

Replication Manager Operations

Date published: 2019-11-07

Date modified: 2024-02-28



Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

CDP Public Cloud Replication Manager.....	4
Support matrix for CDP Public Cloud Replication Manager.....	4
List of features supported by CDP Public Cloud Replication Manager.....	4
Replicate data from CDP Private Cloud Base and CDP Public Cloud source clusters.....	7
Replicate data from CDH and HDP source clusters.....	8
Cloud credentials to use in CDP Public Cloud Replication Manager.....	9
Registering Amazon S3 cloud account in Replication Manager.....	10
Register Azure cloud credentials in Replication Manager.....	12
Registering GCS credentials to use in Replication Manager.....	14
Ports for Replication Manager on CDP Public Cloud.....	19
Ports required for HBase replication policies.....	24
Verify cluster requirements for replication policies.....	24
Configuring SSL/TLS certificate exchange between two Cloudera Manager instances.....	25
 Using HDFS replication policies.....	 26
Preparing to create an HDFS replication policy.....	26
Creating HDFS replication policy.....	30
Manage and monitor HDFS replication policies.....	35
 Using Hive replication policies.....	 38
Preparing to create a Hive replication policy.....	39
Creating Hive replication policy.....	43
Manage and monitor Hive replication policies.....	49
 Using HBase replication policies.....	 51
Preparing to create an HBase replication policy.....	51
Creating HBase replication policy.....	56
Manage and monitor HBase replication policies.....	64
Monitor HBase replication policy job details.....	64
Creating triggers and monitoring replication-related metrics in Cloudera Manager.....	66
Monitor HBase RegionServer replication peer metrics in Replication Manager.....	68
Viewing HBase RegionServer replication peer metrics.....	68
 Troubleshooting replication policies in CDP Public Cloud Replication	
Manager.....	69
Different methods to identify errors related to failed replication policy.....	69
HDFS replication policy fails due to export HTTPS_PROXY environment variable.....	71
Cannot find destination clusters for HBase replication policies.....	71
HBase replication policy fails when Perform Initial Snapshot is chosen.....	72
Optimize HBase replication policy performance when replicating HBase tables with several TB data.....	72
Partition metadata replication takes a long time to complete.....	73
Replicating Hive nested tables.....	73
Target HBase folder is deleted when HBase replication policy fails.....	74
Replicate HBase data in existing and future tables.....	77

CDP Public Cloud Replication Manager

Replication Manager is a service in CDP Public Cloud. You can create replication policies in Replication Manager to copy or migrate HDFS, Hive external tables, and HBase data between on-premises clusters (CDH, CDP Private Cloud Base, or HDP) and CDP Public Cloud (Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS)). Replicating Hive managed tables using Replication Manager from HDP clusters to CDP Public Cloud is a beta feature and is not available for general use.



Important: Before you create replication policies, see [Support matrix for Replication Manager on CDP Public Cloud](#) to verify whether your clusters are supported by Replication Manager.

Before you create a replication policy, you must verify whether the clusters are supported, register the on-premises clusters (CDH or CDP Private Cloud Base) as classic clusters in Management Console, register the cloud credentials in Replication Manager, and verify whether the minimum port configuration is configured. You can also manually set up an SSL/TLS certificate exchange between two Cloudera Manager instances that manage source and target cluster respectively. Replication Manager uses this information to set up the peers for secure data replication.

After you create a replication policy, you can perform various actions on the replication policy. You can monitor the replication jobs, view job history, and perform actions on the replication job.

Related Information

[Introduction to Replication Manager](#)

[Support matrix for Replication Manager on CDP Public Cloud](#)

[Using HDFS replication policy](#)

[Using Hive replication policy](#)

[Using HBase replication policy](#)

[Troubleshooting replication policies in CDP Public Cloud](#)

[CDP CLI for Replication Manager](#)

Support matrix for CDP Public Cloud Replication Manager

You can use Replication Manager or other alternate replication methods to replicate HDFS, Hive external tables, and HBase data between on-premises clusters (CDH clusters, CDP Private Cloud Base clusters, HDP clusters) and CDP Public Cloud (Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS)) clusters. Replication Manager from HDP clusters to CDP Public Cloud Azure is a beta feature and is not available for general use.



Note: Before you create replication policies, you must verify whether the on-premises cluster versions are supported by Replication Manager, register the on-premises clusters as classic clusters in the Management Console, register the cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication.

See the other sections in this topic for the supported cluster and runtime versions.

Related Information

[Introduction to Replication Manager](#)

[CDP Operational Database data replication](#)

List of features supported by CDP Public Cloud Replication Manager

Replication Manager provides replication policies that you can create, edit, and manage to accomplish your data replication goals. You can use other alternate replication methods for scenarios that Replication Manager does not support. Certain features in CDP Public Cloud Replication Manager are available only if the source and target clusters' Cloudera Manager versions support the feature. Verify whether your source and target cluster's Cloudera Manager version support the required feature.

Supported replication policies

You can use the following replication policies in CDP Public Cloud Replication Manager:

- HDFS replication policies replicate HDFS data and metadata from:
 - on-premises clusters (CDH, CDP Private Cloud Base, and HDP) to cloud storage.
 - cloud storage to classic clusters (CDH or CDP Private Cloud Base clusters).

You can choose the frequency during policy creation to replicate the data.

- Hive replication policies support table-level replication and can replicate Hive external tables from on-premises clusters (CDH and CDP Private Cloud Base) to cloud storage and to Data Hubs. They also can:
 - replicate data stored in Hive tables, Hive metadata, data in Hive metastore, and Impala metadata (catalog server metadata) associated with Impala tables registered in the Hive metastore, and



Note: Hive2 managed tables are converted to external tables after replication.

- migrate Sentry permissions to Ranger.





Note: To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

You can choose the frequency during policy creation to replicate the data.

- HBase replication policies replicate HBase data from a source classic cluster (CDH or CDP Private Cloud Base cluster), COD, or Data Hub to a target Data Hub or COD cluster. You can also copy or replicate HBase data between different environments within a Virtual Private Cloud (VPC) using these policies.

Table 1: Supported cluster and runtime versions for HBase replication policies

Source Cluster Type	Lowest Supported Source CDH/CDP Version	Lowest Supported Source Cloudera Manager Version	Target Cluster Type	Lowest Supported Target CDP Version	Lowest Supported Target Cloudera Manager Version
CDP Private Cloud Base	7.1.6*	7.3.1	Data Hub in CDP Public Cloud AWS/Azure	7.2.14	7.6.0
CDH	6.3.3	7.3.1	Data Hub in CDP Public Cloud AWS/Azure	7.2.14	7.6.0
CDH	5.16.2	7.4.4 (patch-5017)	COD (AWS)	7.2.14	-
CDH	5.16.2	<ul style="list-style-type: none"> • 7.6.1 (patch-5610) • 7.6.7 CHF1 and higher 	COD (Azure)	7.2.14	-
COD (AWS/Azure)	7.2.14	-	COD (AWS/Azure)	7.2.14	-
COD (GCP)	7.2.16.1	-	COD (GCP)	7.2.16.1	

Source Cluster Type	Lowest Supported Source CDH/CDP Version	Lowest Supported Source Cloudera Manager Version	Target Cluster Type	Lowest Supported Target CDP Version	Lowest Supported Target Cloudera Manager Version
<p>*CDP Private Cloud Base 7.1.6 and higher clusters must be Kerberos enabled to use them as source classic clusters in an HBase replication policy.</p> <p>Important:</p>  <ul style="list-style-type: none"> You can replicate HBase data in SFT-enabled clusters for target clusters with version CDP 7.2.16 or higher using HBase replication policies. You can replicate Phoenix tables from CDH 5.16.2 and higher using Cloudera Manager 7.4.4 and higher to COD using HBase replication policies. You cannot create HBase replication policies if the target CDP version is 7.2.16, 7.2.16.1, 7.2.16.2, 7.2.16.3, or 7.2.16.200, the source Cloudera Manager version is 7.7.3 or lower, and the source Cloudera Manager API version is v50 or higher. <p> Tip: The endpoint <code>http://[***cm_host***]:[***cm_port***]/api/version</code> shows the API version of Cloudera Manager.</p> <ul style="list-style-type: none"> HBase replication using HBase replication policies is not supported to a CDP Public Cloud cluster destination if the cluster was created using Micro Duty template. 					

HBase replication policies replicate all the data from the specified tables and then continue to replicate the changed data automatically without user intervention.



Note: Alternate replication methods:

- Use the replication plugin for HBase data. For more information, see [COD replication in a Nutshell](#), [Cloudera replication plugin](#), and [HBase data replication](#).
- Contact Cloudera Support for Hive external tables.

Supported features

The following table lists the features and the Cloudera Manager instances that are required for source clusters and target clusters to use the features:

Feature	Lowest supported source Cloudera Manager version	Lowest supported target Cloudera Manager version
Replicate HBase data simultaneously between multiple clusters.	<ul style="list-style-type: none"> 7.9.0-h7 and higher 7.11.0-h2 and higher 7.12.0.0 and higher 	<ul style="list-style-type: none"> 7.9.0-h7 and higher 7.11.0-h2 and higher 7.12.0.0 and higher
Register the GCS credentials to use in Replication Manager on the Cloud Credentials page.	<ul style="list-style-type: none"> 7.9.0-h7 and higher 7.11.0-h3 and higher 7.12.0.0 and higher 	Supports all CDP Cloudera Manager versions.
Specify the YARN queue bandwidth using the Maximum Bandwidth option during the HBase replication policy creation process to export the HBase initial snapshot.	<ul style="list-style-type: none"> 7.9.0-h7 and higher 7.11.0-h3 and higher 7.12.0.0 and higher 	<ul style="list-style-type: none"> 7.9.0-h7 and higher 7.11.0-h3 and higher 7.12.0.0 and higher
Replicate only those HBase tables where the replication scope is already enabled using the Replicate only tables where replication is already enabled option during the HBase replication policy creation process.	Supports all CDP Cloudera Manager versions.	<ul style="list-style-type: none"> 7.9.0-h7 and higher 7.11.0-h3 and higher 7.12.0.0 and higher

Related Information

[Introduction to Replication Manager](#)

[HDFS replication policy](#)

[Using HDFS replication policy](#)

[Hive replication policy](#)

[Using Hive replication policy](#)

[HBase replication policy](#)

[Using HBase replication policy](#)

[Working with cloud credentials](#)

Replicate data from CDP Private Cloud Base and CDP Public Cloud source clusters

Replication Manager replicates HDFS (CDP Private Cloud Base source clusters and CDP Public Cloud storage on AWS and Azure), Hive external tables (CDP Private Cloud Base source clusters), and HBase (CDP Private Cloud Base source clusters) data to CDP Public Cloud (Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS)) clusters. You can use the replication plugin as an alternate replication method to replicate HBase data for scenarios that are not supported by Replication Manager.

The following tables list the minimum source and destination cluster versions, minimum Cloudera Manager versions, supported cloud providers, and supported scenarios:

Replicate data from CDP Private Cloud Base source clusters

Source cluster	Lowest supported source Cloudera Manager version	Lowest supported source Cloudera Runtime version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
CDP Private Cloud Base	7.1.1	7.1.1	Cloud storage in CDP Public Cloud AWS/Azure	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication .
CDP Private Cloud Base	7.1.1	7.1.1	Data Lake in CDP Public Cloud AWS/Azure	Hive external tables	
CDP Private Cloud Base	7.9.0	7.1.1	Data Hub in CDP Public Cloud AWS/Azure	Hive external tables	None
CDP Private Cloud Base	7.3.1	7.1.6	Data Hub in CDP Public Cloud AWS/Azure	HBase	None



Note: Replication Manager converts Hive2 managed tables to external tables after replication.

Replicate data from CDP Public Cloud source clusters

Consider the following limitations while using CDP Public Cloud source and CDP Public Cloud target clusters:

- Replication across cross-cloud providers, that is from AWS to Azure and vice-versa is not supported.
- The source and target clusters must use the same account.

Source cluster	Destination cluster	Supported services on Replication Manager	Services that require alternate replication methods
Cloud storage in CDP Public Cloud AWS* / Azure	CDH 5.x CDH 6.x HDP 2.x HDP 3.x	Not applicable	HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication .
Cloud storage in CDP Public Cloud AWS*	CDH 5.9.0 and higher CDP Private Cloud Base 7.1.7 SP1 and higher	HDFS	None
Cloud storage in CDP Public Cloud Azure	CDH 6.1.0 and higher CDP Private Cloud Base 7.1.7 SP1 and higher	HDFS	None
COD in CDP Public Cloud AWS	AWS	HBase	None

Source cluster	Destination cluster	Supported services on Replication Manager	Services that require alternate replication methods
COD in CDP Public Cloud Azure	Azure	HBase	None
COD in CDP Public Cloud GCP	GCP	HBase	None
*Replication Manager does not support S3 as a source or destination when S3 is configured to use SSE-KMS.			

Replicate data from CDH and HDP source clusters

Replication Manager replicates HDFS data (CDH source clusters and HDP source clusters), Hive external tables (CDH source clusters), and HBase data (CDH 6 source clusters) to CDP Public Cloud (Amazon S3 and Microsoft Azure ADLS Gen2 (ABFS)) clusters. Replication Manager from HDP clusters to CDP Public Cloud Azure is a beta feature and is not available for general use. You can use alternate methods to replicate Hive external tables and HBase data for scenarios that are not supported by Replication Manager.

The following tables list the minimum CDH and HDP source cluster versions, minimum Cloudera Manager versions, supported cloud providers, and supported scenarios:

Table 2: Replicate data from CDH 5 source clusters

Source cluster	Lowest supported source Cloudera Runtime version	Lowest supported source Cloudera Manager version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
CDH 5	5.10	6.3.0	Cloud storage in CDP Public Cloud AWS	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell , Migrating HBase data , and HBase data replication .
CDH 5	5.10	6.3.0	Data Lake in CDP Public Cloud AWS	<ul style="list-style-type: none"> Sentry to Ranger* Hive external tables 	
CDH 5	5.10	6.3.4	Cloud storage in CDP Public Cloud Azure	HDFS	
CDH 5	5.10	6.3.4	Data Lake in CDP Public Cloud Azure	<ul style="list-style-type: none"> Sentry to Ranger* Hive external tables 	
CDH 5	5.10	7.9.0	Data Hub in CDP Public Cloud AWS/Azure	Hive external tables	None
*To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.					

Table 3: Replicate data from CDH 6 source clusters

Source cluster	Lowest supported source Cloudera Runtime version	Lowest supported source Cloudera Manager version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
CDH 6	6.1	6.3.0	Cloud storage in CDP Public Cloud AWS	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell , Migrating HBase data , and HBase data replication .
CDH 6	6.1	6.3.0	Data Lake in CDP Public Cloud AWS	<ul style="list-style-type: none"> Sentry to Ranger* Hive external tables 	
CDH 6	6.1	7.1.1 / 6.3.4	Cloud storage in CDP Public Cloud Azure	HDFS	

Source cluster	Lowest supported source Cloudera Runtime version	Lowest supported source Cloudera Manager version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
CDH 6	6.1	7.1.1 / 6.3.4	Data Lake in CDP Public Cloud Azure	<ul style="list-style-type: none"> Sentry to Ranger* Hive external tables 	
CDH 6	6.1	7.9.0	Data Hub in CDP Public Cloud AWS/Azure	<ul style="list-style-type: none"> Sentry to Ranger* Hive external tables 	
CDH 6	6.3.3	7.3.1	Data Hub in CDP Public Cloud AWS/Azure	HBase	None

*To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.



Note: Replication from HDP clusters to CDP Public Cloud clusters is a technical preview feature and is not available for general use. Before you replicate data from HDP, contact your Cloudera account team.

Table 4: Replicate data from HDP 2 and HDP 3 source clusters

Lowest supported source HDP version	Cloud provider	Supported services on Replication Manager	Services that require alternate replication methods
HDP 2.6.5*	AWS	HDFS	<ul style="list-style-type: none"> HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication. Hive external tables For more information, contact Cloudera Support.
HDP 2.6.5*	Azure	HDFS	HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication .
HDP 3.1.1*	AWS Azure	HDFS	<ul style="list-style-type: none"> HBase To replicate HBase data, see COD replication in a Nutshell and HBase data replication. Hive external tables For more information, contact Cloudera Support.

*No alternate replication methods are available for HDFS, Ranger, and Atlas replication.

Cloud credentials to use in CDP Public Cloud Replication Manager

The Cloud Credentials page shows the registered cloud credentials for Replication Manager. To replicate data to a storage cloud account, you must register the cloud credentials, so that the Replication Manager can access your cloud account. The supported cloud storage accounts are Amazon S3 and Azure Blob Filesystem (ABFS). On the Cloud Credentials page, you can add, update, or delete cloud credentials. Before you register an Amazon S3 cloud account, ensure the cloud bucket requirements are met. Before you add Azure Cloud Credentials in Cloudera Manager, ensure the Blob container requirements are met.

You can perform the following tasks on the Cloud Credentials page to manage cloud credentials:

Add cloud credentials

You can add cloud credentials for your S3 or ABFS account. For information about adding cloud credentials, see [Working with Cloud Credentials](#).



Note: Unregistered credentials can impact the replication process. Credentials associated with a cluster node that do not have updated credentials are called unregistered credentials. For example, if a node is down when the credentials are changed on a bucket or when the node is brought up that has the old credentials.

Update cloud credentials

You can update the cloud credentials based on various factors. When the bucket configuration such as secret or access keys, bucket name or endpoint, and encryption type is changed, it can affect the Replication Manager replication policy run and might require an update to the Replication Manager cloud credentials.

Credential changes are picked up by the next run of the policy. When you change the credentials, the in-progress policy runs might fail but the succeeding runs pick up the changes.

To update a cloud credential, click **Actions Update** option.

Delete cloud credentials

You can delete unwanted credentials from the Replication Manager. When you delete cloud credentials, the replication policies that use the deleted cloud credentials might fail. To avoid failures, delete the Replication Manager cloud policies associated with the deleted credentials and recreate the policies with the new credentials. You can view a list of policies associated with specific credentials on the **Cloud Credentials** page.

To delete a cloud credential, click **Actions Delete** option.

Registering Amazon S3 cloud account in Replication Manager

You must have valid Amazon S3 credentials to register the cloud account with CDP Public Cloud Replication Manager.

Before you begin

Consider the following requirements before you register an Amazon S3 cloud account in Replication Manager:

- You need a cloud bucket with user credentials that you can enter in Replication Manager, so Replication Manager can access the bucket.
- The bucket has to have enough space for the replicated data, and write permissions to copy the data.
- The bucket needs to support cloud storage encryption types supported by Replication Manager (SSE-S3 & SSE-KMS).

About this task

When you add cloud credentials for your Amazon S3 account, you can choose one of the following authentication methods:

- Access secret key. To use this authentication type, you require an AWS Access Key and an AWS Secret key that you obtain from Amazon. Cloudera Manager stores these values securely and does not store them in world-readable locations. The credentials are masked and encrypted in the configurations passed to processes managed by Cloudera Manager, and redacted from the logs.

- IAM role. Amazon Identity and Access Management (IAM) can be used to create users, groups, and roles for use with Amazon Web Services, such as EC2 and Amazon S3. IAM role-based access provides the same level of access to all clients that use the role.

**Important:** IAM role conditions

You can choose the IAM role authentication type only when the following conditions are met:

- The source cluster is hosted on an AWS EC2 infrastructure.
- The source cluster Cloudera Manager and all the nodes in the cluster are running on an EC2 instance.
- The source cluster Cloudera Manager has the same IAM role.

For information about configuring AWS credentials, see [Introduction to role based provisioning credential in AWS](#).

Procedure

1. Go to the Replication Manager Cloud Credentials page, and click Add.
2. In the **Add Cloud Credential** window, perform the following steps:
 - a) Select the Cluster.
 - b) Select S3 as the Cloud Storage Type.
 - c) Name - Provide a unique cloud credential name.
 - d) Authentication Type - Select one of the following authentication types:
 - Select the authentication type as Access Secret Key from the drop-down.
 - Access Key - Enter the valid access key.
 - Secret Key - Enter the valid secret key.
 - Select IAM Role if the *IAM role conditions* mentioned in the **About this task** section are met, and click Save

3. Click Validate.

Note: Using the validation feature is recommended to ensure that the Amazon S3 bucket keys are valid. If the keys are not valid, the Replication Manager policy cannot copy data to the target Amazon S3 bucket.

Add Cloud Credential



Cluster

Select...

Cluster 1 (powqwehugt) X

Cloud Storage Type

S3

Name *

Enter a unique name for the cloud credential

Authentication Type

Access & Secret Key

Access Key *

Enter S3 access key

Secret Key *

Enter S3 secret key



Cancel

Validate

What to do next

Verify whether your credentials are listed on the Cloud Credentials page.

Register Azure cloud credentials in Replication Manager

Currently, registering Azure cloud credentials using the Replication Manager UI does not automatically create the same authorisation rules on the source Cloudera Manager cluster.

You need an ADLS Gen2 storage account and a cloud Blob container with user credentials to use in Replication Manager service. The container has to have enough space for the replicated data and write permissions to copy the data.



Note: The required configurations to be entered by the user for the Azure cloud credentials are currently not available in the Replication Manager UI. When you plan to submit the replication policies with Azure as your cloud storage, Cloudera recommends that you must update cloud credentials in the source cluster Cloudera Manager UI.

Registering ABFS cloud account in Replication Manager

You must have valid ABFS credentials in CDP Public Cloud Replication Manager to register the cloud account.

Before you begin

You must consider the following Blob container requirements to set up ABFS cloud accounts to use in Replication Manager:

- ADLS Gen2 storage account is required. You need a cloud Blob container with user credentials that you can use in Replication Manager service, so that Replication Manager can access the container.
- The container has to have enough space for the replicated data, and write permissions to copy the data.

Procedure

1. In Replication Manager, click **Cloud Credentials Add**.
2. In the **Add Cloud Credential** window, perform the following steps:
 - a) Select the Cluster.
 - b) Select ABFS as the Cloud Storage Type.
 - c) Enter the cloud credential Name.
 - d) Enter your ABFS Storage Client Id.
 - e) Enter your ABFS Storage Client Secret Key.
 - f) Enter your ABFS Storage Tenant Id.
3. Click **Validate**.

After you add the ABFS cloud credentials and you create a replication policy with ABFS as your selected cloud storage for your target cluster, the following error message might appear:

Error



```
java.lang.RuntimeException: com.cloudera.cmf.service.config.ConfigGenException:  
Required account config value not found: adls_tenant_id
```

OK

To resolve this issue, update the ABFS cloud credential values in the source Cloudera Manager instance. For more information, see *Updating Azure cloud credentials in Cloudera Manager*.

What to do next

Verify whether the credentials appear on the Cloud Credentials page.

Updating Azure cloud credentials in Cloudera Manager

To register Azure credentials, you need an ADLS Gen2 storage account and a cloud Blob container with user credentials to use in Replication Manager service. The container has to have enough space for the replicated data and write permissions to copy the data.

About this task



Note: Currently, registering Azure cloud credentials using the Replication Manager UI does not automatically create the same authorisation rules on the source Cloudera Manager cluster. When you plan to submit the replication policies with Azure as your cloud storage, it is recommended that you update cloud credentials in the source cluster Cloudera Manager.

Procedure

1. On the Cloudera Manager Administration Azure Credentials page for the source cluster instance, click Add AD Service Principal.

Cloudera **Manager** Clusters ▾ Hosts ▾ Diagnostics ▾ Audits Charts ▾ Backup ▾ **Administration ▾**

External Accounts AWS Credentials Altus Credentials **Azure Credentials**

Azure credentials allow CDH services and Cloudera tools to securely query and browse data in ADLS. [More Details](#)

Add AD Service Principal

2. Enter the name of the credential.
3. Enter the Client ID for the credential.
4. Provide the Client Secret Key for the credential.
5. Enter the Tenant Identity for the credential.
6. Click Add.

The Azure Cloud credentials are successfully registered with the Cloudera Manager instance.

Add AD Service Principal ✕

Name *
Enter a friendly name to identify this credential.

Client ID *

Client Secret Key *

Tenant ID *

[Cancel](#) **Add**

What to do next

Create the required replication policy.

Registering GCS credentials to use in Replication Manager

You can register Google Cloud Storage (GCS) credentials to use in Replication Manager after you provide access to GCS for the source cluster in the source Cloudera Manager.

Before you begin

Ensure that the following details are available before you register the GCS credentials to use in Replication Manager.

- A valid Google Cloud Platform service account. For more information, see [Configuring access to Google Cloud Storage](#).
- Enough space for the replicated data and the required write permissions to copy the data in the storage of the destination cluster.

Procedure

1. Click **Cloud Credentials Add** in CDP Public Cloud Replication Manager.
2. Select the required source Cluster.
3. Choose GCS as the Cloud Storage Type.
4. Enter a unique Name for the cloud credential.

5. Select one of the following Input Type, and then complete the required steps:
 - a) Select Manual if you have the following details about the GCS service account:
 - Enter your GCS service account Client Email address.
 - Enter your GCS service account Private Key.
 - Enter your GCS service account Private Key ID.

Add Cloud Credential ✕

Cluster

Select... ▼

Cloud Storage Type

GCS ▼

Name ^{*}

Enter a unique name for the cloud credential

Input Type

Manual ▼

Client E-mail ^{*}

Enter your GCS Service Account Client E-mail

Private Key ^{*}

Enter your GCS Service Account Private Key

Private Key ID ^{*}

Enter your GCS Service Account Private Key ID

[Cancel](#) [Save](#)

- b) Select File if you chose to save the service account private key in a file in JSON format. Ensure that you have the required permissions to access and use the file.

Upload File for Replication Manager to use the credentials in it to access data.

Add Cloud Credential ✕

Cluster

Select... ▼

Cloud Storage Type

GCS ▼

Name ^{*}

Enter a unique name for the cloud credential

Input Type

File ▼

Service Account Private Key (json format) ^{*}

Upload File

Cancel

Save

6. Click Save.

Results

The GCS cloud credentials appear on the **Cloud Credentials** page. You can use these credentials when you create the replication policies. The credentials allow Replication Manager to access the cloud data in the source cluster.


Ports for Replication Manager on CDP Public Cloud

Before you create replication policies, you must ensure that the required ports are open and available for data replication. You can verify the mandatory ports using the Replication Manager network security diagram.

HDFS replication policies

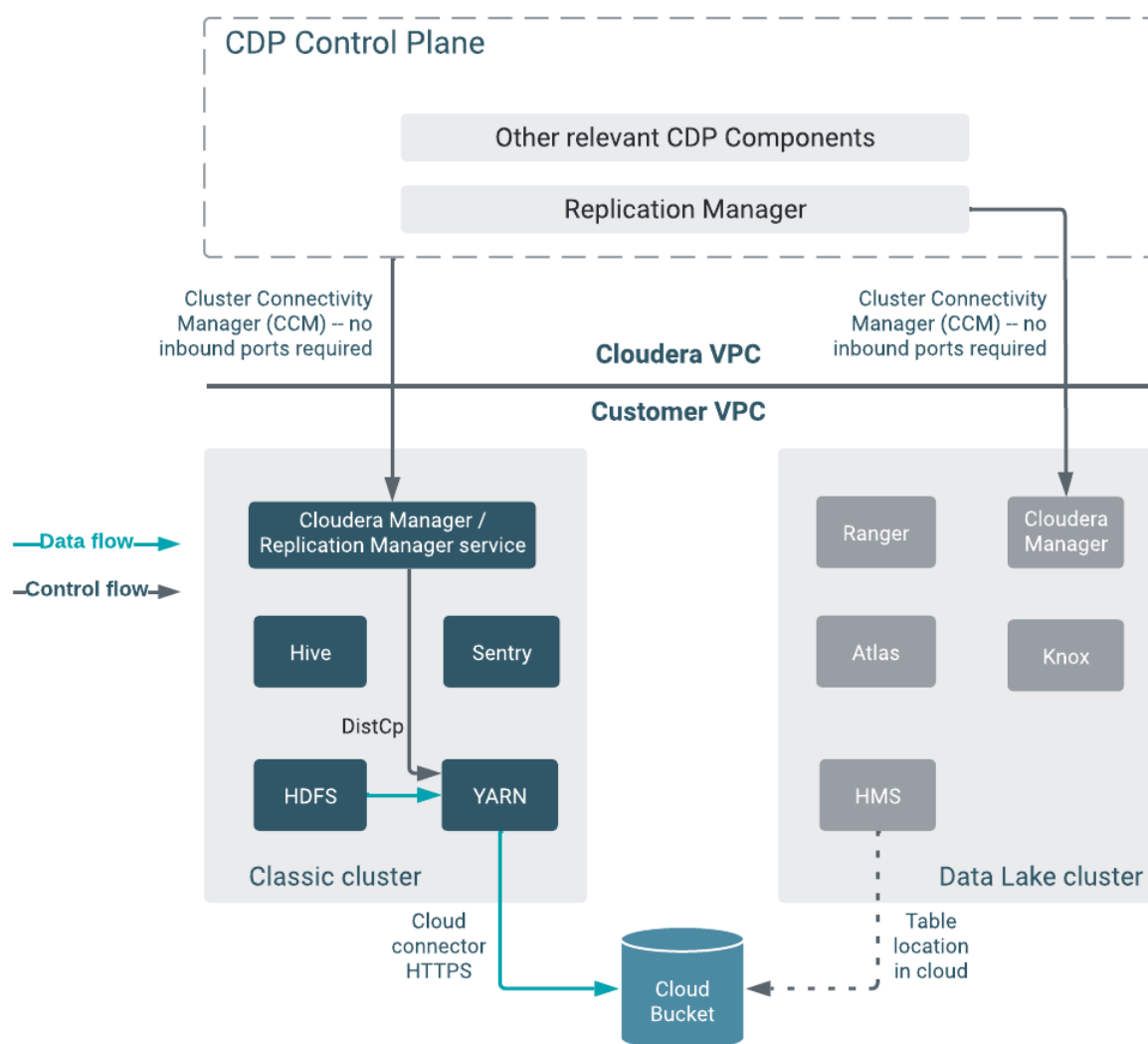
The following ports must be open and available for Replication Manager for HDFS replication policies:

Table 5: Minimum ports required for HDFS replication policies

Connectivity required	Default Port	Type	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound	Outgoing port. All classic cluster nodes must be able to access S3/ADLS Gen2 endpoint.  Note: Before you create replication policies, ensure that the best practices for CDP Public Cloud on Microsoft Azure ADLS Gen2 (ABFS) are complete.
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound	Connecting source classic cluster to the CDP Management Console through Cluster Connectivity Manager (CCM). For more information, see Outbound network access for CCM , and CCM overview .

The following system architecture diagram shows the interaction between components during HDFS replication using HDFS replication policies:


Figure 1: System architecture diagram for HDFS replication in CDP Public Cloud Replication Manager



Hive replication policies

The following ports must be open and available for Replication Manager for Hive replication policies:

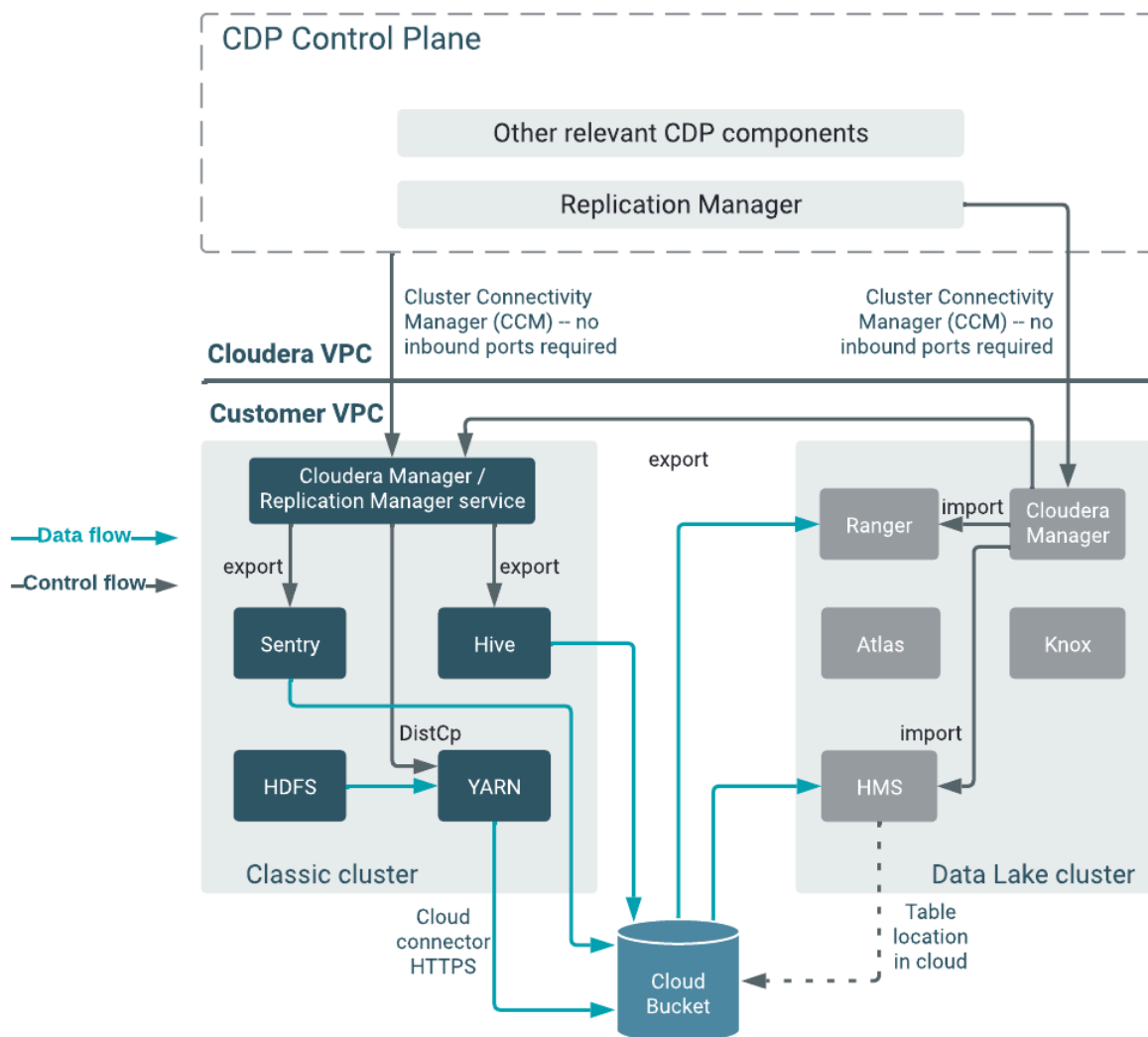
Table 6: Minimum ports required for Hive replication policies

Connectivity required for	Default Port	Port opened toward	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound towards cloud provider public hosts (For example, AWS)	<p>Outgoing port. All classic cluster nodes must be able to access S3/ADLS Gen2 endpoint.</p> <p> Note: Before you create replication policies, ensure that the best practices for CDP Public Cloud on Microsoft Azure ADLS Gen2 (ABFS) are complete.</p>

Connectivity required for	Default Port	Port opened toward	Description
Cloudera Manager Admin Console HTTP	7180 or 7183 (when TLS enabled)	Inbound towards destination CDP Public Cloud Cloudera Manager host	Incoming port. Open on the source cluster to enable the target Cloudera Manager in cloud to communicate to the on-premises Cloudera Manager.
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound towards CDP Public Cloud Control Plane (CCM and cluster proxy)	Connecting the source classic cluster to the CDP Management Console through Cluster Connectivity Manager (CCM) For more information, see Outbound network access for CCM , and CCM overview .

The following system architecture diagram shows the interaction between components during Hive replication using Hive replication policies:

Figure 2: System architecture diagram for Hive replication in CDP Public Cloud Replication Manager



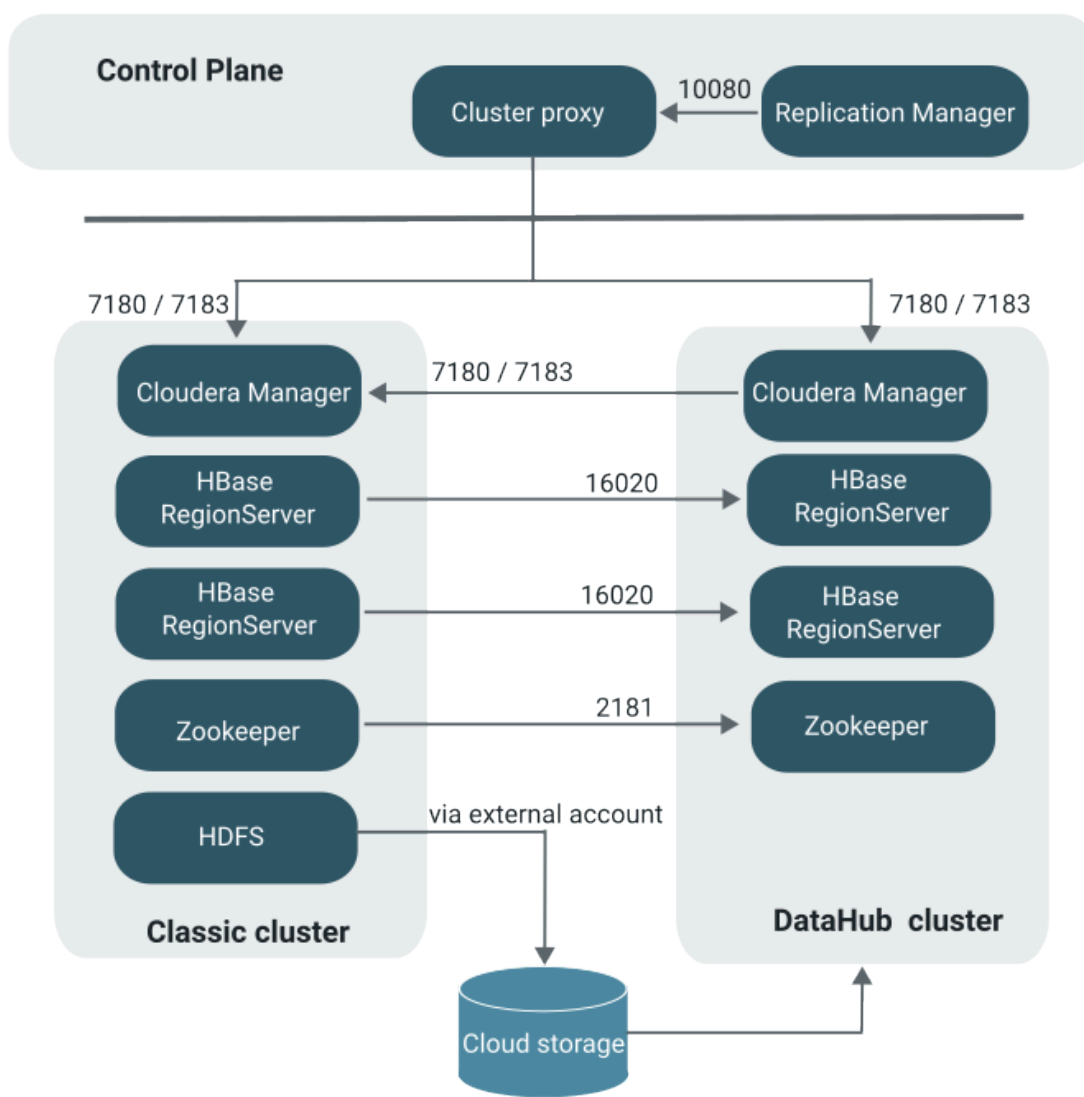
HBase replication policies

The following ports must be open and available for Replication Manager for HBase replication policies:

Table 7: Minimum ports required for HBase replication policies

Service	Ports	Ports opened toward	Description
Destination hosts of the AWS cluster or ADLS cluster (target cluster), and the Cloudera Manager server port on the source cluster	2181 and 16020	Outbound in destination Zookeeper and HBase RegionServer host(s)	Verify whether the ports 16020 for worker security group and 2181 for worker, master, and leader groups are open for connection from the source cluster to the destination cluster on AWS or Azure. This ensures that the source HBase service can communicate with Zookeeper and HBase services on the destination hosts uninterrupted. For more information, see Ports for HBase replication .
HMaster	16000	Outbound in destination HBase Master host(s)	Open the port on the Master Nodes (HBase Master Node and any back-up HBase Master node). Before you select the Validate Replication option during the first HBase replication policy creation between two specific clusters, you must ensure that the port is open on the target cluster.
Cloudera Manager Admin Console HTTP	7180 or 7183	Inbound towards destination CDP Public Cloud Cloudera Manager host	Open on the source cluster to enable Data lake Cloudera Manager to communicate to the on-premises Cloudera Manager. Connects to destination SDX Data Lake Cloudera Manager.
Cluster Connectivity Manager (CCM) for CCMv1	6000-6049	For CDP Public Cloud Control Plane (CCM and Cluster Proxy)	Required for SSL connections to the Control Plane via CCM to communicate with Replication Manager.
Data transfer from secondary node for AWS / ADLS Gen2 for CCMv2	80 or 443	Outbound towards data transfer from secondary node for AWS / ADLS Gen2	Outgoing port. Open on all the HDFS nodes for AWS and ADLS Gen2.  Note: Before you create replication policies, ensure that the best practices for CDP Public Cloud on Microsoft Azure ADLS Gen2 (ABFS) are complete.
Data Lake cluster	8443	On destination CDP Public Cloud Cloudera Manager/Knox host. (applicable when Knox is available on the on-premises source cluster)	Outgoing port. Configure the port on the Data Lake cluster as the outgoing port for CDP Management Console to communicate with Cloudera Manager and Knox.

Figure 3: System architecture diagram for HBase replication in CDP Public Cloud Replication Manager



Best practices

Consider the following best practices while using CDP Public Cloud on Microsoft Azure ADLS Gen2 (ABFS):

- Ensure that the on-premises cluster (port 443) can access the <https://login.microsoftonline.com> endpoint. This is because the Hadoop client in the on-premises cluster (CDH/CDP Private Cloud Base) connects to the endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the [Azure-specific endpoints](#) table.
- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the [Azure-specific endpoints](#) table are complete so that the endpoint connects to the target path successfully.

Related Information

[Preparing to create an HDFS replication policy](#)

[Preparing to create a Hive replication policy](#)

[Preparing to create an HBase replication policy](#)

[Ports required for HBase replication policies](#)

[Support matrix for Replication Manager on CDP Public Cloud](#)

Ports required for HBase replication policies

Open the ports 2181 and 16020 on the source and destination secondary nodes to ensure that the source HBase service can reach Zookeeper and HBase services on the destination hosts. You must also ensure that the Cloudera Manager server port is open on the source cluster.

Use one of the following methods to open the required ports for HBase replication:

- Choose a security group for your environment and open the ports manually. In this method, you choose the security groups that are automatically created for the environment. By default, the security groups do not have any rules for Zookeeper and HBase ports, therefore, you must open the required ports manually after you create a Data Hub.

After you open the ports, the required security groups are assigned to the nodes when the nodes are autoscaled. This is a one-time process that you must perform when you create a Data Hub.

- Define a security group with the required ports open, and assign it to the new Data Hub environment. In this method, you define a security group for a VPC that contains inbound rules to open the required ports which include Zookeeper and HBase ports. When you create an environment, you assign this security group to it. If required, you can assign different security groups to the gateway node and other nodes.

This method allows you to reuse the security groups in other new Data Hubs. Security issues do not appear because the nodes in the same security group do not access each other by default. However, if required, you can add a separate rule to impose this restriction. Sharing the same security group for inbound and outbound network access rules remains as strict as having separate security groups for each environment, but the extra rules for Zookeeper and HBase ports do not need to be added at each environment creation.

The following use cases illustrate the situations where a requirement for autoscaling nodes during HBase replication might appear:

- You replicate HBase data to another CDP account or region in the same cloud provider. In this use case, ensure that the VPC/VNET peering is complete before you open the ports to establish connection over private networks.
- You replicate HBase data to COD or Data Hub using a direct connection. In this use case, you ensure that public IPs and Zookeeper ports are not open to the internet.

Verify cluster requirements for replication policies

You must complete the prerequisites before you create a replication policy in CDP Public Cloud Replication Manager.

The following topics list the prerequisites that must be completed before you create a replication policy:

- [HDFS replication policy prerequisites](#)
- [Hive replication policy prerequisites](#)
- [HBase replication policy prerequisites](#)

Related Information

[Adding a CDH cluster](#)

[Introduction to Replication Manager](#)

[Role-based credential on AWS](#)

[App-based credential on Azure](#)

[Management console](#)

[How to configure AWS credentials](#)

[IAM role-based authentication](#)

[Support matrix for Replication Manager on CDP Public Cloud](#)

Configuring SSL/TLS certificate exchange between two Cloudera Manager instances

The Replication Manager configures replication peers between two clusters before running the replication job. You can manually set up an SSL/TLS certificate exchange between two Cloudera Manager instances that manage source and target cluster respectively. Replication Manager uses this information to set up the peers for secure data replication.

About this task



Important: The sample commands use the open-jdk-11 java version. Use the java version that you use in CDP clusters in these commands.

Procedure

1. Go to the truststore location in *source* Cloudera Manager, and perform the following steps:

- a) List the contents of the keystore file and password using the `[***keytool path***] -list -keystore [***truststore JKS file location ***] -storepass [***truststore password***]` command.

For example, `/usr/lib/jvm/java-openjdk-11/bin/keytool -list -keystore /var/lib/cloudera-scm-agent/agent-cert/cm-auto-global_truststore.jks -storepass [***truststore password***]`



Tip:

- The keytool path can be located in various locations including the keytool itself. For example, it can be located in `/usr/lib/jvm/java-openjdk-11/bin/keytool` or `/usr/java/default/bin/keytool`.
- You can locate the truststore password using the `cat /etc/hadoop/conf/ssl-client.xml` command. You can enter the SSL password for the `/etc/hadoop/conf/ssl-client.xml` file when prompted.
- Alternatively, you can also run the following commands instead of the command in Step a:

```
export JAVA_HOME=[***keytool location***]

export TRUSTSTORE_JKS=[***truststore JKS file location***]

export TRUSTSTORE_PASSWORD=[***password in the ssl-client.xml
file***]
$JAVA_HOME/keytool -list -keystore
$TRUSTSTORE_JKS -storepass
$TRUSTSTORE_PASSWORD
```

- b) Export the certificate contents in the host to a file using the `[***keytool***] -exportcert -keystore [***truststore JKS file location ***] -alias [***cm_alias_on_src_cm***] -file ./[***TXT file, for example: source-cert.txt***] -storepass [***truststore_password***]` command.

For example, `/usr/java/default/bin/keytool -exportcert -keystore /var/lib/cloudera-scm-agent/agent-cert/cm-auto-global_truststore.jks -alias cmrootca-0 -file ./source-cert.txt -storepass [***truststore_password***]`

- c) Copy the text file to all the hosts of the *target* cluster Cloudera Manager securely using the `scp -i [***PEM file***] [***TXT file - source-cert.txt***] root@[***host_ip***]:/home/` command.
- d) Import the certificate into the keystore file on all the hosts of the *target* cluster Cloudera Manager using the `[***keytool***] -importcert -noprompt -v -trustcacerts -keystore [***truststore JKS file`

```
location ***] -alias [***cm_alias_on_dest_cm***] -file ./[***TXT file - source-cert.txt***] --storepass [***truststore_password***] command.
```

For example, `/usr/java/default/bin/keytool -importcert -noprompt -v -trustcacerts -keystore /var/lib/cloudera-scm-agent/agent-cert/cm-auto-global_truststore.jks -alias cmrootca-1 -file ./source-cert.txt --storepass [***truststore_password***]`

2. Go to the truststore location in *target* Cloudera Manager, and perform the following steps:

- List the contents of the keystore file and password using the `[***keytool path***] -list -keystore [***truststore JKS file location ***] -storepass [***truststore password***]` command.
- Export the certificate contents in the host to a file using the `[***keytool***] -exportcert -keystore [***truststore JKS file location ***] -alias [***cm_alias_on_dest_cm***] -file ./[***TXT file, for example: dest-cert.txt***] -storepass [***truststore_password***]` command.
- Copy the text file to all the hosts of the *source* cluster Cloudera Manager securely using the `scp -i [***PEM file***] [***TXT file - dest-cert.txt***] root@[***host_ip***]:/home/` command.
- Import the certificate into the keystore file on all the hosts of the *source* Cloudera Manager using the `[***keytool***] -importcert -noprompt -v -trustcacerts -keystore [***truststore JKS file location ***] -alias [***cm_alias_on_src_cm***] -file ./[***TXT file - dest-cert.txt***] --storepass [***truststore_password***]` command.

Using HDFS replication policies

You can use the HDFS replication policies in CDP Public Cloud Replication Manager to replicate HDFS data. The HDFS replication policies can replicate HDFS data and metadata from classic clusters (CDH, CDP Private Cloud Base, and HDP) to CDP Public Cloud storage buckets such as S3 and ABFS, and from cloud storage to classic clusters (CDH or CDP Private Cloud Base clusters). To use an on-premises cluster (CDH or CDP Private Cloud Base cluster) in the replication policy, you must register it as a classic cluster in the Management Console. To use the cloud storage for data replication, you must register the cloud credentials in Replication Manager so that the Replication Manager service can access the cloud storage. You must also verify cluster access and configure minimum ports for replication before you create HDFS replication policies.



Important: Before you create replication policies, see *Support matrix for Replication Manager on CDP Public Cloud* to verify whether your clusters are supported by Replication Manager.

You can also use CDP CLI commands to create HDFS replication policies. The CDP CLI commands for Replication Manager are under the `replicationmanager CDP` CLI option. For more information, see *CDP CLI for Replication Manager*.

Related Information

[HDFS replication policy](#)

[Support matrix for Replication Manager on CDP Public Cloud](#)

[CDP CLI for Replication Manager](#)

Preparing to create an HDFS replication policy

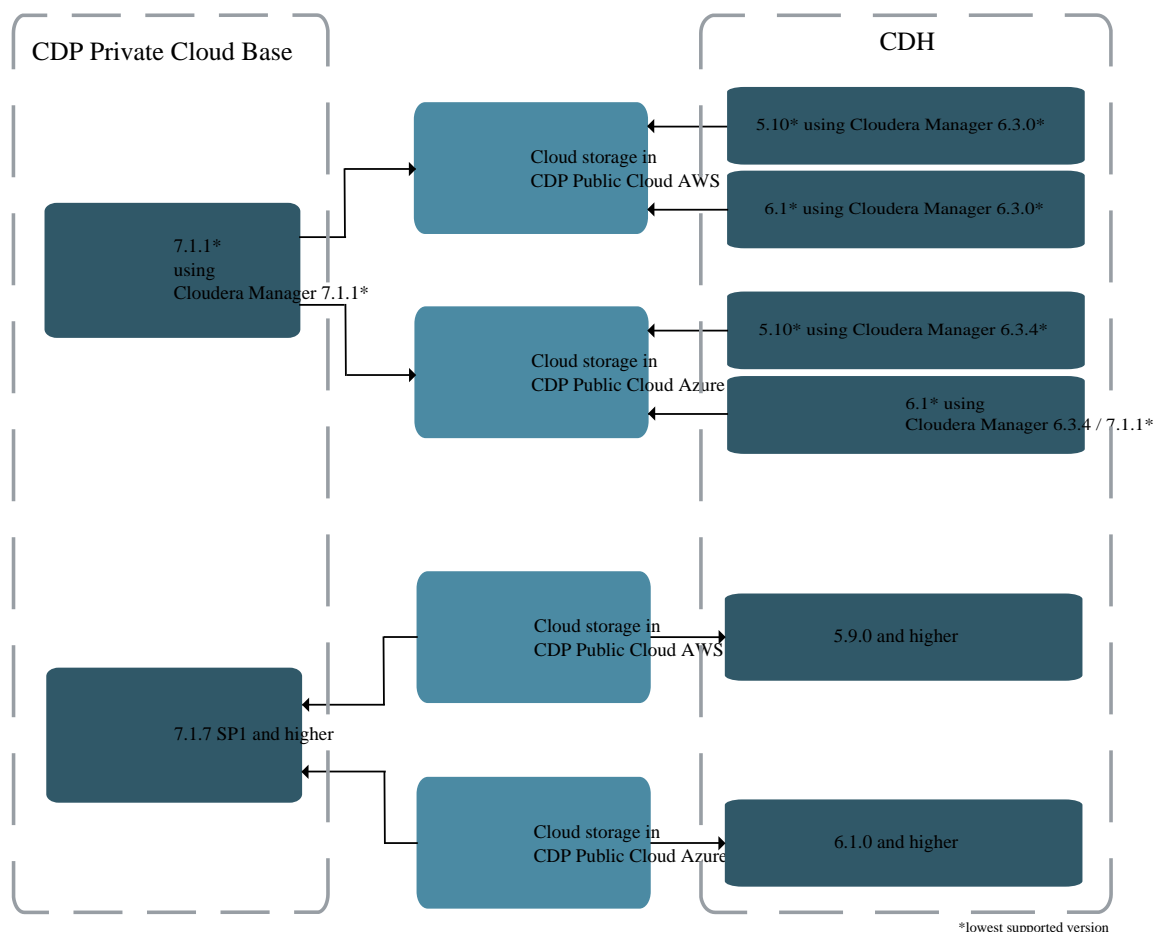
Before you create the HDFS replication policies in CDP Public Cloud Replication Manager to replicate HDFS data, register the on-premises cluster (CDH or CDP Private Cloud Base) as a classic cluster in Management Console, register cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication.

Procedure

- Do the source cluster and target cluster meet the requirements to create an HDFS replication policy?

The following image shows a high-level view of the support matrix for HDFS replication policies, you must consult the [Support matrix for Replication Manager on CDP Public Cloud](#) for the complete list of supporting clusters and scenarios:

Figure 4: High-level replication scenarios supported by HDFS replication policies



- Is the required on-premises cluster (CDH cluster or CDP Private Cloud Base cluster) registered as a classic cluster on the Management Console?

CDH clusters and CDP Private Cloud Base clusters are managed by Cloudera Manager. To enable these on-premises clusters for Replication Manager, you must register them as classic clusters on the Management Console. After registration, you can use them for data migration purposes.



Important: When you register a CDP Private Cloud Base cluster as a classic cluster, ensure that you use only the Cloudera Manager IP Address and Cloudera Manager Port options and not the Register KNOX endpoint (Optional) option.

For information about registering an on-premises cluster as a classic cluster, see [Adding a CDH cluster](#) and [Adding a CDP Private Cloud Base cluster](#).

- Is an external account available in the Cloudera Manager instance that has access to the bucket or container that you are using in the HDFS replication policy?

For more information, see [Role-based credential on AWS](#), [App-based credential on Azure](#), and Cloudera Manager documentation.

- Do you have the required cluster access to create replication policies?

Power users, the user who onboarded the source and target clusters, and users with ClassicClusterAdmin or ClassicClusterUser resource roles can create replication policies on clusters for which they have access. For more information, see [Understanding account roles and resource roles](#).



Tip: Ensure that you have *Replication Administrator* or Full Administrator role on the source on-premises cluster.

- Do you have the required cluster access to view the replication policies?

Existing HDFS replication policies are visible to users who have access to the source cluster in the replication policy. A warning appears if you do not have access to the source cluster.

If you can view the policies, you can perform other actions on the policy including policy update and policy delete operations.



Note: A Replication Manager user can browse, within Replication Manager, the folder structure of any cluster enabled for Replication Manager. Therefore, permitted users can view folders, files, and databases in Replication Manager that they might not have access to in HDFS. The users cannot view the content of files on the source or destination clusters, nor can they modify or delete folders or files that are viewable from Replication Manager.

- Is the required cloud credential that you want to use in the replication policy registered with the Replication Manager service?

For more information, see [Working with cloud credentials](#).

- Do you need to replicate data securely? If so, ensure that the SSL/TLS certificate exchange between two Cloudera Manager instances that manage source and target clusters respectively is configured. For more information, see [Configuring SSL/TLS certificate exchange between two Cloudera Manager instances](#).
- Are the following ports open and available for Replication Manager?

Table 8: Minimum ports required for HDFS replication policies

Connectivity required	Default Port	Type	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound	Outgoing port. All classic cluster nodes must be able to access S3/ADLS Gen2 endpoint.
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound	Connecting source classic cluster to the CDP Management Console through Cluster Connectivity Manager (CCM). For more information, see Outbound network access for CCM , and CCM overview .

Consider the following best practices while using CDP Public Cloud on Microsoft Azure ADLS Gen2 (ABFS):

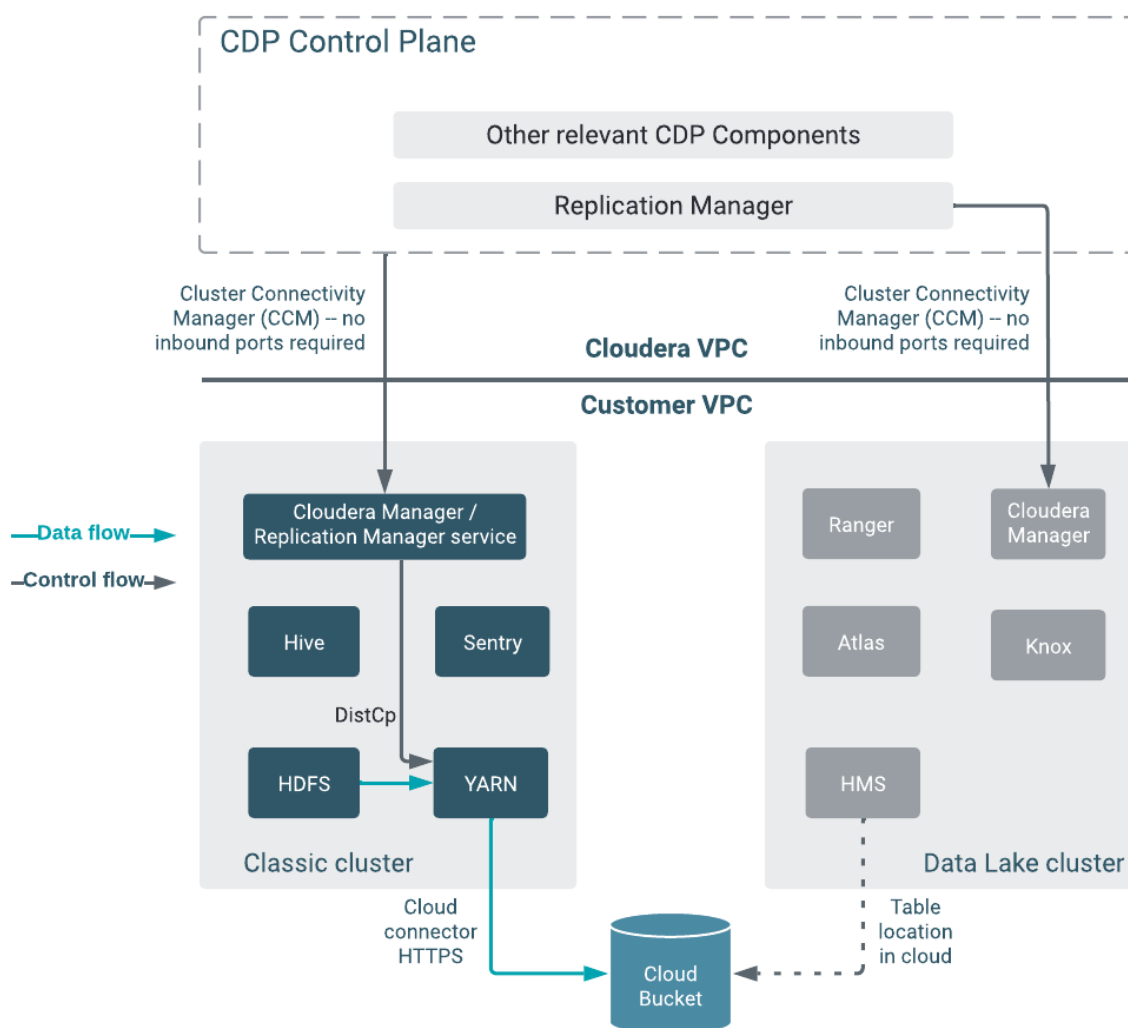
- Ensure that the on-premises cluster (port 443) can access the <https://login.microsoftonline.com> endpoint. This is because the Hadoop client in the on-premises cluster (CDH/CDP Private Cloud Base) connects to the

endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the [Azure-specific endpoints](#) table.

- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the [Azure-specific endpoints](#) table are complete so that the endpoint connects to the target path successfully.

The following system architecture diagram shows the interaction between components during HDFS replication using HDFS replication policies:

Figure 5: System architecture diagram for HDFS replication in CDP Public Cloud Replication Manager



What to do next

After the clusters and cloud storage requirements are met, you can create an HDFS replication policy.

Related Information

[Support matrix for Replication Manager on CDP Public Cloud](#)

[Adding a CDH cluster](#)

[Adding a CDP Private Cloud Base cluster](#)

[Accessing Data Lake services](#)

[Role-based credential on AWS](#)

[App-based credential on Azure](#)

[Understanding account roles and resource roles](#)

[Working with cloud credentials](#)

[Configuring SSL/TLS certificate exchange between two Cloudera Manager instances](#)

[Ports for Replication Manager in CDP Public Cloud](#)

Creating HDFS replication policy

After you register the classic clusters in Management Console and register the cloud credentials in CDP Public Cloud Replication Manager, you can create an HDFS replication policy to replicate HDFS data. An HDFS replication policy can replicate HDFS data from a classic cluster (CDH, CDP Private Cloud Base, and HDP clusters) to cloud storage, and from cloud storage to classic clusters (CDH or CDP Private Cloud Base clusters). You can replicate data on-premises to cloud storage account with a single cluster.

Before you begin

Ensure that the [HDFS replication policy prerequisites](#) are complete.

Procedure

1. On the Management Console Replication Manager Replication Policies page, click Add Policy.
The **Create Replication Policy** wizard appears.
2. On the General page, choose or enter the following information:

Option	Description
HDFS	Creates an HDFS replication policy.
Policy Name	Enter a unique name for the replication policy.
Description	Optional. Enter a brief description about the replication policy.

The following image shows a sample General page in the **Create Replication Policy** wizard:

Create Replication Policy

The screenshot shows the 'Create Replication Policy' wizard in the Cloudera Management Console. The 'General' step is active, indicated by a blue circle with the number 1. The wizard consists of five steps: General, Select Source, Select Destination, Schedule, and Additional Settings. In the 'General' section, the 'Policy Name' field contains 'hdfs_test' and has a red asterisk indicating it is required. The 'Description' field contains the text 'Use a HDFS replication policy to replicate HDFS data.' Below these fields is the 'Type' section, which has three radio button options: 'Hive', 'HDFS', and 'HBase'. The 'HDFS' option is selected and highlighted with an orange rectangular box. At the bottom of the wizard, there is a blue 'Next' button with a right-pointing arrow.

3. Click Next.
4. On the Select Source page, the options change depending on whether you choose a classic cluster or cloud storage as the source for data to be replicated.
 - a) If you are replicating from a classic cluster, choose or enter the following information:

Option	Description
Type	Choose CLUSTER to select an on-premises cluster as the source cluster.
Source Cluster	Select a classic cluster.
Source path	Choose one of the following methods to determine the directory where source data resides on the source cluster: <ul style="list-style-type: none"> Enter the complete directory path. Click File Browser to view and navigate the existing directory list on the selected cluster. Select the required directory that you want to replicate.
Run As Username (on source)	Optional. The replication policy uses the Default username to replicate HDFS data. If you are using a kerberized cluster, enter the required username. The replication policy uses this username to replicate the data in the kerberized cluster.

- b) If the source of your replication is cloud storage, choose or enter the following information:

Option	Description
Type	Choose S3 or ABFS to select your cloud provider.
Cloud Credential on Source	Choose the required cloud credential. The cloud credentials that you register for Replication Manager on the Cloud Credentials page appear in this field.
Path	Based on the cloud provider, enter the path to the source data in the required format. <ul style="list-style-type: none"> For an S3 bucket, provide the path of the directory in the <code>[***bucket_name***]/[***path***]</code> format. For an ABFS container, provide the path of the directory in the <code>[***filesystem***]@[***storage_account***]/[***location***]</code> format.

5. Click Next.
6. On the Select Destination page, the options change depending on whether you choose a classic cluster or cloud storage as the destination for the replicated data:
 - a) If you are replicating from a classic cluster to cloud storage, choose or enter the following information:

Option	Description
Type	Choose S3 or ABFS to select your cloud provider.
Cloud Credential on Source	Choose the required cloud credential. The cloud credentials that you register for the selected source cluster on the Cloud Credentials page appear in this field. You can also add cloud credentials using the Add Cloud Credential link.
Path	Based on the cloud provider, enter the target path, where data is replicated to, in the required format. <ul style="list-style-type: none"> For an S3 bucket, provide the path of the directory in the <code>[***bucket_name***]/[***path***]</code> format. For an ABFS container, provide the path of the directory in the <code>[***filesystem***]@[***storage_account***]/[***location***]</code> format.

b) If you are replicating from cloud storage to a classic cluster, choose or enter the following information:




Option	Description
Type	Choose Classic Cluster.
Destination Cluster	Choose a classic cluster as the target cluster.
Destination Path	Enter the directory on the target cluster to which the replication policy replicates HDFS data.
Run As Username	Optional. The replication policy uses the Default username to replicate HDFS data. Enter another username if you want the replication policy to use it to replicate data.

7. Click Validate Policy.

Replication Manager verifies whether the details provided are correct.

8. Click Next.

9. On the **Schedule** page, choose or enter the following information:


Option	Description
Run Now	Starts to replicate the existing HDFS data after the replication policy creation is complete. Choose the frequency to replicate data periodically.
Schedule Run	Runs the replication policy to replicate data at a later time. Choose the date and time for the first run, and then choose the frequency to replicate data periodically.  Tip: On the Replication Policies page, click  to change the timezone.
Frequency	Choose one of the following options: <ul style="list-style-type: none"> Does Not Repeat Custom - In the Custom Recurrence dialog box, choose the time, date, and the frequency to run the policy. Replication Manager ensures that the same number of seconds elapse between the runs. For example, if you choose the Start Time as January 19, 2022 11.06 AM and Interval as 1 day, Replication Manager runs the replication policy for the first time at the specified time in the timezone the replication policy was created in, and then runs it exactly after 1 day that is, after 24 hours or 86400 seconds.  Note: Ensure that the frequency in a schedule enables a job to finish before the next job starts. Also, ensure that the jobs based on the same policy do not overlap. If a job is not completed before another job starts, the second job does not run and the job status appears as Skipped. If a job is consistently skipped, you might need to modify the frequency of the job.


10. Click Next.

11. On the Additional Settings page, enter or choose the values as necessary:

Option	Description
YARN Queue Name	Enter the name of the YARN queue for the cluster to which the replication job is submitted if you are using Capacity Scheduler queues to limit resource consumption. The default value for this field is default.

Option	Description
Maximum Maps Slots	Set the maximum number of map tasks (simultaneous copies) per replication job. The default value is 20.
Maximum Bandwidth	<p>Adjust this setting so that each map task is throttled to consume only the specified bandwidth.</p> <p>Each map task ((simultaneous copy) is restricted to consume only the specified bandwidth. This is not always exact. The map throttles back its bandwidth consumption during a copy in such a way that the net bandwidth used tends towards the specified value. You can adjust this setting so that each map task is throttled to consume only the specified bandwidth so that the net bandwidth used tends towards the specified value. The default value for the bandwidth is 100MB per second for each mapper.</p>
Path Exclusion	Enter one or more regular expressions separated by comma. Replication Manager does not copy the subdirectories or files from the source that matches one of the specified regular expressions to the target cluster.
Replication Strategy	<p>Choose one of the following replication strategies to determine whether the file replication tasks should be distributed among the mappers statically or dynamically.</p> <ul style="list-style-type: none"> • Static distributes file replication tasks among the mappers up front to achieve an uniform distribution based on the file sizes. • Dynamic distributes the file replication tasks in small sets to the mappers, and as each mapper completes its tasks, it dynamically acquires and processes the next unallocated set of tasks. <p>The default replication strategy is Dynamic.</p>
MapReduce Service	Choose the MapReduce or YARN service to use.
Log Path	Enter an alternate path for the logs, if required.

Option	Description
Error Handling	<p>Select the following options as necessary:</p> <ul style="list-style-type: none"> • Skip Checksum Checks - Determines whether to skip checksum checks on the copied files. If selected, checksums are not validated. Checksums are checked by default. <p> Note: You must skip checksum checks to prevent replication failure due to non-matching checksums in the following cases:</p> <ul style="list-style-type: none"> • Replications from an encrypted zone on the source cluster to an encrypted zone on a destination cluster. • Replications from an encryption zone on the source cluster to an unencrypted zone on the destination cluster. • Replications from an unencrypted zone on the source cluster to an encrypted zone on the destination cluster. <p>Checksums are used for two purposes:</p> <ul style="list-style-type: none"> • To skip replication of files that have already been copied. If Skip Checksum Checks is selected, the replication job skips copying a file if the file lengths and modification times are identical between the source and destination clusters. Otherwise, the job copies the file from the source to the destination. • To redundantly verify the integrity of data. However, checksums are not required to guarantee accurate transfers between clusters. HDFS data transfers are protected by checksums during transfer and storage hardware also uses checksums to ensure that data is accurately stored. These two mechanisms work together to validate the integrity of the copied data. <ul style="list-style-type: none"> • Skip Listing Checksum Checks - Whether to skip checksum check when comparing two files to determine whether they are same or not. If skipped, the file size and last modified time are used to determine if files are the same or not. Skipping the check improves performance during the mapper phase. Note that if you select the Skip Checksum Checks option, this check is also skipped. • Abort on Error - Whether to abort the job on an error. If selected, files copied up to that point remain on the destination, but no additional files are copied. Abort on Error is not selected by default. • Abort on Snapshot Diff Failures - If a snapshot diff fails during replication, the replication policy uses a complete copy to replicate data. If you select this option, the policy aborts the replication when it encounters an error instead.

Option	Description
Preserve	<p>Choose the required options to preserve the block size, replication count, permissions (including ACLs), and extended attributes (XAttrs) as they exist on the source file system, or to use the settings as configured on the destination file system. By default source system settings are preserved.</p> <p>When Permission is selected, and both the source and destination clusters support ACLs, replication preserves ACLs. Otherwise, ACLs are not replicated. When Extended attributes is selected, and both the source and destination clusters support extended attributes, replication preserves them. (This option only displays when both source and destination clusters support extended attributes.)</p> <p>If you select one or more of the Preserve options and you are replicating to S3 or ADLS, the values all of these items are saved in metadata files on S3 or ADLS. When you replicate from S3 or ADLS to HDFS, you can select which of these options you want to preserve.</p> <p> Note: To preserve permissions to HDFS, you must be running as a superuser on the destination cluster. Use the Run As Username option to set the username.</p>
Delete Policy	<p>Choose the required options to determine whether the files that were deleted on the source should also be deleted from the destination directory. This policy also determines the handling of files in the destination location that are unrelated to the source. Options include:</p> <ul style="list-style-type: none"> • Keep Deleted Files - Retains the destination files even when they no longer exist at the source. This is the default option. • Delete to Trash - If the HDFS trash is enabled, files are moved to the trash folder. This is not supported when replicating to S3 or ADLS. • Delete Permanently - Uses the least amount of space; use with caution.
Alerts	<p>Choose to generate alerts for various state changes in the replication workflow. You can choose to generate an alert On Failure, On Start, On Success, or On Abort of the replication job.</p> <p>You can configure alerts to be delivered by email or sent as SNMP traps. If alerts are enabled for events, you can search for and view the alerts on the Events tab, even if you do not have email notification configured. For example, if you choose Command Result that contains the Failed filter on the Diagnostics Events page, the alerts related to the On Failure alert for all the replication policies for which you have set the alert appear. For more information, see Managing Alerts and Configuring Alert Delivery.</p>

12. Click Create.

What to do next

You can track the replication policy job status on the **Replication Policies** page.

Manage and monitor HDFS replication policies



After you create an HDFS replication policy in CDP Public Cloud Replication Manager, you can perform and monitor various tasks related to the replication policy. You can view the job progress and replication logs. You can edit the advanced options to optimize a job run. You can suspend a job and also activate a suspended job. You can edit the replication policy as necessary.

About this task

On the **Replication Policies** page, you can perform the following actions and tasks on a replication policy and its jobs:

Procedure

- When you click Actions for an HDFS replication policy, the following actions appear:

Action	Description
Edit*	<p>Change the replication policy options as required for non-expired policies that are in active or suspended state. Based on the schedule you choose, the replication policy replicates data.</p> <p>You can edit the replication policies to better align with changing requirements. For example, you might want to change the frequency of a policy depending on the data size and importance of the data being replicated.</p> <p> Note: A replication policy is associated with a cluster or a cluster pair, therefore you cannot change the clusters in the policy.</p> <p>Optionally, expand a replication policy on the Replication Policies page to edit the replication policy options which include frequency (start time cannot be modified if the policy has already started), queue name, maximum bandwidth, and maximum map slots.</p> <p> Tip: To optimize the replication policy performance, you can configure the queue name, maximum bandwidth, and maximum map slots as necessary.</p>
Delete	Deletes the replication policy permanently.
Suspend	Suspends a running replication policy. Activate the replication policy, if required.
View Log	<p>Download, copy, or open the log. The log shows a brief output of the stdout and stderr logs of a single step of the latest replication policy job run.</p> <p>You can also view the current job status in the Replication Manager Overview Issues & Updates Job Status column. If the job failed, click Failed to view the log details about the job.</p>
Collect diagnostic bundle	<p>Generates a diagnostic bundle for the replication policy. You can download the bundle as a ZIP file to your machine.</p> <p>Ensure that you are logged into the Cloudera Manager instances for both the source and target clusters before you download the bundle in Replication Manager.</p>
<p>*</p> <p>To view and use the replication policies with an empty name in Replication Manager, you must understand the following implementation:</p> <ul style="list-style-type: none"> If the Cloudera Manager API version is lower than 51, an existing replication policy with an empty name can be used and updated. However, if you edit the replication policy and provide a name for the replication policy in versions higher or equal to 51, you must ensure that the name conforms to the validation rules. If the Cloudera Manager API version is higher or equal to 51, it is mandatory that you provide a unique name to the replication policy to continue using it. This is because API version 51 and higher enforces the validation rules on all the replication policies. <p>To pass the replication policy name validation, you must ensure that the replication policy name is unique. The name can contain letters, numbers, and the _ / - characters. You must also ensure that it does not contain the characters % . ; \ nor any character that is not ASCII printable, which includes the ASCII characters less than 32 and the ASCII characters that are greater than or equal to 127.</p>	

- When you expand the policy details, the **Job History** panel appears.

You can view the following details on the panel:

- Previous jobs, current job, and one future scheduled job if any.
- Job details which include:

Job details	Description
Started	Timestamp when the job started.
Ended	Timestamp when the job ended.
Duration	Time taken to complete the job.

Job details	Description
Progress	Current status of a running job.
Expected	Remaining number of files and bytes expected to be copied for a running job.
Copied	Number of files and bytes copied for a running job and completed job.
Failed	Number of files and bytes that failed to be copied for a completed job.
Deleted	Number of files deleted for a completed job.
Skipped	Remaining number of files and bytes skipped from copying for a running job and complete job.

c) Click Actions to:

- Abort the job.
 - Re-run an aborted or failed job.
 - View Log for the job. You can download, copy, or open it to track the job and to troubleshoot any issues for the job.
- When you click a job on the **Job History** panel, the following tabs appear:

Tab	Description
General	Shows the following job details: <ul style="list-style-type: none"> • Started at timestamp • Duration to complete the job • HDFS Replication Report to download the job statistics in CSV format • Job status Message
Command Details	Shows the steps that Replication Manager ran for the job along with the timestamp.

- You can download the following CSV reports from the **General HDFS Replication Report** field to track the replication jobs and to troubleshoot issues:

Report	Description
Listing	Lists all the files and directories copied during the replication job.
Status	Shows the complete status report of each file as: <ul style="list-style-type: none"> • an Error occurred and the file was not copied. • a Deleted file. • an up-to-date file for which the replication was Skipped.
Error Status	Status report of all the copied files with errors. Each file shows the status, path, and message for the copied files with errors.
Skipped Status	Status report of all skipped files. Each file lists the status, path, and message for the databases and tables that were skipped.
Deleted Status	Status report of all deleted files. Each file lists the status, path, and message for the databases and tables that were deleted.
Performance	Summary report about the performance of the running replication job which includes the last performance sample for each mapper that is working on the replication job.
Full Performance	Performance report of the job which includes the samples taken for all mappers during the replication job.



Note: The reports are generated based on the source Cloudera Manager response. If the Cloudera Manager response is interrupted or is not handled as expected, corresponding error messages appear in HTML format in the reports.

Using Hive replication policies

To create a Hive replication policy in CDP Public Cloud Replication Manager, you must configure the required Ranger policy in Ranger, register the on-premises cluster (CDH or CDP Private Cloud Base) as a classic cluster in Management Console, register cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication. The replication load happens on the source on-premises cluster. You can replicate data on-premises to the cloud with a single cluster if the Metastore is running on the cloud.

These policies support table-level replication and can replicate Hive external tables from on-premises clusters (CDH and CDP Private Cloud Base) to cloud storage such as S3 and ABFS and to Data Hubs. They also can:

- replicate data stored in Hive tables, Hive metadata, data in Hive metastore, and Impala metadata (catalog server metadata) associated with Impala tables registered in the Hive metastore, and



Note: Hive2 managed tables are converted to external tables after replication.

- migrate Sentry permissions to Ranger.



Note: To perform the Sentry policy replication, you must be running the Sentry service on CDH 5.12 or higher, or any CDH 6.x version.

Hive metadata replication involves multiple entities. Replication Manager supports replication of external tables in Hive. Hive supports replication of external tables to the target cluster and it retains all the properties of external tables. The data files permission and ownership are preserved so that the relevant external processes can continue to write in it even after failover.

You can also use CDP CLI commands to create Hive replication policies. The CDP CLI commands for Replication Manager are under the replicationmanager CDP CLI option. For more information, see *CDP CLI for Replication Manager*.



Important: Before you create Hive replication policies, you must ensure that the required Ranger policy is set in Ranger and see [Support matrix for Replication Manager on CDP Public Cloud](#) to verify whether your clusters are supported by Replication Manager.

The Apache Ranger access policy model consists of the following components:

- Specification of the resources that you can apply to a replication policy which includes the HDFS files and directories; Hive databases, tables, and columns; and HBase tables, column-families, and columns.
- Specification of access conditions for specific users and groups.

You must set the Ranger policy for the hdfs user on the target cluster to perform all operations on all databases and tables. The same user role is used to import Hive Metastore. The hdfs user should have access to all Hive datasets, including all operations. Otherwise, Hive import fails during the replication process.

On the target cluster, the hive user must have Ranger admin privileges. The same hive user performs the metadata import operation.

For more information about Hive replication policies to replicate data from CDH clusters to CDP Public Cloud, see [Migrate Hive data from CDH to CDP Public Cloud](#) blog.

Related Information

[Hive replication policy](#)

[Support matrix for Replication Manager on CDP Public Cloud](#)

[CDP CLI for Replication Manager](#)

Preparing to create a Hive replication policy

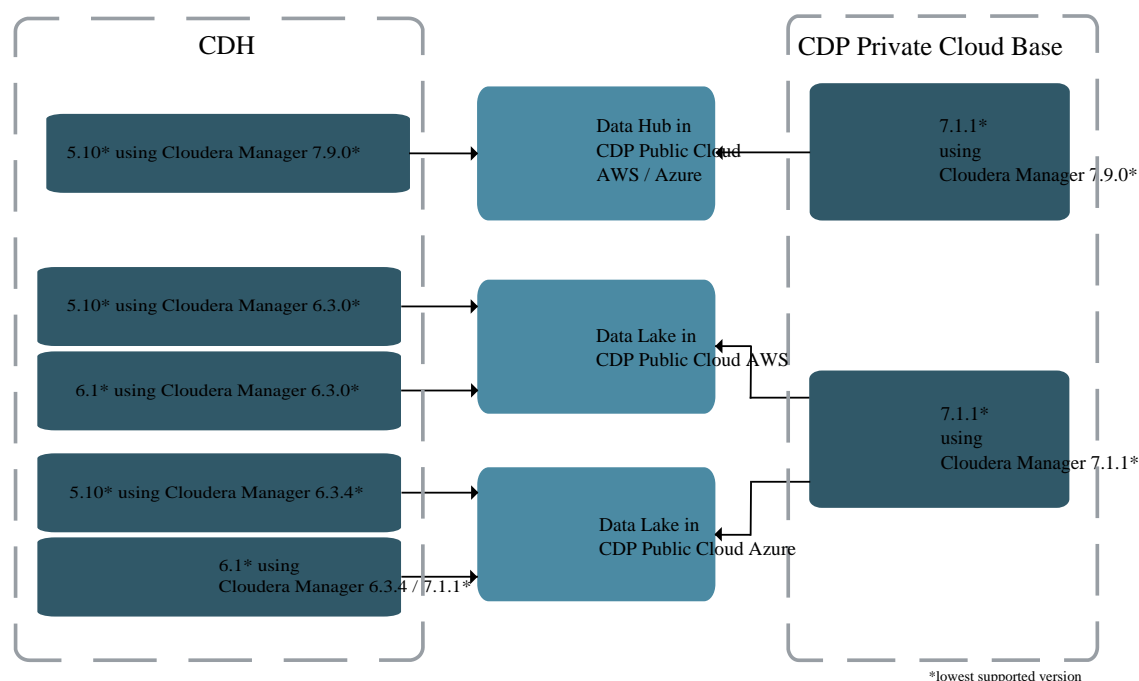
Before you create the Hive replication policies in CDP Public Cloud Replication Manager, you must prepare the clusters and verify cluster access and cloud credentials.

Procedure

- Do the source cluster and target cluster meet the requirements to create an Hive replication policy?

The following image shows a high-level view of the support matrix for HDFS replication policies, you must consult the [Support matrix for Replication Manager on CDP Public Cloud](#) for the complete list of supporting clusters and scenarios:

Figure 6: High-level replication scenarios supported by Hive replication policies



- Is the source CDH cluster or source CDP Private Cloud Base cluster registered as a classic cluster on the Management Console?

CDH clusters and CDP Private Cloud Base clusters are managed by Cloudera Manager. To enable these on-premises clusters for Replication Manager, you must register them as classic clusters on the Management Console. After registration, you can use them for data migration purposes.



Important: When you register a CDP Private Cloud Base cluster as a classic cluster, ensure that you use only the Cloudera Manager IP Address and Cloudera Manager Port options and not the Register KNOX endpoint (Optional) option.

For information about registering an on-premises cluster as a classic cluster, see [Adding a CDH cluster](#) and [Adding a CDP Private Cloud Base cluster](#).

- Does the target Data Hub use Cloudera Manager 7.9.0 or higher? If not, upgrade Cloudera Manager to version 7.9.0 or higher.

- Have you configured the **all-database, table, column** Ranger policy for the hdfs user on the source and target cluster to perform all the operations on all databases and tables?

The hdfs user role is used to import Hive Metastore and must have access to all Hive datasets, including all operations. Otherwise, Hive import fails during the replication process. On the target cluster, the hive user must have Ranger admin privileges. The same hive user performs the metadata import operation.

To provide access, navigate to the Ranger Admin UI Service Manager Hadoop_SQL Policies Access section, and provide hdfs user permission to the **all-database, table, column** policy name.

Policy ID	Policy Name	Policy Labels	Status	Audit Logging	Roles	Groups	Users	Action
7	all - global	--	Enabled	Enabled	cdrep-global_admin	--	rangerlookup, hive, beacon, dpprofiler + More...	View, Edit, Delete
8	all - database, table, column	--	Enabled	Enabled	cdrep-global_admin	--	rangerlookup, hive, beacon, dpprofiler, hdfs, admin, impala, hdfs, [OWNER] - Less...	View, Edit, Delete
9	all - database, table	--	Enabled	Enabled	--	--	hive, beacon, dpprofiler, hdfs + More...	View, Edit, Delete
10	all - database	--	Enabled	Enabled	--	public	hive, beacon, dpprofiler, hdfs + More...	View, Edit, Delete
11	all - hiveservice	--	Enabled	Enabled	cdrep-global_admin	--	rangerlookup, hive, beacon, dpprofiler + More...	View, Edit, Delete

- Is an external account configured on the source CDH cluster's Cloudera Manager which allows the CDH cluster to access CDP cloud storage?



Tip: The external account can be configured on the Cloudera Manager Administration page and the account has the access key / secret key pair that you can use to access CDP cloud storage.

- Do you have the required cluster access to create replication policies?

Power users, the user who onboarded the source and target clusters, and users with ClassicClusterAdmin or ClassicClusterUser resource roles can create replication policies on clusters for which they have access. For more information, see [Understanding account roles and resource roles](#).



Tip: Ensure that you have *Replication Administrator* or Full Administrator role on the source on-premises cluster.

- Do you have the required cluster access to view the replication policies?

Existing Hive replication policies are visible to users who have access to the source cluster in the replication policy. A warning appears if you do not have access to the source cluster.

If you can view the policies, you can perform other actions on the policy including policy update and policy delete operations.



Note: A Replication Manager user can browse, within Replication Manager, the folder structure of any cluster enabled for Replication Manager. Therefore, permitted users can view folders, files, and databases in Replication Manager that they might not have access to in HDFS. The users cannot view the content of files on the source or destination clusters, nor can they modify or delete folders or files that are viewable from Replication Manager.

- Is the required cloud credential that you want to use in the replication policy registered with the Replication Manager service?

For more information, see [Working with cloud credentials](#).

- Do you need to replicate data securely? If so, ensure that the SSL/TLS certificate exchange between two Cloudera Manager instances that manage source and target clusters respectively is configured. For more information, see [Configuring SSL/TLS certificate exchange between two Cloudera Manager instances](#).

- Are the following ports open and available for Replication Manager?

Table 9: Minimum ports required for Hive replication policies

Connectivity required for	Default Port	Type	Description
Data transfer from classic cluster hosts to cloud storage	80 or 443 (TLS)	Outbound	Outgoing port. All classic cluster nodes must be able to access S3/ADLS Gen2 endpoint.
Cloudera Manager Admin Console HTTP	7180 or 7183 (when TLS enabled)	Inbound	Incoming port. Open on the source cluster to enable the target Cloudera Manager in cloud to communicate to the on-premises Cloudera Manager.
Classic cluster	6000-6049 for CCMv1 443 for CCMv2	Outbound	Connecting the source classic cluster to the CDP Management Console through Cluster Connectivity Manager (CCM) For more information, see Outbound network access for CCM , and CCM overview .

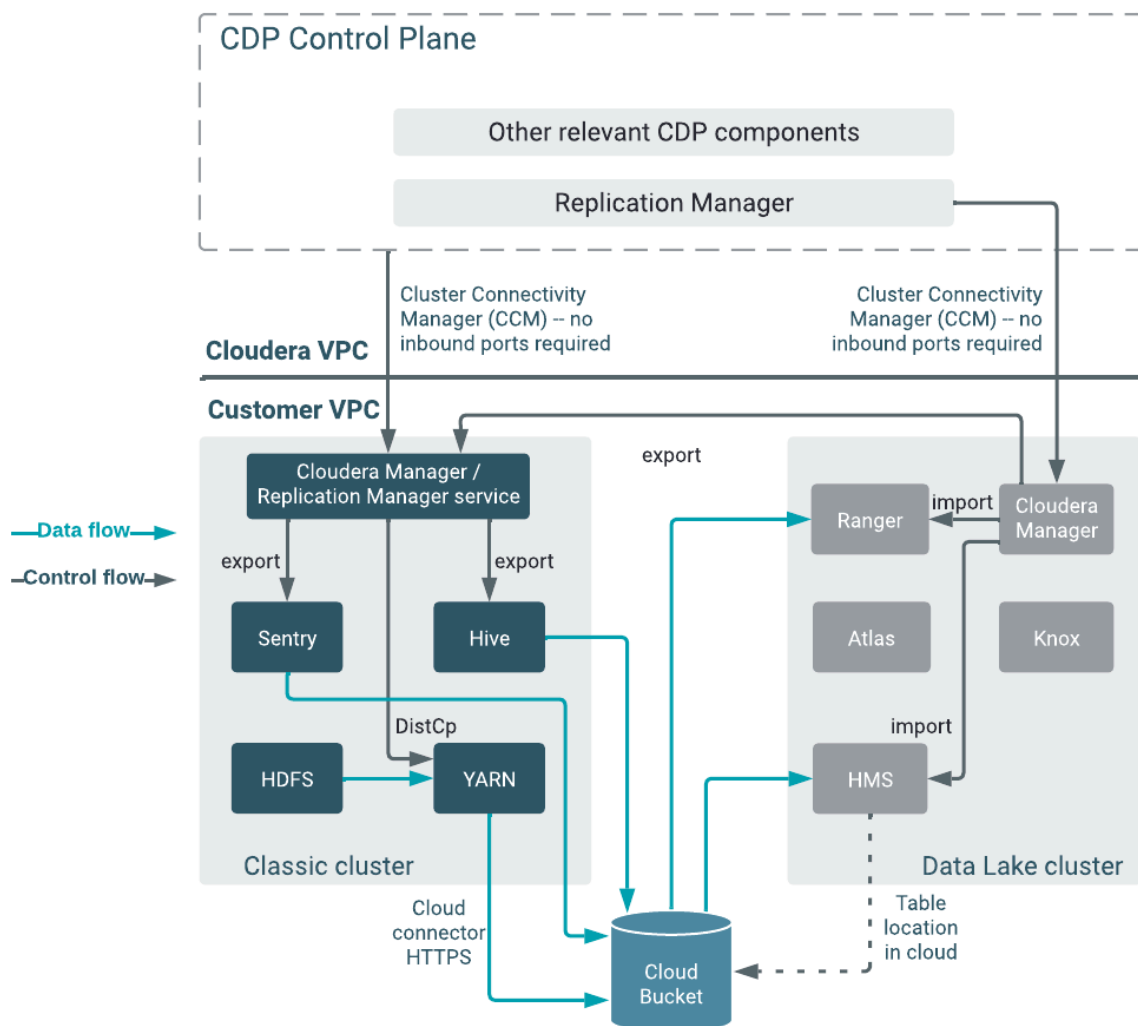
Consider the following best practices while using CDP Public Cloud on Microsoft Azure ADLS Gen2 (ABFS):

- Ensure that the on-premises cluster (port 443) can access the <https://login.microsoftonline.com> endpoint. This is because the Hadoop client in the on-premises cluster (CDH/CDP Private Cloud Base) connects to the endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the [Azure-specific endpoints](#) table.

- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the [Azure-specific endpoints](#) table are complete so that the endpoint connects to the target path successfully.

The following system architecture diagram shows the interaction between components during Hive replication using Hive replication policies:

Figure 7: System architecture diagram for Hive replication in CDP Public Cloud Replication Manager



What to do next

After the clusters and cloud storage requirements are met, you can create a Hive replication policy.

Related Information

[Support matrix for Replication Manager on CDP Public Cloud](#)

[Adding a CDH cluster](#)

[Adding a CDP Private Cloud Base cluster](#)

[Accessing Data Lake services](#)

[Understanding account roles and resource roles](#)

[Working with cloud credentials](#)

[Configuring SSL/TLS certificate exchange between two Cloudera Manager instances](#)

Creating Hive replication policy

Before you create a Hive replication policy in CDP Public Cloud Replication Manager, you must ensure that the prerequisites are complete and the required Ranger policy is set in Ranger.

Before you begin

Ensure that the [Hive replication policy prerequisites](#) are complete.

Procedure

1. On the Management Console Replication Manager Replication Policies page, click Add Policy.

The **Create Replication Policy** wizard appears.

2. On the General page, choose or enter the following information:

Option	Description
Hive	Creates a Hive replication policy.
Policy Name	Enter a unique name for the replication policy.
Description	Optional. Enter a brief description about the replication policy.


The following image shows a sample General page in the **Create Replication Policy** wizard:

Create Replication Policy

The screenshot shows the 'Create Replication Policy' wizard in the 'General' step. On the left, a vertical list of steps is shown: 1. General (active), 2. Select Source, 3. Select Destination, 4. Schedule, and 5. Additional Settings. The main content area is titled 'General' and contains three sections: 'Policy Name' with a red asterisk and a text input field containing 'hive_test'; 'Description' with a text area containing 'Use a Hive replication policy to replicate Hive external tables. ACID tables and managed tables in Hive are converted to external tables after replication.'; and 'Type' with three radio button options: 'Hive' (selected and highlighted with an orange border), 'HDFS', and 'HBase'. At the bottom right, there is a blue 'Next' button with a right arrow.



3. Click Next.
4. On the Select Source page, enter or choose the options as required:

Option	Description
Source cluster	Choose a classic cluster as a source cluster.

Option	Description
Source Databases and Tables	Enter the database name and table name. Click Add to enter more databases and tables as necessary.
Run As Username (on source)	<p>Optional. The replication policy uses the Default username to replicate data.</p> <p>Enter another username if you want the replication policy to use it to replicate data. Ensure that the user has the necessary permissions to replicate data.</p> <p> Note: You must provide a username for Kerberized clusters.</p>

5. Click Next.

6. On the Select Destination page, enter or choose the options as required:

Option	Description
Destination Data Lake or Data Hub	<p>Choose a Data Lake or Data Hub as the destination cluster.</p> <p>The Managed Warehouse Path and the Hive External Table Base Directory Path for the Data Lake appears.</p> <p> Note: Administrators can edit the Hive External Table Base Directory field to add another path to override the default storage location for replicated Hive external tables. Before you add another path to override the default storage location, ensure that the following steps are complete in the Ranger UI:</p> <ul style="list-style-type: none"> Alter the ranger policy Default: Hive warehouse locations in cm_s3 service to allow the Hive service to access the updated locations of S3 bucket path. Manually update the Ranger and Sentry permissions. <p> Note: The Data Hubs, using Cloudera Manager version 7.9.0 and higher, for which you have access permissions appear automatically. If you want to use a Data Hub that appears disabled, you must upgrade the Cloudera Manager version to 7.9.0 or higher for the Data Hub.</p>
Cloud Credential on Source	Choose the required cloud credential. The cloud credentials that you register for the selected source cluster on the Cloud Credentials page appear in this field. You can also add cloud credentials using the Add Cloud Credential link.
Run as Username	<p>Optional. The replication policy uses the Default username to replicate data.</p> <p>Enter another username if you want the replication policy to use it to replicate data.</p>





7. Click Validate Policy.

Replication Manager verifies whether the details provided are correct.

8. Click Next.

9. On the **Schedule** page, choose or enter the following information:


Option	Description
Run Now	Starts to replicate the existing HDFS data after the replication policy creation is complete. Choose the frequency to replicate data periodically.


Option	Description
Schedule Run	<p>Runs the replication policy to replicate data at a later time. Choose the date and time for the first run, and then choose the frequency to replicate data periodically.</p> <p> Tip: On the Replication Policies page, click   to change the timezone.</p>
Frequency	<p>Choose one of the following options:</p> <ul style="list-style-type: none"> • Does Not Repeat • Custom - In the Custom Recurrence dialog box, choose the time, date, and the frequency to run the policy. <p>Replication Manager ensures that the same number of seconds elapse between the runs. For example, if you choose the Start Time as January 19, 2022 11.06 AM and Interval as 1 day, Replication Manager runs the replication policy for the first time at the specified time in the timezone the replication policy was created in, and then runs it exactly after 1 day that is, after 24 hours or 86400 seconds.</p> <p> Note: Ensure that the frequency in a schedule enables a job to finish before the next job starts. Also, ensure that the jobs based on the same policy do not overlap. If a job is not completed before another job starts, the second job does not run and the job status appears as Skipped. If a job is consistently skipped, you might need to modify the frequency of the job.</p>


10. Click Next.

11. On the Additional Settings page, enter or choose the values as necessary:

Option	Description
YARN Queue Name	Enter the name of the YARN queue for the cluster to which the replication job is submitted if you are using Capacity Scheduler queues to limit resource consumption. The default value for this field is default.
Maximum Maps Slots	Set the maximum number of map tasks (simultaneous copies) per replication job. The default value is 20.
Maximum Bandwidth	<p>Adjust this setting so that each map task is throttled to consume only the specified bandwidth.</p> <p>Each map task ((simultaneous copy) is restricted to consume only the specified bandwidth. This is not always exact. The map throttles back its bandwidth consumption during a copy in such a way that the net bandwidth used tends towards the specified value. You can adjust this setting so that each map task is throttled to consume only the specified bandwidth so that the net bandwidth used tends towards the specified value. The default value for the bandwidth is 100MB per second for each mapper.</p>
Number of concurrent HMS connections	<p>Enter the number of concurrent Hive Metastore connections. The connections are used to concurrently import and export metadata from Hive. Increase the number of threads to improve Replication Manager performance. By default, a new replication policy uses 4 connections.</p> <ul style="list-style-type: none"> • If you set the value to 1 or more, Replication Manager uses multi-threading with the number of connections specified. • If you set the value to 0 or fewer, Replication Manager uses single threading and a single connection.

Option	Description
Sentry Permissions	Choose Include Sentry Permissions with Metadata to migrate Sentry permissions during the replication job. Otherwise, choose Exclude Sentry Permissions from Metadata to not migrate Sentry permissions during the replication job.
Skip URI Privileges	<p>Select this option if you do not want to include URI privileges when you migrate Sentry permissions.</p> <p>During migration, the URI privileges are translated to point to an equivalent location in S3. If the resources have a different location in Amazon S3, do not migrate the URI privileges because the URI privileges might not be valid.</p>
Replication Option	Choose Metadata and Data to replicate metadata and data in files and directories. Otherwise, choose Metadata only to replicate the metadata of files and directories
Directory for metadata file	Enter / or a valid folder path in the target cluster to save the metadata file. If the field is empty or if the specified folder does not exist, Replication Manager creates a new folder.
Force Overwrite	<p>Select to overwrite data in the destination metastore if incompatible changes are detected.</p> <p>For example, if the destination metastore was modified, and a new partition was added to a table, this option forces deletion of that partition, overwriting the table with the version found on the source. If you do not choose the option and the Hive replication process detects incompatible changes on the source cluster, Hive replication fails. This sometimes occurs with recurring replications, where the metadata associated with an existing database or table on the source cluster changes over time.</p>
Invalidate Impala Metadata on Destination	<p>Choose the option to run the Impala INVALIDATE METADATA statement per table on the destination cluster after completing the replication. The statement purges the metadata of the replicated tables and views within the destination cluster's Impala upon completion of replication, allowing other Impala clients at the destination to query these tables successfully with accurate results.</p> <p>You must run the INVALIDATE METADATA statement manually for the following scenarios:</p> <ul style="list-style-type: none"> • If the destination Cloudera Manager does not have an Impala service. • If the source contains User Defined Functions (UDF). <p> Warning: However, this operation is potentially unsafe if DDL operations are being performed on any of the replicated tables or views while the replication is running. In general, directly modifying replicated data/metadata on the destination is not recommended. Ignoring this can lead to unexpected or incorrect behavior of applications and queries using these tables or views.</p>
Replication Strategy	<p>Choose one of the following replication strategies to determine whether the file replication tasks should be distributed among the mappers statically or dynamically.</p> <ul style="list-style-type: none"> • Static distributes file replication tasks among the mappers up front to achieve an uniform distribution based on the file sizes. • Dynamic distributes the file replication tasks in small sets to the mappers, and as each mapper completes its tasks, it dynamically acquires and processes the next unallocated set of tasks. <p>The default replication strategy is Dynamic.</p>
MapReduce Service	Choose the MapReduce or YARN service to use.
Log Path	Enter an alternate path for the logs, if required.

Option	Description
Error Handling	<p>Select the following options as necessary:</p> <ul style="list-style-type: none"> • Skip Checksum Checks - Determines whether to skip checksum checks on the copied files. If selected, checksums are not validated. Checksums are checked by default. <p> Note: You must skip checksum checks to prevent replication failure due to non-matching checksums in the following cases:</p> <ul style="list-style-type: none"> • Replications from an encrypted zone on the source cluster to an encrypted zone on a destination cluster. • Replications from an encryption zone on the source cluster to an unencrypted zone on the destination cluster. • Replications from an unencrypted zone on the source cluster to an encrypted zone on the destination cluster. <p>Checksums are used for two purposes:</p> <ul style="list-style-type: none"> • To skip replication of files that have already been copied. If Skip Checksum Checks is selected, the replication job skips copying a file if the file lengths and modification times are identical between the source and destination clusters. Otherwise, the job copies the file from the source to the destination. • To redundantly verify the integrity of data. However, checksums are not required to guarantee accurate transfers between clusters. HDFS data transfers are protected by checksums during transfer and storage hardware also uses checksums to ensure that data is accurately stored. These two mechanisms work together to validate the integrity of the copied data. <ul style="list-style-type: none"> • Skip Listing Checksum Checks - Whether to skip checksum check when comparing two files to determine whether they are same or not. If skipped, the file size and last modified time are used to determine if files are the same or not. Skipping the check improves performance during the mapper phase. Note that if you select the Skip Checksum Checks option, this check is also skipped. • Abort on Error - Whether to abort the job on an error. If selected, files copied up to that point remain on the destination, but no additional files are copied. Abort on Error is not selected by default. • Abort on Snapshot Diff Failures - If a snapshot diff fails during replication, the replication policy uses a complete copy to replicate data. If you select this option, the policy aborts the replication when it encounters an error instead.

Option	Description
Preserve	<p>Choose the required options to preserve the block size, replication count, permissions (including ACLs), and extended attributes (XAttrs) as they exist on the source file system, or to use the settings as configured on the destination file system. By default source system settings are preserved.</p> <p>When Permission is selected, and both the source and destination clusters support ACLs, replication preserves ACLs. Otherwise, ACLs are not replicated. When Extended attributes is selected, and both the source and destination clusters support extended attributes, replication preserves them. (This option only displays when both source and destination clusters support extended attributes.)</p> <p>If you select one or more of the Preserve options and you are replicating to S3 or ADLS, the values all of these items are saved in metadata files on S3 or ADLS. When you replicate from S3 or ADLS to HDFS, you can select which of these options you want to preserve.</p> <p> Note: To preserve permissions to HDFS, you must be running as a superuser on the destination cluster. Use the Run As Username option to set the username.</p>
Delete Policy	<p>Choose the required options to determine whether the files that were deleted on the source should also be deleted from the destination directory. This policy also determines the handling of files in the destination location that are unrelated to the source. Options include:</p> <ul style="list-style-type: none"> • Keep Deleted Files - Retains the destination files even when they no longer exist at the source. This is the default option. • Delete to Trash - If the HDFS trash is enabled, files are moved to the trash folder. This is not supported when replicating to S3 or ADLS. • Delete Permanently - Uses the least amount of space; use with caution.
Alerts	<p>Choose to generate alerts for various state changes in the replication workflow. You can choose to generate an alert On Failure, On Start, On Success, or On Abort of the replication job.</p> <p>You can configure alerts to be delivered by email or sent as SNMP traps. If alerts are enabled for events, you can search for and view the alerts on the Events tab, even if you do not have email notification configured. For example, if you choose Command Result that contains the Failed filter on the Diagnostics Events page, the alerts related to the On Failure alert for all the replication policies for which you have set the alert appear. For more information, see Managing Alerts and Configuring Alert Delivery.</p>

12. Click Create.

What to do next

After the replication policy runs successfully, you can view the replication job status on the **Replication Policies** page. Verify whether the job starts and runs as expected.

The administrator can verify whether the job starts and runs as expected using the following steps:

- For HDFS data replication, check the cloud storage path (for example, S3 bucket path) to verify whether the data was successfully copied in the specified bucket.
- For Hive metadata replication, verify whether the specified source database, along with tables, partitions, UDFs and column stats are available in the Data Lake HMS instance. For this, the administrator can use a Data Hub cluster and run the corresponding queries using Hue or beeline.
- For Ranger policies, query the Ranger policies to ensure that the Sentry policies are properly mapped as Ranger policies for the right users and groups.



Note: If the CDH source database contains functions, you must explicitly run the `reload` function command to view the migrated replication functions in the target location.

Manage and monitor Hive replication policies




After you create a Hive replication policy in CDP Public Cloud Replication Manager, you can perform and monitor various tasks related to the replication policy. You can view the job progress and replication logs. You can edit the advanced options to optimize a job run. You can suspend a job and also activate a suspended job. You can edit the replication policy as necessary.

About this task

On the **Replication Policies** page, you can perform the following actions and tasks on a replication policy and its jobs.

Procedure

- When you click Actions for a Hive replication policy, the following actions appear:

Action	Description
Edit*	<p>Change the replication policy options as required for non-expired policies that are in active or suspended state. Based on the schedule you choose, the replication policy replicates data.</p> <p>You can edit the replication policies to better align with changing requirements. For example, you might want to change the frequency of a policy depending on the data size and importance of the data being replicated.</p> <p> Note: A replication policy is associated with a cluster or a cluster pair, therefore you cannot change the clusters in the policy.</p> <p>Optionally, expand a replication policy on the Replication Policies page to edit the replication policy options which include frequency (start time cannot be modified if the policy has already started), queue name, maximum bandwidth, and maximum map slots.</p> <p> Tip: To optimize the replication policy performance, you can configure the queue name, maximum bandwidth, and maximum map slots as necessary.</p>
Delete	Deletes the replication policy permanently.
Suspend	Suspends a running replication policy. Activate the replication policy, if required.
View Log	<p>Download, copy, or open the log. The log shows a brief output of the stdout and stderr logs of a single step of the latest replication policy job run.</p> <p>You can also view the current job status in the Replication Manager Overview Issues & Updates Job Status column. If the job failed, click Failed to view the log details about the job.</p>
View Command Details	<p>Opens the latest Hive replication policy job page. The steps and substeps appear in a tree view. The failed steps are expanded by default, showing the last 15 lines of the log.</p> <p>You can also view the command details for a Hive replication policy on the Overview Issues & Updates panel.</p> <p> Tip: To view the complete log for all the jobs, go to the target cluster Cloudera Manager Running Commands page.</p>
Collect diagnostic bundle	<p>Generates a diagnostic bundle for the replication policy. You can download the bundle as a ZIP file to your machine.</p> <p>Ensure that you are logged into the Cloudera Manager instances for both the source and target clusters before you download the bundle in Replication Manager.</p>

Action	Description
*	<p>To view and use the replication policies with an empty name in Replication Manager, you must understand the following implementation:</p> <ul style="list-style-type: none"> If the Cloudera Manager API version is lower than 51, an existing replication policy with an empty name can be used and updated. However, if you edit the replication policy and provide a name for the replication policy in versions higher or equal to 51, you must ensure that the name conforms to the validation rules. If the Cloudera Manager API version is higher or equal to 51, it is mandatory that you provide a unique name to the replication policy to continue using it. This is because API version 51 and higher enforces the validation rules on all the replication policies. <p>To pass the replication policy name validation, you must ensure that the replication policy name is unique. The name can contain letters, numbers, and the <code>_</code> / <code>-</code> characters. You must also ensure that it does not contain the characters <code>%</code> <code>:</code> <code>;</code> <code>\</code> nor any character that is not ASCII printable, which includes the ASCII characters less than 32 and the ASCII characters that are greater than or equal to 127.</p>

- When you expand the policy details, the **Job History** panel appears.

You can view the following details on the panel:

- Previous jobs, current job, and one future scheduled job if any.
- Job details which include:

Job details	Description
Started	Timestamp when the job started.
Ended	Timestamp when the job ended.
Duration	Time taken to complete the job.
Tables	Number of imported or exported tables.
Progress	Current status of a running job.
Expected	Remaining number of files and bytes expected to be copied for a running job.
Copied	Number of files and bytes copied for a running job and completed job.
Failed	Number of files and bytes that failed to be copied for a completed job.
Deleted	Number of files deleted for a completed job.
Skipped	Remaining number of files and bytes skipped from copying for a running job and complete job.

- Click Actions to:

- Abort the job.
- Re-run an aborted or failed job.
- View Log for the job. You can download, copy, or open it to track the job and to troubleshoot any issues for the job.

- When you click a job on the **Job History** panel, the following tabs appear:

Tab	Description
General	<p>Shows the following job details:</p> <ul style="list-style-type: none"> Started at timestamp Duration taken to complete the job HDFS Replication Report to download the job statistics in CSV format Hive Replication Report to download the job statistics in CSV format Hive Export/Import is the number of external tables exported or imported using Hive replication. Number of Errors encountered during the replication job. Impala UDFs is the number of tables exported or imported using Impala. Job status Message.

Tab	Description
Command Details	Shows the details about the commands that ran on the source Cloudera Manager for the job, along with the timestamp.
Setup Error	Shows the stack trace for the commands that ran on the source Cloudera Manager for the failed job.

- You can download the following CSV reports from the **General HDFS Replication Report** field to track the replication jobs and to troubleshoot issues:

Report	Description
Listing	Lists all the files and directories copied during the replication job.
Status	Shows the complete status report of each file as: <ul style="list-style-type: none"> an Error occurred and the file was not copied. a Deleted file. an up-to-date file for which the replication was Skipped.
Error Status	Status report of all the copied files with errors. Each file shows the status, path, and message for the copied files with errors.
Skipped Status	Status report of all skipped files. Each file lists the status, path, and message for the databases and tables that were skipped.
Deleted Status	Status report of all deleted files. Each file lists the status, path, and message for the databases and tables that were deleted.
Performance	Summary report about the performance of the running replication job which includes the last performance sample for each mapper that is working on the replication job.
Full Performance	Performance report of the job which includes the samples taken for all mappers during the replication job.



Note: The reports are generated based on the source Cloudera Manager response. If the Cloudera Manager response is interrupted or is not handled as expected, corresponding error messages appear in HTML format in the reports.

- You can download the following CSV reports from the **General Hive Replication Report** field to track the replication jobs and to troubleshoot issues:

Report	Description
Hive Result	List of replicated tables.
Hive Performance	Performance report for Hive replication.

Using HBase replication policies

To create an HBase replication policy in CDP Public Cloud Replication Manager, you must register the on-premises cluster (CDH or CDP Private Cloud Base) as a classic cluster in Management Console, register cloud account credentials in the Replication Manager service, verify cluster access, and configure minimum ports for replication.

Related Information

[HBase replication policy](#)

[Support matrix for Replication Manager on CDP Public Cloud](#)

[CDP CLI for Replication Manager](#)

Preparing to create an HBase replication policy

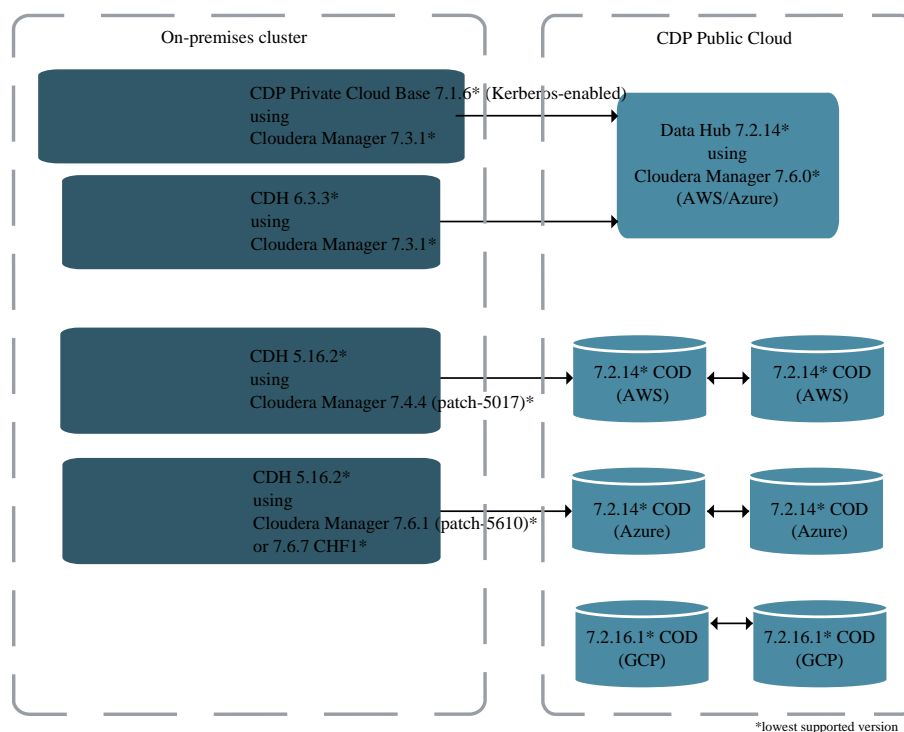
Before you create HBase replication policies in CDP Public Cloud Replication Manager, you must prepare the clusters, register cloud storage in Replication Manager, and verify cluster access.

Procedure

- Do the source cluster and target cluster meet the requirements to create an HBase replication policy?

The following image shows a high-level view of the support matrix for HBase replication policies, you must consult the [Support matrix for Replication Manager on CDP Public Cloud](#) for the complete list of supporting clusters and scenarios:

Figure 8: High-level replication scenarios supported by HBase replication policies



- Is the source CDH cluster or source CDP Private Cloud Base cluster registered as a classic cluster on the Management Console?

CDH clusters and CDP Private Cloud Base clusters are managed by Cloudera Manager. To enable these on-premises clusters for Replication Manager, you must register them as classic clusters on the Management Console. After registration, you can use them for data migration purposes.



Important: When you register a CDP Private Cloud Base cluster as a classic cluster, ensure that you use only the Cloudera Manager IP Address and Cloudera Manager Port options and not the Register KNOX endpoint (Optional) option.

For information about registering an on-premises cluster as a classic cluster, see [Adding a CDH cluster](#) and [Adding a CDP Private Cloud Base cluster](#).



Note:

- CDP Private Cloud Base 7.1.6 and higher clusters must be Kerberos enabled to use them as source classic clusters.
- You must have *Full Administrator* role on the source on-premises cluster.

- Are the following steps complete on the source CDP Private Cloud Base cluster or source CDH cluster (these steps are not required for COD sources)?
 1. Complete Step 1 of [Migrating HBase data from CDH/HDP to COD CDP Public Cloud](#) to install the HBase replication plugin parcel in the CDH source clusters.

This step is applicable for CDH versions 7.2.x that are lower than 7.2.2, versions 7.1.x that are lower than 7.1.5, and for versions lower than 7.x.

2. Create the /user/hbase folder for the hbase user in HDFS in the source cluster using the following commands:

```
sudo -u hdfs hdfs dfs -mkdir /user/hbase
sudo -u hdfs hdfs dfs -chown hbase:hbase /user/hbase
```

These commands allow the HBase replication policy to replicate the existing data in the source cluster.

This step is applicable for Cloudera Manager versions 7.4.3 or lower; Cloudera Manager 7.4.4 only if the API version is lower than v45. The endpoint `http://[***cm_host***]:[***cm_port***]/api/version` shows the API version of Cloudera Manager.

- Do you have the necessary permission to run the HBase replication jobs on YARN?

To verify whether you have the necessary permission, perform the following steps:

1. Go to the source Cloudera Manager *YARN service* Configuration tab.
2. Ensure one of the following conditions is met:
 - The Allowed System Users property must have *hbase*. Otherwise, add hbase to the existing property value.

This property lists the users permitted (or allowed) to run containers. Note that the users with IDs lower than the Minimum User ID property might be permitted (or allowed) to run containers.

- The Minimum User ID property is set to a value that is lower than the HBase user's ID.

To check the HBase user's ID, SSH into a cluster node and run the `id hbase` command.

The following sample snippet shows the HBase user's ID when you run the `id hbase` command:

```
# id hbase
uid=39993(hbase) gid=39993(hbase) groups=39993(hbase)
```

- Is the required cloud credential that you want to use in the replication policy registered with the Replication Manager service?

For more information, see [Working with cloud credentials](#).

You can also add the following advanced configuration settings to use Google Cloud, Amazon S3, and ADLS accounts in Replication Manager:

1. Go to the source Cloudera Manager Clusters *HDFS service* Configuration tab.
2. Locate the HDFS Client Advanced Configuration Snippet (Safety Valve) for `hdfs-site.xml` property.
3. Add the following key-value pairs to register a Google account to use in Replication Manager:

- `fs.gs.impl=com.google.cloud.hadoop.fs.gcs.GoogleHadoopFileSystem`
- `fs.gs.project.id=[***enter the project ID***]`
- `fs.gs.system.bucket=[***enter the bucket name***]`
- `fs.gs.working.dir=/`
- `fs.gs.auth.service.account.enable=true`
- `fs.gs.auth.service.account.email=[***enter the service principal email ID***]`
- `fs.gs.auth.service.account.keyfile=[***Enter the local path of the p12 file***]`

You can add these key-value pairs to register a Google account to use in Replication Manager from CDP Public Cloud 7.2.16.1 and higher versions.

4. Add the following key-value pairs to register an S3 account to use in Replication Manager:
 - `fs.s3a.access.key=[***enter the session access key***]`
 - `fs.s3a.secret.key=[***enter the session secret key***]`
 5. Add the following key-value pairs to register an ADLS account to use in Replication Manager:
 - `fs.azure.account.oauth2.client.id=[***enter the ABFS storage client ID***]`
 - `fs.azure.account.oauth2.client.secret=[***enter the ABFS storage client secret key***]`
 6. Save and restart the HDFS service for the changes to take effect.
- Do you have the required cluster access to create or view replication policies?
 - Have you assigned the managed identity of source roles, Storage Blob Data Owner or Storage Blob Data Contributor, to the destination storage data container and vice versa for bidirectional replication when you are using COD on Microsoft Azure?
- The roles allow writing a snapshot in the destination cluster container.
- Is the required target cluster (Data Hub or COD) available and healthy?
 - Does DNS resolution work as expected between the source and destination clusters?




Tip: If the destination cluster is not reachable from the source RegionServer hosts, add the hostname and IP address of the destination hosts to the `/etc/hosts` file on the RegionServers of the source cluster.

- Is the outgoing SSH port open on the Cloudera Manager host?
- Do you need to replicate data securely? If so, ensure that the SSL/TLS certificate exchange between two Cloudera Manager instances that manage source and target clusters respectively is configured. For more information, see *Configuring SSL/TLS certificate exchange between two Cloudera Manager instances*.

- Are the following ports open and available for Replication Manager?

Table 10: Minimum ports required for HBase replication policies

Ports	Service	Description
2181 and 16020	Destination hosts of the AWS cluster or ADLS cluster (target cluster), and the Cloudera Manager server port on the source cluster	Verify whether the ports 16020 for worker security group and 2181 for worker, master, and leader groups are open for connection from the source cluster to the destination cluster on AWS or Azure. This ensures that the source HBase service can communicate with Zookeeper and HBase services on the destination hosts uninterrupted. For more information, see Ports for HBase replication .
16000	HMaster	<p>Open the port on the Master Nodes (HBase Master Node and any back-up HBase Master node).</p> <p>Before you select the Validate Replication option during the first HBase replication policy creation between two specific clusters, you must ensure that the port is open on the target cluster.</p> <p> Note: Irrespective of whether this port is open or not on the Master nodes, Replication Manager displays a warning message to inform you that this port should be open on the target cluster (to communicate with the source cluster) when you choose Validate Replication on the Select Destination page during the HBase replication policy creation process.</p>
7180 or 7183	Cloudera Manager Admin Console HTTP	Open on the source cluster to enable Data lake Cloudera Manager to communicate to the on-premises Cloudera Manager. Connects to destination SDX Data Lake Cloudera Manager.
9000	Cloudera Manager Agent	Open on the source and target cluster to retrieve diagnostic and log information.
6000-6049	Cluster Connectivity Manager (CCM)	Required for SSL connections to the Control Plane via CCM to communicate with Replication Manager.
80 or 443	Data transfer from secondary node for AWS / ADLS Gen2	Outgoing port. Open on all the HDFS nodes for AWS and ADLS Gen2.
8443	Data Lake cluster	Outgoing port. Configure the port on the Data Lake cluster as the outgoing port for CDP Management Console to communicate with Cloudera Manager and Knox.
8032	YARN Resource Manager	Open on the source and target cluster to access the YARN ResourceManager.

Consider the following best practices while using CDP Public Cloud on Microsoft Azure ADLS Gen2 (ABFS):

- Ensure that the on-premises cluster (port 443) can access the <https://login.microsoftonline.com> endpoint. This is because the Hadoop client in the on-premises cluster (CDH/CDP Private Cloud Base) connects to the endpoint to acquire the access tokens before it connects to Azure ADLS storage. For more information, see the *General Azure guidelines* row in the [Azure-specific endpoints](#) table.

- Ensure that the steps mentioned in the *General Azure guidelines* and *Azure Data Lake Storage Gen 2* rows in the [Azure-specific endpoints](#) table are complete so that the endpoint connects to the target path successfully.
- (Optional) Complete the steps mentioned in the [Optimize HBase replication policy performance when replicating HBase tables with several TB data](#) FAQ if you choose Perform Initial Snapshot during HBase replication policy creation to replicate HBase tables with several TB data.

What to do next

After the clusters and cloud storage requirements are met, you can create an HBase replication policy.

Related Information

[Support matrix for Replication Manager on CDP Public Cloud](#)

[Adding a CDH cluster](#)

[Adding a CDP Private Cloud Base cluster](#)

[Accessing Data Lake services](#)

[Migrating HBase data from CDH/HDP to COD CDP Public Cloud](#)

[Working with cloud credentials](#)

[Configuring SSL/TLS certificate exchange between two Cloudera Manager instances](#)

[Ports for Replication Manager in CDP Public Cloud](#)

Creating HBase replication policy

You can replicate HBase data from a source classic cluster (CDH or CDP Private Cloud Base cluster), COD, or Data Hub to a target Data Hub or COD cluster in CDP Public Cloud Replication Manager.

Before you begin

Ensure that the [HBase replication policy prerequisites](#) are complete.

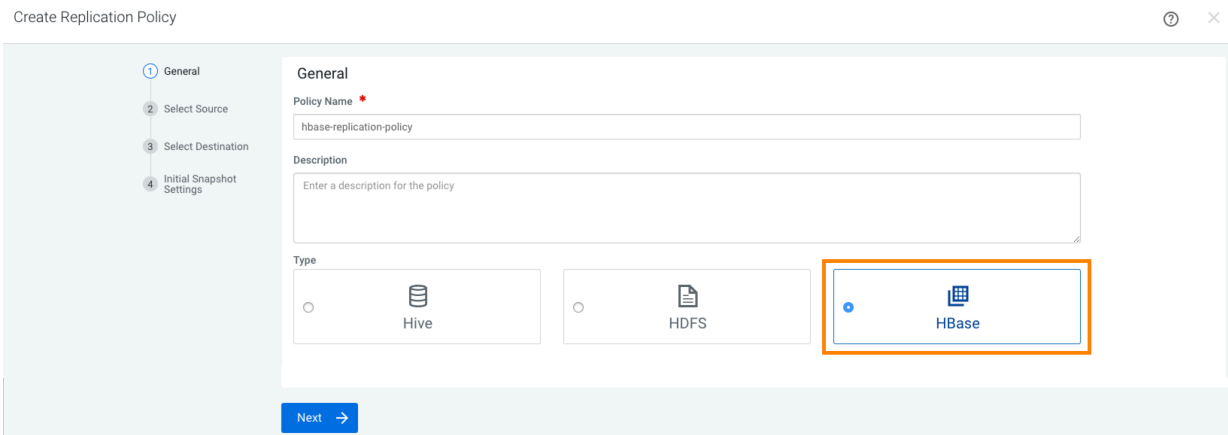
Procedure

1. On the Management Console Replication Manager Replication Policies page, click Add Policy.
The **Create Replication Policy** wizard appears.
2. On the General page, choose or enter the following information:

Option	Description
HBase	Creates an HBase replication policy.
Policy Name	Enter a unique name for the replication policy.


Option	Description
Description	Optional. Enter a brief description about the replication policy.

The following image shows a sample General page in the **Create Replication Policy** wizard:



- 3. Click Next.
- 4. On the Select Source page, enter or choose the options as required:

Option	Action
Source Cluster or Database	Choose a source cluster.
Source Tables	Enter a table name that you want to replicate. Click the Add icon to add more table names.
Perform Initial Snapshot	Select the option to replicate existing data.
Credentials are available in source cluster HDFS service configuration setting	You can choose this option when you want to use a Google Cloud account. You can use this option for S3 and ADLS accounts as well. Before you use this option, ensure that the advanced configuration settings in Preparing to create an HBase replication policy are configured.
Credentials from External Account	Choose this option for S3 and ADLS storage options. This option appears when you choose a CDP Private Cloud Base cluster or CDH cluster as the source cluster and you choose Perform Initial Snapshot. Click Add Cloud Credential. In the Add Cloud Credential dialog box, enter a unique name for the cloud credential. Click Save after you choose one of the following cloud storage types and enter the required options: <ul style="list-style-type: none">S3 - Choose an authentication type, enter an access key and secret key.ADLS - Enter the client ID, tenant ID, and secret key.

Option	Action
Replicate Database	<p>Replication Manager replicates:</p> <ul style="list-style-type: none"> current and future data from the existing tables if you choose the "Perform Initial Snapshot" option. Otherwise, only the future data is replicated. data from the future tables that are created after policy creation. <p>To replicate data from the future tables successfully, you must create similar empty tables on the target cluster. You can perform this action when you create or add a table to the database on the source cluster.</p> <p>You can choose the Replicate Database option only if the following conditions are true:</p> <ul style="list-style-type: none"> Target Cloudera Manager version is 7.11.0 or higher. Source cluster version is CDH 6.x or higher. <p>CDH 5.16.2 and higher versions also support the Replicate Database option after you upgrade the source cluster Cloudera Manager.</p> <ul style="list-style-type: none"> No existing HBase replication policies exist between the source and target clusters. <p> Tip: If you want to replicate the new tables that are created after the replication policy creation is complete, you must configure the replication scope to "1" for those tables on the source cluster.</p> <p>To configure the replication scope for a table on the master cluster, run the alter <code>[***table_name***], {NAME => [***column-family***], REPLICATION_SCOPE => 1}</code> command for each column family that must be replicated. REPLICATION_SCOPE is a column-family level attribute, where the value '0' means replication is disabled, and '1' means replication is enabled.</p>
Replicate all user tables	<p>Appears after you choose the Replicate Database option. Choose this option to replicate all the HBase tables in the database. This action sets the replication scope to 1 for all the tables and then replicates the tables to the target cluster.</p>

Option	Action
Replicate only tables where replication is already enabled	<p>Appears after you choose the Replicate Database option. Choose this option to replicate only those tables for which the replication scope is already set to 1. This provides you a choice to replicate only the required tables in a database.</p> <p>This option is supported only if the target cluster CDP version is 7.2.17.300 using Cloudera Manager 7.11.0-h3 or higher versions or CDP version 7.2.16.500 using Cloudera Manager 7.9.0-h7 or higher versions, or CDP version 7.12.0.0.</p>

The following sample image shows the Select Source page in the Create Replication Policy wizard:

The screenshot shows the 'Create Replication Policy' wizard with the 'Select Source' step active. On the left, a progress bar shows four steps: General (checked), Select Source (active), Select Destination, and Initial Snapshot Settings. The main area is titled 'Select Source' and contains the following fields:


- Source Cluster or Database ***: A dropdown menu showing 'Cluster 1 (hbase-repl-633)'.
- Cloud Credential ***: A dropdown menu showing 'azure-cred' with an 'Add Cloud Credential' button next to it.
- Source Tables**: Two input fields containing 't1' and 'test.t2', each with minus and plus icons for removal and addition.
- Perform Initial Snapshot**: A checked checkbox.

At the bottom are 'Back' and 'Next' buttons. On the right, a 'Summary' panel shows 'Type: HBase' and 'Policy Name: pol-hbase-cdh633-to-COD'.



5. Click Next.

6. On the Select Destination page, enter or choose the options as required:

Option	Description
Destination Data Hub or COD	Choose a Data Hub cluster or COD.
Set HBase Replication Machine User	<p>Optional. Choose the option and then enter the username and password. Ensure that you enter the correct password for an existing user because if the password is incorrect, the data is not replicated even though the policy is created successfully.</p> <p>Based on the username and password that you enter, one of the following possible scenarios is implemented by Replication Manager:</p> <ul style="list-style-type: none"> If Set HBase Replication Machine User is not selected, an HBase replication machine user is created automatically with an auto-generated username. If Set HBase Replication Machine User is selected and Create User If Does Not Exist is not selected, ensure that the username you enter exists in the CDP User Management System (UMS), otherwise an error message appears. If Set HBase Replication Machine User is selected and Create User If Does Not Exist is selected and the username does not exist in UMS, the username is created.

Option	Description
Sync Replication User	<p>Optional. Replication Manager validates the existing username with the UMS and synchronizes the new username and password to the destination cluster's environment (and to the source's as well if the source is COD).</p> <p> Note: Error appears after you click Sync Replication User if you entered a non-existent username and did not choose Create User If Does Not Exist.</p>

The following options appear if the source Cloudera Manager is 7.6.0 or higher:

Option	Description
Rolling HBase Service Restart on Source	[Appears if you select COD or Data Hub as the source cluster] Select to enable automatic rolling restart* of HBase service on the source cluster after the HBase replication policy first-time setup steps are complete. Otherwise, Cloudera Manager performs an automatic full restart* of the service.
Rolling HBase Service Restart on Destination	Select to enable automatic rolling restart* of HBase service on the target cluster as a rolling restart* after the HBase replication policy first-time setup steps are complete. Otherwise, Cloudera Manager performs an automatic full restart of the service.
I want to force the setup of this HBase replication policy	<p>Choose to run the first-time setup configuration between the selected source and destination clusters.</p> <p>This option appears when the selected source or target cluster is part of an existing cluster pair, and one of the following is true about the cluster pair:</p> <ul style="list-style-type: none"> • No HBase replication policies exist between them. • The other cluster in the pair is currently unreachable. <p>When you select the option, you acknowledge that the existing pairing for the selected source or target cluster will be cleared and the first-time setup will be initiated with the chosen new source or destination cluster.</p>
Validate Replication	<p>Select the option to notify Replication Manager to verify the provided details so that the replication is initiated after the policy creation is complete.</p> <p> Note: This option is available for target clusters using Cloudera Manager version 7.6.2 and higher.</p> <p>Before you select the Validate Replication option during the first HBase replication policy creation between two specific clusters, you must ensure that the 16000 port is open on the target cluster.</p> <p> Note: Irrespective of whether this port is open or not on the Master nodes, Replication Manager displays a warning message to inform you that this port should be open on the target cluster (to communicate with the source cluster) when you choose Validate Replication on the Select Destination page during the HBase replication policy creation process.</p>

Option	Description
*During rolling restart, one node is restarted at a time and this continues until all the nodes in the cluster are restarted. This type of restart ensures that there is no disruption of service. During full restart, all the nodes are shut down at once and restarted simultaneously.	

The following sample image shows the Select Destination page in the Create Replication Policy wizard:

Create Replication Policy

General

Select Source

Select Destination

Initial Snapshot Settings

Select Destination

Type *
Cloudera Data Platform - Data Hub Cluster

Some clusters may not be viewed by the current logged in user if role ClassicClusterUser is not assigned.

Destination Data Hub or COD *
dmx-tgt-valid (dmx-dev-15)

These clusters have not been set up as peers yet, please take care of the following

You can skip these steps if you have already done so.

Steps on destination cluster [cod-dz1xnfgq1ri](#)

1.1 Make sure the ZooKeeper and HBase Servers network ports on proper destination hosts accept incoming connections.

[Learn more on how to set up HBase replication.](#)

After policy creation

Beware that HBase service on both clusters will be automatically restarted after setup.

☒ Rolling HBase Service Restart on Source

☒ Rolling HBase Service Restart on Destination

☒ Validate Replication

☐ Set HBase Replication Machine User

Next →

Summary

Type
HBase

Policy Name
hbase-test4

Source
dmx-src-valid
r1

7. Click Next.
8. On the Initial Snapshot Settings page, configure the following options for the source cluster:

Option	Description
YARN Queue Name	Enter the name of the YARN queue for the cluster to which the replication job is submitted only if you are using Capacity Scheduler queues to limit resource consumption. The default value for this field is default.
Maximum Maps Slots	Configure the maximum number of map tasks (simultaneous copies) per replication job. The default value is 20.

Option	Description
Maximum Bandwidth	<p>Adjust this setting so that each map task is throttled to consume only the specified bandwidth.</p> <p>Each map task ((simultaneous copy) is restricted to consume only the specified bandwidth. This is not always exact. The map throttles back its bandwidth consumption during a copy in such a way that the net bandwidth used tends towards the specified value. You can adjust this setting so that each map task is throttled to consume only the specified bandwidth so that the net bandwidth used tends towards the specified value. The default value for the bandwidth is 100MB per second for each mapper.</p> <p>You can adjust the setting only if the source and destination Cloudera Manager instances support this option.</p>

The following sample image shows the Initial Snapshot Settings page in the Create Replication Policy wizard:

Create Replication Policy

☒ General
☒ Select Source
☒ Select Destination
☒ 4 Initial Snapshot Settings

Initial Snapshot Settings

YARN Queue Name [?](#)

default

Maximum Maps Slots [?](#)

20

Maximum Bandwidth [?](#)


100 MB/s (per mapper)

[< Back](#)
[Create >](#)

9. Click Create.

10. Restart the HBase service on the on-premises source cluster when the policy status on the **Replication Policies** page shows Manual restart (src) / restarting (dest) or Manual HBase restart needed on source. After the service

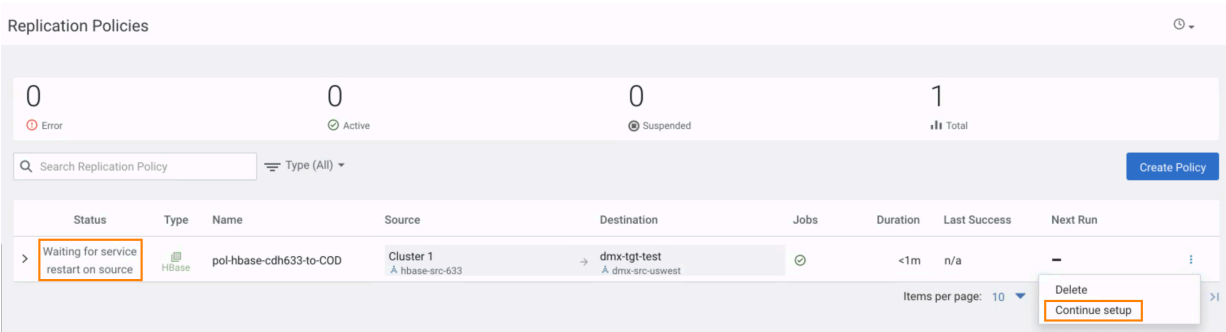
restart is complete, the setup continues automatically for the replication policy. You do not need to restart the HBase service if the source is COD or Data Hub.

- **Important:** If the source cluster Cloudera Manager version is 7.6.0 or lower and you are using an on-premises source cluster, you must perform the following steps to complete the HBase replication policy setup:
- a.

Restart the HBase service on the on-premises source cluster when the policy status on the **Replication Policies** page shows **Waiting for ‘Continue Setup’ action call**.
- b.


Click Continue setup for the policy on the **Replication Policies** page after the service restart is complete. This action informs Replication Manager to continue the replication policy setup.

The following image shows the Continue setup option for the HBase replication policy on the Replication Policies page:



Results

After you create the first replication policy between a source cluster and target cluster (policy that is in setup/service restart state), Replication Manager creates and runs two schedules or jobs. The first schedule shows the service configuration and service restart progress, and the second schedule shows the policy creation progress. Subsequent replication policies between the same source cluster and target cluster creates only one job. Replication Manager restarts the HBase services on both the clusters if they are COD clusters.

- **Note:** The first HBase replication policy between a source cluster and target cluster takes more time to complete because of the first-time setup configuration process that runs in the background. You can continue creating more HBase replication policies while the setup process of the in-progress policies is running in the background. The subsequent replication policies for the same source and target cluster do not run the first-time setup configuration.

What to do next

To verify whether a replication policy is running, you can either click the replication policy on the Replication Policies page, or click Running Commands in Cloudera Manager.

The following sample image shows the Job History page of a HBase replication policy between COD clusters:

Manage and monitor HBase replication policies

About this task

Procedure

- ## Monitor HBase replication policy job details

Replication policy job details

64

- **Type** of job. Click the job to view the completed and in-progress steps that Replication Manager takes to complete the job. Here, you can verify whether the first-time setup configuration between the source and destination clusters was successful.





Tip: You can also view the progress and results of the first-time setup on the [Cloudera Manager Running Commands](#) page for the source and destination clusters.

- Timestamp when the job **Started**.
- Timestamp when the job **Ended**.
- **Duration** taken to complete the job.
- **Progress** of the job run.

Actions menu

On the **Actions** menu, you can perform the following tasks on the HBase replication policy:

Option	Description
Edit*	Enables you to edit the replication policy name and description, or delete one or more tables for the HBase replication policy. You can also edit Source to delete the tables in the HBase replication policy.
Delete	Removes the HBase replication policy permanently. You can Force Delete an HBase replication policy after Replication Manager fails to delete the replication policy.
Suspend	Pauses an active HBase replication policy. This option appears for target clusters using Cloudera Manager 7.9.0 (Cloudera Manager API** version 51) and higher.  Warning: When you suspend an HBase replication policy, Replication Manager suspends all the HBase replication policies between the same source and destination cluster. You can suspend only the specified replication policy for target clusters using Cloudera Manager with API version lower than 45. However, the Suspend option is not available for clusters using Cloudera Manager with API versions 45 through 50.
Activate	Resumes a suspended HBase replication policy.  Warning: When you activate a suspended HBase replication policy, Replication Manager activates all the suspended HBase replication policies between the same source and destination cluster. You can activate only the specified suspended policy for target clusters using Cloudera Manager with API** version lower than 51.
View command details	Opens the latest HBase replication policy job page. The steps and substeps appear in a tree view. The failed steps are expanded by default, showing the last 15 lines of the log. You can also view the command details for a Hive replication policy on the Overview Issues & Updates panel. To view the complete log for all the jobs, go to the target cluster Cloudera Manager Running Commands page.

Option	Description
Retry First Time Setup	<p>Runs the first-time setup configuration between the source and destination clusters if the first-time setup has failed. This option is available only if the first-time setup configuration fails.</p> <p>After you click Retry Create, you can choose:</p> <ul style="list-style-type: none"> to set an Set HBase Replication Machine User and Sync Replication User to run the first-time setup. to opt for Rolling HBase Service Restart on Source and Rolling HBase Service Restart On Destination after the first-time setup is complete. <p>For more information about these options, see <i>Creating HBase replication policy</i>.</p>
Retry Failed Snapshots	<p>Reruns the failed initial snapshots (and only the failed ones) in the replication policy if the Replication Manager failed to replicate the existing data of some tables.</p> <p>This option appears if you selected Select Source Perform Initial Snapshot during policy creation and Replication Manager failed to replicate the existing data. In this scenario, the policy status shows Active with a <i>Snapshot Failure</i> warning message on the Replication Policies page for the HBase replication policy.</p>
Collect diagnostic bundle	<p>Generates a diagnostic bundle for the replication policy. You can download the bundle as a ZIP file to your machine.</p> <p>Ensure that you are logged into the Cloudera Manager instances for both the source and target clusters before you download the bundle in Replication Manager.</p>
<p>*</p> <p>To view and use the replication policies with an empty name in Replication Manager, you must understand the following implementation:</p> <ul style="list-style-type: none"> If the Cloudera Manager API version is lower than 51, an existing replication policy with an empty name can be used and updated. However, if you edit the replication policy and provide a name for the replication policy in versions higher or equal to 51, you must ensure that the name conforms to the validation rules. If the Cloudera Manager API version is higher or equal to 51, it is mandatory that you provide a unique name to the replication policy to continue using it. This is because API version 51 and higher enforces the validation rules on all the replication policies. <p>To pass the replication policy name validation, you must ensure that the replication policy name is unique. The name can contain letters, numbers, and the <code>_</code> / <code>-</code> characters. You must also ensure that it does not contain the characters <code>%</code> <code>.</code> <code>;</code> <code>\</code> nor any character that is not ASCII printable, which includes the ASCII characters less than 32 and the ASCII characters that are greater than or equal to 127.</p> <p>**</p> <p>The endpoint <code>http://[***cm_host***]:[***cm_port***]/api/version</code> shows the API version of the Cloudera Manager.</p>	

Creating triggers and monitoring replication-related metrics in Cloudera Manager

After you create an HBase replication policy between two Cloudera Operational Database (COD) clusters in CDP Public Cloud Replication Manager, you can set up the triggers in Cloudera Manager to monitor specific replication-related metrics for the HBase replication policy. When the condition specified in the trigger is met, Cloudera Manager triggers an alert, and you can view the metrics for the replication policy in CDP Public Cloud Replication Manager and in Cloudera Manager.

Procedure

1. Go to the **source** Cloudera Manager **Hosts All Hosts** page.
2. Click a link in the **Name** column to open the host status page.
3. Click **Create Trigger** in the **Health Tests** section.

For information about creating a trigger, see [Cloudera Manager Trigger Use Cases](#). For information about the list of supported HBase RegionServer replication peer metrics, see [HBase RegionServer Replication Peer Metrics](#).

4. Enter the required values and the following expression to create a trigger for a metric with a certain threshold value on the **Create New Trigger** page:

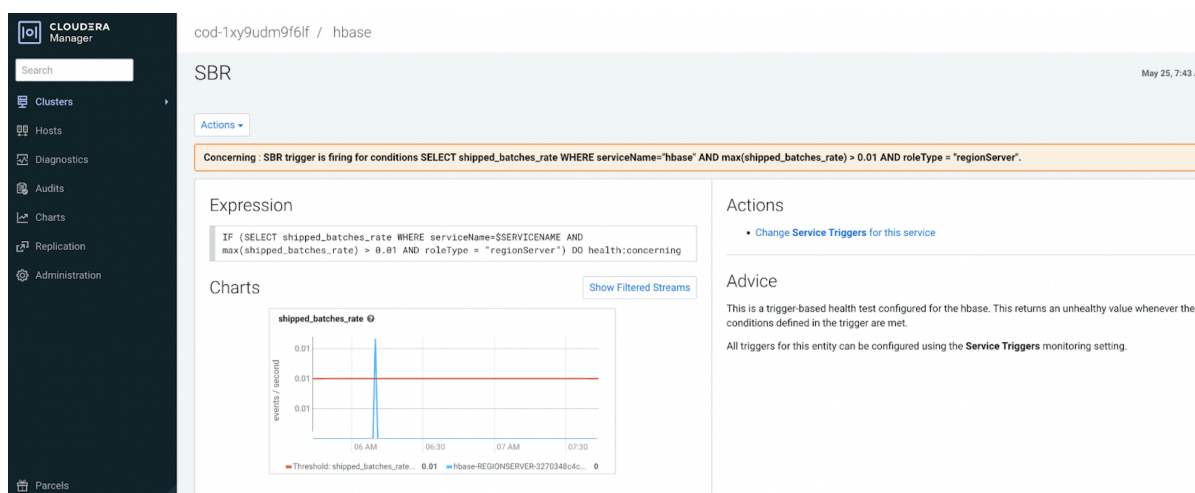
IF (SELECT [***enter the metric name***] WHERE serviceName=\$SERVICENAME AND max([***enter the metric name***]) > [***enter the threshold value***] AND roleType=RegionServer) DO health:concerning

The following examples illustrate how you can use triggers to monitor the replication metrics in Cloudera Manager:

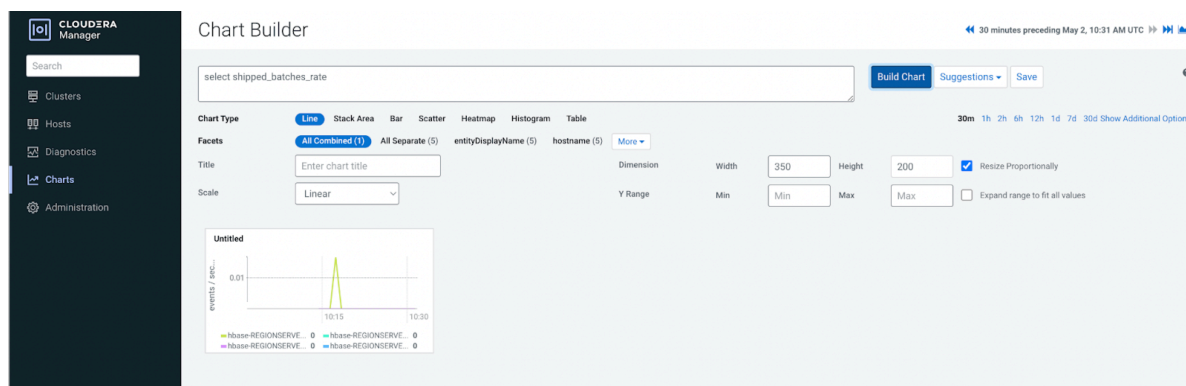
- a) The IF (SELECT log_queue_size WHERE serviceName=\$SERVICENAME AND max(log_queue_size) > 10 AND roleType=RegionServer) DO health:concerning trigger statement triggers the health alert when the value of the log_queue_size is greater than 2.

This trigger condition is met when the processing of the queue becomes slow which results in the logs being queued or the replication process being halted.

- b) The following sample image shows the chart that appears on the source Cloudera Manager Charts page when the health alert is triggered for the following trigger expression: IF (SELECT shipped_batches_rate WHERE serviceName=\$SERVICENAME AND max(shipped_batches_rate) > 0.01 AND roleType=RegionServer) DO health:concerning



- c) The following image shows the **Chart Builder** where you can configure the options as required and save it for troubleshooting purposes or for future references.



For information about monitoring the metrics related to replication and HBase health on your COD environment, see [Monitor COD metrics](#).

What to do next

You can view and monitor the metrics in Cloudera Manager and on the **Charts** page in CDP Public Cloud Replication Manager.

Monitor HBase RegionServer replication peer metrics in Replication Manager

When you click a replication policy on the "Replication Policies" page, the "Charts" tab appears. On this tab, when you choose the HBase replication peer metric, aggregate rollup level, start time, and end time, a graphical representation of HBase RegionServer metrics specific to the replication peer appears.

Cloudera Manager monitors several metrics which includes performance metrics for the HBase RegionServer. Replication Manager leverages this capability to display the HBase RegionServer metrics specific to a replication peer as a graph on the **Charts** tab for a replication policy, and stores it in the source cluster of the replication policy. You can use these metrics to monitor a HBase replication job and to find and diagnose issues related to the HBase replication peer.

You can view these metrics for a replication peer only if the following conditions are true:

- The source Cloudera Manager API version is 42 or higher and the target Cloudera Manager API version is 53 or higher.



Tip: The endpoint `http://[***cm_host***]:[***cm_port***]/api/version` shows the API version of Cloudera Manager.

- The HBase replication policy has been successfully created and is not in an erroneous state.



Note: If the target Cloudera Manager version is 45 or higher, the graphs generated for HBase replication policies are the same if they share the same source and target clusters. If the target Cloudera Manager version is lower than 45, the graph differs from one HBase replication policy to another because one replication peer is created for each HBase replication policy in these API versions.

For more information about the available metrics, see [HBase RegionServer Replication Peer Metrics](#).

Viewing HBase RegionServer replication peer metrics

You can view HBase RegionServer metrics specific to a replication peer as a graph on the "Charts" tab for a replication policy on the "Replication Policies" page.

Procedure

1. Go to the **Replication Policies** page in CDP Public Cloud Replication Manager.
2. Click a successful HBase replication policy that is not in an erroneous state.
3. Go to the **Charts** tab.

4. Choose the following options depending on your requirement:
 - a) Choose one of the following HBase replication peer metric:
 - Age of last shipped operation
 - Age of last shipped operation rate
 - Log edits filtered rate
 - Log edits read rate
 - Log queue size
 - Log read in bytes rate
 - Shipped batches rate
 - Shipped ops rate
 - Shipped size in kb rate
 - Shipped hfiles rate
 - Size of hfile refs queue
 - b) Choose an aggregate rollup level for the metrics:
 - Raw
 - 10 minutes
 - 1 hour
 - 6 hours
 - Daily
 - Weekly
 - c) Choose a Start Time.
 - d) Choose an End Time.
5. Click Load Metrics to view the graphs.

Troubleshooting replication policies in CDP Public Cloud Replication Manager

The troubleshooting scenarios in this topic help you to troubleshoot issues in the Replication Manager service in CDP Public Cloud.

Related Information

[Using HDFS replication policy](#)

[Using Hive replication policy](#)

[Using HBase replication policy](#)

[Support matrix for Replication Manager on CDP Public Cloud](#)

Different methods to identify errors related to failed replication policy

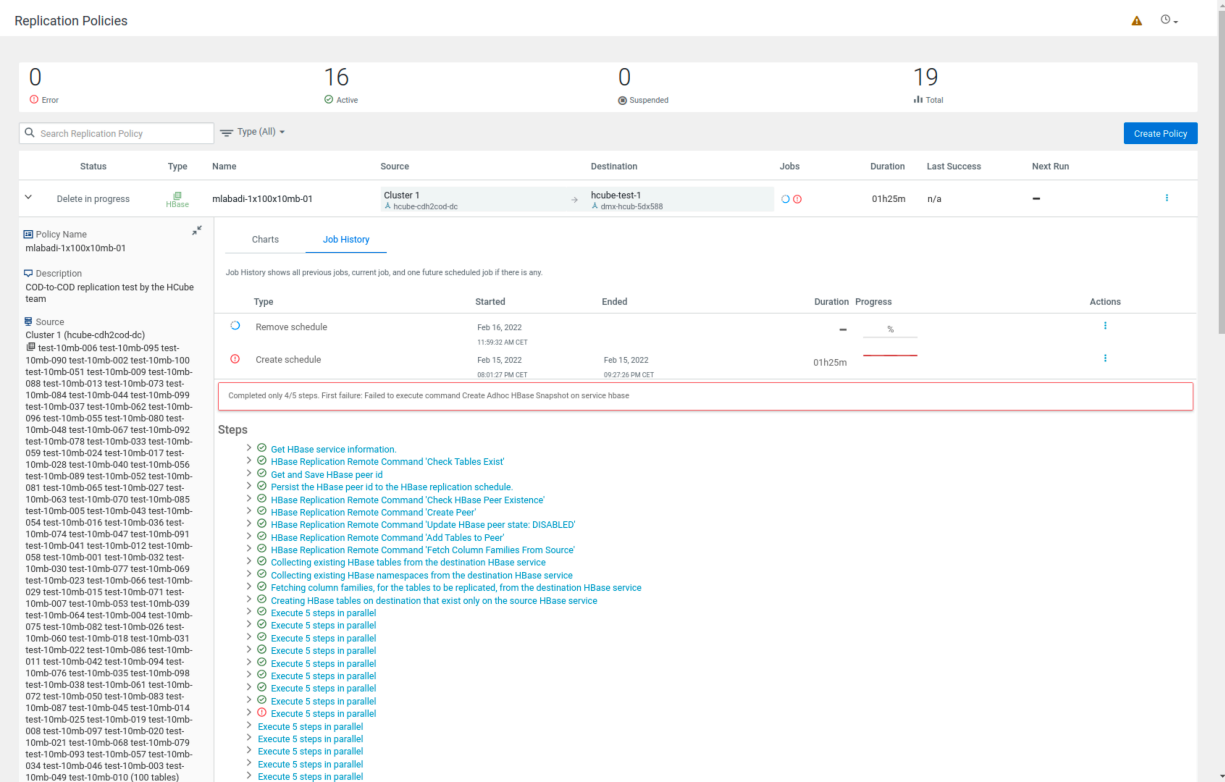
What are the different methods to identify errors while troubleshooting a failed replication policy?

Procedure

You can choose one of the following methods to identify the errors to troubleshoot a job failure:

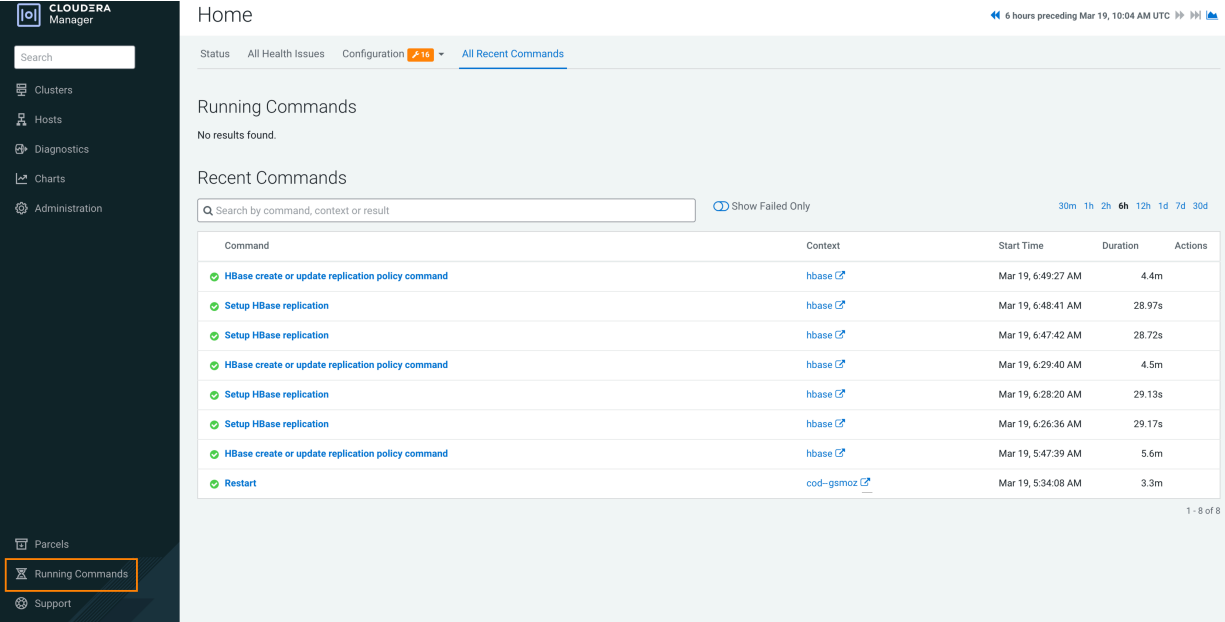
- On the **Replication Policies** page, click the failed job in the **Job History** pane. The errors for the failed job appear.

The following sample image shows the **Job History** pane for a replication policy job:



- In the source and target Cloudera Manager, click **Running Commands** on the left navigation bar. The recent command history shows the failed commands.

The following sample image shows the **Running Commands** page for an HBase replication policy:



- On the source cluster and target cluster, open the service logs to track the errors (For example, HBase service logs).

You can also search on the Cloudera Manager Diagnostics Logs page to view the logs.

HDFS replication policy fails due to export HTTPS_PROXY environment variable

HDFS replication policies fail when the export HTTPS_PROXY environment variable is set to access AWS through proxy servers. How to resolve this issue?

Remedy

Procedure

To resolve this issue, perform the following steps:

1. Open the core-site.xml file on the source cluster.
2. Enter the following properties in the file:

```
<property>
  <name>fs.s3a.proxy.host</name>
  <description>Hostname of the (optional) proxy server for S3 connection
s.</description>
</property>

<property>
  <name>fs.s3a.proxy.port</name>
  <description>Proxy server port. If this property is not set
    but fs.s3a.proxy.host is, port 80 or 443 is assumed (consistent with
    the value of fs.s3a.connection.ssl.enabled).</description>
</property>
```

3. Save and close the file.
4. Restart the source Cloudera Manager.
5. Run the failed HDFS replication policies in Replication Manager.
Replication Manager completes the replication successfully.

Cannot find destination clusters for HBase replication policies

When you ping destination clusters using their host names, the source cluster hosts for HBase replication policies do not find the destination clusters. How to resolve this issue?

Cause

This might occur for on-premises clusters such as CDP Private Cloud Base clusters or CDH clusters because the source clusters are not on the same network as the destination Data Hub. Therefore, hostnames cannot be resolved by the DNS service on the source cluster.

Remedy

Procedure

- Add the destination Region Server and Zookeeper IP to host name mappings in the /etc/hosts files of all the Region Servers on the source cluster.

The following snippet shows the contents in a sample /etc/hosts file:

```
10.115.74.181 dx-7548-worker2.dx-hbas.x2-8y.dev.dr.work
10.115.72.28 dx-7548-worker1.dx-hbas.x2-8y.dev.dr.work
10.115.73.231 dx-7548-worker0.dx-hbas.x2-8y.dev.dr.work
10.115.72.20 dx-7548-master1.dx-hbas.x2-8y.dev.dr.work
```

```
10.115.74.156 dx-7548-master0.dx-hbas.x2-8y.dev.dr.work  
10.115.72.70 dx-7548-leader0.dx-hbas.x2-8y.dev.dr.work
```

HBase replication policy fails when Perform Initial Snapshot is chosen

An HBase replication policy fails for COD on Microsoft Azure when the "Perform Initial Snapshot" option is chosen but data replication is successful when the option is not chosen. How to resolve this issue?

Cause

This issue appears when the required managed identity of source roles are not assigned.

Remedy

Procedure

Assign the managed identity of source roles, Storage Blob Data Owner or Storage Blob Data Contributor, to the destination storage data container and vice versa for bidirectional replication.

The roles allow writing a snapshot in the destination cluster container.

Optimize HBase replication policy performance when replicating HBase tables with several TB data

Can HBase replication policy performance be optimized when replicating HBase tables with several TB of data if the "Perform Initial Snapshot" option is chosen during HBase replication policy creation?

Complete the following manual steps to optimize HBase replication policy performance when replicating several TB of HBase data if you choose the Perform Initial Snapshot option during the HBase replication policy creation process.

Remedy

Procedure

1. Before you create the HBase replication policy, perform the following steps:
 - a) Navigate to the source Cloudera Manager YARN service Configuration tab.
 - b) Search for the `mapreduce.task.timeout` parameter.
 - c) Increase the value or set it to 0 to switch off the timeout.
 - d) Restart the YARN service.
 - e) Navigate to the source Cloudera Manager HBase service Configuration tab.
 - f) Search and configure the following key-value pairs:
 - `hbase.snapshot.master.timeout.millis` = 840000
 - `hbase.client.sync.wait.timeout.msec` = 180000
 - `hbase.client.operation.timeout` = 2400000
 - `hbase.client.procedure.future.get.timeout.msec` = 3000000
 - `hbase.hfilearchiver.thread.pool.max`=100
 - `hbase.snapshot.thread.pool.max`=24
 - g) Restart the HBase service.
 - h) Perform steps e through g on the target Cloudera Manager.

2. When you create the HBase replication policy for the first time using the above configured source cluster, you must increase the Maximum Map Slots value to a higher number on the Advanced Settings page.



Tip: Calculate this number by multiplying the number of nodes on the source cluster and the number of cores in the node.

3. If Store File Tracking (SFT) is enabled in the target COD, perform the steps mentioned in the [COD migration](#) topic after the replication policy creation is complete.



Note: SFT is enabled by default on CDP clusters with COD version 7.2.14.2 and higher using Cloudera Manager versions 7.2.16 and higher.

Partition metadata replication takes a long time to complete

How can partition metadata replication be improved when the Hive tables use several Hive partitions?

Hive metadata replication process takes a long time to complete when the Hive tables use several Hive partitions. This is because the Hive partition parameters are compared during the import stage of the partition metadata replication process and if the exported and existing partition parameters do not match, the partition is dropped and recreated. You can configure a key-value pair to support partition metadata replication.

Procedure

1. Go to the Cloudera Manager Clusters *Hive service* Configuration tab.
2. Search for the Hive Replication Environment Advanced Configuration Snippet (Safety Valve) property.
3. Enter the `HIVE_IGNORED_PARTITION_PARAMETERS=[***comma separated list of Hive partition parameters***]` key-value pair.

For example,

```
HIVE_IGNORED_PARTITION_PARAMETERS=transient_lastDdlTime,totalSize,numRows,COLUMN_STATS_ACCURATE,numFiles
```

The partition parameter names you provide are not compared during the import stage of the partition metadata replication process. Therefore, even if the partition parameters do not match between the exported and existing partitions, the partition is not dropped or recreated. After you configure this key-value pair, the import stage of the partition metadata replication process completes faster.

4. Save the changes, and restart the Hive service.

Replicating Hive nested tables

CDP Public Cloud Replication Manager does not support Hive nested tables. What do I do if there are Hive nested tables in the source cluster?

CDP Public Cloud Replication Manager does not support Hive nested tables for replication. Therefore, it is recommended that you move the nested tables to a different location in HDFS and then replicate Hive external tables. However, if this is not possible, you can perform the following steps in the given order as a workaround.

Solution

Procedure

1. Create a Hive replication policy on the target cluster. Ensure that the Additional Settings Replication Option Metadata only option is selected to replicate the metadata of required files and directories.



Note: If you are using CDP Public Cloud 7.2.15 or lower and if you are using Amazon S3 as the disaster recovery cluster, ensure that you (the administrator) run the following command after the Hive metadata replication is complete:

```
ALTER TABLE table SET LOCATION "[***[s3a://s3_bucket/warehouse_path]***]";
```

2. Create a HDFS replication policy on the source cluster to replicate the table data.

Target HBase folder is deleted when HBase replication policy fails

When the snapshot export fails during the HBase replication policy job run, the target HBase folder in the destination Data Hub or COD gets deleted.



Note: This scenario appears if you are using Cloudera Manager versions that are lower than 7.6.7 CHF8, 7.11.0, or 7.9.0-h6 on the source cluster.

You can either revoke the delete permission for the user, or ensure that you use an access key/role that does not have delete permissions to the required storage component.

The following steps show how to create an access key in AWS and an Azure service principal, which do not have delete permission for the storage component.

Solution in AWS

Procedure

If the Data Hub or COD is on Amazon S3, you can perform the following steps to create an access key that does not have delete permission for the storage component:

1. Login to AWS.
2. Click Create policy on the IAM Policies Create policy page.

The following sample image shows the **Policies** page in the AWS IAM section to create a policy.

Policy name	Type	Used as	Description
AllowBillingReaper	Customer managed	Permissions policy (1)	Allow management of reser
AllowROBilling	Customer managed	Permissions policy (1)	Allow RO access to billing i
asinghal-cb	Customer managed	Permissions policy (2)	
aws-cb-policy	Customer managed	Permissions policy (1)	
besi-dev-assume-full-bucket-role	Customer managed	Permissions policy (1)	

3. Choose the S3 Service, and then choose the following permissions for generic access (assign resources as required):
 - *List/ListBucket*
 - *Read/GetBucketLocation*
 - *Read/GetObject*
 - *Write/AbortMultipartUpload*
 - *Write/PutObject*
 - *Permissions management/PutBucketPublicAccessBlock*
4. Add *Delete/DeleteObject* permission to the target COD cluster's snapshot temporary folder. For example, the target COD cluster's snapshot temporary folder might be located in *[***target COD S3 path***/hbase/.hbase-snapshot/.tmp/*]*.
5. Enter a Name for the policy, add tags, and click Create policy.
6. Click Add Users to create a user on the IAM Users page.
7. Enter a Name, and click Next.
8. Choose the Attach policies directly option on the **Set permission** page, and then assign the previously created policy to the user.
9. Optionally, add tags and create the user.
10. Click Create access key to create an access key for the user on the IAM Users *[***new user***]* page.
11. On the Security Credentials Access keys page, choose Application running outside AWS. Click Next.
12. Optionally, attach the tags, create and save the access key. This access key is used as an external account for replication.

How do I verify whether the target HBase folder in the destination Data Hub or COD does not get deleted if the snapshot export fails during the HBase replication policy job run?

Perform the following steps to verify if the delete operation is allowed for the access key that you previously created:

- a. Run the `aws configure --profile delete-test` command to setup the credentials in AWS CLI.
- b. Delete an arbitrarily created temporary file from the account using the `aws s3 --profile delete-test rm --recursive s3://[***account name***/delete-testing/` command.

The delete operation is not allowed.

Solution in Microsoft Azure

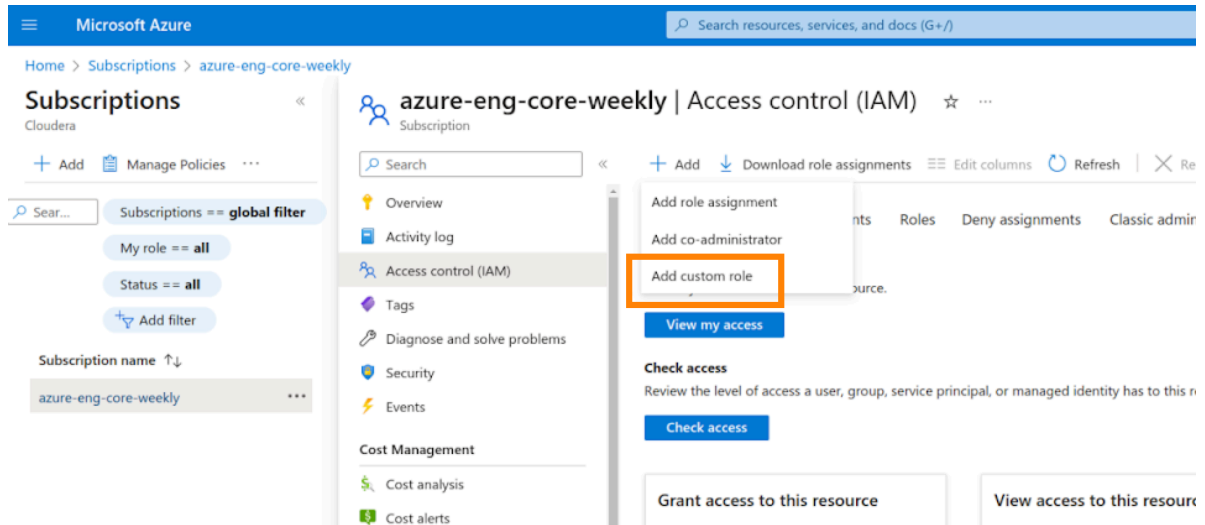
Procedure

If the Data Hub or COD is on ABFS, you can perform the following steps to create an Azure service principal that does not have delete permission for the storage component:

1. Login to Microsoft Azure.

2. Click **Add Add a custom role** on the **Subscriptions** Select the subscription **Access Control (IAM)** page in Microsoft Azure, and complete the following steps:
 - a) On the **Basics** tab, provide a name for the role, select **Clone a role** for the **Baseline permissions** field, and choose **Storage Blob Data Contributor** for the **Role to clone** field. Click **Next**.
 - b) On the **Permissions** tab, remove the *Microsoft.Storage/storageAccounts/blobServices/containers/delete* and *Microsoft.Storage/storageAccounts/blobServices/containers/blobs/delete* permissions.
 - c) Click **Review + create**.

The following sample image shows the Access control (IAM) page in Microsoft Azure:



3. Click **Add Add role assignment** on the **Storage accounts** Select the account where snapshots will be written **Access Control (IAM)** page, and complete the following steps:
 - a) On the **Role** tab, select the custom role previously created. Click **Next**.
 - b) On the **Members** tab, select **User, group, or service principal** for **Assign access to** field, and select the required service principal.
 - c) Click **Review + assign**.
 - d) Click **Review + assign** on the **Conditions (optional)** tab.
4. Click **Add principal** on the **Storage accounts** ****account name**** ****container where snapshots are to be written**** **Manage ACL** page.
 - a) Select the required service principal.
 - b) Choose the **Execute** permission for the required container, and click **Save**.



Important: Ensure that no other role containing delete access to the given storage account is assigned to the new service principal on the **Access Control (IAM)** page.

How do I verify whether the target HBase folder in the destination Data Hub or COD does not get deleted if the snapshot export fails during the HBase replication policy job run?

To verify if the delete operation is allowed on the service principal that you previously created, perform the following steps:

- a. Open the Azure Cloud Shell terminal.
- b. Login using the service principal that you created previously using the `az login --service-principal -u ***client id*** -p ***client secret*** --tenant ***tenant id***` command.
- c. Delete an arbitrarily created temporary file from the account using the `az storage fs file delete --path ***temporary file*** -f data --account-name ***account name*** --auth-mode login` command.

The delete operation is not allowed.

Replicate HBase data in existing and future tables

Errors might appear when you try to replicate HBase data from existing tables and future tables in a database using the “Replicate Database” option during the HBase replication policy creation process. These errors appear when there are compatibility issues.

The following list shows a few errors that might appear and how to mitigate these issues:

Procedure

- The error *Destination cluster doesn't support replication of all HBase tables. Please change the destination cluster or specify the tables that should be replicated.* appears during HBase replication policy creation process. How to mitigate this issue?

This error appears if you choose the `Select Source Replicate Database` option during the HBase replication policy creation process for unsupported cluster versions.

To mitigate this issue, ensure that the target Cloudera Manager version is 7.11.0 or higher and the source cluster version is CDH 6.x or higher before you choose the `Select Source Replicate Database` option during the HBase replication policy creation process.

- The error *Exception in thread "main" java.lang.IllegalArgumentException: Cannot add a peer with id=_repl__f1907207cd4a528777bb4a316_ba23f09f7328494bbcbf81f40b because that id already exists* appears after creating subsequent HBase replication policies.

This error appears if you created an HBase peer manually using the *hbase shell* to replicate all column families on a source cluster using CDH 5.x.

You can replicate the HBase data (existing tables and future tables) in a database only if the target Cloudera Manager version is 7.11.0 or higher and the source cluster version is CDH 6.x or higher. Therefore, even if you create a peer using *hbase shell* in the source cluster version CDH 5.16.2 or any other unsupported version, errors might appear.