

How To Back Up and Restore HDFS Data Using Cloudera Enterprise BDR

[\(#howto backup restore hdfs\)](#)

Cloudera Enterprise Backup and Disaster Recovery (BDR) uses replication schedules to copy data from one cluster to another, enabling the second cluster to provide a backup for the first. In case of any data loss, the second cluster—the backup—can be used to restore data to production.

This tutorial shows you how to create a replication schedule to copy HDFS data from one cluster to another for a backup, and how to create and test a replication schedule that you can use to restore data when needed in the future.

Creating replication schedules for backup and restore requires:

- A license for Cloudera Enterprise. Cloudera Enterprise BDR is available from the Backup menu of Cloudera Manager Admin Console when licensed for Enterprise.
- The **BDR Administrator** or **Full Administrator** role on the clusters involved (typically, a production cluster and a backup cluster).

This tutorial includes the following information:

- [Best Practices for Back Up and Restore](#)
- [About the Example Clusters](#)
- [Backing Up and Restoring HDFS Data](#)
 - [Backing Up HDFS Files](#)
 - [Step 1: Establish a Peer Relationship to the Production Cluster](#)
 - [Step 2: Configure the Replication Schedule for the Backup](#)
 - [Step 3: Verify Successful Replication](#)
 - [Configuring a Restore Replication Schedule](#)
 - [Step 1: Establish a Peer Relationship to the Backup Cluster](#)
 - [Step 2: Configure Replication Schedule to Test the Restore](#)
 - [Step 3: Disable the Replication Schedule](#)
 - [Step 4: Test the Restore Replication Schedule](#)
 - [Step 5: Verify Successful Data Restoration](#)
 - [Recovering from Catastrophic Data Loss](#)
 - [Step 1: Disable the Backup Replication Schedule](#)
 - [Step 2: Edit the Existing Replication Schedule](#)
 - [Step 3: Run the Restore Replication Schedule](#)
 - [Step 4: Return the Restore Replication Schedule to a Disabled State](#)
 - [Step 5: Re-enable the Backup Replication Schedule](#)

Best Practices for Back Up and Restore [\(#general_guidelines\)](#)

When configuring replication schedules for HDFS back up and restore, follow these guidelines:

- Make sure that the time-frames configured for replication schedules allow for the replication process to complete.
- Create a restore replication schedule in advance but leave it disabled, as shown in this tutorial.
- Test your replication schedules for both back up and restore before relying on them in a production environment.
- Enable only one replication schedule for the same dataset at the same time. That means you must first disable the backup replication schedule before enabling or creating a restore replication schedule for the same dataset, and vice versa.
- Enable snapshots on the HDFS file system. Snapshots ensure consistency if changes are still being made to data during the replication process. See [Using Snapshots with Replication \(cm_bdr_snap_repl.html#concept_cqn_sqb_kt\)](#) for details.

About the Example Clusters [\(#about_the_example_clusters\)](#)

This tutorial uses the two example clusters listed in the table. The nodes shown are the master nodes for the clusters.

Production cluster	Backup cluster
cloudera-bdr-src-{1..4}.vpc.cloudera.com	cloudera-bdr-tgt-{1..4}.vpc.cloudera.com
http://cloudera-bdr-src-1.vpc.cloudera.com	http://cloudera-bdr-tgt-1.vpc.cloudera.com
Source cluster for a backup replication schedule.	Destination cluster for a backup replication schedule.
Destination for a restore replication schedule.	Source for a restore replication schedule.

Production cluster	Backup cluster
For restore, set peer relationship from this cluster.	For backup, set peer relationship from this cluster. To create an initial backup, set Schedule to Immediate.

The example clusters are not configured to use Kerberos, nor do they use external accounts for cloud storage on Amazon Web Services (AWS).

The example production cluster contains nine HDFS files in the `/user/cloudera` path:

```
hdfs@cloudera-bdr-src-1:~
[hdfs@cloudera-bdr-src-1 ~]$ hadoop fs -ls /user/cloudera
Found 9 items
-rw-r--r-- 3 hdfs supergroup 16106127360 2016-09-01 14:17 /user/cloudera/large.0
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:29 /user/cloudera/medium.0
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:33 /user/cloudera/medium.1
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:35 /user/cloudera/medium.2
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:36 /user/cloudera/small.0
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:36 /user/cloudera/small.1
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:47 /user/cloudera/small.2
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:48 /user/cloudera/small.3
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:48 /user/cloudera/small.4
[hdfs@cloudera-bdr-src-1 ~]$
```

The example backup cluster has not been used as the destination of a replication schedule yet, so the HDFS file system has no `/user/cloudera` directory:

```
hdfs@cloudera-bdr-tgt-1:~
[hdfs@cloudera-bdr-tgt-1 ~]$ hadoop fs -ls /user/cloudera
ls: '/user/cloudera': No such file or directory
[hdfs@cloudera-bdr-tgt-1 ~]$
```

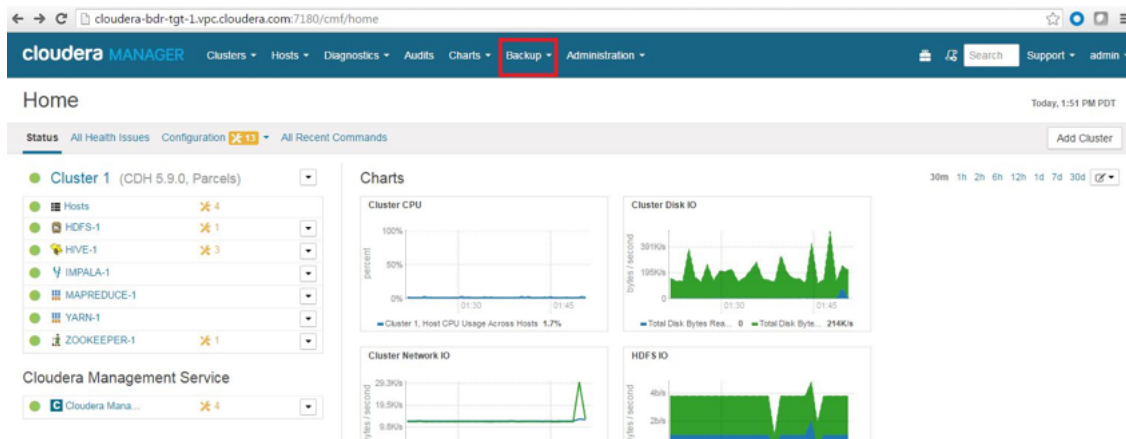
Backing Up and Restoring HDFS Data ([#backup and restore using rep scheds](#))

This tutorial steps through the processes of creating and running a backup replication schedule, and creating a restore replication schedule designed for future use.

Preparing for disaster recovery includes these three major tasks:

- [Backing Up HDFS Files \(cm_bdr_howto_hdfs.html#backup_hdfs_from_prod_to_backup\)](#)
- [Configuring a Restore Replication Schedule \(cm_bdr_howto_hdfs.html#replicate-data-from-backup-cluster\)](#)
- [Recovering from Catastrophic Data Loss \(cm_bdr_howto_hdfs.html#catastrophic_data_loss\)](#)

Backup and restore are both configured and managed using Replication Schedules, available from the Backup menu on Cloudera Manager Admin Console:



Note: Screenshots in this guide show version 5.9 of the Cloudera Manager Admin Console.

The backup and restore processes are configured, managed, and executed using replication schedules. Each replication schedule identifies a *source* and a *destination* for the given replication process. The replication process uses a *pull* model. When the replication process runs, the configured destination cluster accesses the given source cluster and transparently performs all tasks needed to copy the HDFS files to the destination cluster.

The destination cluster handles configuration and running the schedule. Typically, creating a backup replication schedule takes place on the backup cluster and creating a restore replication schedule takes place on the production cluster. Thus, as shown in this tutorial, the example production cluster, `cloudera-bdr-src-{1..4}.vpc.cloudera.com`, is the source for the backup replication schedule and the destination for the restore replication schedule.

Backing Up HDFS Files ([#backup_hdfs_from_prod_to_backup](#))

The backup process begins at the Cloudera Manager Admin Console on the cluster designated as the backup, and includes these steps:

- [Step 1: Establish a Peer Relationship to the Production Cluster \(cm_bdr_howto_hdfs.html#establish_peer_from_backup_to_source\)](#)
- [Step 2: Configure the Replication Schedule for the Backup \(cm_bdr_howto_hdfs.html#configure_hdfs_replication_schedule_for_backup\)](#)
- [Step 3: Verify Successful Replication \(cm_bdr_howto_hdfs.html#validate_the_backup\)](#)

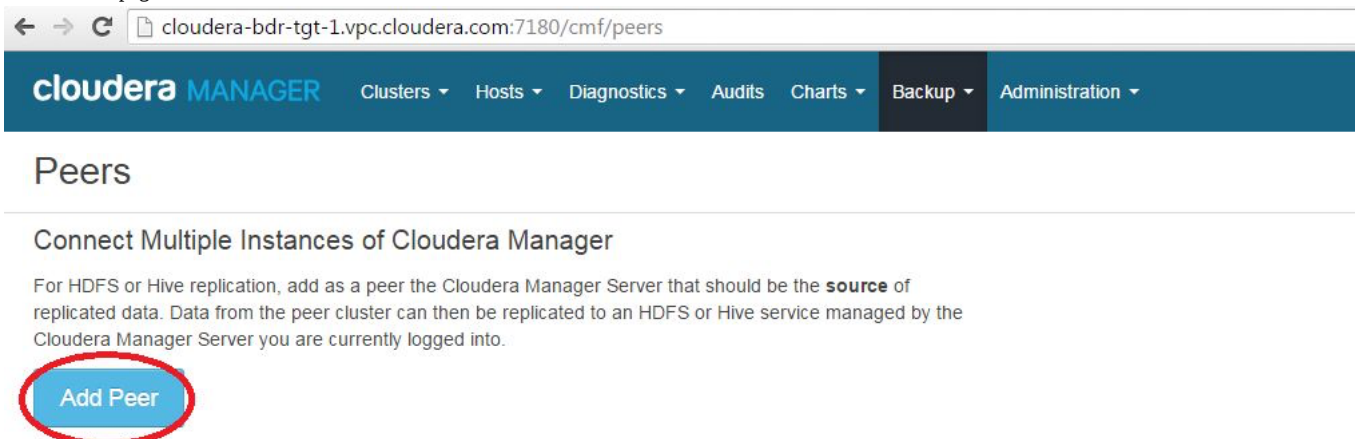
Step 1: Establish a Peer Relationship to the Production Cluster ([#establish_peer_from_backup_to_source](#))

For a backup, the destination is the backup cluster, and the source is the production cluster.

The cluster establishing the peer relationship gains access to the source cluster and can run the export command, list HDFS files, and read files for copying them to the destination cluster. These are all the actions performed by the replication process whenever the defined schedule goes into action.

Defining the replication starts from the Cloudera Manager Admin Console on the backup cluster.

1. Log in to Cloudera Manager Admin Console on the backup cluster.
2. Click the Backup tab and select Peers from the menu.
3. On the Peers page, click Add Peer:



4. On the Add Peer page:

- a. Peer Name - Enter a meaningful name for the cluster that you want to back up. This peer name becomes available in the next step, to be selected as the source for the replication.

- b. Peer URL - Enter the URL for the Cloudera Manager Admin Console running on the master node of the cluster.
- c. Peer Admin Username - Enter the administrator user account for the production cluster.
- d. Peer Admin Password - Enter the password for the administrator account for the production cluster.

Add Peer
✕

Peer Name

SourceCluster

Peer URL

http://cloudera-bdr-src-1.vpc.cloudera.com:7180/

TLS/SSL should be used if possible.

Peer Admin Username

admin

Peer Admin Password

Add Peer

Cancel

5. Click Add Peer to save your settings, connect to the production cluster, and establish this peer relationship.

After the system establishes and verifies the connection to the peer, the Peers page re-displays, showing the Status column as Connected (note the green check-mark):

← → ↻ cloudera-bdr-tgt-1.vpc.cloudera.com:7180/cm/peers

cloudera MANAGER
Clusters ▾ Hosts ▾ Diagnostics ▾ Audits Charts ▾ Backup ▾ Administration ▾

Peers

Add Peer

Name	URL	Status
SourceCluster	http://cloudera-bdr-src-1.vpc.cloudera.com:7180/	<div style="display: flex; align-items: center;"> <div style="border: 2px solid #0070c0; border-radius: 50%; width: 20px; height: 20px; display: flex; align-items: center; justify-content: center; margin-right: 5px;"> ✓ </div> <div>Connected</div> <div style="margin-left: 10px;"> <div style="border: 1px solid #ccc; padding: 2px 5px; font-size: 0.8em;">Test Connectivity</div> </div> </div>

With the peer relationship established from destination to source, create a schedule to replicate HDFS files from the source (production cluster) to the destination (backup cluster).

Step 2: Configure the Replication Schedule for the Backup [\(#configure_hdfs_replication_schedule_for_backup\)](#)

From the Cloudera Manager Admin Console on the backup cluster:

1. Click the Backup tab and select Replication Schedules from the menu.
2. On the Replication Schedules page, select Create Schedule > HDFS Replication.
3. On the Create HDFS Replication page, click the General tab to display the default schedule options:
 - a. Source - Make sure the cluster node selected in the drop-down is the production cluster (the cluster to be backed up).
 - b. Source Path - Specify the directory name on the production cluster holding the files to back up. Use an asterisk (*) on the directory name to specify that only the explicit directory and no others should be created on the destination. Without the asterisk, directories may be nested inside a containing directory on the destination.
 - c. Destination - Select the cluster to use as the target of the replication process, typically, the cluster to which you have logged in and the cluster to which you want to backup HDFS data.
 - d. Destination Path - Specify a directory name on the backup cluster.
 - e. Schedule - Immediate.

https://www.cloudera.com/documentation/enterprise/5-9-x/topics/cm_bdr_howto_hdfs.html#establish_peer_from_backup_to_so... 4/13

f. Run As Username - Leave as Default.

g. Scheduler Pool - Leave as Default.

4. Click Save Schedule.

Create HDFS Replication

General Resources Advanced

Source: HDFS-1 (Cluster 1 @ SourceCluster) ▼
Use [External Accounts](#) to add cloud replication sources.
 Use [Peers](#) to add more replication sources.

Source Path: /user/cloudera/*

Destination: HDFS-1 (Cluster 1) ▼
Use [External Accounts](#) to add cloud replication destinations.

Destination Path: /user/cloudera

Schedule: Immediate ▼

Run As Username: Default

Scheduler Pool: Default

Save Schedule Cancel

The files are replicated from the source cluster to the backup cluster immediately.

When the process completes, the Replication Schedules page re-displays, showing a green check-mark and time-stamp in the Last Run column.

Step 3: Verify Successful Replication ([#validate the backup](#))

Verify that the HDFS files are now on the backup cluster by using the HDFS File Browser or the `hadoop fs -ls` command (shown below):

```
hdfs@cloudera-bdr-tgt-1:~$ hadoop fs -ls /user/cloudera
Found 9 items
-rw-r--r-- 3 hdfs supergroup 16106127360 2016-09-01 14:17 /user/cloudera/large.0
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:29 /user/cloudera/medium.0
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:33 /user/cloudera/medium.1
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:35 /user/cloudera/medium.2
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:36 /user/cloudera/small.0
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:36 /user/cloudera/small.1
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:47 /user/cloudera/small.2
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:48 /user/cloudera/small.3
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:48 /user/cloudera/small.4
hdfs@cloudera-bdr-tgt-1:~$
```

The HDFS files are now on both the production cluster and the backup cluster.

You can now use the backup as a source for a restore onto the production system when needed.

Configuring a Restore Replication Schedule ([#replicate-data-from-backup-cluster](#))

To restore data from a backup, configure a Restore Schedule on the cluster to which you want the data pulled. For this example, the replication schedule is created on the example production cluster and designed to pull HDFS files from the backup cluster to a different path on the production cluster, to enable testing the restore process in advance of any failure.

Note: This approach lets you step through and test the process prior to using the schedule in a production environment.

Setting up a schedule for a restore follows the same pattern as setting up the backup, but with all actions initiated using the Cloudera Manager Admin Console on the production cluster.

To set up and test a replication schedule to restore HDFS from an existing backup copy, follow these steps:

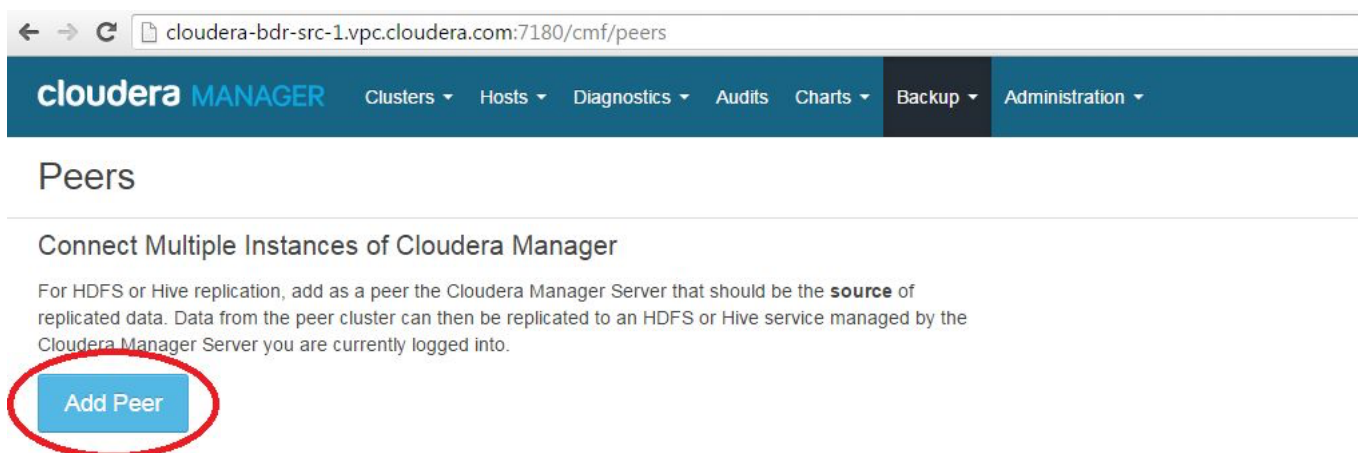
- [Step 1: Establish a Peer Relationship to the Backup Cluster \(cm_bdr_howto_hdfs.html#est-peer-relationship-from-prod-to-backup\)](#)
- [Step 2: Configure Replication Schedule to Test the Restore \(cm_bdr_howto_hdfs.html#configure_replication_schedule_for_restore\)](#)
- [Step 3: Disable the Replication Schedule \(cm_bdr_howto_hdfs.html#disable_replication_schedule_for_restore\)](#)
- [Step 4: Test the Restore Replication Schedule \(cm_bdr_howto_hdfs.html#test_replication_sched_for_restore_to_alt_dir\)](#)
- [Step 5: Verify Successful Data Restoration \(cm_bdr_howto_hdfs.html#verify-restoration-from-backup\)](#)

Step 1: Establish a Peer Relationship to the Backup Cluster ([#est-peer-relationship-from-prod-to-backup](#))

To restore HDFS files on the production cluster, establish a Peer relationship from the destination to the source for the restore.

Log in to the Cloudera Manager Admin Console on the production cluster.

1. Click the Backup tab and select Peers from the menu.
2. On the Peers page, click Add Peers.



3. On the Add Peer page:
 - a. Peer Name - Enter a meaningful name for the cluster that you want to restore data from, for example, *BackupCluster*.
 - b. Peer URL - Enter the URL for the Cloudera Manager Admin Console running on the master node of the cluster.
 - c. Peer Admin Username - Enter the administrator user name for the peer cluster.
 - d. Peer Admin Password - Enter the password for the administrator user account.

Add Peer

Peer Name

Peer URL

TLS/SSL should be used if possible.

Peer Admin Username

Peer Admin Password

Add Peer

Cancel

4. Click Add Peer to save the settings, connect to the backup cluster, and establish the peer relationship.


Once the system connects and tests the peer relationship, the Peers page lists its name, URL, and Status (Connected):

cloudera MANAGER

Clusters ▾ Hosts ▾ Diagnostics ▾ Audits Charts ▾ Backup ▾ Administration ▾

Peers

Add Peer

Name	URL	Status
BackupCluster	http://cloudera-bdr-tgt-1.vpc.cloudera.com:7180/	 Connected <div>Test Connectivity</div>

The backup cluster is now available as a peer, for use in a Replication Schedule.

Step 2: Configure Replication Schedule to Test the Restore ([#configure replication schedule for restore](#))

The goal in these steps is to create a replication schedule that can be used when needed, in the future, but to leave it in a disabled state. However, because Replication Schedules cannot be created in a disabled state, you initially set the date far into the future and then disable the schedule in a subsequent step.

From the Cloudera Manager Admin Console on the production cluster:

- Click the Backup tab and select Replication Schedules from the menu.
- On the Replication Schedules page, select Create Schedule > HDFS Replication.
- On the General settings tab of the Create HDFS Replication page:
 - Source - Make sure the cluster node selected in the drop-down is the backup cluster.
 - Source Path - Specify the directory name that you want to back up. Use an asterisk (*) on the directory name to specify that only the explicit directory and no others should be created on the destination.
 - Destination - Select the cluster to use as the target for the replication process.
 - Destination Path - Specify a directory name on the backup cluster.
Important: The path is a new directory on the example production cluster.
 - Schedule - Set to a time far in the future, so that the schedule does not run as soon as it is saved.
 - Run As Username - Leave as Default.
 - Scheduler Pool - Leave as Default.

Create HDFS Replication

General

Resources

Advanced

Source

HDFS-1 (Cluster 1 @ BackupCluster)

Use [External Accounts](#) to add cloud replication sources.

Use [Peers](#) to add more replication sources.

Source Path

/user/cloudera/*

Destination

HDFS-1 (Cluster 1)

Use [External Accounts](#) to add cloud replication destinations.

Destination Path

/user/cloudera-restored

Schedule

Once

Start Time

09/30/2016 15:00

PDT

Run As Username

Default

Schedule Pool

Default

Save Schedule

Cancel

4. Click Save Schedule.

The Replication Schedule is saved and displays in the Replication Schedules list, with the future date listed in the Next Run column.

Before continuing, immediately disable this newly created Replication Schedule.

Step 3: Disable the Replication Schedule ([#disable_replication_schedule_for_restore](#))

1. On the Replication Schedules page, select the replication schedule.
2. From the Actions drop-down menu, select Disable. When the page refreshes, Disabled displays in the Next Run column:

← → ↺

cloudera-bdr-src-1.vpc.cloudera.com:7180/cm/bdr2/replication

cloudera MANAGER

Clusters ▾

Hosts ▾

Diagnostics ▾

Audits

Charts ▾

Backup ▾

Administration ▾

Replication Schedules

Filters

▼ STATUS

Failed

0

Succeeded

0

Running

0

Disabled

1

Dry-run

0

▼ TYPE

HDFS

1

HDFS-S3

0

Hive

0

Search

Actions for Selected ▾

Create Schedule ▾

<input type="checkbox"/>	ID	Type	Source	Destination	Last Run
<input checked="" type="checkbox"/>	3	HDFS	HDFS-1 Cluster 1 @ BackupCluster	HDFS-1 Cluster 1	● N

Message: —

From: /user/cloudera/* To: /user/cloudera-restored

You can leave restore Replication Schedules pre-configured and disabled in this way so they are ready to use in the event of a catastrophic data loss. Before relying on this approach, test the schedule.

Step 4: Test the Restore Replication Schedule ([#test replication sched for restore to alt dir](#))

The Replication Schedule defined in the example restores data to a specific directory path identified for the purpose of restoration (/user/cloudera-restored) rather than targeting the original source directory path.

From the Cloudera Manager Admin Console on the production cluster, with Replication Schedules page displayed:

1. On the Replication Schedules page, select the disabled replication schedule.
2. Select Actions > Run Now.

← → ↺ cloudera-bdr-src-1.vpc.cloudera.com:7180/cm/bdr2/replication

clouderaMANAGER Clusters ▾ Hosts ▾ Diagnostics ▾ Audits Charts ▾ Backup ▾ Administration ▾

Replication Schedules

Filters

▼ STATUS

Failed0

Succeeded0

Running1

Disabled1

Dry-run0

▼ TYPE

HDFS1

HDFS-S30

Hive0

Search

Actions for Selected ▾ Create Schedule ▾

<input type="checkbox"/>	ID	Type	Source	Destination	L
<input checked="" type="checkbox"/>	3	HDFS	HDFS-1 Cluster 1 @ BackupCluster	HDFS-1 Cluster 1	

Message: 0 of 9 file(s) processed... [Command Details](#) [Performance Report](#)
From: /user/cloudera/* To: /user/cloudera-restored

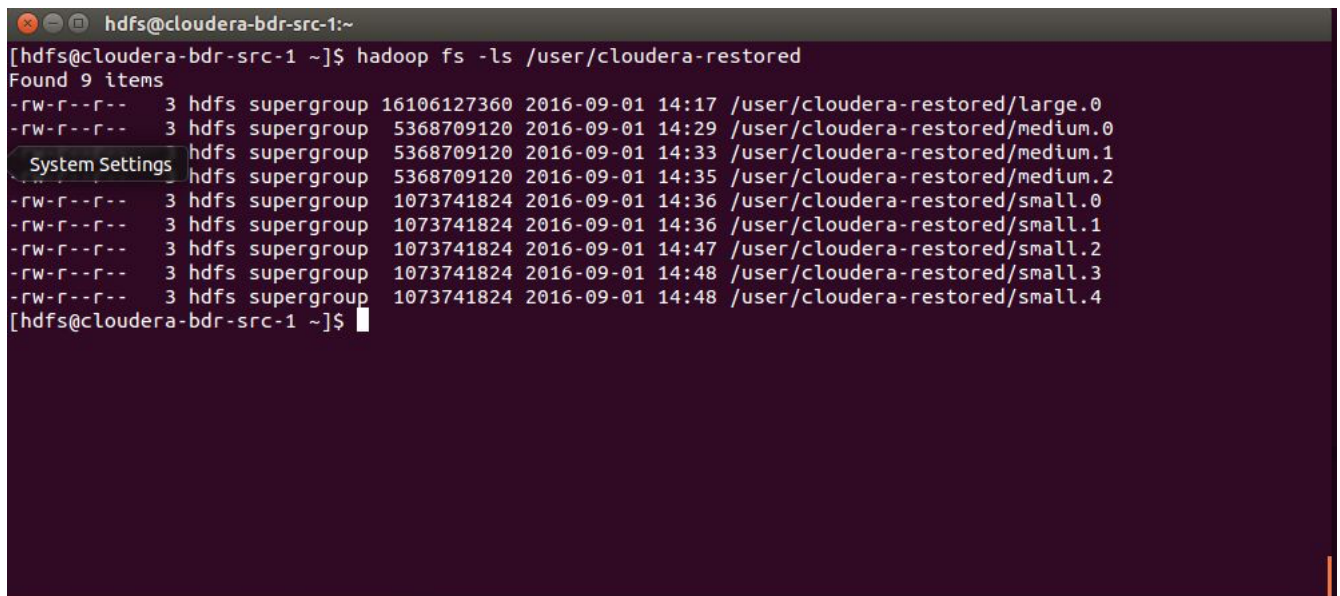
When the replication process completes, disable the Replication Schedule once again:

1. On the Replication Schedules page, select the replication schedule.
2. Select Actions > Disable .

You can now verify that the files have been replicated to the destination directory path.

Step 5: Verify Successful Data Restoration ([#verify-restoration-from-backup](#))

To manually verify that your data has been restored to the source cluster, you can use the HDFS File Browser or the `hadoop` command-line, as shown here:



```

hdfs@cloudera-bdr-src-1:~$ hadoop fs -ls /user/cloudera-restored
Found 9 items
-rw-r--r-- 3 hdfs supergroup 16106127360 2016-09-01 14:17 /user/cloudera-restored/large.0
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:29 /user/cloudera-restored/medium.0
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:33 /user/cloudera-restored/medium.1
-rw-r--r-- 3 hdfs supergroup 5368709120 2016-09-01 14:35 /user/cloudera-restored/medium.2
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:36 /user/cloudera-restored/small.0
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:36 /user/cloudera-restored/small.1
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:47 /user/cloudera-restored/small.2
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:48 /user/cloudera-restored/small.3
-rw-r--r-- 3 hdfs supergroup 1073741824 2016-09-01 14:48 /user/cloudera-restored/small.4
hdfs@cloudera-bdr-src-1 ~$

```

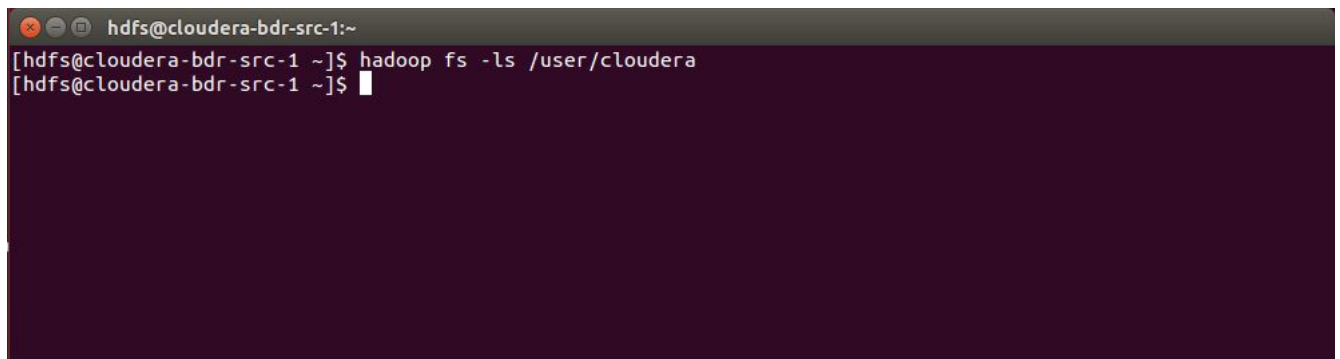
Compare the restored data in `/user/cloudera-restored` to the data in `/user/cloudera` to validate that the restore schedule operates as expected.

At this point, until you need to actually restore production HDFS data files, you can leave the Replication Schedule disabled.

Important: In a production environment, only one Replication Schedule for a given dataset should be active at the same time. In this example, the Replication Schedule that created the backup has not been disabled yet, because the HDFS files from the backup cluster were restored to a different path on the example production cluster.

Recovering from Catastrophic Data Loss ([#catastrophic_data_loss](#))

With backup and restore Replication Schedules set up and validated, you can restore data when production data has been erroneously deleted, as shown in this screenshot:



```

hdfs@cloudera-bdr-src-1:~$ hadoop fs -ls /user/cloudera
hdfs@cloudera-bdr-src-1 ~$

```

In the event of an actual data loss on a production cluster, you should first disable any existing replication schedules for the affected datasets *before* activating a replication schedule for the restore, to avoid overwriting existing replicas on the backup cluster with defective files.

- [Step 1: Disable the Backup Replication Schedule \(cm_bdr_howto_hdfs.html#always-disable-backup-schedule-before-restore\)](#)
- [Step 2: Edit the Existing Replication Schedule \(cm_bdr_howto_hdfs.html#reconfigure-restore-sched\)](#)
- [Step 3: Run the Restore Replication Schedule \(cm_bdr_howto_hdfs.html#run-the-schedule-one-time-from-action-menu\)](#)
- [Step 4: Return the Restore Replication Schedule to a Disabled State \(cm_bdr_howto_hdfs.html#edit-restore-schedule-and-disable-again\)](#)
- [Step 5: Re-enable the Backup Replication Schedule \(cm_bdr_howto_hdfs.html#re-enable-backup-rep-sched-after-restore\)](#)

Step 1: Disable the Backup Replication Schedule ([#always-disable-backup-schedule-before-restore](#))

Disabling any existing replication schedule for HDFS backups can help prevent the replication of lost or corrupted data files over existing backups.

At the Cloudera Manager Admin Console on the backup cluster:

1. Select Backup > Replication Schedules.

2. On the Replication Schedules page, select the schedule.
3. From the Actions drop-down menu, select Disable:

When the Replication Schedules page refreshes, you see Disabled in the Next Run column for the schedule.

Step 2: Edit the Existing Replication Schedule ([#reconfigure-restore-sched](#))

With the replication schedule disabled, you can edit the replication schedule verified previously ([Step 4: Test the Restore Replication Schedule \(cm_bdr_howto_hdfs.html#test_replication_sched_for_restore_to_alt_dir\)](#)) and restore the data to the production cluster.

From the Cloudera Manager Admin Console on the production cluster:

1. Click the Backup tab and select Replication Schedules from the menu.
2. On the Replication Schedules page, select Create Schedule > HDFS Replication.
3. On the General settings tab of the Create HDFS Replication page:
 - a. Source - Name of the backup cluster from which to pull the data. For this example, the source is the backup cluster.
 - b. Source Path - The path on the backup cluster that contains the data you want to restore. Use the asterisk (*) at the end of the directory name to prevent extraneous sub-directories being created on the destination.
 - c. Destination - The name of the cluster on which to restore the data, in which case, the example production cluster.
 - d. Destination Path - Directory in which to restore the HDFS data files, in this case, the directory on the example production system.
 - e. Schedule - Once.
 - f. Start Time - Leave set to the future date and time that you originally defined in [Step 2: Configure Replication Schedule to Test the Restore \(cm_bdr_howto_hdfs.html#configure_replication_schedule_for_restore\)](#).
 - g. Run as Username - Leave set as Default.

Edit HDFS Replication

General Resources Advanced

Source HDFS-1 (Cluster 1 @ BackupCluster) ▼
 Use External Accounts to add cloud replication sources.
 Use Peers to add more replication sources.

Source Path /user/cloudera/*

Destination HDFS-1 (Cluster 1) ▼
 Use External Accounts to add cloud replication destinations.

Destination Path /user/cloudera

Schedule Once ▼

Start Time 09/30/2016 15:45 PDT

Run As Username Default

Save Schedule Cancel

4. Click Save Schedule.

The settings for the Replication Schedule are saved, and the page refreshes. Because this replication schedule is currently disabled, you must actively run the schedule to restore the data.

Step 3: Run the Restore Replication Schedule ([#run-the-schedule-one-time-from-action-menu](#))

While still on the Replication Schedules page:

1. Select the edited replication schedule.

2. From the Actions drop-down menu, select Run Now.

The replication schedule executes and restores the HDFS files to the original location.

When the process completes, the Replication Schedules page displays the time-stamp and check-mark in the Last Run column for the schedule.

Step 4: Return the Restore Replication Schedule to a Disabled State ([#edit-restore-schedule-and-disable-again](#))

You can now disable the schedule, and after verifying that HDFS data has been successfully restored to the production cluster, you can re-enable the backup schedule. While still displaying the Replication Schedules page on the production cluster:

1. Select the replication schedule used for the restore.
2. Edit its configuration again, to point to a non-production directory.
3. Select Actions > Disable.

After confirming that the schedule has been disabled—you see Disabled in the Next Run column for this schedule—but before re-enabling the backup schedule, verify that the HDFS files have been restored to the production cluster.

Step 5: Re-enable the Backup Replication Schedule ([#re-enable-backup-rep-sched-after-restore](#))

With data restored to the production cluster and the replication schedule on the production cluster disabled, you can re-enable the replication schedule on the backup cluster.

1. At the backup cluster, log in to the Cloudera Manager Admin Console.
2. Select Backup > Replication Schedules.
3. On the Replication Schedules page, select the appropriate replication schedule to back up the production cluster.
4. From the Actions drop-down menu, select Enable.

This concludes the tutorial. In an actual production environment, you should configure replication schedules to regularly backup production systems. For restoring files from any backup, you can create and test a replication schedules in advance, as shown in this tutorial.

Alternatively, you can create a replication schedule to restore data specifically when needed. See [How To Back Up and Restore Apache Hive Data Using Cloudera Enterprise BDR \(cm_bdr_howto_hive.html#howto_backup_restore_hive_db\)](#) for details.

See [Backup and Disaster Recovery \(cm_bdr_about.html#cmug_topic_1\)](#) and [BDR Tutorials \(cm_bdr_tutorials.html#bdr_tutorials\)](#) for more information about Cloudera Enterprise BDR.

Categories: [BDR \(../categories/hub_bdr.html\)](#) | [Backups \(../categories/hub_backups.html\)](#) | [Cloudera Manager \(../categories/hub_cloudera_manager.html\)](#) | [HDFS \(../categories/hub_hdfs.html\)](#) | [Restore \(../categories/hub_restore.html\)](#) | [All Categories \(../categories/hub.html\)](#)

- [About Cloudera \(http://www.cloudera.com/about-cloudera.html\)](http://www.cloudera.com/about-cloudera.html)
- [Resources \(http://www.cloudera.com/resources.html\)](http://www.cloudera.com/resources.html)
- [Contact \(http://www.cloudera.com/contact-us.html\)](http://www.cloudera.com/contact-us.html)
- [Careers \(http://www.cloudera.com/about-cloudera/careers.html\)](http://www.cloudera.com/about-cloudera/careers.html)
- [Press \(http://www.cloudera.com/about-cloudera/press-center.html\)](http://www.cloudera.com/about-cloudera/press-center.html)
- [Documentation \(http://www.cloudera.com/documentation.html\)](http://www.cloudera.com/documentation.html)

United States: +1 888 789 1488

Outside the US: +1 650 362 0488

© 2017 Cloudera, Inc. All rights reserved. [Apache Hadoop \(http://hadoop.apache.org\)](http://hadoop.apache.org) and associated open source project names are trademarks of the [Apache Software Foundation \(http://apache.org\)](http://apache.org). For a complete list of trademarks, [click here. \(http://www.cloudera.com/legal/terms-and-conditions.html#trademarks\)](http://www.cloudera.com/legal/terms-and-conditions.html#trademarks)

- <https://www.linkedin.com/company/cloudera>
- <https://www.facebook.com/cloudera>
- <https://twitter.com/cloudera>
- <http://www.cloudera.com/contact-us.html>

[Terms & Conditions \(http://www.cloudera.com/legal/terms-and-conditions.html\)](http://www.cloudera.com/legal/terms-and-conditions.html) | [Privacy Policy \(http://www.cloudera.com/legal/policies.html\)](http://www.cloudera.com/legal/policies.html)

Page generated November 22, 2017.

