NetBackup™ for Hadoop Administrator's Guide

UNIX, Windows, and Linux

Release 10.3

VERITAS™

NetBackup™ for Hadoop Administrator's Guide

Last updated: 2024-03-07

Legal Notice

Veritas Technologies LLC
2625 Augustine Drive
Santa Clara, CA 95054

http://www.veritas.com

Technical Support

Technical Support maintains support centers globally. All support services will be delivered in accordance with your support agreement and the then-current enterprise technical support policies. For information about our support offerings and how to contact Technical Support, visit our website:

https://www.veritas.com/support

You can manage your Veritas account information at the following URL:

https://my.veritas.com

If you have questions regarding an existing support agreement, please email the support agreement administration team for your region as follows:

Worldwide (except Japan)	CustomerCare@veritas.com
Japan	CustomerCare_Japan@veritas.com

Documentation

Make sure that you have the current version of the documentation. Each document displays the date of the last update on page 2. The latest documentation is available on the Veritas website:

https://sort.veritas.com/documents

Documentation feedback

Your feedback is important to us. Suggest improvements or report errors or omissions to the documentation. Include the document title, document version, chapter title, and section title of the text on which you are reporting. Send feedback to:

NB.docs@veritas.com

You can also see documentation information or ask a question on the Veritas community site:

http://www.veritas.com/community/

Veritas Services and Operations Readiness Tools (SORT)

Veritas Services and Operations Readiness Tools (SORT) is a website that provides information and tools to automate and simplify certain time-consuming administrative tasks. Depending on the product, SORT helps you prepare for installations and upgrades, identify risks in your datacenters, and improve operational efficiency. To see what services and tools SORT provides for your product, see the data sheet:

https://sort.veritas.com/data/support/SORT_Data_Sheet.pdf

Contents

Chapter 1

Introduction

This chapter includes the following topics:

Protecting Hadoop data using NetBackup

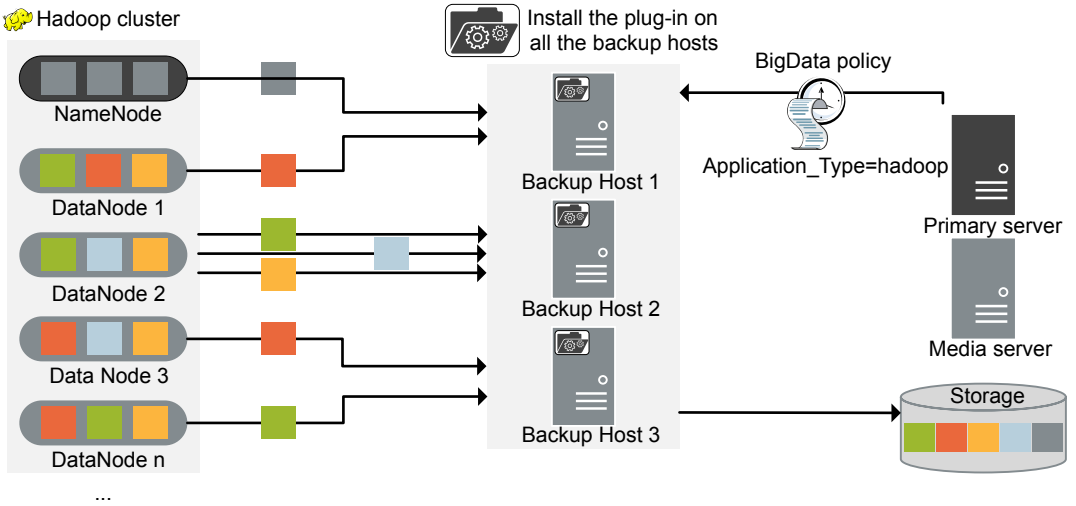
Using the NetBackup Parallel Streaming Framework (PSF), Hadoop data can now be protected using NetBackup.

The following diagram provides an overview of how Hadoop data is protected by NetBackup.

Also, review the related terms for Hadoop.

See “NetBackup for Hadoop terms” on page 11.

Figure 1-1 Architectural overview



As illustrated in the diagram:

- The data is backed up in parallel streams wherein the DataNodes stream data blocks simultaneously to multiple backup hosts. The job processing is accelerated due to multiple backup hosts and parallel streams.
- The communication between the Hadoop cluster and the NetBackup is enabled using the NetBackup plug-in for Hadoop.
  The plug-in is installed as part of the NetBackup installation.
- For NetBackup communication, you need to configure a BigData policy and add the related backup hosts.
- You can configure a NetBackup media server, client, or primary server as a backup host. Also, depending on the number of DataNodes, you can add or remove backup hosts. You can scale up your environment easily by adding more backup hosts.
- The NetBackup Parallel Streaming Framework enables agentless backup wherein the backup and restore operations run on the backup hosts. There is no agent footprint on the cluster nodes. Also, NetBackup is not affected by the Hadoop cluster upgrades or maintenance.

For more information:

- See “Backing up Hadoop data” on page 9.
- See “Restoring Hadoop data” on page 10.

- See “Limitations” on page 13.
- For information about the NetBackup Parallel Streaming Framework (PSF) refer to the *NetBackup Administrator's Guide, Volume I*.

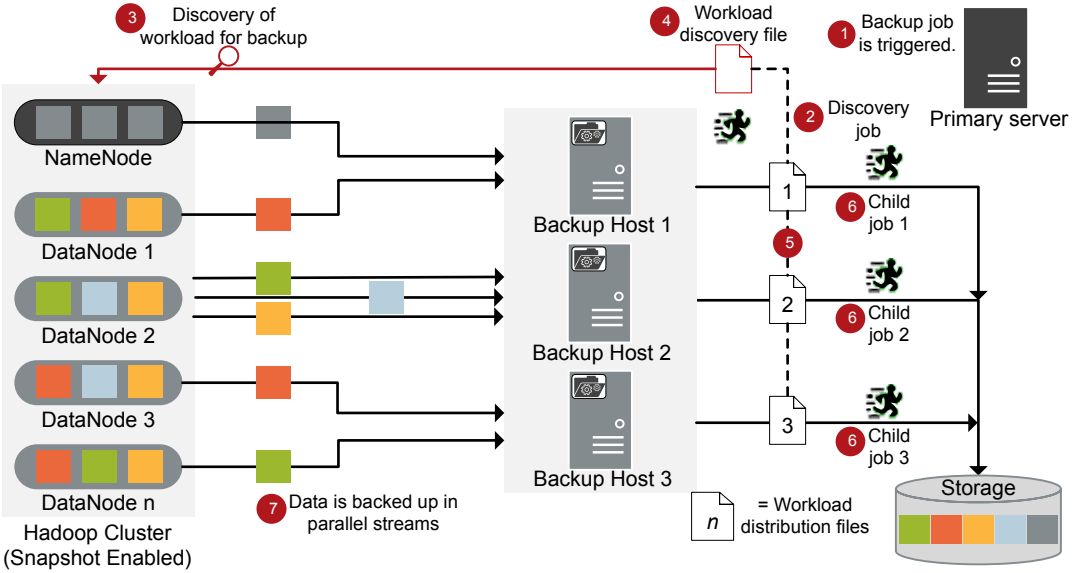
Backing up Hadoop data

Hadoop data is backed up in parallel streams wherein Hadoop DataNodes stream data blocks simultaneously to multiple backup hosts.

Note: All the directories specified in Hadoop backup selection must be snapshot-enabled before the backup.

The following diagram provides an overview of the backup flow:

Figure 1-2 Backup flow



As illustrated in the following diagram:

1. A scheduled backup job is triggered from the primary server.
2. Backup job for Hadoop data is a compound job. When the backup job is triggered, first a discovery job is run.
3. During discovery, the first backup host connects with the NameNode and performs a discovery to get details of data that needs to be backed up.

4. A workload discovery file is created on the backup host. The workload discovery file contains the details of the data that needs to be backed up from the different DataNodes.
5. The backup host uses the workload discovery file and decides how the workload is distributed amongst the backup hosts. Workload distribution files are created for each backup host.
6. Individual child jobs are executed for each backup host. As specified in the workload distribution files, data is backed up.
7. Data blocks are streamed simultaneously from different DataNodes to multiple backup hosts.

The compound backup job is not completed until all the child jobs are completed. After the child jobs are completed, NetBackup cleans all the snapshots from the NameNode. Only after the cleanup activity is completed, the compound backup job is completed.

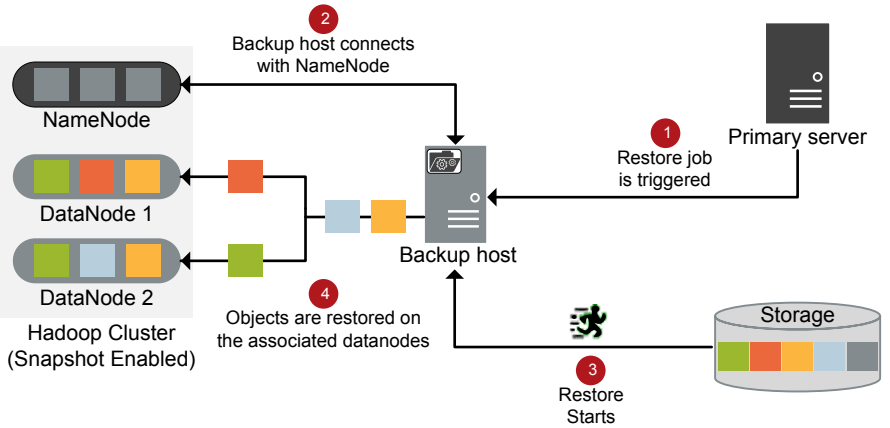
See "About backing up a Hadoop cluster" on page 40.

Restoring Hadoop data

For restore only one backup host is used.

The following diagram provides an overview of the restore flow.

Figure 1-3 Restore flow



As illustrated in the diagram:

1. The restore job is triggered from the primary server.

2. The backup host connects with the `NameNode`. Backup host is also the destination client.
3. The actual data restore from the storage media starts.
4. The data blocks are restored on the `DataNodes`.

See "About restoring a Hadoop cluster" on page 42.

NetBackup for Hadoop terms

The following table defines the terms you will come across when using NetBackup for protecting Hadoop cluster.

Table 1-1 NetBackup terminologies

Terminology	Definition
Compound job	A backup job for Hadoop data is a compound job. ■ The backup job runs a discovery job for getting information of the data to be backed up. ■ Child jobs are created for each backup host that performs the actual data transfer. ■ After the backup is complete, the job cleans up the snapshots on the NameNode and is then marked complete.
Discovery job	When a backup job is executed, first a discovery job is created. The discovery job communicates with the NameNode and gathers information of the block that needs to be backed up and the associated DataNodes. At the end of the discovery, the job populates a workload discovery file that NetBackup then uses to distribute the workload amongst the backup hosts.
Child job	For backup, a separate child job is created for each backup host to transfer data to the storage media. A child job can transfer data blocks from multiple DataNodes.
Workload discovery file	During discovery, when the backup host communicates with the NameNode, a workload discovery file is created. The file contains information about the data blocks to be backed up and the associated DataNodes.
Workload distribution file	After the discovery is complete, NetBackup creates a workload distribution file for each backup host. These files contain information of the data that is transferred by the respective backup host.

Table 1-1 NetBackup terminologies *(continued)*

Terminology	Definition
Parallel streams	The NetBackup parallel streaming framework allows data blocks from multiple DataNodes to be backed up using multiple backup hosts simultaneously.
Backup host	The backup host acts as a proxy client. All the backup and restore operations are executed through the backup host. You can configure media servers, clients, or a primary server as a backup host. The backup host is also used as destination client during restores.
BigData policy	The BigData policy is introduced to: ■ Specify the application type. ■ Allow backing up distributed multi-node environments. ■ Associate backup hosts. ■ Perform workload distribution.
Application server	Namenode is referred to as a application server in NetBackup.
Primary NameNode	In a high-availability scenario, you need to specify one NameNode with the BigData policy and with the `tpconfig` command. This NameNode is referred as the primary NameNode.
Fail-over NameNode	In a high-availability scenario, the NameNodes other than the primary NameNode that are updated in the `hadoop.conf` file are referred as fail-over NameNodes.

Table 1-2 Hadoop terminologies

Terminology	Definition
NameNode	NameNode is also used as a source client during restores.
DataNode	DataNode is responsible for storing the actual data in Hadoop.

Table 1-2 Hadoop terminologies *(continued)*

Terminology	Definition
Snapshot-enabled directories (snapshottable)	Snapshots can be taken on any directory once the directory is snapshot-enabled. ■ Each snapshot-enabled directory can accommodate 65,536 simultaneous snapshots. There is no limit on the number of snapshot-enabled directories. ■ Administrators can set any directory to be snapshot-enabled. ■ If there are snapshots in a snapshot-enabled directory, it can cannot be deleted or renamed before all the snapshots are deleted. ■ A directory cannot be snapshot-enabled if one of its ancestors or descendants is a snapshot-enabled directory.

Limitations

Review the following limitations before you deploy the Hadoop plug-in:

- Only RHEL and SUSE platforms are supported for backup hosts. For platforms supported for Hadoop clusters, see the NetBackup Database and Application Agent Compatibility List.
- Delegation Token authentication method is not supported for Hadoop clusters.
- Hadoop plug-in does not capture Extended Attributes (xattrs) or Access Control Lists (ACLs) of an object during backup and hence these are not set on the restored files or folders.
- For highly available Hadoop cluster, if fail-over happens during a backup or restore operation, the job fails.
- If you cancel a backup job manually while the discovery job for a backup operation is in progress, the snapshot entry does not get removed from the Hadoop web graphical user interface (GUI).
- If the CRL expires during the backup of an HTTPS-based Hadoop cluster, the backup runs partially.
- If you have multiple CRL-based Hadoop clusters, ensure that you add different backup hosts for every cluster.
- Backup and restore operations are not supported with Kerberos authentication if `NB_FIPS_MODE` is enabled at the `bp.conf`.

Note: To perform backup with Kerberos authentication, deploy a new backup host with `NB_FIPS_MODE=0` or disabled.

Chapter 2

Prerequisites and best practices for the Hadoop plug-in for NetBackup

This chapter includes the following topics:

About deploying the Active Directory plug-in

The Active Directory plug-in is installed with NetBackup. Review the following topics to complete the deployment.

Table 2-1 Deploying the Active Directory plug-in

Task	Reference
Prerequisites and requirements	See "Prerequisites for the Hadoop plug-in" on page 15.
Preparing the Active Directory cluster	See "Preparing the Hadoop cluster" on page 15.
Best practices	See "Best practices for deploying the Hadoop plug-in" on page 16.

Table 2-1 Deploying the Active Directory plug-in *(continued)*

Task	Reference
Configuring	See "About configuring NetBackup for Hadoop" on page 17.

Prerequisites for the Hadoop plug-in

Ensure that the following prerequisites are met before you use the Hadoop plug-in:

- See "Operating system and platform compatibility" on page 15.
- See "License for Hadoop plug-in for NetBackup" on page 15.

Operating system and platform compatibility

With this release, RHEL and SUSE platforms are supported for Hadoop clusters and NetBackup backup hosts.

For more information, see the NetBackup Primary Compatibility List.

License for Hadoop plug-in for NetBackup

Backup and restore operations using the Hadoop plug-in for NetBackup, require the Application and Database pack license.

More information is available on how to add licenses.

See the NetBackup Administrator's Guide, Volume I

Preparing the Hadoop cluster

Perform the following tasks to prepare the Hadoop cluster for NetBackup:

- Ensure that the Hadoop directory is snapshot-enabled.
  To make a directory snapshottable, run the following command on the NameNodes:

  ```
  hdfs dfsadmin -allowSnapshot directory_name
  ```

  Note: A directory cannot be snapshot-enabled if one of its ancestors or descendants is a snapshot-enabled directory.

  For more information, refer to the Hadoop documentation.

- Update firewall settings (ensure that the correct port is added along with the Hadoop credentials) so that the backup hosts can communicate with the Hadoop cluster.
- Add the entries of all the NameNodes and DataNodes to the `/etc/hosts` file on all the backup hosts. You must add the hostname in FQDN format.
  Or
  Add the appropriate DNS entries in the `/etc/resolv.conf` file.
- Ensure that `webhdfs` service is enabled on the Hadoop cluster.

Best practices for deploying the Hadoop plug-in

Consider the following when you deploy Hadoop plug-in and configure NetBackup for Hadoop:

- Use consistent conventions for hostnames of backup hosts, media servers, and primary server. For example, if you are using the host name as **hadoop.veritas.com** (FQDN format) use the same everywhere.
- Add the entries of all the NameNodes and DataNodes to the `/etc/hosts` file on all the backup hosts. You must add the hostname in FQDN format.
  Or
  Add the appropriate DNS entries in the `/etc/resolv.conf` file.
- Always specify the NameNode and DataNodes in FQDN format.
- Ping all the nodes (use FQDN) from the backup hosts.
- Hostname and port of the NameNode must be the same as you have specified with the http address parameter in the `core-site.xml` of the Hadoop cluster.
- Canceling a parent job in a compound restore job does not cancel the child restore job. You must manually cancel the child restore jobs.
- Ensure the following for a Hadoop cluster that is enabled with SSL (HTTPS):
  - A valid certificate exists on the backup host that contains the public keys from all the nodes of the Hadoop cluster.
  - For a Hadoop cluster that uses CRL, ensure that the CRL is valid and not expired.

Chapter 3

Configuring NetBackup for Hadoop

This chapter includes the following topics:

About configuring NetBackup for Hadoop

Table 3-1 Configuring NetBackup for Hadoop

Task	Reference
Adding backup hosts	See "Managing backup hosts" on page 18. If you want to use NetBackup client as a backup host, you need to include the NetBackup client on the primary server allowed list. See "Including a NetBackup client on NetBackup primary server allowed list" on page 20.

Table 3-1 Configuring NetBackup for Hadoop *(continued)*

Task	Reference
Adding Hadoop credentials in NetBackup	See "Adding Hadoop credentials in NetBackup" on page 21.
Configuring the Hadoop plug-in using the Hadoop configuration file	See "Configuring the Hadoop plug-in using the Hadoop configuration file" on page 22. See "Configuring NetBackup for a highly-available Hadoop cluster" on page 24. See "Configuring number of threads for backup hosts" on page 27. See "Configuring distribution algorithm and golden ratio for backup hosts" on page 28. See "Configuring number of streams for backup hosts" on page 28.
Configuring the backup hosts for Hadoop clusters that use Kerberos	See "Configuration for a Hadoop cluster that uses Kerberos" on page 36.
Configuring NetBackup policies for Hadoop plug-in	

Managing backup hosts

A backup host acts as a proxy client which hosts all the backup and restore operations for Hadoop clusters. For the Hadoop plug-in for , the backup host performs all the backup and restore operations and does that require that a separate agent be installed on the Hadoop cluster.

The backup host must be a Linux computer. 10.3 release supports only RHEL and SUSE platforms as a backup host.

The backup host can be a client or a media server or a primary server. recommends that you have a media server as a backup host.

Consider the following before adding a backup host:

- For backup operations, you can add one or more backup hosts.
- For restore operations, you can add only one backup host.
- A primary, media, or client can perform the role of a backup host.

- Hadoop plug-in for is installed on all the backup hosts.

Add a backup host

To add a backup host

1. Open the web UI.
2. Create a **BigData** policy.

   See “Create a BigData policy for Hadoop clusters” on page 37.
3. In the **Backup selections** tab, click **Add** and add the backup host in the following format:

   Backup_Host=IP_address or hostname

   Alternatively, you can also add a backup host using the following command:

   For Windows:

   ```
   <install_path>\NetBackup\bin\admincmd\bpplinclude PolicyName -add
   Backup_Host=IP_address or hostname
   ```

   For UNIX:

   ```
   /usr/openv/var/global/bin/admincmd/bpplinclude PolicyName -add
   Backup_Host=IP_address or hostname
   ```
4. As a best practice, add the entries of all the NameNodes and DataNodes to the `/etc/hosts` file on all the backup hosts. You must add the host name in FQDN format.

   OR

   Add the appropriate DNS entries in the `/etc/resolv.conf` file.

Remove a backup host

To remove a backup host

1. In the **Backup Selections** tab, select the backup host that you want to remove.
2. Right click the selected backup host and click **Delete**.

   Alternatively, you can also remove a backup host using the following command:

   For Windows:

   ```
   Install_Path\NetBackup\bin\admincmd\bpplinclude PolicyName -delete
   Backup_Host=IP_address or hostname
   ```

   For UNIX:

   ```
   /usr/openv/var/global/bin/admincmd/bpplinclude PolicyName -delete
   'Backup_Host=IP_address or hostname'
   ```

Including a NetBackup client on NetBackup primary server allowed list

To use the NetBackup client as a backup host, you must include it on the allowed list. Perform the Allowed list procedure on the NetBackup primary server .

Allowlisting is a security practice used for restricting systems from running software or applications unless these have been approved for safe execution.

To place a NetBackup client on NetBackup primary server on the allowed list

Run the following command on the NetBackup primary server:

- For UNIX
  The directory path to the command:
  ```
  /usr/openv/var/global/bin/admincmd/bpsetconfig
  bpsetconfig -h primaryserver
  bpsetconfig> APP_PROXY_SERVER = clientname.domain.org
  bpsetconfig>
  UNIX systems: <ctl-D>
  ```
- For Windows
  The directory path to the command:
  ```
  <Install_Path>\NetBackup\bin\admincmd\bpsetconfig
  bpsetconfig -h primaryserver
  bpsetconfig> APP_PROXY_SERVER = clientname1.domain.org
  bpsetconfig> APP_PROXY_SERVER = clientname2.domain.org
  bpsetconfig>
  Windows systems: <ctl-Z>
  ```

This command sets the *APP_PROXY_SERVER = clientname* entry in the backup configuration (bp.conf) file.

For more information about the *APP_PROXY_SERVER = clientname*, refer to the *Configuration options for NetBackup clients* section in *NetBackup Administrator's Guide, Volume I*

Veritas NetBackup Documentation

Configure a NetBackup Appliance as a backup host

Review the following articles if you want to use NetBackup Appliance as a backup host:

- Using NetBackup Appliance as the backup host of Hadoop with Kerberos authentication
  For details, contact Veritas Technical Support and have the representative refer to article 100039992.
- Using NetBackup Appliance as the backup host with highly-available Hadoop cluster
  For details, contact Veritas Technical Support and have the representative refer to article 100039990.

Adding Hadoop credentials in NetBackup

To establish a seamless communication between Hadoop clusters and NetBackup for successful backup and restore operations, you must add and update Hadoop credentials to the NetBackup primary server.

Use the `tpconfig` command to add Hadoop credentials in NetBackup primary server.

For information on parameters to delete and update the credentials using the `tpconfig` command, see the NetBackup Commands Reference Guide.

Consider the following when you add Hadoop credentials:

- For a highly-available Hadoop cluster, ensure that the user for the primary and fail-over NameNode is the same.
- Use the credentials of the application server that you will use when configuring the BigData policy.
- For a Hadoop cluster that uses Kerberos, specify "**kerberos**" as `application_server_user_id` value.

- Hostname and port of the NameNode must be same as you have specified with the http address parameter in the `core-site.xml` of the Hadoop cluster.
- For password, provide any random value. For example, Hadoop.

To add Hadoop credentials in NetBackup

1. Run `tpconfig` command from the following directory paths:

   On UNIX systems, `/usr/openv/volmgr/bin/`

   On Windows systems, `install_path\Volmgr\bin\`

2. Run the `tpconfig --help` command. A list of options which are required to add, update, and delete Hadoop credentials is displayed.

3. Run the `tpconfig -add -application_server` *`application_server_name`* `-application_server_user_id` *`user_ID`* `-application_type` *`application_type`* `-requiredport` *`IP_port_number`* `[-password` *`password`* `[-key` *`encryption_key`*`]]` command by providing appropriate values for each parameter to add Hadoop credentials.

   For example, if you want to add credentials for Hadoop server which has *application_server_name* as *hadoop1*, then run the following command using the appropriate *<user_ID>* and *<password>* details.

   ```
   tpconfig -add -application_server hadoop1 -application_type hadoop
   -application_server_user_id Hadoop -requiredport 50070 -password
   Hadoop
   ```

   Here, the value `hadoop` specified for `-application_type` parameter corresponds to Hadoop.

4. Run the `tpconfig -dappservers` command to verify if the NetBackup primary server has the Hadoop credentials added.

Configuring the Hadoop plug-in using the Hadoop configuration file

The backup hosts use the `hadoop.conf` file to save the configuration settings of the Hadoop plug-in. You need to create a separate file for each backup host and copy it to the `/usr/openv/var/global/`. You need to manually create the `hadoop.conf` file in JSON format. This file is not available by default with the installer.

Note: You must not provide a blank value for any of the parameters, or the backup job fails.

Ensure that you configure all the required parameters to run the backup and restore operations successfully.

With this release, the following plug-in settings can be configured:

- See "Configuring NetBackup for a highly-available Hadoop cluster" on page 24.
- See "Configuring a custom port for the Hadoop cluster" on page 26.
- See "Configuring number of threads for backup hosts" on page 27.
- See "Configuring communication between NetBackup and Hadoop clusters that are SSL-enabled (HTTPS)" on page 29.

Following is an example of the `hadoop.conf` file.

Note: For non-HA environment, the fail-over parameters are not required.

```
{
 "application_servers":
 {
  "hostname_of_the_primary_namenode":
  {
   "failover_namenodes":
   [
    {
     "hostname":"hostname_of_failover_namenode",
     "port":port_of_the_failover_namenode
    }
   ],
   "port":port_of_the_primary_namenode
   "distro_algo": distribution_algorithm,
   "num_streams": number_of_streams
   }
  }
 },
  "number_of_threads":number_of_threads
}
```

Configuring NetBackup for a highly-available Hadoop cluster

To protect a highly-available Hadoop cluster, when you configure NetBackup for Hadoop cluster:

- Specify one of the NameNodes (primary) as the client in the BigData policy.
- Specify the same NameNode (primary and fail-over) as application server when you execute the `tpconfig` command.
- Create a `hadoop.conf` file, update it with the details of the NameNodes (primary and fail-over), and copy it to all the backup hosts. The `hadoop.conf` file is in JSON format.
- Hostname and port of the NameNode must be same as you have specified with the http address parameter in the `core-site.xml` of the Hadoop cluster.
- User name of the primary and fail-over NameNode must be same.
- Do not provide a blank value for any of the parameters, or the backup job fails.

To update the hadoop.conf file for highly-available Hadoop cluster

1 Update the `hadoop.conf` file with the following parameters:

```
{
 "application_servers":
 {
  "hostname_of_primary_namenode1":
  {
   "failover_namenodes":
   [
    {
     "hostname": "hostname_of_failover_namenode1",
     "port": port_of_failover_namenode1
    }
   ],
   "port":port_of_primary_namenode1
  }
 }
}
```

2 If you have multiple Hadoop clusters, use the same `hadoop.conf` file to update the details. For example,

```
{
 "application_servers":
 {
  "hostname_of_primary_namenode1":
  {
   "failover_namenodes":
   [
    {
     "hostname": "hostname_of_failover_namenode1",
     "port": port_of_failover_namenode1
    }
   ],
   "port"::port_of_primary_namenode1
  },
  "hostname_of_primary_namenode2":
  {
   "failover_namenodes":
   [
    {
     "hostname": "hostname_of_failover_namenode2",
     "port": port_of_failover_namenode2
    }
   ],
   "port":port_of_primary_namenode2
  }
 }
}
```

3 Copy this file to the following location on all the backup hosts:

`/usr/openv/var/global/`

Configuring a custom port for the Hadoop cluster

You can configure a custom port using the Hadoop configuration file. By default, NetBackup uses port 50070.

To configure a custom port for the Hadoop cluster

1. Update `hadoop.conf` file with the following parameters:

```
{
  "application_servers": {
   "hostname_of_namenode1":{

     "port":port_of_namenode1
  }
}
```

2. Copy this file to the following location on all the backup hosts:

```
/usr/openv/var/global/
```

Configuring number of threads for backup hosts

To enhance to the backup performance, you can configure the number of threads (streams) that each backup host can allow. You can improve the backup performance either by adding more number of backup hosts or by increasing the number of threads per backup host.

To decide the number threads consider the following:

- The default value is 4.
- You can set minimum 1 and maximum 32 threads for each backup host.
- Each backup host can have different number of threads configured.
- When you configure the number of threads, consider the number of cores that are available and the number of cores you want to use. As a best practice, you should configure 1 thread per core. For example, if 8 cores are available and you want to use 4 cores, configure 4 threads.

`/usr/openv/var/global/`**To update the hadoop.conf file for configuring number of threads**

1. Update the `hadoop.conf` file with the following parameters:

```
{
 "number_of_threads": number_of_threads
}
```

2. Copy this file to the following location on the backup host:

```
/usr/openv/var/global/
```

Configuring number of streams for backup hosts

To enhance to the backup performance, you can configure the number of streams that each backup host can allow. You can improve the backup performance either by adding more number of backup hosts or by increasing the number of streams per backup host.

To decide the number streams consider the following:

- The default value is 1.
- Number of parallel streams is based on tunable parameters.

To update the hadoop.conf file for configuring number of streams

1. Update the `hadoop.conf` file with the following parameters:

```
{
 "num_of_streams": number_of_streams
}
```

2. Copy this file to the following location on the backup host:

```
/usr/openv/var/global/
```

Note: If you increase number of streams, update the maximum number of jobs per client, update the `stu` setting for multiple threads, and client timeout to avoid abrupt failures.

Configuring distribution algorithm and golden ratio for backup hosts

To enhance the backup performance, you can configure the distribution algorithm and golden ratio based on the tunable parameters. You can improve the backup performance by Performance fine tuning of these algorithms is possible via combination of distribution algorithm and golden ratio.

To decide the distribution algorithm and golden ratio, consider the following:

- If you have `small number of large sized files` in your data set: Use distribution algorithm 1 and change in golden ratio is not honored.
- If you have `large number of small sized files` in your data set: Use distribution algorithm 2 and change in golden ratio is not honored.
- If you have `small number of very large sized files and large number of small sized files` in your data set: Use distribution algorithm 4 or 5 and

golden ratio that fits your deployment. Golden ratio supported range is from 1 to 100. If not provided default is considered as 75.

Note: Adjusting this value can change performance drastically.

`/usr/openv/var/global/`**To update the hadoop.conf file for configuring algorithm and golden ratio**

1. Update the `hadoop.conf` file with the following parameters:

```
{
"distro_algo": distribution_algorithm and
"golden_ratio":godlen_ratio
}
```

2. Copy this file to the following location on the backup host:

```
/usr/openv/var/global/
```

Configuring communication between NetBackup and Hadoop clusters that are SSL-enabled (HTTPS)

To enable communication between NetBackup and Hadoop clusters that are SSL-enabled (HTTPS), complete the following steps:

- Update the `hadoop.conf` file that is located in the `/usr/openv/var/global/` directory on the backup host using the `use_ssl` parameter in the following format:

```
{
  "application_servers":
  {
   "hostname_of_namenode1":
    {
       "use_ssl":true
    }
  }
}
```

Configuration file format for SSL and HA:

```
{
   "application_servers":
   {
    "primary.host.com":
```

```
    {
      "use_ssl":true,
     "failover_namenodes":
     [
      {
       "hostname":"secondary.host.com",
       "use_ssl":true,
       "port":11111
      }
     ]
    }
   }
}
```

By default, the value is set to `false`.

If you use multiple backup hosts, the backup host in that has defined the `use_ssl` parameter in the `hadoop.conf` file is used for communication.

You must define the `use_ssl` parameter in the `hadoop.conf` file for every Hadoop cluster.

- Use the `nbsetconfig` command to configure the following NetBackup configuration options on the access host:
  For more information on the configuration options, refer to the *NetBackup Administrator's Guide*.

`ECA_TRUST_STORE_PATH`

Specifies the file path to the certificate bundle file that contains all trusted root CA certificates.

If you have already configured this external CA option, append the Hadoop CA certificates to the existing external certificate trust store.

If you have not configured the option, add all the required Hadoop server CA certificates to the trust store and set the option.

See "ECA_TRUST_STORE_PATH for NetBackup servers and clients" on page 31.

ECA_CRL_PATH	Specifies the path to the directory where the certificate revocation lists (CRL) of the external CA are located. If you have already configured this external CA option, append the Hadoop server CRLs to the CRL cache. If you have not configured the option, add all the required CRLs to the CRL cache and then set the option. See "ECA_CRL_PATH for NetBackup servers and clients" on page 33.
HADOOP_SECURE_CONNECT_ENABLED	This option affects Hadoop secure communication. Set this value to YES when you have set the use_ssl as true in the hadoop.conf file. The single value is applicable to all Hadoop clusters when use_ssl is set to true. For Hadoop, secure communication is enabled by default. This option lets you skip the security certificate validation. See "HADOOP_SECURE_CONNECT_ENABLED for servers and clients" on page 34.
HADOOP_CRL_CHECK	Lets you validate the revocation status of the Hadoop server certificate against the CRLs. The single value is applicable to all Hadoop clusters when use_ssl is set to true. By default, the option is disabled. See "HADOOP_CRL_CHECK for NetBackup servers and clients" on page 35.

ECA_TRUST_STORE_PATH for NetBackup servers and clients

The ECA_TRUST_STORE_PATH option specifies the file path to the certificate bundle file that contains all trusted root CA certificates.

This certificate file should have one or more certificates in PEM format.

Do not specify the ECA_TRUST_STORE_PATH option if you use the Windows certificate store.

The trust store supports certificates in the following formats:

- PKCS #7 or P7B file having certificates of the trusted root certificate authorities that are bundled together. This file may either be PEM or DER encoded.

- A file containing the PEM encoded certificates of the trusted root certificate authorities that are concatenated together.

This option is mandatory for file-based certificates.

The root CA certificate in Cloudera distribution can be obtained from the Cloudera administrator. It may have a manual TLS configuration or an Auto-TLS enabled for the Hadoop cluster. For both cases, NetBackup needs a root CA certificate from the administrator.

The root CA certificate from the Hadoop cluster can validate the certificates for all nodes and allow NetBackup to run the backup and restore process in case of the secure (SSL) cluster. This root CA certificate is a bundle of certificates that has been issued to all such nodes.

Certificate from root CA must be configured under `ECA_TRUST_STORE_PATH` in case of self-signed, third party CA or Local/Intermediate CA environments. For example: In case of AUTO-TLS enabled Cloudera environments, you can typically find the root CA file named with `cm-auto-global_cacerts.pem` at path `/var/lib/cloudera-scm-agent/agent-cert`. For more details, refer Cloudera documentation.

Table 3-2 ECA_TRUST_STORE_PATH information

Usage	Description
Where to use	On NetBackup servers or clients. If certificate validation is required for VMware, Red Hat Virtualization servers, or Nutanix AHV, this option must be set on the NetBackup primary server and respective access hosts, irrespective of the certificate authority that NetBackup uses for host communication (NetBackup CA or external CA).
How to use	Use the `nbgetconfig` and the `nbsetconfig` commands to view, add, or change the option. For information about these commands, see the NetBackup Commands Reference Guide. Use the following format: `ECA_TRUST_STORE_PATH = Path to the external CA certificate` For example: `c:\rootCA.pem` If you use this option on a Flex Appliance application instance, the path must be `/mnt/nbdata/hostcert/`.
Equivalent UI property	No equivalent exists.

ECA_CRL_PATH for NetBackup servers and clients

The ECA_CRL_PATH option specifies the path to the directory where the Certificate Revocation Lists (CRL) of the external certificate authority (ECA) are located.

These CRLs are copied to NetBackup CRL cache. Revocation status of the external certificate is validated against the CRLs from the CRL cache.

CRL in the CRL cache is periodically updated with the CRL on the location that is specified for ECA_CRL_PATH based on the ECA_CRL_PATH_SYNC_HOURS option.

If the ECA_CRL_CHECK or HADOOP_CRL_CHECK option is not set to DISABLE (or 0) and the ECA_CRL_PATH option is not specified, NetBackup downloads the CRLs from the URLs that are specified in the CRL distribution point (CDP) and uses them to verify revocation status of the peer host's certificate.

Note: For validating the revocation status of a virtualization server certificate, the VIRTUALIZATION_CRL_CHECK option is used.

For validating the revocation status of a Hadoop server certificate, the HADOOP_CRL_CHECK option is used.

Table 3-3 ECA_CRL_PATH information

Usage	Description
Where to use	On NetBackup servers or clients. If certificate validation is required for VMware, Red Hat Virtualization servers, Nutanix AHV, or Hadoop, this option must be set on the NetBackup primary server and respective access or backup hosts, irrespective of the certificate authority that NetBackup uses for host communication (NetBackup CA or external CA). If certificate validation is required for VMware, Red Hat Virtualization servers, or Hadoop, this option must be set on the NetBackup primary server and respective access or backup hosts, irrespective of the certificate authority that NetBackup uses for host communication (NetBackup CA or external CA).

Table 3-3 ECA_CRL_PATH information *(continued)*

Usage	Description

How to use	Use the `nbgetconfig` and the `nbsetconfig` commands to view, add, or change the option.
	For information about these commands, see the NetBackup Commands Reference Guide.
	Use the following format to specify a path to the CRL directory:
	`ECA_CRL_PATH = Path to the CRL directory`
	For example:
	`ECA_CRL_PATH = /usr/eca/crl/eca_crl_file.crl`
	If you use this option on a Flex Appliance application instance, the path must be `/mnt/nbdata/hostcert/crl`.

Equivalent UI property	No equivalent exists.

HADOOP_SECURE_CONNECT_ENABLED for servers and clients

The `HADOOP_SECURE_CONNECT_ENABLED` option enables the validation of Hadoop server certificates using its root or intermediate certificate authority (CA) certificates.

Table 3-4 HADOOP_SECURE_CONNECT_ENABLED information

Usage	Description
Where to use	On all backup hosts.
How to use	Use the `nbgetconfig` and the `nbsetconfig` commands to view, add, or change the option. For information about these commands, see the NetBackup Commands Reference Guide. By default, the `HADOOP_SECURE_CONNECT_ENABLED` is set to `YES`. Use the following format to enable certificate validation for Hadoop: `HADOOP_SECURE_CONNECT_ENABLED = YES`
Equivalent UI property	No equivalent exists.

HADOOP_CRL_CHECK for NetBackup servers and clients

The HADOOP_CRL_CHECK option lets you specify the revocation check level for external certificates of the Hadoop server. Based on the check, revocation status of the Hadoop server certificate is validated against the certificate revocation list (CRL) during host communication.

By default, the HADOOP_CRL_CHECK option is disabled. If you want to validate the revocation status of the Hadoop server certificate against certificate revocation list (CRL), set the option to a different value.

You can choose to use the CRLs from the directory that is specified for the ECA_CRL_PATH configuration option or the CRL distribution point (CDP).

See "ECA_CRL_PATH for NetBackup servers and clients" on page 33.

Table 3-5 HADOOP_CRL_CHECK information

Usage	Description
Where to use	On all backup hosts.
How to use	Use the nbgetconfig and the nbsetconfig commands to view, add, or change the option. For information about these commands, see the NetBackup Commands Reference Guide. Use the following format: HADOOP_CRL_CHECK = *CRL check* You can specify one of the following: ■ DISABLE (or 0) - Revocation check is disabled. Revocation status of the certificate is not validated against the CRL during host communication. This is the default value. ■ LEAF (or 1) - Revocation status of the leaf certificate is validated against the CRL. ■ CHAIN (or 2) - Revocation status of all certificates from the certificate chain are validated against the CRL.
Equivalent UI property	No equivalent exists.

Example values for the parameters in the bp.conf file

Here is an example of values added in the bp.conf file for a CRL-based Hadoop cluster that has SSL enabled (HTTPS):

```
ECA_TRUST_STORE_PATH=/tmp/cacert.pem
ECA_CRL_PATH=/tmp/backuphostdirectory
```

```
HADOOP_SECURE_CONNECT_ENABLED=YES/NO
HADOOP_CRL_CHECK=DISABLE / LEAF / CHAIN
```

Configuration for a Hadoop cluster that uses Kerberos

For a Hadoop cluster that uses Kerberos, perform the following tasks on all the backup hosts:

- Ensure that the Kerberos package is present on all the backup hosts.
  - krb5-workstation package for RHEL
  - krb5-client for SUSE
- Acquire the `keytab` file and copy it to a secure location on the backup host.
- Ensure that the `keytab` has the required principal.
- Manually update the `krb5.conf` file with the appropriate KDC server and realm details.

  Note: Enure that `default_ccache_name` parameter is not set to the **KEYRING:persistent:%{uid}** value. You can comment the parameter to use the default or you can specify a file name such as, **FILE:/tmp/krb_file_name:%{uid}**.

- When you add Hadoop credentials in NetBackup, specify "**kerberos**" as `application_server_user_id` value. See “Adding Hadoop credentials in NetBackup” on page 21.
- To run backup and restore operations for a Hadoop cluster that uses Kerberos authentication, Hadoop needs a valid Kerberos ticket-granting ticket (TGT) to authenticate with the Hadoop cluster. See “Prerequisites for running backup and restore operations for a Hadoop cluster with Kerberos authentication” on page 41.
- To use Kerberos, the user must be a super user with full access and ownership of the HDFS. A valid token is required with the user on the backup host.

`Hadoop.conf` configuration for parallel restore

TBD

```
"application_servers": {
"punnbuucsm5b29-v14.vxindia.veritas.com": {
"port": 9000,
"distro_algo": 4,
"num_streams": 2,
"golden_ratio": 80,
"additionalBackupHosts": ["bh1.vxindia.veritas.com", "bh2.vxindia.veritas.com
  }
 },
"number_of_threads": 10
}
------------
```

`num_stream`: To enhance the restore performance, you can configure the number of streams that each backup host can allow. Default value is 1.

`additionalBackupHosts`: To enhance the restore performance, you can configure additional backup host details. You can specify the hostname of additional backup hosts.

Notes:

- You must keep `additionalBackupHosts` empty, if no additional backup hosts are available.
- The `hadoop.conf` configuration must be same on all the backup hosts.
- The `num_stream` configuration must be same for backup and restore process.
- Hadoop setups and NetBackup setups must be in the same timezone.
- If you increase streams, adjust the maximum number of jobs per client, update the `stu` setting for multiple threads, and update the client timeout to avoid abrupt failures.

Create a BigData policy for Hadoop clusters

Backup policies provide the instructions that follows to back up clients. To configure backup policies for the Hadoop plug-in for , use the **BigData > Hadoop** type as the **Policy type**.

Note: The host name and port of the NameNode must be the same as the values that you specified with the HTTP address parameter in the `core-site.xml` of the Hadoop cluster.

To create a BigData policy for Hadoop clusters

1. Open the web UI.
2. On the left, click **Protection > Policies**.
3. On the **Policies** tab, click **Add**.
4. On the **Attributes** tab, for the **Policy type** select **BigData**.
5. On the **Schedules** tab, click **Add** to create a new schedule.

   You can create a schedule for a **Full backup**, **Differential incremental backup**, or **Cumulative incremental backup** for your BigData policy. After you set the schedule, Hadoop data is backed up automatically as per the set schedule without any further user intervention.
6. On the **Clients** tab, enter the IP address or the host name of the `NameNode`.
7. On the **Backup selections** tab, enter the following parameters and their values as shown:
   - *Application_Type=hadoop*
     The parameter values are case-sensitive.
   - *Backup_Host=IP_address or hostname*
     The backup host must be a Linux computer. The backup host can be a NetBackup client or a media server.
     You can specify multiple backup hosts.
   - File path or the directory to back up.
     You can specify multiple file paths.

   Note: The directory or folder that is specified for the backup selection when you define a BigData Policy with Application_Type=hadoop must not contain a space or a comma in their names.
8. Click **Create**.

For more information on using NetBackup for BigData applications, refer to the Veritas NetBackup documentation page.

Disaster recovery of a Hadoop cluster

For disaster recovery of the Hadoop cluster, perform the following tasks:

Table 3-6 Performing disaster recovery

Task	Description
After the Hadoop cluster and nodes are up, prepare the cluster for operations with NetBackup.	Perform the following tasks: Update firewall settings so that the backup hosts can communicate with the Hadoop cluster. Ensure that webhdfs service is enabled on the Hadoop cluster. See "Preparing the Hadoop cluster" on page 15.
To establish a seamless communication between Hadoop clusters and NetBackup for successful backup and restore operations, you must add and update Hadoop credentials to NetBackup primary server.	Use `tpconfig` command to add Hadoop credentials in NetBackup primary server. See "Adding Hadoop credentials in NetBackup" on page 21.
The backup hosts use the `hadoop.conf` file to save the configuration settings of the Hadoop plug-in. You need to create separate file for each backup host and copy it to `/usr/openv/var/global/`. You need to create the `hadoop.conf` file in JSON format.	With this release, the following plug-in settings can be configured ■ See "Configuring NetBackup for a highly-available Hadoop cluster" on page 24. ■ See "Configuring number of threads for backup hosts" on page 27.
Update the BigData policy with the original NameNode name.	

Chapter 4

Performing backups and restores of Hadoop

This chapter includes the following topics:

About backing up a Hadoop cluster

Use the **NetBackup web UI** to manage backup operations.

Table 4-1 Backing up Hadoop data

Task	Reference
Process understanding	See "Backing up Hadoop data" on page 9.
(Optional) Complete the prerequisites for Kerberos	See "Prerequisites for running backup and restore operations for a Hadoop cluster with Kerberos authentication" on page 41.
Backing up a Hadoop cluster	See "Backing up a Hadoop cluster" on page 42.
Best practices	See "Best practices for backing up a Hadoop cluster" on page 41.

Table 4-1 Backing up Hadoop data *(continued)*

Task	Reference
Troubleshooting tips	For discovery and cleanup related logs, review the following log file on the first backup host that triggered the discovery. `/usr/openv/var/global/logs/nbaapidiscv` For data transfer related logs, search for corresponding backup host (using the hostname) in the log files on the primary server. See "Troubleshooting backup issues for Hadoop data" on page 51.

Prerequisites for running backup and restore operations for a Hadoop cluster with Kerberos authentication

To run backup and restore operations for a Hadoop cluster that uses Kerberos authentication, Hadoop needs a valid Kerberos ticket granting-ticket (TGT) to authenticate with the Hadoop cluster.

Note: During the backup and restore operations, the TGT must be valid. Thus, specify the TGT validity accordingly or renew it when required during the operation.

Run the following command to generate the TGT:

```
kinit -k -t /keytab_file_location/keytab_filename principal_name
```

For example,

```
kinit -k -t /usr/openv/var/global/nbusers/hdfs_mykeytabfile.keytab
hdfs@MYCOMPANY.COM
```

Also review the configuration-related information. See "Configuration for a Hadoop cluster that uses Kerberos" on page 36.

Best practices for backing up a Hadoop cluster

Before backing up a Hadoop cluster, consider the following:

- To backup an entire Hadoop file system provide "/" as the backup selection and ensure that "/" is snapshot enabled.
- Before you execute a backup job, ensure for a successful ping response from the backup hosts to hostname (FQDN) of all the nodes.
- Update the firewall settings so that the backup hosts can communicate with the Hadoop cluster.

- Ensure that the local time on the HDFS nodes and the backup host are synchronized with the NTP server.
- Ensure that you have valid certificates for a Hadoop cluster that is enabled with SSL (HTTPS).

Backing up a Hadoop cluster

You can either create a policy for a backup or run the backup manually.

See "Create a BigData policy for Hadoop clusters" on page 37.

An overview of the backup process is available.

See "Backing up Hadoop data" on page 9.

The backup process comprises of the following stages:

1. Pre-processing: In the pre-processing stage, the first backup host that you have configured with the BigData policy, triggers the discovery. At this stage, a snapshot of the complete backup selection is generated. The snapshot details are visible on the NameNode web interface.
2. Data transfer: During the data transfer process, one child job is created for each backup host.
3. Post-processing: As part of the post-processing, NetBackup cleans up the snapshots on NameNode.

About restoring a Hadoop cluster

Use the NetBackup web UI to manage restore operations.

Table 4-2 Restoring Hadoop data

Task	Reference
Process understanding	See "Restoring Hadoop data" on page 10.
Complete the prerequisites for Kerberos	See "Prerequisites for running backup and restore operations for a Hadoop cluster with Kerberos authentication" on page 41.
Restoring Hadoop data on the same NameNode or Hadoop cluster	■ See "Restore Hadoop data on the same Hadoop cluster" on page 44. ■

Table 4-2 Restoring Hadoop data *(continued)*

Task	Reference
Restoring Hadoop data to an alternate NameNode or Hadoop cluster This task can be performed only using the `bprestore` command.	See "Restoring Hadoop data on an alternate Hadoop cluster" on page 45.
Best practices	See "Best practices for restoring a Hadoop cluster" on page 43.
Troubleshooting tips	See "Troubleshooting restore issues for Hadoop data" on page 56.

Best practices for restoring a Hadoop cluster

When restoring a Hadoop cluster, consider the following points:

- Before you run a restore job, ensure that there is sufficient space on the cluster to complete the restore job.
- Update the firewall settings so that the backup hosts can communicate with the Hadoop cluster.
- Ensure that you have the valid certificates all the cluster nodes for a Hadoop cluster that is enabled with SSL (HTTPS).
- Ensure that you have the valid PEM certificate file on the backup host.
- Ensure that correct parameters are added in the `hadoop.conf` file for HTTP or HTTPS-based clusters.
- Ensure that the backup host contains a valid CRL that is not expired.
- Application-level or file system-level encryption is not supported for Hadoop. You must be a Hadoop superuser to ensure that the restore works correctly.

Restoring Hadoop data on the same Hadoop cluster

To restore Hadoop data on the same Hadoop cluster, consider following:

- Use the NetBackup web UI to initiate Hadoop data restore operations. This interface lets you select the NetBackup server from which the objects are restored

and the client whose backup images you want to browse. Based upon these selections, you can browse the backup image history, select individual items and initiate a restore.

- The restore browser is used to display Hadoop directory objects. A hierarchical display is provided where objects can be selected for restore. The objects (Hadoop directory or files) that make up a Hadoop cluster are displayed by expanding an individual directory.
- An administrator can browse for and restore Hadoop directories and individual items. Objects that users can restore include Hadoop files and folders.

Restore Hadoop data on the same Hadoop cluster

This topic describes how to restore Hadoop data on the same Hadoop cluster.

To restore Hadoop data on the same Hadoop cluster

1. Open the web UI.
2. On the left, select **Recovery**.
3. On the **Regular recovery** card, click **Start recovery**.
4. On the **Basic properties** tab, enter the following:
   - For the **Policy type** select **BigData > Hadoop**.
   - Specify the Hadoop application server as the source for which you want to perform the restore operation.
     From the **Source client** list, select the required Application server.
   - Specify the backup host as the destination client.
     From the **Destination client** list, select the required backup host. Restore is faster if the backup host is the media server that had backed up the node.
   - Click **Next**.
5. On the **Recovery details** tab, do the following:
   - Select the appropriate date range to restore the complete data set or go to **Use backup history** and select the backup images that you want to restore.
   - From left directory hierarchy, select the files and folders for restore.

     Note: All the subsequent files and folders under the directory are displayed in the right pane.

   - Click **Next**.
6. On the **Recovery options** tab, do the following:

- Select **Restore everything to original location** if you want to restore your files to the same location where you performed your backup.
- Select **Restore everything to a different location** if you want to restore your files to a location which is not the same as your backup location. Provide the path.
- Select **Restore individual directories and files to different locations** if you want to restore files and directories to separate locations. Edit and add file path.
- In **Recovery options**, select the appropriate options.
- Click **Next**.

7 On the **Review** tab, verify the details and click **Start recovery**.

Restoring Hadoop data on an alternate Hadoop cluster

NetBackup lets you restore Hadoop data to another NameNode or Hadoop cluster. This type of restore method is also referred to as redirected restores.

Note: Make sure that you have added the credentials for the alternate NameNode or Hadoop cluster in NetBackup primary server and also completed the allowlisting tasks on NetBackup primary server. For more information about how to add Hadoop credentials in NetBackup and whitlelisting procedures, See “Adding Hadoop credentials in NetBackup” on page 21. See “Including a NetBackup client on NetBackup primary server allowed list” on page 20.

To perform redirected restore for Hadoop

1 Modify the values for *rename_file* and *listfile* as follows:

Parameter	Value
rename_file	Change `/<source_folder_path>` to `/<destination_folder_path> ALT_APPLICATION_SERVER=<alternate name node>`
listfile	List of all the Hadoop files to be restored

2 Run the `bprestore -S primary_server -D backup_host -C client -R rename_file -t 44 -L progress log -f listfile` command on the NetBackup primary server using the modified values for the mentioned parameters in step 1.

Where,

`-S primary_server`

Specifies the name of the NetBackup primary server.

`-D backup host`

Specifies the name of the backup host.

`-C client`

Specifies a NameNode as a source to use for finding backups or archives from which to restore files. This name must be as it appears in the NetBackup catalog.

`-f listfile`

Specifies a file (listfile) that contains a list of files to be restored and can be used instead of the file names option. In listfile, list each file path must be on a separate line.

`-L progress_log`

Specifies the name of allowlisted file path in which to write progress information.

`-t 44`

Specifies BigData as the policy type.

`-R rename_file`

Specifies the name of a file with name changes for alternate-path restores.

Use the following form for entries in the rename file:

```
change backup_filepath to restore_filepath
ALT_APPLICATION_SERVER=<Application Server Name>
```

The file paths must start with / (slash).

Note: Ensure that you have allowlisted all the file paths such as `<rename_file_path>`, `<progress_log_path>` that are already not included as a part of NetBackup install path.

Best practice for improving performance during backup and restore

Performance issues such as slow throughput and high CPU usage are observed during the backup and recovery of Hadoop using the SSL environment (HTTPS). The issue is caused if the internal communications in Hadoop are not encrypted. The HDFS configurations must be tuned correctly in the HDFS cluster to improve the internal communication and performance in Hadoop, which can also improve the backup and recovery performance.

- For a better backup and restore performance, NetBackup recommended to follow the Hadoop configuration recommendations from Apache or Hadoop distributions in use.
- If you have Hadoop encryption turned on within the cluster, follow the recommendations from Apache or Hadoop distributions in use to select the right cipher and bit length for data transfer within Hadoop cluster.
- NetBackup performs better during backup and recovery when AES 128 is used for data encryption during the block data transfer.
- You can also increase the number of backup hosts in case of backup to get a better performance; when you have more than one folder to be backed up in the Hadoop cluster. You can have maximum one backup host per folder in the Hadoop cluster to get the maximum benefit.
- You can also increase the number of threads per backup host that are used to fetch data from the Hadoop cluster by NetBackup during backup operation. If you have files with the size in the range of tens of GBs, then you can increase the number of threads for better performance. The default number for threads is 4.
- You can also increase the number of streams per backup host that are used for parallel streaming.
- You can choose any one of the data distribution algorithms best suited for your deployment:
  - For small number of large files in your data set, use distribution algorithm 1.
  - For large number of small sized files in your data set, use distribution algorithm 2.
  - For a mix of small number of very large sized files and large number of small sized files in your data set, use the appropriate combination of distribution algorithm and golden ratio. See the example below:

Table 4-3 Example for large number of small files and small number of large file case

Data size	Number of backup hosts	Number of threads	Number of streams	Distribution algorithm	Golden ratio
Upto 1 TB	4	16	5	4	80
Upto 50TB	5	32	5	4	80
>50TB	6	32	5	4	80

For more details, refer **Apache Hadoop documentation for secure mode**.

Additionally for optimal performance, ensure the following:

- Primary server is not used as a backup host.
- In case of multiple policies scheduled to be triggered in parallel:
  - Avoid using the same discovery host in all policies.
- The last Backup_Host entry is different for these policies.

Note: Discovery host is the last entry in the Backup_Host list.

Chapter 5

Troubleshooting

This chapter includes the following topics:

About troubleshooting NetBackup for Hadoop issues

Table 5-1 Troubleshooting NetBackup for Hadoop issues

Area	References
General logging and debugging	See "About NetBackup for Hadoop debug logging" on page 51.
Backup issues	See "Troubleshooting backup issues for Hadoop data" on page 51.
Restore issues	See "Troubleshooting restore issues for Hadoop data" on page 56.
To avoid issues also review the best practices	See "Best practices for deploying the Hadoop plug-in" on page 16. See "Best practices for backing up a Hadoop cluster" on page 41. See "Best practices for restoring a Hadoop cluster" on page 43.

About NetBackup for Hadoop debug logging

NetBackup maintains process-specific logs for the various processes that are involved in the backup and restore operations. Examining these logs can help you to find the root cause of an issue.

These log folders must already exist in order for logging to occur. If these folders do not exist, you must create them.

The log folders reside on the following directories

- On Windows: `install_path\NetBackup\logs`
- On UNIX or Linux: `/usr/openv/var/global/logs`

Table 5-2 NetBackup logs related to Hadoop

Log Folder	Messages related to	Logs reside on
`install_path/NetBackup/logs/bpVMutil`	Policy configuration	Primary server
`install_path/NetBackup/logs/nbaapidiscv`	BigData framework, discovery, and Hadoop configuration file logs	Backup host
`install_path/NetBackup/logs/bpbrm`	Policy validation, backup, and restore operations	Media server
`install_path/NetBackup/logs/bpbkar`	Backup	Backup host
`install_path/NetBackup/logs/tar`	Restore and Hadoop configuration file	Backup host

For more details, refer to the NetBackup Logging Reference Guide.

Troubleshooting backup issues for Hadoop data

Review the following topics:

- See "About NetBackup for Hadoop debug logging" on page 51.
- See "Backup operation fails with error 6609" on page 52.
- See "Backup operation failed with error 6618" on page 52.
- See "Backup operation fails with error 6647" on page 52.

- See "Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop" on page 53.
- See "Backup operation fails with error 6654" on page 54.
- See "Backup operation fails with bpbrm error 8857" on page 54.
- See "Backup operation fails with error 6617" on page 54.
- See "Backup operation fails with error 6616" on page 55.

Backup operation fails with error 6609

This error is encountered during the following scenarios:

1. The Hadoop plug-in files are deleted or missing from any of the backup hosts (single or multiple).

   Workaround:

   Download and install the Hadoop plug-in.

2. The `Application_Type` details are incorrect.

   Workaround:

   Use `hadoop` instead of `Hadoop` while specifying `Application_Type`.

Backup operation failed with error 6618

Backup operation failed with error 6618 wherein the following error is displayed:

```
NetBackup cannot find the file to complete the operation.(6618)
```

This error is encountered if you have provided an invalid directory as backup selection.

Workaround:

Provide a valid directory as backup selection in the BigData policy.

Backup operation fails with error 6647

Backup operation fails with error 6647 wherein the following error is displayed:

```
Unable to create or access a directory or a path. (6647)
```

This error is encountered in one of the following scenarios:

- Directory is not snapshot-enabled.

- Policy is configured to take snapshot of the root folder as backup selection, whereas one of the child folder is already snapshot-enabled.
- Policy is configured to take snapshot of a child folder as backup selection, whereas one of the parent folder is already snapshot-enabled.
- Policy is configured to take snapshot of a file as backup selection

Workaround:

Nested snapshot-enabled directories are not allowed in Hadoop. If the parent directory is already a snapshot-enabled, than any other child directory under the parent directory cannot be enabled for snapshot. For backup selection in Bigdata policy type, only snapshot-enabled directory must be selected for backup and any other child directories must not be selected.

Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop

Extended attributes allow user applications to associate additional metadata with a file or directory in Hadoop. By default, this is enabled on Hadoop Distributed File System (HDFS).

Access Control Lists provide a way to set different permissions for specific named users or named groups, in addition to the standard permissions. By default, this is disabled on HDFS.

Hadoop plug-ins do not capture extended attributes or Access Control Lists (ACLs) of an object during backup and hence these are not set on the restored files or folders.

Workaround:

If the extended attributes are set on any of the files or directories that is backed up using the BigData policy with `Application_Type = hadoop`, then, you have to explicitly set the extended attributes on the restored data.

Extended attributes can be set using the Hadoop shell commands such as `fs -getfattr` and `hadoop fs -setfattr`.

If the Access Control Lists (ACLs) are enabled and set on any of the files or directories that is backed up using the BigData policy with `Application_Type = hadoop`, then, you have to explicitly set the ACLs on the restored data.

ACLs can be set using the Hadoop shell commands such as `hadoop fs -getfacl` and `hadoop fs -setfacl`.

Backup operation fails with error 6654

This error is encountered during the following scenarios:

- If Hadoop credentials are not added in NetBackup primary server.
  Workaround:
  Ensure that the Hadoop credentials are added in NetBackup primary server. Use the `tpconfig` command. For more information, See "Adding Hadoop credentials in NetBackup" on page 21.
- If Hadoop plug-in files are not installed on backup host.
  Workaround:
  Ensure that the Hadoop plug-in files are installed on all backup hosts before you begin backup operation.
- If a NetBackup client that is used as a backup host is not allowlisted.
  Workaround:
  Ensure that the NetBackup client that is used as a backup host is allowlisted before you begin backup operation.
  See "Including a NetBackup client on NetBackup primary server allowed list" on page 20.

Backup operation fails with bpbrm error 8857

This error is encountered if you have not included the NetBackup client on NetBackup primary server allowed list.

Workaround:

You must perform the allowlisting procedure on NetBackup primary server if you want to use the NetBackup client as the backup host. For more information, See "Including a NetBackup client on NetBackup primary server allowed list" on page 20.

Backup operation fails with error 6617

Backup operation failed with error 6617 wherein the following error is displayed:

```
A system call failed.
```

Verify that the backup host has valid Ticket Granting Ticket (TGT) in case of Kerberos enabled Hadoop cluster.

Workaround:

Renew the TGT.

Backup operation fails with error 6616

Backup operation fails with error 6616 wherein the following error is logged:

```
hadoopOpenConfig: Failed to Create Json Object From Config File.
```

Workaround:

Verify the `hadoop.conf` file to ensure that blank values or incorrect syntax is not used with the parameter values.

Backup operation fails with error 84

Backup operation failed with error 84 media write error.

Workaround:

- Run a backup using valid media server.
- Stop one of the media server storage.
- Run full backup again.

NetBackup configuration and certificate files do not persist after the container-based NetBackup appliance restarts

The NetBackup configuration files like `hadoop.conf` or `hbase.conf` or SSL certificate and CRL paths do not persist after the container-based NetBackup Appliance restarts for any reason. This issue is applicable where container-based NetBackup Appliance is used as a backup host to protect the Hadoop or HBase workload.

Reason:

In the NetBackup Appliance environments the files that are available in the docker host's persistent location are retained after restart operation. The `hadoop.conf` and `hbase.conf` files are custom configuration files and are not listed in the persistent location.

The configuration files are used for defining values like HA (high availability) nodes during a failover and number of threads for backup. If these files get deleted, backups use the default values for both HA and number of threads that are Primary Name Node and 4 respectively. Backup fails only if the primary node goes down in such a case as plug-in fails to find secondary server.

If the SSL certificates and CRL path files are stored at a location that is not persistent the appliance restart, the backups and restore operations fail.

Workaround:

If custom configuration files for Hadoop and HBase get deleted after a restart, you can manually create the files at the following location:

- Hadoop:`/usr/openv/var/global/hadoop.conf`
- HBase:`/usr/openv/var/global/hbase.conf`

You can store the CA certificate that has signed the Hadoop or HBase SSL certificate and CRL at the following location:

`/usr/openv/var/global/`

Unable to see incremental backup images during restore even though the images are seen in the backup image selection

This issue occurs when you try to restore incremental backup images and the Backup Selections list in the backup policy has Backup Selection(s) in a subfolder of `/`.

For example:

```
/data/1
/data/2
```

Workaround

To view the available data that can be restored from an incremental backup image, select the related full backup images along with the incremental backup images.

One of the child backup jobs goes in a queued state

One of the child backup jobs goes in a queued state for a scenario with multiple backup hosts and it keeps waiting for the media server.

Reason:

This issue is seen in the NetBackup Appliance environment where multiple backup hosts are used and the media server goes in an inactive state.

Workaround:

Open the web UI. On the left, click **Storage > Media servers**. Locate and select the media server. Then click **Activate**.

Troubleshooting restore issues for Hadoop data

- See "Restore fails with error code 2850" on page 57.
- See "NetBackup restore job for Hadoop completes partially" on page 57.

- See "Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop" on page 53.
- See "Restore operation fails when Hadoop plug-in files are missing on the backup host" on page 58.
- See "Restore fails with bpbrm error 54932" on page 58.
- See "Restore operation fails with bpbrm error 21296" on page 58.

Restore fails with error code 2850

This error is encountered in the following scenarios:

- Error:`2850 "errno = 62 - Timer expired"`
  Workaround:
  Update firewall settings so that the backup hosts can communicate with the Hadoop cluster.
- Requested files are not recovered.
  Workaround:
  Verify that the backup host has valid Ticket Granting Ticket (TGT) in case of Kerberos enabled Hadoop cluster.
  Renew the TGT.
- Incorrect values and invalid credentials for the application server.
  Workaround:
  Ensure that you have correctly entered hostname of destination Hadoop cluster during restore. This should be same as provided in `tpconfig` command.

NetBackup restore job for Hadoop completes partially

A restore job completes partially if the restore data is more than the space available on the Hadoop cluster.

Workaround:

Clean up space on the Hadoop cluster.

Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop

For more information about this issue, See "Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop" on page 53.

Restore operation fails when Hadoop plug-in files are missing on the backup host

When a restore job is triggered on a backup host which does not have Hadoop plug-in files installed, the restore operation fails with the following error:

```
client restore EXIT STATUS 50: client process aborted
```

Workaround: Download and install the Hadoop plug-in.

Restore fails with bpbrm error 54932

This error is encountered if the files that you want to restore are not backed up successfully.

Workaround:

Before you begin the restore operation, make sure that the backup is completed successfully.

Alternatively, on **Activity Monitor** menu, click **Job Status** tab to locate the specific Job ID and review the error message details.

Restore operation fails with bpbrm error 21296

This error is encountered if you have provided incorrect values for `<application_server_name>` while adding Hadoop credentials to NetBackup primary server.

Workaround:

Verify if the details provided for `<application_server_name>` are correct.

Hadoop with Kerberos restore job fails with error 2850

A Hadoop with Kerberos restore job fails with error 2850. This issue arises if the HDFS owner does not set ownership for files and directories or if there are issues with Kerberos configuration.

Workaround: Before restoring, ensure the following.

- Ensure that the HDFS owner user is used for Kerberos backup.
- Ensure that with the current Kerberos user, it is possible to set the owners/ACLS manually using HDFS commands, such as `chown` and `setfacl`.

For more information, see the NetBackup for Hadoop Administrator's Guide.

Configuration file is not recovered after a disaster recovery

When you use NetBackup primary server as a backup host for high availability with a Hadoop cluster or a Hadoop cluster that is SSL-enabled (HTTPS) and run a full catalog recovery, the `hadoop.conf` configuration file is not recovered.

Create the configuration file manually. Use the following format for the configuration file:

```
{
   "application_servers":
   {
    "primary.host.com":
    {
      "use_ssl":true
     "failover_namenodes":
     [
      {
       "hostname":"secondary.host.com",
       "use_ssl":true
       "port":11111
      }
     ],
     "port":11111
    }
   },
   "number_of_threads":5
}
```

Index