all settings of the environment that you backed up, using the CLI.

Procedure

1. SSH into a host on your cluster from which you can access the CDP Private Cloud Data Services cluster.

2. Run the following command to restore the cluster:

```
cdp dw restore-cluster --cluster-id [***CLUSTER-ID***] --data [***DATA-
STRING-FROM-BACKUP***]
```

Replace [***CDW-CLUSTER-ID***] with the actual cluster ID of your environment. The cluster ID is a unique CDW environment identifier.

Replace [***DATA-STRING-FROM-BACKUP***] with the actual data string from the output of the backup.

[--cli-input-json <value>] and [--generate-cli-skeleton] are optional parameters.

To specify the –cli-input-json parameter, you must obtain the skeleton of the JSON file by running the following command:

```
cdp dw backup-cluster --generate-cli-skeleton
```

The output of this command is a JSON object as follows:

```
{
    "clusterId": ""
```

```
}
```

You can now use the JSON string as a parameter for the --cli-input-json command option as follows:

```
cdp dw backup-cluster --cli-input-json '{"clusterId":"[***CDW-CLUSTER-
ID***]"}'
```

The output contains the following information:

- clusterId: The ID of the cluster, a unique identifier of the CDW environment.
- operationId: The ID of the backup operation. You can use the operation ID to query the backup execution details using the CLI.
- action: The action to be taken on the cluster. Possible actions are *Create* and *Skip*.
- message: The description of the cluster action.
- dbcRestorePlans: Information about the restore-plan of the Database Catalogs.

    - item: Status of the entity after the restore operation.
    - ref: The reference of the entity in the backup data.
    - id: The ID of the entity.
    - action: The action associated with the plan. Possible actions are *Create*, *Update*, and *Skip*.
    - message: The description of the plan.
- hueRestorePlans: Information about Hue's restore plan.

    - item: Status of the entity after the restore operation.
    - ref: The reference of the entity in the backup data.
    - id: The ID of the entity.
    - action: The action associated with the plan. Possible actions are *Create*, *Update*, and *Skip*.
    - message: The description of the plan.
- hiveRestorePlans: Information about Hive Virtual Warehouses' restore plan.

    - item: Status of the entity after the restore operation.
    - ref: The reference of the entity in the backup data.
    - id: The ID of the entity.
    - action: The action associated with the plan. Possible actions are *Create*, *Update*, and *Skip*.
    - message: The description of the plan.
- impalaRestorePlans: Information about Impala Virtual Warehouses' restore plan.

    - item: Status of the entity after the restore operation.
    - ref: The reference of the entity in the backup data.
    - id: The ID of the entity.
    - action: The action associated with the plan. Possible actions are *Create*, *Update*, and *Skip*.
    - message: The description of the plan.
- vizRestorePlans: Information about the restore plan of the Data Visualization Apps.

    - item: Status of the entity after the restore operation.
    - ref: The reference of the entity in the backup data.
    - id: The ID of the entity.
    - action: The action associated with the plan. Possible actions are *Create*, *Update*, and *Skip*.
    - message: The description of the plan.

Results

After several minutes, the environment is activated, and the Virtual Warehouses are created in the new cluster and attached to the Database Catalog. The Virtual Warehouse and Data Visualization IDs will be changed.

Note: If you have enabled the Use deterministic namespace names option from the **Advanced Configuration** page, then the Virtual Warehouse and Data Visualization IDs do not change after the restore operation.

Database Catalog's ID will change regardless of whether you have enabled the Use deterministic namespace names option.

What to do next
Adjust the Data Visualization connection settings to point to the new Virtual Warehouse(s) if the IDs have changed.
Related Information
CDP CLI documentation: restore-cluster