The software described in this book is furnished under a license agreement and may be used only in accordance with the terms of the agreement.

Documentation version: 1.0

Product version: 1.0

Legal Notice

Copyright © 2012 Symantec Corporation. All rights reserved.

Symantec and the Symantec Logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.

This Symantec product may contain third party software for which Symantec 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. Please see the Third Party Legal Notice Appendix to this Documentation or TPIP ReadMe File accompanying this Symantec product for more information on the Third Party Programs.

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 Symantec Corporation 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. SYMANTEC CORPORATION 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, "Rights in Commercial Computer Software or Commercial Computer Software Documentation", as applicable, and any successor regulations. 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.
Technical Support

Symantec Technical Support maintains support centers globally. Technical Support’s primary role is to respond to specific queries about product features and functionality. The Technical Support group also creates content for our online Knowledge Base. The Technical Support group works collaboratively with the other functional areas within Symantec to answer your questions in a timely fashion. For example, the Technical Support group works with Product Engineering and Symantec Security Response to provide alerting services and virus definition updates.

Symantec’s support offerings include the following:

- A range of support options that give you the flexibility to select the right amount of service for any size organization
- Telephone and/or Web-based support that provides rapid response and up-to-the-minute information
- Upgrade assurance that delivers software upgrades
- Global support purchased on a regional business hours or 24 hours a day, 7 days a week basis
- Premium service offerings that include Account Management Services

For information about Symantec’s support offerings, you can visit our Web site at the following URL:

**www.symantec.com/business/support/**

All support services will be delivered in accordance with your support agreement and the then-current enterprise technical support policy.

Contacting Technical Support

Customers with a current support agreement may access Technical Support information at the following URL:

**www.symantec.com/business/support/**

Before contacting Technical Support, make sure you have satisfied the system requirements that are listed in your product documentation. Also, you should be at the computer on which the problem occurred, in case it is necessary to replicate the problem.

When you contact Technical Support, please have the following information available:

- Product release level
- Hardware information
- Available memory, disk space, and NIC information
- Operating system
- Version and patch level
- Network topology
- Router, gateway, and IP address information
- Problem description:
  - Error messages and log files
  - Troubleshooting that was performed before contacting Symantec
  - Recent software configuration changes and network changes

### Licensing and registration

If your Symantec product requires registration or a license key, access our technical support Web page at the following URL:


### Customer service

Customer service information is available at the following URL:


Customer Service is available to assist with non-technical questions, such as the following types of issues:

- Questions regarding product licensing or serialization
- Product registration updates, such as address or name changes
- General product information (features, language availability, local dealers)
- Latest information about product updates and upgrades
- Information about upgrade assurance and support contracts
- Information about the Symantec Buying Programs
- Advice about Symantec's technical support options
- Nontechnical presales questions
- Issues that are related to CD-ROMs, DVDs, or manuals
Support agreement resources

If you want to contact Symantec regarding an existing support agreement, please contact the support agreement administration team for your region as follows:

- Asia-Pacific and Japan: customercare_apac@symantec.com
- Europe, Middle-East, and Africa: semea@symantec.com
- North America and Latin America: supportsolutions@symantec.com
Technical Support .................................................................................................................................. 4

Chapter 1  Introducing Symantec® Enterprise Solution for Hadoop ......................................................................... 9

About Hadoop ...................................................................................................................................... 9
About Symantec® Enterprise Solution for Hadoop .............................................................................. 10
Symantec® Enterprise Solution for Hadoop description ...................................................................... 11

Chapter 2  Pre-installation .......................................................................................................................... 13

About planning for SFHDCFS installation ............................................................................................... 13
Hardware compatibility list (HCL) ............................................................................................................ 13
CFS requirement ...................................................................................................................................... 14
Java requirement ..................................................................................................................................... 14
Hadoop requirements .............................................................................................................................. 14
Supported operating system .................................................................................................................... 15
Disk space requirements .......................................................................................................................... 15
Discovering product versions .................................................................................................................... 15
About installation and configuration methods ........................................................................................ 15
Licensing requirements ............................................................................................................................ 16

Chapter 3  Installation and uninstallation ....................................................................................................... 17

Installation preparation overview .......................................................................................................... 17
Setting up environment variables .......................................................................................................... 18
About using ssh or rsh with the installer .................................................................................................. 18
Downloading Symantec® Enterprise Solution for Hadoop ..................................................................... 18
Installing Symantec® Enterprise Solution for Hadoop using script-based installer .................................. 19
About the Symantec® Enterprise Solution for Hadoop installer ................................................................ 19
Installing SFHDCFS ................................................................................................................................ 19
Verifying product installation .................................................................................................................... 22
Installation log files ................................................................................................................................ 23
Uninstalling Symantec® Enterprise Solution for Hadoop using script-based installer .................................. 23
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uninstalling SFHDCFS</td>
<td>23</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Using Symantec® Enterprise Solution for Hadoop</td>
</tr>
<tr>
<td>Starting and stopping Symantec® Enterprise Solution for Hadoop</td>
<td>27</td>
</tr>
<tr>
<td>Deploying Symantec® Enterprise Solution for Hadoop</td>
<td>29</td>
</tr>
<tr>
<td>Using Veritas File System compression and de-duplication</td>
<td>30</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Troubleshooting issues</td>
</tr>
<tr>
<td>General troubleshooting tips</td>
<td>31</td>
</tr>
<tr>
<td>The <code>start-mapred-symc.sh</code> script fails to start JobTracker</td>
<td>32</td>
</tr>
<tr>
<td>JobTracker fails to start</td>
<td>34</td>
</tr>
<tr>
<td>JobTracker dies with an exception</td>
<td>36</td>
</tr>
<tr>
<td>TaskTracker fails to start</td>
<td>37</td>
</tr>
<tr>
<td>Problem in accessing CFS files</td>
<td>37</td>
</tr>
<tr>
<td>Submitting jobs as non Hadoop user</td>
<td>38</td>
</tr>
<tr>
<td>Starting and stopping map-reduce daemons as non Hadoop user</td>
<td>39</td>
</tr>
<tr>
<td>Modifying a SFHDCFS cluster to run map-reduce jobs over HDFS</td>
<td>39</td>
</tr>
<tr>
<td>Appendix A</td>
<td>Differences between Symantec Hadoop and Open Source Hadoop</td>
</tr>
<tr>
<td>Introduction</td>
<td>41</td>
</tr>
<tr>
<td>Disabled scripts</td>
<td>41</td>
</tr>
<tr>
<td>Added scripts</td>
<td>42</td>
</tr>
<tr>
<td>Setting up a non-HA cluster</td>
<td>43</td>
</tr>
<tr>
<td>Disabled Hadoop commands</td>
<td>44</td>
</tr>
<tr>
<td>Modified behavior of some Hadoop <code>fs</code> commands</td>
<td>45</td>
</tr>
<tr>
<td>Appendix B</td>
<td>How to run HortonWorks HDP v1.0 with Symantec® Enterprise Solution for Hadoop</td>
</tr>
<tr>
<td>Introduction</td>
<td>47</td>
</tr>
<tr>
<td>Running HortonWorks HDP v1.0 with SFHDCFS</td>
<td>48</td>
</tr>
<tr>
<td>Index</td>
<td>57</td>
</tr>
</tbody>
</table>
This chapter includes the following topics:

- About Hadoop
- About Symantec® Enterprise Solution for Hadoop
- Symantec® Enterprise Solution for Hadoop description

About Hadoop

Apache Hadoop is an open source distributed storage and compute framework. It allows certain applications to run in parallel and also lets them use the computational resources on a large number of independent machines while handling a huge amount of data.

Hadoop consists of two core components:

- Hadoop Distributed File System (HDFS)
  It provides a replicated distributed storage for applications to store and use huge amount of data.

- Hadoop Map Reduce
  It provides the framework for distributed computing to run analysis on this data and derive business value out of it.

The biggest reason why enterprises are looking at Hadoop and adopting it is because of its promise of unlocking value from the vast amount of data that is generated and stored within the enterprise. There are a few very well publicized
use cases which have shown incremental revenue generation due to using Hadoop and other Big data technologies. Each of the organizations looking at Hadoop is striving to leverage the power of data analytics to drive the revenues.

About Symantec® Enterprise Solution for Hadoop

While enterprises are looking at Hadoop to drive revenues, there are several challenges that Hadoop brings into the enterprise environment.

- Increased storage and server sprawl:
  The Hadoop cluster is built with numerous commodity servers, each with individually attached storage. The failure rate in such an environment is very high and the number of physical objects are significantly more. Hence, the cost of managing such an environment, keeping all the nodes up-to-date with the right software, and replacing the failed components is significant.

- Wastage of storage capacity:
  To mitigate the risk of data loss due to higher failure rate of the commodity hardware, Hadoop stores three copies of all the data. This approach mean that the storage required to store any data is three times the size of the data. This not only results in wastage of storage capacity, but also in poor CPU utilization as the system gets overprovisioned with additional servers to provide the 3x storage.

- Ingestion and extraction of data:
  Hadoop does not understand the data stored in the existing POSIX file systems. Therefore the data need to be moved into Hadoop before any kind of analysis can be done on it. Similarly, the results of analysis need to be extracted out of Hadoop. This adds to the cost and complexity since Hadoop does not support any of the standard file system commands and tools.

- 24 x 7 Availability:
  The data is distributed across multiple nodes but there is only one NameNode in the cluster, Hadoop’s metadata server, that knows this data distribution. All applications must go through this single NameNode to access data. This makes NameNode both a performance bottleneck and a single point of failure.

- Backup and DR Support:
  Recommendations for mitigating the costs and complexity of data moves include using Hadoop cluster as the primary store for the data. The issue is that a reliable backup solution for Hadoop cluster does not exist. Hadoop’s way of storing three copies of data is not the same as taking a backup. It does not provide archiving or a point in time recovery.

SFHDCFS uses the world’s leading Cluster File System - the Veritas Cluster File System product to overcome the above mentioned shortcomings of the Hadoop
and makes it suitable for deployment in enterprise data centers. This solution allows the enterprises to use their existing server and storage to run Big data analytics using the Hadoop platform.

Running Hadoop on Cluster File System enables:

- **High Availability** by providing access to meta-data from all the nodes in the cluster, thus eliminating a single point of failure.
- **Workload distribution** by providing high-performance access to data from all the nodes in the cluster, and preserving task distribution of map-reduce.
- **Efficient storage utilization** through de-duplication, compression and most importantly avoiding three copies problem.
- **Commoditizing storage hardware** through its support for most storage arrays in the market.
- **Simpler backups** by providing snapshots.

To summarize, the Cluster File System based Hadoop solution enables enterprises to take advantage of Big Data analytics without the trade-offs.

**Symantec® Enterprise Solution for Hadoop description**

The SFHDCFS solution works within the confines of the modular architecture of Apache Hadoop, using only the public APIs. This allows it to deliver an enterprise ready Big data solution without creating yet another competing Hadoop distribution. This works with Open Source Apache Hadoop and enables the customers to run the standard analytics applications that they would run on Apache Hadoop.

- **Replacing HDFS class:**
  The SFHDCFS solution replaces the Hadoop Distributed File System with the Cluster file system. This is achieved by implementing a CFS specific version of the file system specific classes in the Hadoop Java Class Hierarchy. This implementation talks to the layers above and below these classes only through well defined APIs and therefore, seamlessly supports the rest of the Hadoop stack (such as map-reduce).

- **Works on existing CFS installations:**
  The SFHDCFS solution requires an existing SFCFS implementation and works as-is on it. This makes it possible to run analytics along with other applications on the same set of server and storage without extracting, transforming and loading data to a separate Hadoop Cluster.

- **Single package install:**
With the SFHDCFS solution, adding analytics is as easy as installing a single package and running simple configuration scripts. It includes all the JAR files and configuration scripts in a single package. Only a Cluster file system, Apache Hadoop distribution, and Java runtime environment are needed.

- High Availability:
The SFHDCFS solution addresses every single point of failure that exists in the open source Hadoop deployment. It addresses the HDFS NameNode availability issue by making sure that each of the node has all the information about the file system metadata. It also address the availability and continuance of map-reduce jobs by providing monitoring and failover through Veritas Cluster Server (VCS). All these features together ensure that the analytics application continues to run as long as there is at least one working node in the cluster.

- Easy ingestion and extraction of data:
Any POSIX command (like, `cp`, `mv`) can be used to copy in or copy out data. In addition, this can be done over Network File System (NFS) as well as locally.
Pre-installation

This chapter includes the following topics:

- About planning for SFHDCFS installation
- Hardware compatibility list (HCL)
- CFS requirement
- Java requirement
- Hadoop requirements
- Supported operating system
- Disk space requirements
- Discovering product versions
- About installation and configuration methods
- Licensing requirements

About planning for SFHDCFS installation

This installation guide is designed for system administrators who already have knowledge of basic Linux system (such as `tar`, `mkdir`, and simple shell scripting) and network administration. Basic familiarity with the specific platform and operating system where SFHDCFS will be installed is also required.

Hardware compatibility list (HCL)

The Symantec® Enterprise Solution for Hadoop is built on top of Veritas Storage Foundation Cluster File System High Availability (SFCFS-HA) and inherits the
same hardware compatibility list as the underlying Veritas Storage Foundation Cluster File System High Availability product. For details, refer to the Veritas Storage Foundation™ Cluster File System High Availability Installation Guide.

CFS requirement

Symantec® Enterprise Solution for Hadoop requires a working SFCFS-HA cluster on all the cluster nodes. Please ensure that all the cluster nodes have the following:

- Veritas Storage Foundation Cluster File System High Availability product installed and configured.
- There should be at least one CFS mount point present on each of the nodes and all of them should mount the same Cluster File System.
- The name of the mountpoint should be same on all the cluster nodes.
- The mount point should be mounted read-write on all the cluster nodes.

Java requirement

Symantec® Enterprise Solution for Hadoop requires Sun Java version 1.6 or above installed on all the cluster nodes.

To check Java version

- From shell, type `java -version`. For example:

```
# java -version
java version "1.6.0_30"
Java(TM) SE Runtime Environment (build 1.6.0_30-b12) (rhel-1.41.1.10.4.el6-x86_64)
Java HotSpot(TM) 64-Bit Server VM (build 20.5-b03, mixed mode)
```

Hadoop requirements

The current version of SFHDCFS supports only Hadoop version 0.20.203.0, 1.0.2, and 1.0.3.

You can download Apache Hadoop from the following URL:

http://hadoop.apache.org/common/releases.html
Supported operating system

Symantec® Enterprise Solution for Hadoop release 1.0 is supported on the following operating system environments:

- Red Hat Enterprise Linux 5 Update 5, 6 & 7
- Red Hat Enterprise Linux 6 Update 1 & 2

Disk space requirements

Symantec® Enterprise Solution for Hadoop requires around 200 MB of disk space to install Hadoop.

- If Hadoop is installed on a CFS mount point, then 200 MB free disk space is required under that mount point.
- If Hadoop is installed locally on each of the node, then 200 MB of disk space is required on each of the node.

Discovering product versions

You can check the version of the currently installed Symantec® Enterprise Solution for Hadoop by checking the version of the installed SFHDCFS package.

Enter the following command to check the version:

```
# rpm -q SYMChdcfs
```

About installation and configuration methods

You can install and configure Symantec® Enterprise Solution for Hadoop using Command-line Product Installer (CPI) method for installing both Hadoop and SFHDCFS from scratch.

The installer performs the following functions:

- Installation of Hadoop binaries from a local tarball
- Installation of SFHDCFS solution
- Configuration of Hadoop for use with SFHDCFS and JobTracker HA
Licensing requirements

Symantec® Enterprise Solution for Hadoop is built on top of SFCFS-HA solution and is not separately licensed. The licenses installed for SFCFS-HA are adequate for this solution.
Installation and uninstallation

This chapter includes the following topics:

- Installation preparation overview
- Setting up environment variables
- About using ssh or rsh with the installer
- Downloading Symantec® Enterprise Solution for Hadoop
- Installing Symantec® Enterprise Solution for Hadoop using script-based installer
- Verifying product installation
- Uninstalling Symantec® Enterprise Solution for Hadoop using script-based installer

Installation preparation overview

Table 3-1 provides an overview of an installation using the product installer.

<table>
<thead>
<tr>
<th>Installation task</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download Hadoop Tarball from Apache Hadoop</td>
<td>You can download Hadoop Tarball from the following URL:</td>
</tr>
<tr>
<td></td>
<td><a href="http://hadoop.apache.org/common/releases.html">http://hadoop.apache.org/common/releases.html</a></td>
</tr>
<tr>
<td>Download the software.</td>
<td>See “Downloading Symantec® Enterprise Solution for Hadoop” on page 18.</td>
</tr>
</tbody>
</table>
## Table 3-1  Installation overview  (continued)

<table>
<thead>
<tr>
<th>Installation task</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set environment variables.</td>
<td>See “Setting up environment variables” on page 18.</td>
</tr>
<tr>
<td>Configure the Secure Shell (ssh) on all nodes.</td>
<td>See “About using ssh or rsh with the installer” on page 18.</td>
</tr>
<tr>
<td>Check that sufficient disk space is available.</td>
<td>See “Disk space requirements” on page 15.</td>
</tr>
<tr>
<td>Use the installer to install the product.</td>
<td>See “About the Symantec® Enterprise Solution for Hadoop installer” on page 19.</td>
</tr>
</tbody>
</table>

## Setting up environment variables

Most of the commands that are used in the installation are either in the /sbin or in the /usr/sbin directory. Add these directories to your PATH environment variable as necessary.

## About using ssh or rsh with the installer

The SFHDCFS installer uses passwordless secure shell (ssh) or remote shell (rsh) communications among systems. The installer uses the ssh or rsh daemon that comes bundled with the operating system.

During an installation, you choose the communication method that you want to use. Then provide the installer with the superuser passwords for the systems where you plan to install. The ssh or rsh communication among the systems is removed when the installation process completes, unless the installation abruptly terminates. If installation is terminated abruptly, use the installation script's -comcleanup option to remove the ssh or rsh configuration from the systems.

## Downloading Symantec® Enterprise Solution for Hadoop

You can download the SFHDCFS software from the following SORT URL: https://sort.symantec.com/bigdata
Installing Symantec® Enterprise Solution for Hadoop using script-based installer

This section provides information about installing Symantec® Enterprise Solution for Hadoop.

About the Symantec® Enterprise Solution for Hadoop installer

The installer enables you to install and configure the product, verify the pre-installation requirements, and view the description of the product. Use the product installation script to install the product.

During the installation, you can type the following characters for performing various actions:

<table>
<thead>
<tr>
<th>Character</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>b (back)</td>
<td>Returns to a previous section of the installation procedure. The back feature of the installation scripts is context-sensitive, so it returns to the beginning of a grouped section of questions.</td>
</tr>
<tr>
<td>q</td>
<td>Quits the installer.</td>
</tr>
<tr>
<td>?</td>
<td>Displays help information.</td>
</tr>
<tr>
<td>Enter</td>
<td>Accepts a default response.</td>
</tr>
<tr>
<td>--help</td>
<td>Displays a list of options available with the installer program.</td>
</tr>
</tbody>
</table>

Installing SFHDCFS

Use the installer program to install SFHDCFS on your system.

The following sample procedure installs SFHDCFS on a single system.

To install SFHDCFS

1. Download the SFHDCFS software from the following URL https://sort.symantec.com/bigdata and untar it into the directory.

2. In the untarred directory structure, locate the directory corresponding to the platform on which the product needs to be installed. Enter that directory.

   # cd rhel6_x86_64
3 From this directory, type the following command to start the installer program.

```bash
# ./installer
```

4 When the installer prompts you, enter the systems that you want to be a part of the SYMChdcfs cluster. A default input for system names consists of all the nodes of the underlying SF-CFSHA cluster.

```
Enter the 64 bit RHEL6 system names separated by spaces: [q,?]
(node1 node2 node3 node4)
```

5 Enter the mount point where the CFS storage to be used by SYMChdcfs is mounted.

```bash
Enter the CFS mount point to be used as storage for Hadoop:[q,?]
/cfsmount
```

6 Enter the local file path of Hadoop tarball.

```bash
Enter the full path of the Hadoop tar image that needs to be installed : [q,?]
/tmp/hadoop-1.0.2.tar.gz
```

7 Enter the location for Hadoop installation.

```bash
Enter the path at which Hadoop should be installed [q,?]
/cfsmount/install
```

---

**Note:** Hadoop can be installed either on a CFS mount point and accessible from the cluster nodes or in a local directory on each of the node.

8 Enter the user and group name for Hadoop user.

```bash
Enter the username and group (user:group) for Hadoop:[q,?]
(hduser:hdgroup)
```
9 Enter the virtual IP address which is to be used for JobTracker failover.

Enter the Virtual IP to be used for JobTracker Communication & Failover:[q,?] 10.209.134.153

**Note:** The Virtual IP is a floating IP address that always points to the node where the Hadoop JobTracker daemon is currently online. When the node failure occurs, the IP address moves to next available node in the cluster on which the same daemon is made available. You can use any free IP address as the Virtual IP. The address will be associated with a VCS service group by the installer which will handle the fail-over.

10 Enter the network mask for the virtual IP.

Enter the network mask for the Virtual IP to be used for JobTracker communication and Failover:[q,?] 255.255.252.0

11 Enter the JobTracker port.

Enter the port to be used for JobTracker communication:[q,?] (8021)

12 Enter the public interface for JobTracker.

Enter the public network interface on which JobTracker should be configured for system dfs1xel: (eth0)

13 Enter y or n as appropriate.

- Enter y if the NIC is same for all the devices.
  
  Do you want to use the same NIC device for all systems? [y,n,q] (y)

- Enter n if the NIC is different in different devices.
  
  Do you want to use the same NIC device for all systems? [y,n,q] (n)

The installer will ask for different NICs.

14 After taking these inputs, the installer checks for various factors like communication between the nodes, release compatibility, installed product, platform version, version of installed Java, product licensing, and prechecks. If the passwordless ssh communication between the nodes is not already present, installer creates it for installation purpose only.
15 Installer displays a list of Symantec® Enterprise Solution for Hadoop packages which are to be installed on all systems. For example, Apache Hadoop and SYMChdcfs.

Press Enter to continue with the installation. The installer installs the product packages.

16 Specify whether you want to send your installation information to Symantec.

Would you like to send the information about this installation to Symantec to help improve installation in the future? [y,n,q,?] (y)

Note: The information that is sent to Symantec helps to improve the installer software.

17 The installer completes the installation and starts the SFHDCFS processes.

If required check the log files to confirm the installation.

Installation log files, response and summary file are saved at:

/opt/VRTS/install/logs/installer-****

18 Installer prompts to specify whether you want to view the summary files.

Would you like to view summary file? [y,n,q,?] (n)

After installation, the installer populates the Hadoop configuration files with the required parameters like IP address, CFS mount path, etc. The installer also adds the configuration for JobTracker HA in the already running CFSHA cluster and so JobTracker becomes highly available as soon as you start it.

**Verifying product installation**

Use the following command to check that the required package SYMChdcfs has been installed.

```
# rpm -qa | grep SYM
```

Also verify that the Hadoop tar file has been extracted at the desired location.

You can verify the version of the installed product using the following command:

```
# rpm -q SYMChdcfs
```
Installation log files

After every Symantec® Enterprise Solution for Hadoop installation, the installer creates two types of log files:

- Installation log file
- Summary file

The location of installation logs directory is displayed at the end of a product installation. This directory is always located in the /opt/VRTS/install/logs directory.

You are recommended to keep the files for auditing, debugging, and future use.

Using the installation log file

The installation log file contains all commands executed during the procedure, their output, and errors generated by the commands. This file is for debugging installation problems and can be used for analysis by Symantec Support.

Using the summary file

The summary file contains the results of the installation by the installer or product installation scripts. The summary includes the list of packages, and their status (success or failure).

Uninstalling Symantec® Enterprise Solution for Hadoop using script-based installer

This section provides information about uninstalling Symantec® Enterprise Solution for Hadoop.

Uninstalling SFHDCFS

Use the following procedure to uninstall SFHDCFS:
To uninstall SFHDCFS

1. In the untarred directory structure, locate the directory corresponding to the platform on which the product needs to be uninstalled. Enter that directory.

   ```
   # cd rhel6_x86_64
   ```

2. Run the uninstaller program.

   ```
   # ./uninstaller
   ```

3. When the installer prompts you, indicate the systems that are part of the SYMChdcfs cluster. A default input for system names consists of all the nodes of the underlying SF-CFSHA cluster.

   Enter the 64 bit RHEL6 system names separated by spaces: [q,?] (node1 node2 node3 node4)

4. The installer program checks the system. It then asks you if you want to delete the Hadoop installation as well during the uninstallation.

   Would you like to delete the hadoop installation [y,n,q,?] (y)

5. The installer prompts you to enter the location where Hadoop is installed.

   Enter the path at which Hadoop is installed [q,?] /cfsmount/install

6. The installer then asks you if you want to delete the Hadoop user.

   Would you like to delete the hadoop user? [y,n,q,?] (y)

7. If you respond (y) to the question in step 6, installer prompts you to enter the user or group name for Hadoop.

   Enter the username and group (user:group) for Hadoop : [q,?] (hduser:hdgroup)

8. After taking these inputs, the installer checks for various factors like communication between the nodes, release compatibility, installed product, platform version, version of installed Java, product licensing, and prechecks.

9. Installer displays a list of Symantec® Enterprise Solution for Hadoop packages which are installed on all systems. For example, Apache Hadoop and SYMChdcfs.
10 Press **Enter** to continue with the uninstallation.

Symantec® Enterprise Solution for Hadoop packages are uninstalled. Specify whether you want to send your installation information to Symantec.

Symantec Enterprise Solution for Hadoop Uninstall completed successfully

Would you like to send the information about this installation to Symantec to help improve installation in the future? [y,n,q,?] (y) (n)

**Note:** The information that is sent to Symantec helps to improve the installer software.

11 After the uninstallation is completed, the installer displays the location of the summary, response, and log files. If required, view the files to confirm the status of uninstallation.

uninstaller log files, summary file, and response file are saved at:

/opt/VRTS/install/logs/uninstaller-****

Would you like to view the summary file? [y,n,q] (n)
Installation and uninstallation

Uninstalling Symantec® Enterprise Solution for Hadoop using script-based installer
Using Symantec® Enterprise Solution for Hadoop

This chapter includes the following topics:

■ Starting and stopping Symantec® Enterprise Solution for Hadoop
■ Deploying Symantec® Enterprise Solution for Hadoop
■ Using Veritas File System compression and de-duplication

Starting and stopping Symantec® Enterprise Solution for Hadoop

After the installation is successfully completed, you should have a working Hadoop directory configured for HDCFS which is named as HADOOP_HOME. This directory is created inside the directory which is given to the installer as Hadoop Installation Directory. Enter the directory and confirm the existence of the bin/hadoop executable.

Note: The installer ensures the existence of the Hadoop user in all the nodes. By default, the Hadoop user has the user name hduser.

After installation and configuration is complete, ensure the following:

■ The Hadoop user has appropriate permissions to store the temporary as well as the output data.

■ The CFS mountpoint that is configured as Hadoop storage has read-write permissions for the Hadoop user. If write permissions cannot be granted at the mount point level, create a tmp directory under the CFS mountpoint and
change the ownership and permissions for this directory to Hadoop user. This
directory is required for JobTracker and TaskTracker to store temporary state
data. It is highly recommended to change the ownership of this directory to
the Hadoop user.

- The directory which contains the input for a map-reduce application should
  have read permissions for the Hadoop user. Similarly, the directory where the
  output of a map-reduce application is written should have write permissions
for the Hadoop user.

You can bring up the map-reduce daemons by entering the following command
as Hadoop user:

```bash
$ bin/start-mapred-symc.sh
Starting Hadoop JobTracker and TaskTrackers using VCS.
You will be prompted for root credentials.
Password:
VCS NOTICE V-16-1-50735 Attempting to online group on system node1
......done
Password:
VCS NOTICE V-16-1-50735 Attempting to online group on system node1
VCS NOTICE V-16-1-50735 Attempting to online group on system node2
.....done
```

You are ready to run map-reduce jobs now.

---

**Note:** VCS requires the root password to start processes under its control. Enter
the root password to start the daemons through VCS, when prompted by VCS.

---

For example, if you want to run the pi value calculator job, enter the following
command as Hadoop user:

```bash
$ bin/hadoop jar hadoop-examples-*-jar pi 2 500000
```

This calculates the value of pi up to six places after decimals using two maps.
The map-reduce deemons should be stopped prior to uninstalling the Hadoop
solution.

---

**Note:** The root password is required by VCS to bring down the resources.

---

You can stop the map-reduce daemons using the following command as Hadoop
user:
$ bin/stop-mapred-symc.sh
Stopping Hadoop JobTracker and TaskTrackers using VCS.
You will be prompted for root credentials.
Password:
VCS NOTICE V-16-1-50733 Attempting to offline group on system node1
VCS NOTICE V-16-1-50733 Attempting to offline group on system node2
..Password:
VCS NOTICE V-16-1-50733 Attempting to offline group on system node1
..Done

You can add the $HADOOP_HOME/bin directory to the PATH environment variable for convenience.

Deploying Symantec® Enterprise Solution for Hadoop

SFHDCFS can be deployed in a variety of configurations and the following points should be considered while deploying and configuring SFHDCFS:

- Directory for installing Hadoop binaries and configuration files:
  SFHDCFS provides the flexibility to install the Hadoop binaries and configuration files either on a Cluster File System mounted on all the participating nodes or in the local filesystem of each of the node. The recommended deployment practice is to install Hadoop binaries and configuration files in a directory on a Cluster File System mounted on all the participating nodes. This not only reduces the space requirement for the Hadoop installation, but also protects against accidental configuration drifts between various nodes as only one shared copy of binaries and configuration is maintained. The installer for SFHDCFS automatically detects whether the install directory provided for Hadoop installation is on CFS or on local mount and installs a single or multiple instances of Hadoop appropriately.

- Directory for intermediate output from map tasks:
  The intermediate output from map tasks of a MR application is stored in the /tmp directory on each of the nodes. The space requirement for this directory depends on the map-reduce application and the size of the input data. It is generally a good practice to make sure that there is adequate space in the /tmp directory on each of the nodes before configuring and running the MR application so that application failure due to ENOSPC error can be avoided.

- Directory for Hadoop logs:
  The logs from JobTracker and TaskTracker are stored in the /var/SYMChdcfs directory on each of the node. Since all nodes are simultaneously writing to this directory, it is recommended to not put this directory on CFS and keep it local to each of the node.
Block size of the underlying VxFS and CFS File System:
Hadoop is generally used to process large files. It is recommended that underlying VxFS and CFS file systems be created with 8 K block size when starting the deployment on a fresh environment without any pre-existing data.

Caution: Creating and re-creating a file system to change the block size will erase all the data on that file system. Do not re-create the filesystem if you want to use the existing data on the filesystem.

Hadoop file system block size: The default block size for the Hadoop file system is set to 128 MB. This is also the recommended value for SFHDCFS. However this value can be changed based on the characteristics of the environment in which this solution is being used.

Using Veritas File System compression and de-duplication

Symantec® Enterprise Solution for Hadoop allows a map-reduce analytics application running on Hadoop to use inputs which has been compressed or de-duplicated using the features of the Veritas File System (release version 6.0 or later). When an application reads an input file which has been compressed, the file system uncompreses it in the memory before returning the data to the application. Similarly, de-duplicated data can be referenced seamlessly through multiple files while occupying the same blocks on the disk.

Compression and de-duplication features are available with Storage Foundation for Cluster File System - HA product (release version 6.0 and later).

Please refer to the Storage Foundations Advanced Solutions guide for more information on these features.
Troubleshooting issues

This chapter includes the following topics:

■ General troubleshooting tips
■ The start-mapred-symc.sh script fails to start JobTracker
■ JobTracker fails to start
■ JobTracker dies with an exception
■ TaskTracker fails to start
■ Problem in accessing CFS files
■ Submitting jobs as non Hadoop user
■ Starting and stopping map-reduce daemons as non Hadoop user
■ Modifying a SFHDCFS cluster to run map-reduce jobs over HDFS

General troubleshooting tips

Ensure the following:

■ All the nodes have the same filesystem mount on the CFS mount point being used for Hadoop Storage.
■ All the nodes have read write permission on the CFS mount point which is being used for Hadoop Storage.
■ The log and the pid directories in /var/SYMChdcfs have write permission for the Hadoop user on all the nodes.
■ The time is synchronized for all the cluster nodes.
The virtual IP address used for JobTracker HA is not in use for any other service.

Java version 1.6.0 or above is installed on all nodes.

A supported Hadoop version is used for installation.
For more information See “Hadoop requirements” on page 14.

The ports that are used by JobTracker and TaskTracker processes are not already in use.

The minimum space requirement for the installer is met.
For more information See “Hadoop requirements” on page 14.

The start-mapred-symc.sh script fails to start JobTracker

One of the following messages is displayed when the start-mapred-symc.sh script fails to start JobTracker:

Case 1:

$ bin/start-mapred-symc.sh
Starting Hadoop JobTracker and TaskTrackers using VCS.
You will be prompted for root credentials.
Password:

VCS NOTICE V-16-1-50997 Failover group (Hadoop_JT_SG) is online in cluster.
done
dfs1x1: tasktracker running as process 17831. Stop it first.
dfs1x3: tasktracker running as process 21227. Stop it first.
dfs1x4: tasktracker running as process 17204. Stop it first.
dfs1x2: tasktracker running as process 21783. Stop it first.

Suggested solution:
The message “Failover group (Hadoop_JT_SG) is online in cluster” indicates that the JobTracker is already running on one of the cluster nodes. You do not need to do anything in this case.

Case 2:

$ bin/start-mapred-symc.sh
Starting Hadoop JobTracker and TaskTrackers using VCS.
You will be prompted for root credentials.
Password:
Unable to online group Hadoop_JT_SG on system dfslx1.
Group is not completely offline on system

Unable to online group Hadoop_JT_SG on system dfslx2.
Group is not completely offline on system

Unable to online group Hadoop_JT_SG on system dfslx3.
Group is not completely offline on system

Unable to online group Hadoop_JT_SG on system dfslx4.
Group is not completely offline on system

Suggested solution:
Check the state of the VCS service group Hadoop_JT_SG using the following command:

```
# /opt/VRTSvcs/bin/hagrp -state Hadoop_JT_SG
```

If the state is **[OFFLINE][FAULTED]**, examine the reason for the failure of the JobTracker in the logs and correct issues, if any. Clear the VCS state using the following command:

```
# /opt/VRTSvcs/bin/hagrp -clear Hadoop_JT_SG
```

Normal execution of the `start-mapred-symc.sh` script should give the following output:

```
$ bin/start-mapred-symc.sh
Starting Hadoop JobTracker and TaskTrackers using VCS.
You will be prompted for root credentials.
Password:
VCS NOTICE V-16-1-50735 Attempting to online group on system dfslx1
......done
dfslx2: starting tasktracker, logging to /var/SYMChdcfs/logs/
hadoop-hduser-tasktracker-dfslx2.vxindia.veritas.com.out
dfslx3: starting tasktracker, logging to /var/SYMChdcfs/logs/
hadoop-hduser-tasktracker-dfslx3.vxindia.veritas.com.out
dfslx4: starting tasktracker, logging to /var/SYMChdcfs/logs/
hadoop-hduser-tasktracker-dfslx4.vxindia.veritas.com.out
dfslx1: starting tasktracker, logging to /var/SYMChdcfs/logs/
hadoop-hduser-tasktracker-dfslx1.vxindia.veritas.com.out
```
JobTracker fails to start

One of the following error messages is displayed when the JobTracker fails to start:

- **Case 1:**

  ```java
  java.lang.UnsatisfiedLinkError: no VeritasFileSystem in java.library.path
      at java.lang.ClassLoader.loadLibrary(ClassLoader.java:1681)
      at java.lang.Runtime.loadLibrary0(Runtime.java:840)
      at java.lang.System.loadLibrary(System.java:1047)
      at com.symantec.hadoop.fs.vxfs.LibVxFS.<clinit>(LibVxFS.java:25)
      at com.symantec.hadoop.fs.vxfs.VeritasFileSystem.setPermission
          (VeritasFileSystem.java:492)
      at com.symantec.hadoop.fs.vxfs.VeritasFileSystem.mkdirs
          (VeritasFileSystem.java:126)
      at org.apache.hadoop.mapred.JobHistory.init
          (JobHistory.java:510)
      at org.apache.hadoop.mapred.JobTracker$2.run
          (JobTracker.java:2325)
      at org.apache.hadoop.mapred.JobTracker$2.run
          (JobTracker.java:2322)
  ```

  **Suggested solution:**
  
  This error message indicates that JVM is not able to find the native libVeritasFileSystem.so library in the path. Include the library in the path and try again.

- **Case 2:**

  ```java
      at java.util.Properties$LineReader.readLine(Properties.java:435)
      at java.util.Properties.load0(Properties.java:354)
      at java.util.Properties.load(Properties.java:342)
      at com.symantec.hadoop.fs.vxfs.VxFSFileOutputStream.<init>
          (VxFSFileOutputStream.java:50)
      at com.symantec.hadoop.fs.vxfs.VeritasFileSystem.create
          (VeritasFileSystem.java:262)
      at com.symantec.hadoop.fs.vxfs.VeritasFileSystem.create
          (VeritasFileSystem.java:241)
      at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:408)
      at org.apache.hadoop.fs.FileSystem.create(FileSystem.java:287)
  ```
updateRestartCount(JobTracker.java:1721)
at org.apache.hadoop.mapred.JobTracker.offerService
(JobTracker.java:2562)
at org.apache.hadoop.mapred.JobTracker.main(JobTracker.java:5033)

Suggested Solution:
This error message indicates that JVM is not able to find the vxfs.properties file in the path. Copy the file from the /opt/SYMChdcfs/conf/ directory to the $HADOOP_HOME/conf directory.

■ Case 3:

2012-08-02 02:05:11,329 WARN org.apache.hadoop.mapred.JobTracker: Manually delete the mapred.system.dir (vxfs://localhost:9003/tmp/hadoop-hduser/mapred/system) and then start the JobTracker.
2012-08-02 02:05:11,331 WARN org.apache.hadoop.mapred.JobTracker: Bailing out ...
org.apache.hadoop.security.AccessControlException: The systemdir vxfs://localhost:9003/tmp/hadoop-hduser/mapred/system is not owned by hduser
at org.apache.hadoop.mapred.JobTracker.<init>(JobTracker.java:2370)
at org.apache.hadoop.mapred.JobTracker.<init>(JobTracker.java:2180)
at org.apache.hadoop.mapred.JobTracker.startTracker(JobTracker.java:307)
at org.apache.hadoop.mapred.JobTracker.startTracker(JobTracker.java:298)
at org.apache.hadoop.mapred.JobTracker.main(JobTracker.java:5042)

Suggested Solution:
This error message indicates insufficient permissions for the hduser to write to the <CFSMOUNT>/tmp directory. This directory should be owned by the Hadoop user (hduser by default) and should have write permissions for hduser.

Note: Error messages are generated in the JobTracker log file. This file is located inside the Hadoop logs directory. The default path is /var/SYMChdcfs/logs.
JobTracker dies with an exception

If the JobTracker dies with an exception similar to:

(Ljava/lang/String;Ljava/lang/String;)I
   at com.symantec.hadoop.fs.vxfs.LibVxFS.setFilePermission (Native Method)
   at com.symantec.hadoop.fs.vxfs.LibVxFS.setFilePermission (LibVxFS.java:154)
   at com.symantec.hadoop.fs.vxfs.VeritasFileSystem.setPermission (VeritasFileSystem.java:492)
   at com.symantec.hadoop.fs.vxfs.VeritasFileSystem.mkdirs (VeritasFileSystem.java:126)
   at org.apache.hadoop.mapred.JobHistory.init (JobHistory.java:510)
   at org.apache.hadoop.mapred.JobTracker$2.run (JobTracker.java:2325)
   at org.apache.hadoop.mapred.JobTracker$2.run (JobTracker.java:2322)
   at java.security.AccessController.doPrivileged (Native Method)
   at javax.security.auth.Subject.doAs (Subject.java:416)
   at org.apache.hadoop.security.UserGroupInformation.doAs (UserGroupInformation.java:1059)
   at org.apache.hadoop.mapred.JobTracker.<init> (JobTracker.java:2322)
   at org.apache.hadoop.mapred.JobTracker.<init> (JobTracker.java:2180)
   at org.apache.hadoop.mapred.JobTracker.startTracker (JobTracker.java:307)
   at org.apache.hadoop.mapred.JobTracker.startTracker (JobTracker.java:298)
   at org.apache.hadoop.mapred.JobTracker.main (JobTracker.java:5032)

Suggested Solution:

This means that VM is able to find the native libVeritasFileSystem.so library in the path, but the library is not correct. You can copy the libVeritasFileSystem.so file from the
/opt/SYMChdcfs/lib/$HADOOP_VERSION/native/Linux-i386-32/ or
/opt/SYMChdcfs/lib/$HADOOP_VERSION/native/Linux-amd64-64/ to the
$HADOOP_HOME/lib/native/Linux-amd64-64 directory depending on whether you are using 32-bit or 64-bit java for Hadoop..
TaskTracker fails to start

The following error message is displayed in the TaskTracker log file when the TaskTracker fails to start:

```
2012-08-02 02:05:16,004 INFO org.apache.hadoop.mapred.TaskTracker: Starting tasktracker with owner as hduser
2012-08-02 02:05:16,006 WARN org.apache.hadoop.mapred.TaskTracker: TaskTracker local directory is not writable: /tmp/hadoop-hduser/mapred/local
2012-08-02 02:05:16,006 ERROR org.apache.hadoop.mapred.TaskTracker: Can not start task tracker because org.apache.hadoop.util.DiskChecker$DiskErrorException: all local directories are not writable
   at org.apache.hadoop.mapred.TaskTracker.checkLocalDirs(TaskTracker.java:3403)
   at org.apache.hadoop.mapred.TaskTracker.initialize(TaskTracker.java:625)
   at org.apache.hadoop.mapred.TaskTracker.<init>(TaskTracker.java:1328)
   at org.apache.hadoop.mapred.TaskTracker.main(TaskTracker.java:3430)
```

2012-08-02 02:05:16,008 INFO org.apache.hadoop.mapred.TaskTracker: SHUTDOWN_MSG:

Suggested Solution:

This error indicates insufficient permissions to the /tmp/hadoop-<HADOOP_USER> directory. This directory should be owned by the Hadoop user (hduser) by default and should have write permission.

Problem in accessing CFS files

The bin/hadoop dfs -ls command fails to list any file system contents or the FileNotFound exception occurs for any of the files in CFS.

Suggested Solution:

Such errors occur when the underlying CFS file system is not mounted on one or more systems.

In this case, do the following:

- Check that the file system is mounted on all the nodes.
The files are accessible from each of the nodes.

---

**Submitting jobs as non Hadoop user**

Symantec® Enterprise Solution for Hadoop recommends that you start and stop the map-reduce daemons as the designated Hadoop user. Only the Hadoop user can submit jobs. If any other user tries to submit a job, the job fails with the following exception:

```
org.apache.hadoop.ipc.RemoteException: java.io.IO
Exception: java.io.FileNotFoundException: File [...] xml does not exist.
```

**Suggested solution:**

If you require multiple users to be able to run map-reduce jobs, you need to set up root as the designated Hadoop user.

You can modify the SFHDCFS setup using the following steps:

1. Stop all the map-reduce daemons that are currently running.
   
   ```
   $ stop-mapred-symc.sh
   ```

2. Change the Hadoop user to get root access on the system.
   
   ```
   $ su
   ```

3. Change the VCS configurations.
   
   ```
   # /opt/VRTSvcs/bin/haconf -makerw
   # /opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker User "root"
   # /opt/VRTSvcs/bin/haconf -dump -makero
   # /opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker User "root"
   ```

4. Set up keyless ssh for root.
   
   If you omit this step, every time you try to start or stop the map-reduce daemons, you will be prompted to provide root passwords for all the nodes.

5. Start the map-reduce daemons as root.
   
   ```
   # start-mapred-symc.sh
   ```
Starting and stopping map-reduce daemons as non Hadoop user

The map-reduce daemons (JobTracker and TaskTracker) are started and stopped using the `start-mapred-symc.sh` and `stop-mapred-symc.sh` scripts respectively. These scripts should always be invoked by the Hadoop user only. A non Hadoop user cannot start or stop the daemons. If by mistake, these commands are fired by a user other than the hadoop user (say, root) then not only does that attempt fail, but all subsequent attempts to start the mapred daemons, even as a Hadoop user, may fail due to permission issues.

You can rectify the permission issues to re-enable the Hadoop user to start the daemons again.

To rectify the permission issues to re-enable the Hadoop user to start the daemons:

1. `# chown -R hduser:hdgroup $CFSMOUNT/.symchdcfs`
2. `# chown -R hduser:hdgroup $CFSMOUNT/tmp`
3. `# chown -R hduser:hdgroup /tmp/hadoop-hduser`
4. `# chown -R hduser:hdgroup /var/SYMChdcfs/`

where `CFSMOUNT` is the cfs mount point used for Hadoop. Now, you can invoke the `start-mapred-symc.sh` and `stop-mapred-symc.sh` scripts as Hadoop user to start and stop the map-reduce daemons respectively.

Modifying a SFHDCFS cluster to run map-reduce jobs over HDFS

You may modify a standard Symantec® Enterprise Solution for Hadoop to run map-reduce jobs over HDFS (the default Apache Hadoop behavior). You need to do the following:

- Replace the Hadoop scripts to their original versions. While modifying scripts inside the `$HADOOP_HOME/bin` according to our need, we keep backup copies of the original scripts in the `$HADOOP_HOME/bin/SYMCHDCFS_BACKUP` directory. You can use them to replace the modified scripts with their original versions.
- Modify the configuration carefully.
Replace all the configuration parameters set in core-site.xml and mapred-site.xml with the standard Apache Hadoop configurations. Additionally, you may need to modify hdfs-site.xml. These files can be located in the $HADOOP_HOME/conf directory. You need to modify mapred-site.xml as well because the configuration parameters set in here by SFHDCFS assumes standard SFHDCFS setup.
Differences between Symantec Hadoop and Open Source Hadoop

This appendix includes the following topics:

- Introduction
- Disabled scripts
- Added scripts
- Setting up a non-HA cluster
- Disabled Hadoop commands
- Modified behavior of some Hadoop fs commands

Introduction

There are a few start-up and shutdown scripts, Hadoop commands, and command line options which are not applicable in the Symantec® Enterprise Solution for Hadoop (SYMChdcfs) context. There are also some additional commands that support the extra functionality provided by SFHDCFS. This appendix provides a comprehensive list of all these modified behaviors.

Disabled scripts

Some of the scripts in the $HADOOP_HOME/bin are not applicable in the Symantec® Enterprise Solution for Hadoop (SYMChdcfs) context. This is because these scripts
manage the HDFS part of open source Hadoop. Entering these scripts generates the following error message:

This functionality is not applicable for Symantec® Enterprise Solution for Hadoop. Please see the product documentation for more information.

List of disabled scripts:

- start-all.sh
- start-balancer.sh
- start-dfs.sh
- start-mapred.sh
- stop-all.sh
- stop-balancer.sh
- stop-dfs.sh
- stop-mapred.sh

These scripts are disabled by the Symantec® Enterprise Solution for Hadoop installer. Backups of the original scripts are kept in the following directory:

$HADOOP_HOME/bin/SYMC_HDCFS_BACKUP directory.

Added scripts

The following new scripts are available to support the extra functionality that has been provided:

<table>
<thead>
<tr>
<th>Product installation script</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>apache_hadoop</td>
<td>It is a copy of the original Hadoop executable.</td>
</tr>
<tr>
<td>start-mapred-symc.sh</td>
<td>Starts the map-reduce daemons in HA mode.</td>
</tr>
<tr>
<td>stop-mapred-symc.sh</td>
<td>Stops the map-reduce daemons in HA mode.</td>
</tr>
<tr>
<td>start-jobtracker.sh</td>
<td>Starts only the JobTracker daemon. Also needed for JobTracker HA.</td>
</tr>
<tr>
<td>stop-jobtracker.sh</td>
<td>Stops only the JobTracker daemon. Also needed for JobTracker HA.</td>
</tr>
</tbody>
</table>
Setting up a non-HA cluster

Instead of the regular start-up and shut-down scripts, Symantec Enterprise Solution for Hadoop requires the use of Symantec provided `start-mapred-symc.sh` and `stop-mapred-symc.sh` commands to start and stop the map-reduce daemons. Both these scripts require a root password for starting the JobTracker daemon in HA mode.

Alternatively, you can set up a non-HA cluster which does not require root access.

To set up a non-HA cluster which does not require root access

1. Prepare a non-HA cluster:
   - Disable VCS monitoring. This is an one time operation and requires root access. Enter the following commands:
     
     ```
     # /opt/VRTS/bin/haconf -makerw
     # /opt/VRTS/bin/hagrp -modify Hadoop_JT_SG ManageFaults NONE
     # /opt/VRTS/bin/hagrp -modify Hadoop_TT_SG ManageFaults NONE
     # /opt/VRTS/bin/haconf -dump -makero
     ```
   - Edit `$HADOOP_HOME/conf/mapred-site.xml` file to mention the hostname of the node where you want to run the JobTracker daemon.
     
     ```
     # vi $HADOOP_HOME/conf/mapred-site.xml
     [...]
     <property>
     <name>mapred.job.tracker</name>
     <value>hostname of JobTracker node:8021</value>
     <description>
     The ip address and the port number for the JobTracker.
     <description>
     ```

---

**Table A-1** *(continued)*

<table>
<thead>
<tr>
<th>Product installation script</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>start-tasktracker.sh</td>
<td>Starts only the TaskTracker daemons in all the nodes. It is also needed for TaskTracker HA.</td>
</tr>
<tr>
<td>stop-tasktracker.sh</td>
<td>Stops only the TaskTracker daemons in all the nodes.</td>
</tr>
</tbody>
</table>
2 Start the map-reduce daemons in non-HA mode:
Log in to the system where you want to run the non-HA daemon and enter the following commands in the same sequence.

$ start-jobtracker.sh

$ start-tasktracker.sh

3 Stop the map-reduce daemons in non-HA mode:

$ stop-tasktracker.sh

$ stop-jobtracker.sh

4 Revert back to the HA clustering:
To get the HA cluster back, you can do the following:

- Edit $HADOOP_HOME/conf/mapred-site.xml file to put the JobTracker virtual IP back in place.
- Re enable VCS monitoring. This is a one time operation and requires root access.

# /opt/VRTS/bin/haconf -makerw
# /opt/VRTS/bin/hagrp -modify Hadoop_JT_SG ManageFaults ALL
# /opt/VRTS/bin/hagrp -modify Hadoop_TT_SG ManageFaults ALL
# /opt/VRTS/bin/haconf -dump -makero

Disabled Hadoop commands

A few Hadoop commands (invoked as $ hadoop <command>) are disabled. These commands are not applicable in the Symantec® Enterprise Solution for Hadoop context because they manage the HDFS part of open source Hadoop. Entering these commands results in the following error message:

This option is not supported with Symantec® Enterprise Solution for Hadoop. Please refer to product documentation for more details.

List of disabled Hadoop commands:
- namenode -format
- secondarynamenode
- namenode
- datanode
- dfsadmin
- fsck
- balancer
- fetchdt

**Note:** If you need the above functionalities, use the
$HADOOP_HOME/bin/apache_hadoop executable. It is a copy of the original Hadoop executable created by Symantec® Enterprise Solution for Hadoop.

### Modified behavior of some Hadoop `fs` commands

The behavior of some commands under `hadoop fs` and `hadoop dfs` (invoked as `hadoop fs <-command>`) have been modified. These commands mostly deal with HDFS specific functionalities.

**Table A-2**

<table>
<thead>
<tr>
<th>Commands</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hadoop fs -setrep</td>
<td>Set replication is not applicable if the underlying storage is CFS.</td>
</tr>
<tr>
<td>hadoop fs -ls [path]</td>
<td>If no path is specified, the contents of &lt;SYMChdcfs Mount Point&gt; will be listed.</td>
</tr>
<tr>
<td>hadoop fs -rm -skipTrash</td>
<td>As SYMChdcfs does not have Trash functionality, skipping Trash is not applicable if the underlying storage is CFS. This option is ignored.</td>
</tr>
<tr>
<td>hadoop fs -rmr -skipTrash</td>
<td>As SYMChdcfs does not have Trash functionality, skipping Trash is not applicable if the underlying storage is CFS. This option is ignored.</td>
</tr>
</tbody>
</table>
Modified behavior of some Hadoop fs commands

Table A-2  (continued)

<table>
<thead>
<tr>
<th>Commands</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hadoop fs -expunge</td>
<td>As SYMChdfs does not have Trash functionality, expunging Trash is not applicable if the underlying storage is CFS.</td>
</tr>
</tbody>
</table>
How to run HortonWorks HDP v1.0 with Symantec® Enterprise Solution for Hadoop

This appendix includes the following topics:

■ Introduction
■ Running HortonWorks HDP v1.0 with SFHDCFS

Introduction

To make the HortonWorks HDP v1.0 work with SFHDCFS, you have to first ensure that HDP v1.0 is installed, configured, and is working properly. You can run a small map-reduce example program to confirm that the set up is running. Ensure that you stop the HDP cluster before executing any steps. You can use the HMC (web GUI) to stop all the services or call the /usr/lib/hadoop/bin/stop-all.sh script.

Note: With Hortonworks HDP v1.0, Symantec's Jobtracker failover functionality works in a reduced functionality mode. If a job fails due to JobTracker failover, it is resubmitted and started from the beginning and not from the point where the job was aborted.
Running HortonWorks HDP v1.0 with SFHDCFS

**Note:** Perform the following steps on each node in the cluster unless otherwise specified.

To make the HortonWorks HDP v1.0 work with SFHDCFS

1. **Install SF-CFS.**

   Symantec® Enterprise Solution for Hadoop uses Storage Foundation Cluster File System (SF-CFS) as a basic file system for Hadoop storage, replacing the standard HDFS used in the HDP’s Hadoop solution.

   **Note:** It is assumed that you have already installed and configured the SF-CFS product on the cluster on which you would like to run the HDP solution and that one CFS file system with adequate storage capacity (needed to host the input and output data for your map-reduce applications) is mounted on all nodes. It is also assumed that SF-CFS is tested with basic installation of Hortonworks HDP v1.0 consisting of MapReduce service only and that other services like HBase, Hive, Pig and Zookeeper have not been tested with SFHDCFS.

   For details, refer to the *Veritas Storage Foundation™ Cluster File System High Availability Installation Guide.*

   Make a note of the mount point and the name of the Veritas Cluster Server (VCS) service group that is being used to bring up this CFS mount point on all the nodes as it is referred in the steps below. You can list the existing VCS service groups using the following command:

   ```bash
   # /opt/VRTS/bin/hagrp -state
   ```

2. **Install SYMChdcfs.rpm.**

   - Install the package `SYMChdcfs-1.0.000.000-GA_GENERIC.x86_64.rpm` on all nodes of the cluster.
   This package is found at the following location in the media:

     - `SYMChdcfs-1.0/rhel6_x86_64/rpms`
     - `SYMChdcfs-1.0/rhel5_x86_64/rpms`

   - Copy the appropriate rpm application to your RHEL version on all the nodes using either scp or rcp or any other mechanism.
Enter the following command:

```
rpm -ivh SYMChdcfs-1.0.000.000-GA_GENERIC.x86_64.rpm
```

Verify that the `/opt/SYMChdcfs` directory is credited on each system.

3 Copy files from SFHDCFS package.

**Note**: Backup the HDP configuration files and the `start-*.sh` and `stop-*.sh` scripts before overwriting them.

Assuming you have installed the HDP Hadoop at `/usr/lib/hadoop`, copy the following files on each node:

```
# cp /opt/SYMChdcfs/script/start*.sh /usr/lib/hadoop/bin

# cp /opt/SYMChdcfs/script/stop*.sh /usr/lib/hadoop/bin

# cp /opt/SYMChdcfs/conf/vxfs.properties /usr/lib/hadoop/conf/

# cp /opt/SYMChdcfs/lib/1.0.3/SYMChdcfs.jar /usr/lib/hadoop/lib/

# cp /opt/SYMChdcfs/lib/1.0.3/native/Linux-amd64-64/libVeritasFileSystem.so /usr/lib/hadoop/lib/native/Linux-amd64-64/

# cp /opt/SYMChdcfs/lib/1.0.3/native/Linux-i386-32/libVeritasFileSystem.so /usr/lib/hadoop/lib/native/Linux-i386-32/
```
4 Modify Hadoop configuration files.

Edit the `/usr/lib/hadoop/conf/core-site.xml` file and append the following xml tags at the end of file. Delete the original tags with same name if they already exist. Replace `__SYMC_CFSMNTPT__` with the CFS mount point as mentioned in step 1.

```
<property>
    <name>fs.default.name</name>
    <value>vxfs://localhost:9003</value>
    <description>
        This is the URI of the shared FileSystem.
    </description>
</property>

<property>
    <name>fs.vxfs.impl</name>
    <value>com.symantec.hadoop.fs.vxfs.VeritasFileSystem</value>
    <description>
        This is implementation of VeritasFileSystem Class.
    </description>
</property>

<property>
    <name>vxfs.dir</name>
    <value>__SYMC_CFSMNTPT__</value>
    <description>
        CFS FileSystem to be used as the data store for Hadoop. This mount point must be identical and accessible on all the nodes in the hadoop configuration.
    </description>
</property>

<property>
    <name>ipcclient.connect.max.retries</name>
    <value>60</value>
    <description>
        The maximum number of retries a client will make to establish a server connection. This is needed for JobTracker failover.
    </description>
</property>
```

5 Modify Hadoop configuration files.
Edit the /usr/lib/hadoop/conf/mapred-site.xml file and append the following xml tags to the end of the file. Delete the original tags, if they already exist.

```xml
<property>
  <name>mapred.job.tracker</name>
  <value>__SYMC_JTVIP__:50300</value>
  <description>
    The ip address and the port number for the JobTracker.
  </description>
</property>

<property>
  <name>mapred.job.tracker.http.address</name>
  <value>__SYMC_JTVIP__:50030</value>
  <description>
    No description
  </description>
</property>

<property>
  <name>mapreduce.history.server.http.address</name>
  <value>__SYMC_JTVIP__:51111</value>
  <description>
    Http address of the history server
  </description>
</property>

<property>
  <name>hadoop.job.history.location</name>
  <value>/__SYMC_CFSMNTPT__/symchdcfs/history</value>
  <description>
    Store the JobTracker history file in CFS to aid in recovery and failover.
  </description>
</property>

<property>
  <name>mapred.job.tracker.persist.jobstatus.dir</name>
  <value>/__SYMC_CFSMNTPT__/symchdcfs/controldir</value>
  <description>
    This directory is used to persistently store the job status information and other transient information, so that it is
  </description>
</property>
```
available even after it drops off the memory queue and also across jobtracker restart and failover. This is required for JobTracker HA using SF-CFS/VCS.

In the above file, replace

- __SYMC_JTVIP__ with the IP address that you want the JobTracker process to use. This IP address is used by VCS to provide High Availability (HA) for the JobTracker process. In case the node running the JobTracker process becomes unavailable, VCS migrates this IP address to a new node and brings up the JobTracker process so that it becomes available again on the same IP address as before.

- __SYMC_CFSMNTPT__ with the CFS mount point that you noted down in step 1.

6 Create two directories using the following commands:

```bash
# mkdir -p /var/SYMChdcfs/logs /var/SYMChdcfs/pid
# chown -R mapred:mapred /var/SYMChdcfs
```
Edit /usr/lib/hadoop/conf/hadoop-env.sh and add the following:

```bash
export HADOOP_LOG_DIR=/var/SYMChdcfs/logs
export HADOOP_PID_DIR=/var/SYMChdcfs/pid
```

Undefine the environment variables, if they are already defined in this file.

Create a VCS service group to provide HA service for the JobTracker.

This step should be performed on only one node in the cluster. Create a VCS service group to tell VCS how to start the JobTracker process and monitor it. You need an IP address that the JobTracker process can use to bind to. When the JobTracker process fails at a node because of a process crash or a node crash, this IP address is moved to a new node.

```
/opt/VRTSvcs/bin/haconf -makerw
/opt/VRTSvcs/bin/hagrp -add Hadoop_JT_SG
/opt/VRTSvcs/bin/hagrp -modify Hadoop_JT_SG SystemList
<HOST1> 0 <HOST2> 1 <HOST3> 2 <HOST4> 3

/opt/VRTSvcs/bin/hares -add Hadoop_JobTracker Application Hadoop_JT_SG
/opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker StartProgram
/usr/lib/hadoop/bin/start-jobtracker.sh
/opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker StopProgram
/usr/lib/hadoop/bin/stop-jobtracker.sh
/opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker PidFiles
/var/SYMChdcfs/pid/hadoop-mapred-jobtracker.pid
/opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker User mapred
/opt/VRTSvcs/bin/hares -override Hadoop_JobTracker RestartLimit
/opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker RestartLimit 1
/opt/VRTSvcs/bin/hares -override Hadoop_JobTracker MonitorInterval
/opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker MonitorInterval 10
/opt/VRTSvcs/bin/hares -modify Hadoop_JobTracker Enabled 1
/opt/VRTSvcs/bin/hagrp -add Hadoop_TT_SG
/opt/VRTSvcs/bin/hagrp -modify Hadoop_TT_SG Parallel 1
/opt/VRTSvcs/bin/hagrp -modify Hadoop_TT_SG AutoFailOver 0
/opt/VRTSvcs/bin/hagrp -modify Hadoop_TT_SG SystemList
<HOST1> 0 <HOST2> 1 <HOST3> 2 <HOST4> 3
```

```
/opt/VRTSvcs/bin/hares -add Hadoop_TaskTracker Application Hadoop_TT_SG
/opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker StartProgram
/usr/lib/hadoop/bin/start-tasktracker.sh
/opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker StopProgram
/usr/lib/hadoop/bin/stop-tasktracker.sh
/opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker PidFiles
```
/var/SYMChdcfs/pid/hadoop-mapred-tasktracker.pid
/opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker User mapred
/opt/VRTSvcs/bin/hares -override Hadoop_TaskTracker RestartLimit
/opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker RestartLimit 1
/opt/VRTSvcs/bin/hares -override Hadoop_TaskTracker MonitorInterval
/opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker MonitorInterval 10
/opt/VRTSvcs/bin/hares -modify Hadoop_TaskTracker Enabled 1
/opt/VRTSvcs/bin/hares -add Hadoop_JT_NIC NIC Hadoop_JT_SG
/opt/VRTSvcs/bin/hares -modify Hadoop_JT_NIC Device <NIC>
/opt/VRTSvcs/bin/hares -modify Hadoop_JT_NIC Enabled 1
/opt/VRTSvcs/bin/hares -add Hadoop_JT_VIP IP Hadoop_JT_SG
/opt/VRTSvcs/bin/hares -modify Hadoop_JT_VIP Device <NIC>
/opt/VRTSvcs/bin/hares -modify Hadoop_JT_VIP Address <JTVIP>
/opt/VRTSvcs/bin/hares -modify Hadoop_JT_VIP NetMask <JTVIP_NMASK>
/opt/VRTSvcs/bin/hares -modify Hadoop_JT_VIP Enabled 1
/opt/VRTSvcs/bin/hagrp -link Hadoop_JT_VIP Hadoop_JT <NIC>
/opt/VRTSvcs/bin/hagrp -link Hadoop_JobTracker Hadoop_JT_VIP
/opt/VRTSvcs/bin/hagrp -link Hadoop_JT_SG <CFS_SG_NAME> online local firm
/opt/VRTSvcs/bin/hagrp -link Hadoop_TT_SG Hadoop_JT_SG online global soft
/opt/VRTSvcs/bin/haconf -dump -makero

where

- <HOST1>, <HOST2> ... : Host names to be configured as Hadoop cluster.
- <JTVIP> <JTVIP_NMASK>: Addresses for the JobTracker and its corresponding netmask.
- <NIC>: Network interface on which you would like JTVIP address to be configured (for example, eth0).
- <CFS_SG_NAME>: Service group name associated with this CFS mount point. This can be created using cfsmntadm command, if it is not already done.
  Refer to the cfsmntadm man page.
- <HADOOP_HOME>: Directory where the HDP Hadoop is installed.
  Typically, this is /usr/lib/hadoop.
9 Start the JobTracker and TaskTracker daemons.

Enter the following command:

```
/usr/lib/hadoop/bin/start-mapred-symc.sh
```

This brings up the Hadoop_JT_SG and Hadoop_TT_SG service groups online. You can verify this by checking the status as follows:

```
/opt/VRTS/bin/hagrp -state
```

The output of this command should show that the service group Hadoop_JT_SG is **online** on one of the nodes in the cluster and **offline** on all other nodes. It should also show that the service group Hadoop_TT_SG is **online** on all nodes in the cluster.

Now, the cluster is ready to run map-reduce applications on HDP using SF-CFS as the underlying file system and providing HA functionality to the JobTracker through VCS.
Running HortonWorks HDP v1.0 with SFHDCFS
SFHDCFS (continued)
  stopping 27

T
  troubleshooting tips
    general tips 31
    Job Tracker dies 36
    Job Tracker fails 34
    map-reduce daemons
      starting 39
      stopping 39
    modify SFHDCFS cluster 39
    start-mapred-symc.sh script fails 32
    submit jobs as non Hadoop user 38
    Task Tracker dies 37
    unable to access CFS files 37

U
  uninstallation
    using script-based installer 23