HDFS Snapshots

[About snapshots](#h.l904bdw1r202)

[Enabling Snapshots](#h.azg0eb3jg528)

[Command Line](#h.drvnvugczzqb)

[Cloudera Manager](#h.fjijc4ydox1q)

[Restoring a snapshot](#h.82habod8u7qo)

[Command Line](#h.dsu9duyxgu13)

[Cloudera Manager](#h.i74dbbjf1yrl)

[Deleting snapshots](#h.6lh05tz8vvor)

[Command Line:](#h.lc987o2j8qah)

[Cloudera Manager](#h.o6m9l8fs0hj9)

[Show snapshots/ Diff command](#h.h9aneins67da)

[Get snapshottable directory Listing](#h.82o0ozppsdgt)

[Get snapshots difference report](#h.k2poazcscq1y)

# About snapshots

Cloudera Manager enables the creation of snapshot policies that define the directories or tables to be snapshotted, the intervals at which snapshots should be taken, and the number of snapshots that should be kept for each snapshot interval. For example, you can create a policy that takes both daily and weekly snapshots, and specify that 7 daily snapshots and 5 weekly snapshots should be maintained.

Reasons to use snapshots:  
- Protect against user errors  
- Backup  
- Experimental/Test Setups  
- Disaster Recovery

HDFS directories must be enabled for snapshots in order for snapshots to be created. You cannot specify a directory as part of a snapshot policy unless it has been enabled for snapshotting. Only the namenode is aware of the snapshot, and no block duplication takes place. Generally, data nodes are not aware of the snapshot- or the block ownership behind the scenes. Blocks in datanodes are not copied: the snapshot files record the block list and the file size. There is no data copying.

Snapshots are:

* **Performant and Reliable:** Snapshot creation is atomic and instantaneous, no matter the size or depth of the directory subtree
* **Scalable:** Snapshots do not create extra copies of blocks on the file system. Snapshots are highly optimized in memory and stored along with the NameNode’s file system namespace

**Note**: Cloudera Manager does not support snapshot operations for HDFS paths with encryption-at-rest enabled.

# Enabling Snapshots

Snapshots can be enabled either through the command line or through the snapshot utility in Cloudera Manager.

## Command Line

In order to provide data protection and recovery for data in HDFS, you must specify that snapshots are allowed on the parent directory

|  |
| --- |
| hdfs dfsadmin -allowSnapshot <HDFS directory> |

Snapshots need to be explicitly enabled for directories. This provides system administrators with the level of granular control they need to manage the data. When you are ready to take an actual snapshot of a directory, the following command can be used:

|  |
| --- |
| hdfs dfs -createSnapshot <HDFS directory> [<snapshotName>] |

This will create a snapshot, and give it a default name which matches the timestamp at which the snapshot was created. Users can provide an optional snapshot name instead of the default. With the default name, the created snapshot path will be:

|  |
| --- |
| <HDFS directory>/.snapshot/s20130903-000941.091 |

Users can schedule a CRON job to create snapshots at regular intervals. To view the state of the directory at the recently created snapshot:

|  |
| --- |
| hdfs dfs -ls <HDFS directory>/.snapshot/s20130903-000941.091 |

## Cloudera Manager

You can create snapshots by either navigating to the HDFS Filebrowser, or by selecting the Snapshot policy utility, which is available by clicking the **Backup** tab in the top navigation bar and selecting **Snapshots**.

**Browsing to enable snapshots on an HDFS directory:**

1. From the **Clusters** tab, select the HDFS service.
2. Go to the **File Browser** tab.
3. Verify the Snapshottable Path and click **Enable Snapshots**.

You will see dialog with the command line stats, and stdout for the processes. If the commands are successful, you will now see a button for taking snapshots.

The ‘take snapshot’ function will run dfsadmin commands on the command line, and runs with the credentials that are verified through Kerberos. Example of successful output:

|  |
| --- |
| * **$>hdfs/hdfs.sh ["dfs","-createSnapshot","/user/nventmatt","nventmatt-keep"]**   Thu Mar 17 23:29:36 EDT 2016 JAVA\_HOME=/usr/java/jdk1.8.0\_45 using /usr/java/jdk1.8.0\_45 as JAVA\_HOME using 5 as CDH\_VERSION using /run/cloudera-scm-agent/process/936-hdfs-createSnapshot as CONF\_DIR using as SECURE\_USER using as SECURE\_GROUP unlimited /bin/kinit using hdfs/bos-rd1-cdh-master1.rd1.hq.YourCompanyDomain.com@RD1.HQ.YourCompanyDomain.COM as Kerberos principal using /run/cloudera-scm-agent/process/936-hdfs-createSnapshot/krb5cc\_10005 as Kerberos ticket cache Created snapshot /user/nventmatt/.snapshot/nventmatt-snapshot |

# **Restoring a snapshot**

Data can be restored using the command line or through Cloudera Manager.

## Command Line

To recover from data deletion, data is restored by copying the needed files from the snapshot path:

|  |
| --- |
| hdfs dfs -cp <HDFS directory>/.snapshot/s20130903-000941.091/20130902 <HDFS directory>/ |

This will restore the lost set of files to the working data set.

Since /user/<name>/<dir> has a snapshot, the snapshot will protect from the file blocks being removed from the file system. A deletion will only modify the metadata to remove /user/<name>/<dir> from the working directory.

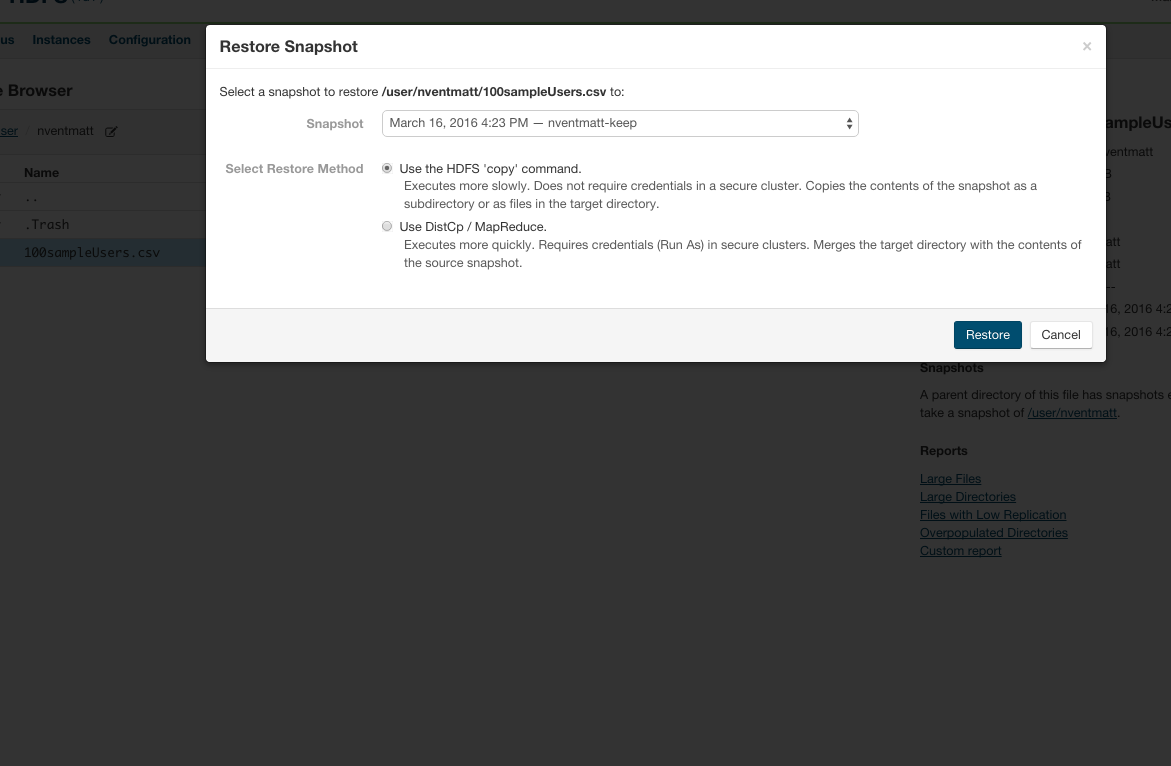
Since snapshots are read-only, HDFS will also protect against user or application deletion of the snapshot data itself. The following operation will fail:

|  |
| --- |
| hdfs dfs -rmdir <HDFS directory>/.snapshot/s20130903-000941.091/20130902 |

## Cloudera Manager

Click the drop-down button and select **Restore**. The dialog gives you 2 options for the restore process. Either using the HDFS copy command, or DistCp as a MapReduce job. Using DistCP allows you to configure options for the snapshot restoration.

* **MapReduce Service** - The MapReduce or YARN service to use.
* **Scheduler Pool** - The scheduler pool to use.
* **Run as** - The user that should run the job. By default this is hdfs. If you want to run the job as a different user, you can enter that here. If you are using Kerberos, you *must* provide a user name here, and it must be one with an ID greater than 1000. Verify that the user running the job has a home directory, /user/<username>, owned by username:supergroup in HDFS.
* **Log path** - An alternative path for the logs.
* **Maximum map slots** and **Maximum bandwidth** - Limits for the number of map slots and for bandwidth per mapper. The defaults are unlimited.
* **Abort on error** - Whether to abort the job on an error (default is not to do so). This means that files copied up to that point will remain on the destination, but no additional files will be copied.
* **Skip Checksum Checks** - Whether to skip checksum checks (the default is to perform them). If checked, checksum validation will not be performed.
* **Remove deleted files** - Whether to remove deleted files from the target directory if they have been removed on the source. When this option is enabled, files deleted from the target directory are sent to trash if HDFS trash is enabled, or are deleted permanently if trash is not enabled. Further, with this option enabled, if files unrelated to the source exist in the target location, then those files will also be deleted.
* **Preserve** - Whether to preserve the block size, replication count, and permissions as they exist on the source file system, or to use the settings as configured on the target file system. The default is to preserve these settings as on the source.



# Deleting snapshots

**Note:** Navigation within the snapshot is possible using regular hdfs commands. The folders and file structure is a point-in-time copy of the original source, although the commands available are for read-only (RO). For example:

|  |
| --- |
| hdfs dfs -ls /user/nventmatt/.snapshot/<saved\_folder> |

Disallowing snapshots for directory

|  |
| --- |
| hdfs dfsadmin -disallowSnapshot <path>  Example:  sudo -u hdfs hdfs dfsadmin -disallowSnapshot /user/nventmatt/ |

## Command Line:

To delete a snapshot

|  |
| --- |
| hdfs dfs -deleteSnapshot <path> <snapshotName>  Example:  hdfs dfs -createSnapshot /user/nventmatt/.snapshot/nventmatt-snapshot |

Since snapshots are read-only, HDFS will also protect against user or application deletion of the snapshot data itself. The following operation will fail:

|  |
| --- |
| hdfs dfs -rmdir /data/weblogs/.snapshot/s20130903-000941.091/20130902 |

## Cloudera Manager

From the **Clusters** tab, select your CDH 5 HDFS service.

1. Go to the **File Browser** tab.
2. Navigate to the directory with the snapshot you want to delete.
3. In the list of snapshots, locate the snapshot you want to delete.
4. Click the drop-down button and select **Delete**.

# Show snapshots/ Diff command

## Get snapshottable directory Listing

|  |
| --- |
| hdfs lsSnapshottableDir  Sample Output: drwxr-xr-x 0 nventmatt nventmatt 0 2016-03-18 01:49 1 65536 /user/nventmatt |

## Get snapshots difference report

|  |
| --- |
| hdfs snapshotDiff <path> <fromSnapshot> <toSnapshot> |

* Arguments:

|  |  |
| --- | --- |
| * path | : The path of the snapshottable directory. |
| * fromSnapshot | : The name of the starting snapshot. |
| * toSnapshot | : The name of the ending snapshot. |

Results:

|  |  |
| --- | --- |
| * + | The file/directory has been created. |
| * - | The file/directory has been deleted. |
| * M | The file/directory has been modified. |
| * R | The file/directory has been renamed. |