Veritas NetBackup for Hadoop Administrator's Guide

UNIX, Windows, and Linux

Release 8.1

VERITAS™

Veritas Hadoop Administartor's Guide

Last updated: 2018-05-10

Document version:NetBackup 8.1

Legal Notice

Copyright © 2018 Veritas Technologies LLC. All rights reserved.

Veritas, the Veritas Logo, and NetBackup are trademarks or registered trademarks of Veritas Technologies LLC or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.

This product may contain third-party software for which Veritas is required to provide attribution to the third party (“Third-party Programs”). Some of the Third-party Programs are available under open source or free software licenses. The License Agreement accompanying the Software does not alter any rights or obligations you may have under those open source or free software licenses. Refer to the Third-party Legal Notices document accompanying this Veritas product or available at:

https://www.veritas.com/about/legal/license-agreements

The product described in this document is distributed under licenses restricting its use, copying, distribution, and decompilation/reverse engineering. No part of this document may be reproduced in any form by any means without prior written authorization of Veritas Technologies LLC and its licensors, if any.

THE DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID. VERITAS TECHNOLOGIES LLC SHALL NOT BE LIABLE FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH THE FURNISHING, PERFORMANCE, OR USE OF THIS DOCUMENTATION. THE INFORMATION CONTAINED IN THIS DOCUMENTATION IS SUBJECT TO CHANGE WITHOUT NOTICE.

The Licensed Software and Documentation are deemed to be commercial computer software as defined in FAR 12.212 and subject to restricted rights as defined in FAR Section 52.227-19 "Commercial Computer Software - Restricted Rights" and DFARS 227.7202, et seq. "Commercial Computer Software and Commercial Computer Software Documentation," as applicable, and any successor regulations, whether delivered by Veritas as on premises or hosted services. Any use, modification, reproduction release, performance, display or disclosure of the Licensed Software and Documentation by the U.S. Government shall be solely in accordance with the terms of this Agreement.

Veritas Technologies LLC
500 E Middlefield Road
Mountain View, CA 94043

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 6

Protecting Hadoop data using NetBackup 6
Backing up Hadoop data 8
Restoring Hadoop data 9
Deploying the Hadoop plug-in 10
NetBackup for Hadoop terminologies 12
Limitations 13

Chapter 2 Installing and deploying Hadoop plug-in for NetBackup 15

About installing and deploying the Hadoop plug-in 15
Pre-requisites for installing the Hadoop plug-in 16
Operating system and platform compatibility 16
License for Hadoop plug-in for NetBackup 16
Best practices for deploying the Hadoop plug-in 16
Preparing the Hadoop cluster 17
Downloading the Hadoop plug-in 17
Installing the Hadoop plug-in 18
Verifying the installation of the Hadoop plug-in 19

Chapter 3 Configuring NetBackup for Hadoop 20

About configuring NetBackup for Hadoop 20
Managing backup hosts 21
Whitelisting a NetBackup client on NetBackup master server 23
Configure a NetBackup Appliance as a backup host 24
Adding Hadoop credentials in NetBackup 24
Configuring the Hadoop plug-in using the Hadoop configuration file 25
Configuring NetBackup for a highly-available Hadoop cluster 26
Configuring a custom port for the Hadoop cluster 28
Configuring number of threads for backup hosts 29
Configuration for a Hadoop cluster that uses Kerberos 30
Configuring NetBackup policies for Hadoop plug-in 30

Creating a BigData backup policy 31
Disaster recovery of a Hadoop cluster 35

Chapter 4 Performing backups and restores of Hadoop 37

About backing up a Hadoop cluster 37
Pre-requisite for running backup and restore operations for a Hadoop cluster with Kerberos authentication 38
Backing up a Hadoop cluster 38
Best practices for backing up a Hadoop cluster 39
About restoring a Hadoop cluster 39
Restoring Hadoop data on the same Hadoop cluster 40
Restoring Hadoop data on an alternate Hadoop cluster 44
Best practices for restoring a Hadoop cluster 48

Chapter 5 Troubleshooting 49

About troubleshooting NetBackup for Hadoop issues 49
About NetBackup for Hadoop debug logging 50
Troubleshooting backup issues for Hadoop data 50
Backup operation for Hadoop fails with error code 6599 51
Backup operation fails with error 6609 51
Backup operation failed with error 6618 51
Backup operation fails with error 6647 52
Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop 52
Backup operation fails with error 6654 53
Backup operation fails with bpbrm error 8857 53
Backup operation fails with error 6617 53
Backup operation fails with error 6616 54
Troubleshooting restore issues for Hadoop data 54
Restore fails with error code 2850 54
NetBackup restore job for Hadoop completes partially 55
Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop 55
Restore operation fails when Hadoop plug-in files are missing on the backup host 55
Restore fails with bpbrm error 54932 55
Restore operation fails with bpbrm error 21296 56

Index 57

Chapter 1

Introduction

This chapter includes the following topics:

- Protecting Hadoop data using NetBackup
- Backing up Hadoop data
- Restoring Hadoop data
- Deploying the Hadoop plug-in
- NetBackup for Hadoop terminologies
- Limitations

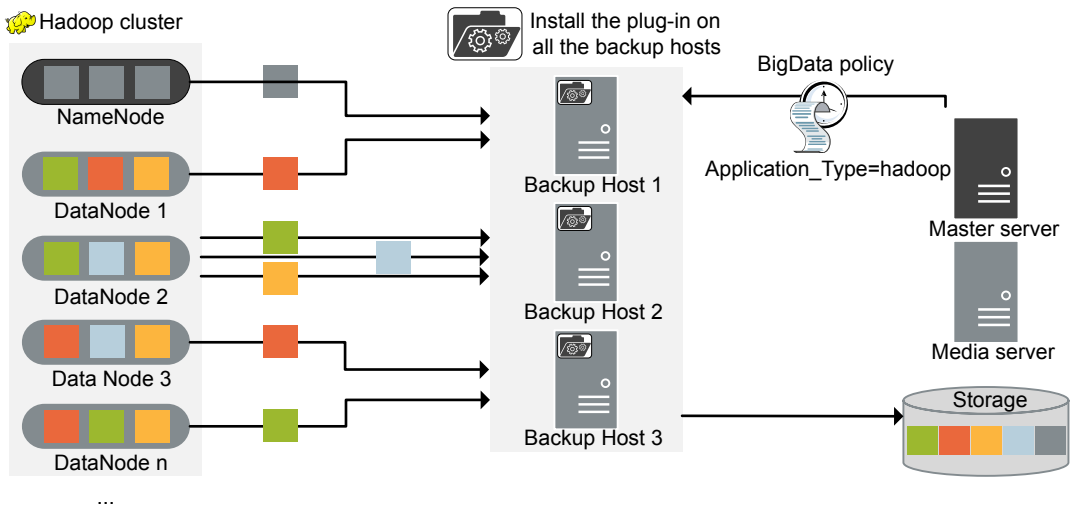
Protecting Hadoop data using NetBackup

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 definitions of terminologies.See “NetBackup for Hadoop terminologies” on page 12.

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 available separately and must be installed on all the backup hosts.
- For NetBackup communication, you need to configure a Big Data policy and add the related backup hosts.
- You can configure a NetBackup media server, client, or master 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 8.

- See “Restoring Hadoop data” on page 9.
- See “Deploying the Hadoop plug-in” 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*.

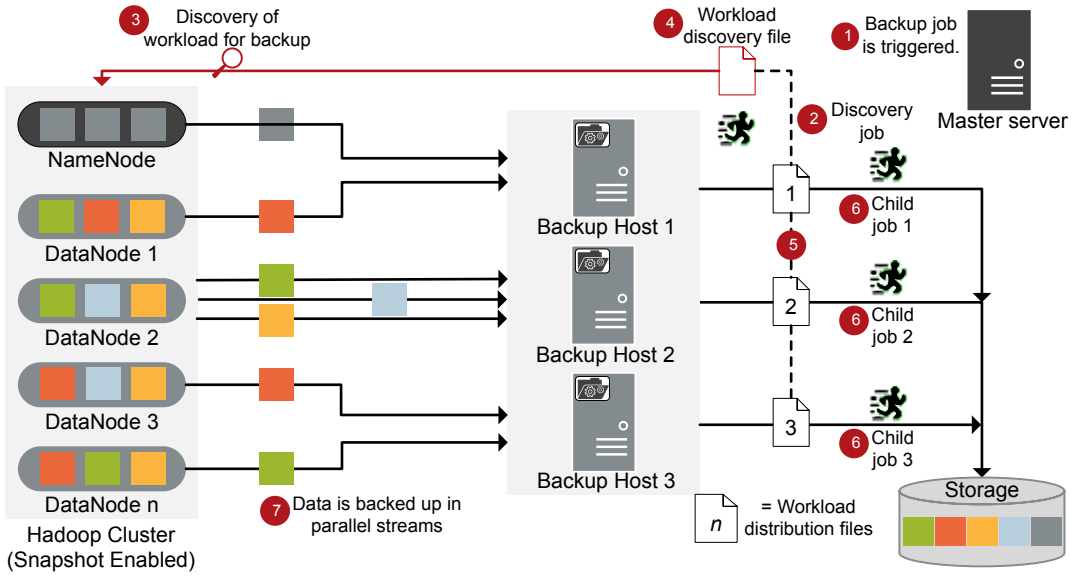
Backing up Hadoop data

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 master 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 untill 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.

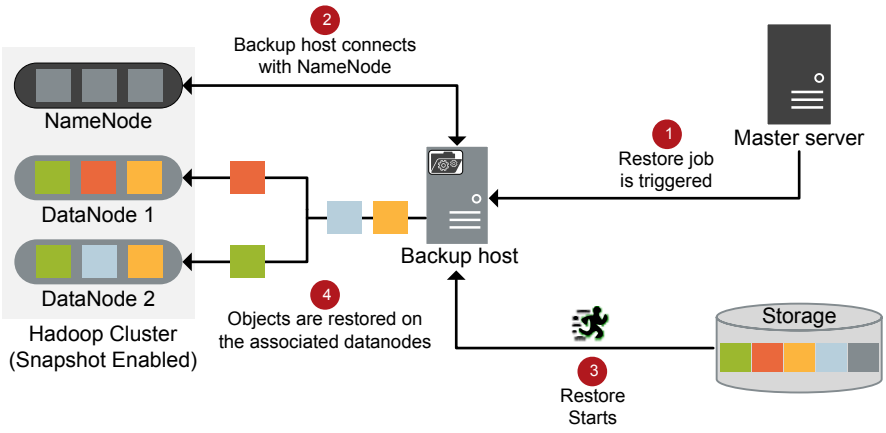
See “About backing up a Hadoop cluster” on page 37.

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 master 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.

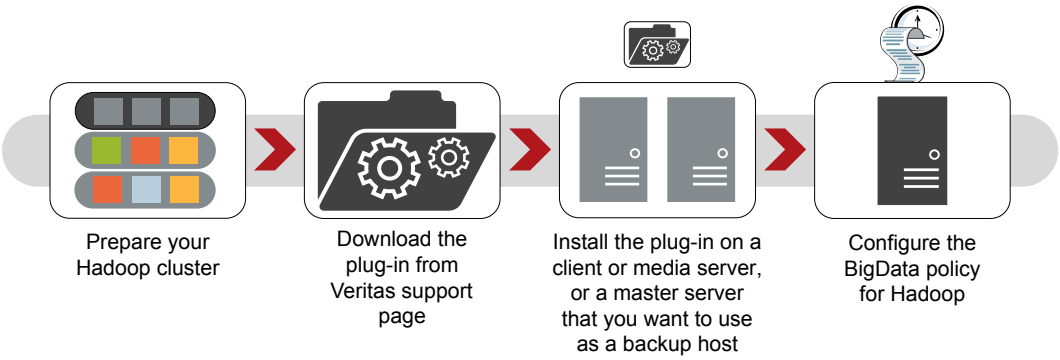
See "About restoring a Hadoop cluster" on page 39.

Deploying the Hadoop plug-in

The availability of the plug-in is not tied to any NetBackup release.

The following diagram provides an overview of how you can deploy the plug-in.

Figure 1-4 Deployment flow



As illustrated in the diagram, to deploy the plug-in:

1. Prepare the Hadoop cluster for NetBackup. All directories specified in Hadoop backup selection must be snapshot enabled.
2. Download the plug-in package from *Veritas Entitlement Management System*.
3. Install the plug-in on all the backup hosts.
4. Configure the BigData policy. This policy is introduced to:
   - Specify the application type.
   - Allow backing up of distributed multi-node environments.
   - Associate backup hosts.
   - Perform workload distribution.

See “About installing and deploying the Hadoop plug-in” on page 15.

See “About configuring NetBackup for Hadoop” on page 20.

See “Limitations” on page 13.

NetBackup for Hadoop terminologies

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.
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 master server as a backup host. The backup host is also used as destination client during restores.

Table 1-1 NetBackup terminologies *(continued)*

Terminology	Definition
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.
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 Hadoop clusters and backup hosts.

- 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).

Chapter 2

Installing and deploying Hadoop plug-in for NetBackup

This chapter includes the following topics:

- About installing and deploying the Hadoop plug-in
- Pre-requisites for installing the Hadoop plug-in
- Best practices for deploying the Hadoop plug-in
- Preparing the Hadoop cluster
- Downloading the Hadoop plug-in
- Installing the Hadoop plug-in
- Verifying the installation of the Hadoop plug-in

About installing and deploying the Hadoop plug-in

Table 2-1 Installing and deploying the Hadoop plug-in

Task	Reference
Pre-requisites and requirements	See "Pre-requisites for installing the Hadoop plug-in" on page 16.
Preparing the Hadoop cluster	See "Preparing the Hadoop cluster" on page 17.

Table 2-1 Installing and deploying the Hadoop plug-in *(continued)*

Task	Reference
Best practices	See "Best practices for deploying the Hadoop plug-in" on page 16.
Downloading the Hadoop plug-in	See "Downloading the Hadoop plug-in" on page 17.
Performing the installation	See "Installing the Hadoop plug-in" on page 18.
Verifying the installation	See "Verifying the installation of the Hadoop plug-in " on page 19.
Configuring	See "About configuring NetBackup for Hadoop" on page 20.

Pre-requisites for installing the Hadoop plug-in

Ensure that the following pre-requisites are met before you install the Hadoop plug-in:

- See "Operating system and platform compatibility" on page 16.
- See "License for Hadoop plug-in for NetBackup" on page 16.

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 Master 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

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 master server. For example, if you are using the hostname 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/resolve.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 same as you have specified with the http address parameter in the `core-site.xml` of the Hadoop cluster.

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 (port 50070 by default) 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/resolve.conf` file.
- Ensure that webhdfs service is enabled on the Hadoop cluster.

Downloading the Hadoop plug-in

You can download the following Hadoop plug-in package for NetBackup from the Veritas Support site.

`NetBackup_PSF_Hadoop_Plugin_8.1-BuildID.tar.gz`

To download the Hadoop plug-in:

1. Go to https://www.veritas.com/support site.
2. Click **Licensing**. You are directed to the **Veritas Account Manager** page to access your Veritas account.
3. Enter your user credentials to access your Veritas account. You are directed to the *Veritas Entitlement Management System* site.
4. On the **Entitlements** menu, use your **Entitlement ID** to locate and download the following file for Hadoop plug-in for NetBackup.

   `NetBackup_PSF_Hadoop_Plugin_8.1-BuildID.tar.gz`

   Alternatively, on the **Downloads** menu, locate

   `NetBackup_PSF_Hadoop_Plugin_8.1-BuildID.tar.gz`

   The list of software or plug-in package that is available to download may vary across user accounts based on the entitlements within each account.
5. In the **Actions** column against the software or plug-in package you want to download, click **Download**.
6. Save the downloaded file in a local directory on the intended backup host.

Installing the Hadoop plug-in

Log on to the Veritas Support site to download the Hadoop plug-in and store it in a local directory.

Use your **Entitlement ID** to locate the PSF plug-ins and download the `NetBackup_PSF_Hadoop_Plugin_8.1-BuildID.tar.gz` file.

See "Downloading the Hadoop plug-in" on page 17.

You must install Hadoop plug-in on all the media servers, clients, or master server you want to use as backup host.

To install Hadoop plug-in on a backup host

1. Login to the backup host with root privileges.
2. Extract the plug-in. Run the following command:

   `tar -xvf package_name`

   Following files are extracted:

   - `plug-in_pkg.linuxR_x86.tar`

- `plug-in_pkg.linuxS_x86.tar`
- `install`
- `LICENSE`

3. Run the installer. Run the following command:

   `./install`

4. Accept the End User License Agreement (EULA).

   Enter **y** for the prompt.

5. Following plugin-in binaries are installed in `/usr/openv/lib/` directory:

   - `libaapihdfs.so`
   - `libaapipgnhadoop.so`

Verifying the installation of the Hadoop plug-in

After you install the Hadoop plug-in, the following files are deployed at the`/usr/openv/lib` directory:

- `libaapihdfs.so`
- `libaapipgnhadoop.so`

File `/usr/openv/tmp/install_trace` contains a trace of the installation. This file can be deleted after you are sure the installation was successful.

Chapter 3

Configuring NetBackup for Hadoop

This chapter includes the following topics:

- About configuring NetBackup for Hadoop
- Managing backup hosts
- Adding Hadoop credentials in NetBackup
- Configuring the Hadoop plug-in using the Hadoop configuration file
- Configuration for a Hadoop cluster that uses Kerberos
- Configuring NetBackup policies for Hadoop plug-in
- Disaster recovery of a Hadoop cluster

About configuring NetBackup for Hadoop

Table 3-1 Configuring NetBackup for Hadoop

Task	Reference
Adding backup hosts	See "Managing backup hosts" on page 21. If you want to use NetBackup client as a backup host, you need to whitelist the NetBackup client on the master server. See "Whitelisting a NetBackup client on NetBackup master server" on page 23.

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

Task	Reference
Adding Hadoop credentials in NetBackup	See "Adding Hadoop credentials in NetBackup" on page 24.
Configuring the Hadoop plug-in using the Hadoop configuration file	See "Configuring the Hadoop plug-in using the Hadoop configuration file" on page 25. See "Configuring NetBackup for a highly-available Hadoop cluster" on page 26. See "Configuring number of threads for backup hosts" on page 29.
Configuring the backup hosts for Hadoop clusters that use Kerberos	See "Configuration for a Hadoop cluster that uses Kerberos" on page 30.
Configuring NetBackup policies for Hadoop plug-in	See "Configuring NetBackup policies for Hadoop plug-in" on page 30.

Managing backup hosts

A backup host acts as a proxy client which hosts all the backup and restore operations for Hadoop clusters. In case of Hadoop plug-in for NetBackup, backup host performs all the backup and restore operations without any separate agent installed on the Hadoop cluster.

The backup host must be a Linux computer. NetBackup 8.1 release supports only RHEL and SUSE platforms as a backup host.

The backup host can be a NetBackup client or a media server or a master server. Veritas recommends that you have 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 master, media, or client can perform the role of a backup host.
- Hadoop plug-in for NetBackup is installed on all the backup hosts.

You can add a backup host while configuring BigData policy using either the NetBackup Administration Console or Command Line Interface.

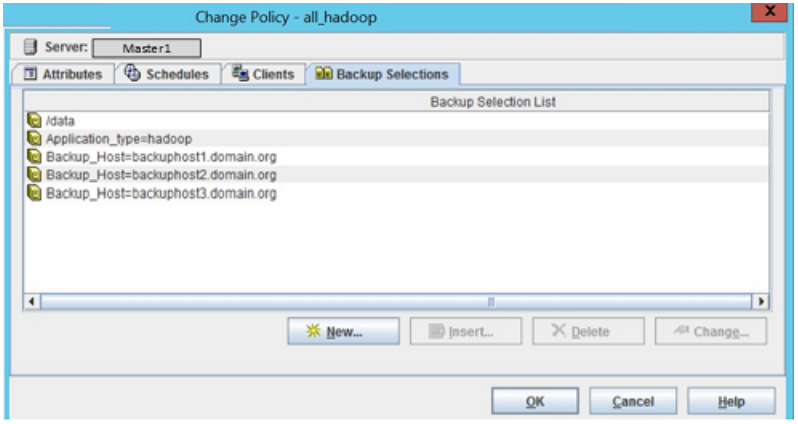
For more information on how to create a policy, see See "Creating a BigData backup policy" on page 31.

To add a backup host

1 In the **Backup Selections** tab, click **New** and add the backup host in the following format:

Backup_Host=<IP_address or hostname>

For more information on how to create a policy, See "Creating a BigData backup policy" on page 31.



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

For Windows:

```
bpplinclude PolicyName -add "Backup_Host=IP_address or hostname"
```

For UNIX:

```
bpplinclude PolicyName -add 'Backup_Host=IP_address or hostname'
```

For more information, See "Using NetBackup Command Line Interface (CLI) to create a BigData policy for Hadoop clusters " on page 33.

2 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/resolve.conf` file.

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:

   ```
   bpplinclude PolicyName -delete "Backup_Host=IP_address or
   hostname"
   ```

   For UNIX:

   ```
   bpplinclude PolicyName -delete 'Backup_Host=IP_address or
   hostname'
   ```

Whitelisting a NetBackup client on NetBackup master server

To use the NetBackup client as a backup host, you must whitelist it. Perform the Whitelisting procedure on the NetBackup master server .

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

To Whitelist a NetBackup client on NetBackup master server

- Run the following command on the NetBackup master server:
  - For UNIX

    ```
    bpsetconfig -h masterserver
    bpsetconfig> APP_PROXY_SERVER = clientname.domain.org
    bpsetconfig>
    UNIX systems: <ctl-D>
    ```

  - For Windows

    ```
    bpsetconfig -h masterserver
    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 master server.

Use the `tpconfig` command to add Hadoop credentials in NetBackup master 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 1
   -application_server_user_id Hadoop -password Hadoop
   ```

   Here, the numeric value `1` specified for `-application_type` parameter corresponds to Hadoop.

4. Run the `tpconfig -dappservers` command to verify if the NetBackup master 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/netbackup/`. 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.

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

- See "Configuring NetBackup for a highly-available Hadoop cluster" on page 26.
- See "Configuring a custom port for the Hadoop cluster" on page 28.

- See "Configuring number of threads for backup hosts" 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
  }
 },
 "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/netbackup/`

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/netbackup/
```

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.

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/netbackup/
```

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 24.
- 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 "Pre-requisite for running backup and restore operations for a Hadoop cluster with Kerberos authentication" on page 38.

Configuring NetBackup policies for Hadoop plug-in

Backup policies provide the instructions that NetBackup follows to back up clients. For configuring backup policies for Hadoop plug-in for NetBackup, use the **BigData** policy as the **Policy Type**.

You can create **BigData** policy using either the **NetBackup Administration Console** or the **Command Line Interface**.

Note: 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 more information on how to create a BigData policy, See "Creating a BigData backup policy" on page 31.

Creating a BigData backup policy

Use the BigData policy to backup big data applications such as Hadoop clusters.

A BigData policy differs from other policies in the following respects:

- You must specify **BigData** as the policy type.
- The entries which are provided in the **Clients** tab and the **Backup Selections** differ based on the application that you want to back up.
- In the **Backup Selections** tab, you must specify certain parameters and their appropriate values.

Creating BigData policy using the NetBackup Administration Console

If you prefer using the **NetBackup Administration Console** for creating BigData policy, you can use either of the following methods:

- Creating a BigData policy using the **Policy Configuration Wizard**
- Creating a BigData policy using the NetBackup **Policies** utility

The easiest method to set up a **BigData** policy is to use the **Policy Configuration Wizard**. This wizard guides you through the setup process by automatically choosing the best values for most configurations. Not all policy configuration options are presented through the wizard. For example, **calendar-based** scheduling and the **Data Classification** setting. After the policy is created, modify the policy in the Policies utility to configure the options that are not part of the wizard.

Using the Policy Configuration Wizard to create a BigData policy for Hadoop clusters

Use the following procedure to create a BigData policy with the Policy Configuration Wizard.

To create a BigData policy with the Policy Configuration Wizard

1. In the **NetBackup Administration Console**, in the left pane, click **NetBackup Management**.
2. In the right pane, click **Create a Policy** to begin the **Policy Configuration Wizard**.
3. Select the type of policy to create:
   - **BigData** policy : A policy to backup **Hadoop** data
4. Select the storage unit type for BigData policy.
5. Click **Next** to start the wizard and follow the prompts.

   Click **Help** on any wizard panel for assistance while running the wizard.

Using the NetBackup Policies utility to create a BigData policy for Hadoop clusters

Use the following procedure to create a BigData policy with the NetBackup Policies utility.

To create a BigData policy with the NetBackup Policies utility

1. In the **NetBackup Administration Console**, in the left pane, expand **NetBackup Management > Policies**.
2. On the **Actions** menu, click **New > Policy**.
3. Type a unique name for the new policy in the **Add a New Policy** dialog box.

   Click **OK**.
4. On the **Attributes** tab, select **BigData** as the policy type.
5. On the **Attributes** tab, select the storage unit for BigData policy type.
6. On the **Schedules** tab, click **New** 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. Once you set the schedule, Hadoop data is backed up automatically as per the set schedule without any further user intervention.
7. On the **Clients** tab, enter the IP address or the host name of the NameNode.
8. 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 specified for backup selection while defining BigData Policy with Application_Type=hadoop must not contain space or comma in their names.

9 Click **OK** to save the changes.

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

Using NetBackup Command Line Interface (CLI) to create a BigData policy for Hadoop clusters

You can also use the CLI method to create a BigData policy for Hadoop.

To create a BigData policy using NetBackup CLI method

1 Log on as an Administrator.

2 Navigate to `/usr/openv/netbackup/bin/admincmd`.

3 Create a new BigData policy using the default settings.

```
bppolicynew policyname
```

4 View the details about the new policy using the `-L` option.

```
bpplinfo policyname -L
```

5 Modify and update the policy type as **BigData**.

```
bpplinfo PolicyName -modify -v -M MasterServerName -pt BigData
```

6 Specify the *Application_Type* as Hadoop.

For Windows:

```
bpplinclude PolicyName -add "Application_Type=hadoop"
```

For UNIX:

```
bpplinclude PolicyName -add 'Application_Type=hadoop'
```

Note: The parameter values for *Application_Type=hadoop* are case-sensitive.

7 Specify the backup host on which you want the backup operations to be performed for Hadoop.

For Windows:

```
bpplinclude PolicyName -add "Backup_Host=IP_address or hostname"
```

For UNIX:

```
bpplinclude PolicyName -add 'Backup_Host=IP_address or hostname'
```

Note: The backup host must be a Linux computer. The backup host can be a NetBackup client or a media server or a master server.

---

8 Specify the Hadoop directory or folder name that you want to backup.

For Windows:

```
bpplinclude PolicyName -add "/hdfsfoldername"
```

For UNIX:

```
bpplinclude PolicyName -add '/hdfsfoldername'
```

Note: Directory or folder used for backup selection while defining BigData Policy with Application_Type=hadoop must not contain space or comma in their names.

---

9 Modify and update the policy storage type for BigData policy.

```
bpplinfo PolicyName -residence STUName -modify
```

10 Specify the IP address or the host name of the NameNode for adding the client details.

For Windows:

```
bpplclients PolicyName -M "MasterServerName" -add
"HadoopServerNameNode" "Linux" "RedHat"
```

For UNIX:

```
bpplclients PolicyName -M 'MasterServerName' -add
'HadoopServerNameNode' 'Linux' 'RedHat'
```

11 Assign a schedule for the created BigData policy as per your requirements.

```
bpplsched PolicyName -add Schedule_Name -cal 0 -rl 0 -st
sched_type -window 0 0
```

Here, *sched_type* value can be specified as follows:

Schedule Type	Description
FULL	Full backup
INCR	Differential Incremental backup
CINC	Cumulative Incremental backup
TLOG	Transaction Log
UBAK	User Backup
UARC	User Archive

The default value for *sched_type* is **FULL**.

Once you set the schedule, Hadoop data is backed up automatically as per the set schedule without any further user intervention.

12 Alternatively, you can also perform a manual backup for Hadoop data.

For performing a manual backup operation, execute all the steps from Step 1 to Step 11.

13 For a manual backup operation, navigate to `/usr/openv/netbackup/bin`

Initiate a manual backup operation for an existing BigData policy using the following command:

```
bpbackup -i -p PolicyName -s Schedule_Name -S MasterServerName
-t 44
```

Here, `-p` refers to policy, `-s` refers to schedule, `-S` refers to master server, and `-t 44` refers to BigData policy type.

Disaster recovery of a Hadoop cluster

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

Table 3-2 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 17.
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 master server.	Use `tpconfig` command to add Hadoop credentials in NetBackup master server. See "Adding Hadoop credentials in NetBackup" on page 24.
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/netbackup/`. 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 26. ■ See "Configuring number of threads for backup hosts" on page 29.
Update the BigData policy with the original NameNode name.	See "Configuring NetBackup policies for Hadoop plug-in" on page 30.

Chapter 4

Performing backups and restores of Hadoop

This chapter includes the following topics:

- About backing up a Hadoop cluster
- About restoring a Hadoop cluster

About backing up a Hadoop cluster

Use the **NetBackup, Backup, Archive, and Restore** console to manage backup operations.

Table 4-1 Backing up Hadoop data

Task	Reference
Process understanding	See "Backing up Hadoop data" on page 8.
(Optional) Complete the pre-requisite for Kerberos	See "Pre-requisite for running backup and restore operations for a Hadoop cluster with Kerberos authentication" on page 38.
Backing up a Hadoop cluster	See "Backing up a Hadoop cluster" on page 38.
Best practices	See "Best practices for backing up a Hadoop cluster" on page 39.

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/netbackup/logs/nbaapidiscv` For data transfer related logs, search for corresponding backup host (using the hostname) in the log files on the master server. See "Troubleshooting backup issues for Hadoop data" on page 50.

Pre-requisite 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/netbackup/nbusers/hdfs_mykeytabfile.keytab
hdfs@MYCOMPANY.COM
```

Also review the configuration-related information. See "Configuration for a Hadoop cluster that uses Kerberos" on page 30.

Backing up a Hadoop cluster

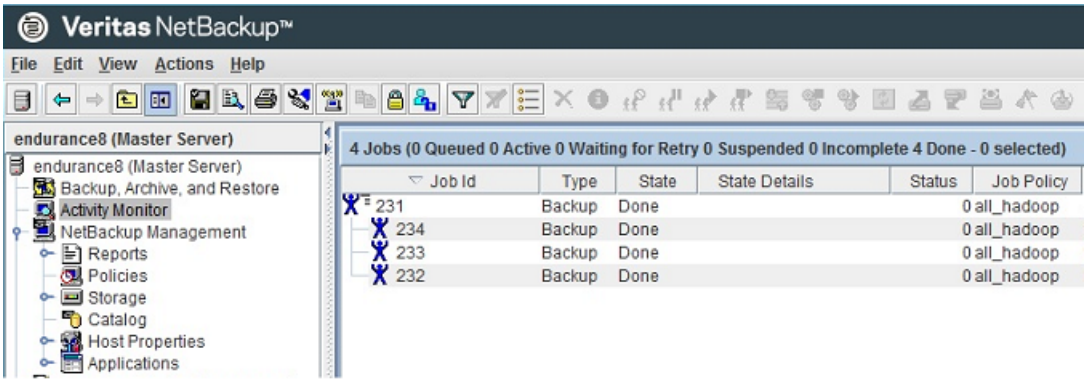
You can either schedule a backup job or run a backup job manually. See, NetBackup Administrator's Guide, Volume I

For overview of the backup process, See "Backing up Hadoop data" on page 8.

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.

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.

About restoring a Hadoop cluster

Use the **NetBackup, Backup, Archive, and Restore** console to manage restore operations.

Table 4-2 Restoring Hadoop data

Task	Reference
Process understanding	See “Restoring Hadoop data” on page 9.

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

Task	Reference
Complete the pre-requisites for Kerberos	See "Pre-requisite for running backup and restore operations for a Hadoop cluster with Kerberos authentication" on page 38.
Restoring Hadoop data on the same NameNode or Hadoop cluster	■ See "Using the Restore Wizard to restore Hadoop data on the same Hadoop cluster" on page 41. ■ See "Using the `bprestore` command to restore Hadoop data on the same Hadoop cluster" on page 42.
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 44.
Best practices	See "Best practices for restoring a Hadoop cluster" on page 48.
Troubleshooting tips	See "Troubleshooting restore issues for Hadoop data" on page 54.

Restoring Hadoop data on the same Hadoop cluster

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

- Use the Backup, Archive, and Restore console 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.

Using the Restore Wizard to restore Hadoop data on the same Hadoop cluster

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

To use the Restore Wizard to perform a restore

1. Open the **Backup, Archive, and Restore** interface.
2. Select the appropriate date range to restore the complete data set.
3. In the **Browse** directory, specify the root directory (“/”) as the path to browse.
4. From the File menu (Windows) or Actions menu (UNIX), choose **Specify NetBackup Machines and Policy Type**.
5. On the **Specify NetBackup Machines and Policy Type** wizard, enter the source and destination details for restore.
   - Specify the Hadoop NameNode as the source for which you want to perform the restore operation.
     From the **Source client for restores** list, select the required NameNode.
   - Specify the backup host as the destination client.
     From the **Destination client for restores** list, select the required backup host.
   - On the **Specify NetBackup Machines and Policy Type** wizard, enter the policy type details for restore.
     From the **Policy type for restores** list, choose **BigData** as the policy type for restore.
     Click **Ok**.
6. Go to the **Backup History** and select the backup images that you want to restore.
7. In the **Directory Structure** pane, expand the **Directory**.

   All the subsequent files and folders under the directory are displayed in the **Contents of Selected Directory** pane.
8. In the **Contents of Selected Directory** pane, select the check box for the Hadoop files that you want to restore.
9. Click **Restore**.
10. In the **Restore Marked Files** dialog box, select the destination for restore as per your requirement.
    - Select **Restore everything to its 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.

11 Click **Start Restore**.

12 Verify the restored files.

Using the `bprestore` command to restore Hadoop data on the same Hadoop cluster

The `bprestore` command lets you restore a backed up or archived file or list of files. You can also name directories to restore. If you include a directory name, `bprestore` restores all files and subdirectories of that directory. You can exclude a file or a directory path that was previously included in the restore by placing an exclamation mark (!) in front of the file or the directory path (does not apply to NDMP restores). For example, the exclude capability is useful if you want to exclude part of a directory from the restore.

To restore Hadoop data on the same location as your backup location

1. Log on as an Administrator or root user based on windows or UNIX system respectively.
2. Run the following command on the NetBackup master server by providing appropriate values:

   ```
   bprestore -S master_server -D backup_host -C client -t 44 -L
   progress log -f listfile
   ```

   Where,

   `-S master_server`

   Specifies the name of the NetBackup master server.

   `-D backup host`

   Specifies the name of the backup host.

   `-C client`

   Specifies a NameNode as a 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 whitelisted file path in which to write progress information.

   `-t 44`

   Specifies BigData as the policy type.

To restore Hadoop data on an alternate location

1. Log on as an Administrator.
2. Run the following command on the NetBackup master server by providing appropriate values:

   ```
   bprestore -S master_server -D backup_host -C client -t 44 -L
   progress log -R rename_file -f listfile
   ```

   Where,

   `-S master_server`

   Specifies the name of the NetBackup master 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 whitelisted 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.

   Change the `/<source_folder_path>` to `/<destination_folder_path>`

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: NetBackup supports redirected restores only using the Command Line Interface (CLI).

Note: Make sure that you have added the credentials for the alternate NameNode or Hadoop cluster in NetBackup master server and also completed the Whitelisting tasks on NetBackup master server. For more information about how to add Hadoop credentials in NetBackup and whitlelisting procedures, See “Adding Hadoop credentials in NetBackup” on page 24. See “Whitelisting a NetBackup client on NetBackup master server” on page 23.

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 master_server -D backup_host -C client -R rename_file -t 44 -L progress log -f listfile` command on the NetBackup master server using the modified values for the mentioned parameters in step 1.

Where,

`-S master_server`

Specifies the name of the NetBackup master 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 whitelisted 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 whitelisted 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 practices for restoring a Hadoop cluster

When restoring a Hadoop cluster, consider the following:

- Before you execute a restore job, ensure that there is sufficient space on the cluster to complete the restore job.
- Update firewall settings so that the backup hosts can communicate with the Hadoop cluster.

Chapter 5

Troubleshooting

This chapter includes the following topics:

- About troubleshooting NetBackup for Hadoop issues
- About NetBackup for Hadoop debug logging
- Troubleshooting backup issues for Hadoop data
- Troubleshooting restore issues for Hadoop data

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 50.
Backup issues	See "Troubleshooting backup issues for Hadoop data" on page 50.
Restore issues	See "Troubleshooting restore issues for Hadoop data" on page 54.
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 39. See "Best practices for restoring a Hadoop cluster" on page 48.

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/netbackup/logs`

Table 5-2 NetBackup logs related to Hadoop

Log Folder	Messages related to	Logs reside on
`install_path/NetBackup/logs/bpVMutil`	Policy configuration	Master 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 50.
- See "Backup operation for Hadoop fails with error code 6599" on page 51.
- See "Backup operation fails with error 6609" on page 51.
- See "Backup operation failed with error 6618" on page 51.

- 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 52.
- See "Backup operation fails with error 6654" on page 53.
- See "Backup operation fails with bpbrm error 8857" on page 53.
- See "Backup operation fails with error 6617" on page 53.
- See "Backup operation fails with error 6616" on page 54.

Backup operation for Hadoop fails with error code 6599

This error is encountered in the following scenarios:

- Backup operation for Hadoop fails if the directory or folder which is used as a backup selection when you define BigData policy with `Application_Type=hadoop` contains a space or comma in its name.
  Workaround:
  When you define backup selection list for BigData policy with `Application_Type=hadoop`, do not use comma or space for folder names.
- Backup host has invalid Ticket Granting Ticket (TGT) in case of Kerberos enabled Hadoop cluster.
  Workaround:
  Renew the TGT.

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:

1. If Hadoop credentials are not added in NetBackup master server

   Workaround:

   Ensure that the Hadoop credentials are added in NetBackup master server. Use the `tpconfig` command. For more information, See "Adding Hadoop credentials in NetBackup" on page 24.

2. 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.

Backup operation fails with bpbrm error 8857

This error is encountered if you have not whitelisted NetBackup client on NetBackup master server.

Workaround:

You must perform the whitelisting procedure on NetBackup master server if you want to use the NetBackup client as the backup host. For more information, See "Whitelisting a NetBackup client on NetBackup master server" on page 23.

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.

Troubleshooting restore issues for Hadoop data

- See "Restore fails with error code 2850" on page 54.
- See "NetBackup restore job for Hadoop completes partially" on page 55.
- See "Extended attributes (xattrs) and Access Control Lists (ACLs) are not backed up or restored for Hadoop" on page 52.
- See "Restore operation fails when Hadoop plug-in files are missing on the backup host" on page 55.
- See "Restore fails with bpbrm error 54932" on page 55.
- See "Restore operation fails with bpbrm error 21296" on page 56.

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 52.

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 master server.

Workaround:

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

Index

A

Adding
- backup host 21

B

Backup 38
- Hadoop 37

backup 8

BigData policy
- Command Line Interface 33
- NetBackup Administration Console 31
- Policies utility 32
- Policy Configuration Wizard 31

C

compatibility
- supported operating system 16

Creating
- BigData backup policy 31

D

deployment 10

disaster recovery 35

Downloading
- Hadoop plug-in for NetBackup 17

H

Hadoop credentials
- adding 24

I

installation 18

Installing
- verifying 19

K

Kerberos
- post installation 30

kerberos
- backup 38
- restore 38

L

License
- Hadoop 16

Limitations 13

N

NetBackup
- debug logging 50

NetBackup Appliance
- backup host 24

O

overview
- backup 6
- configuration 6
- deployment 6
- installation 6
- restore 6

P

parallel streaming framework 6

policies
- configuring 30

Preparing
- Hadoop 17

R

Removing
- backup host 21

Restore
- bprestore command 42
- Hadoop 39

restore 9

Restoring
- alternate NameNode 44

Restoring *(continued)*
Hadoop cluster 40

T

terms 12
Troubleshoot
backup 50
troubleshooting
restore 55

W

Whitelisting
backuphost 23