



Use case 1 - Backing up Hadoop data

NetApp Solutions

NetApp
October 20, 2023

This PDF was generated from <https://docs.netapp.com/us-en/netapp-solutions/data-analytics/hdcs-sh-use-case-1--backing-up-hadoop-data.html> on October 20, 2023. Always check docs.netapp.com for the latest.

Table of Contents

- Use case 1: Backing up Hadoop data 1
 - Scenario 1
 - Requirements and challenges 1
 - Customer’s existing backup solution 1
 - Backup solutions 2

Use case 1: Backing up Hadoop data

[Previous: Overview of Hadoop data protection use cases.](#)

Scenario

In this scenario, the customer has a large on-premises Hadoop repository and wants to back it up for disaster recovery purposes. However, the customer's current backup solution is costly and is suffering from a long backup window of more than 24 hours.

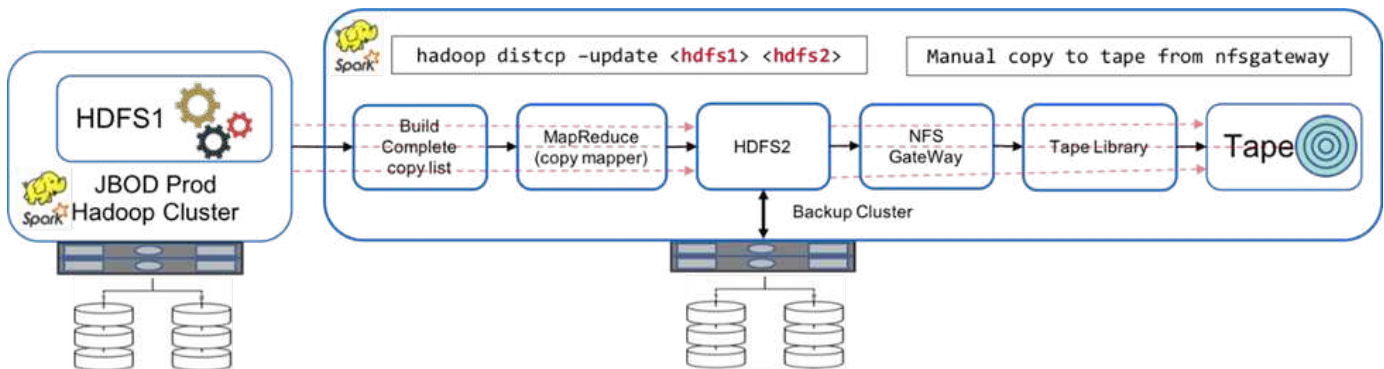
Requirements and challenges

The main requirements and challenges for this use case include:

- Software backward compatibility:
 - The proposed alternative backup solution should be compatible with the current running software versions used in the production Hadoop cluster.
- To meet the committed SLAs, the proposed alternative solution should achieve very low RPOs and RTOs.
- The backup created by the NetApp backup solution can be used in the Hadoop cluster built locally in the data center as well as the Hadoop cluster running in the disaster recovery location at the remote site.
- The proposed solution must be cost effective.
- The proposed solution must reduce the performance effect on the currently running, in-production analytics jobs during the backup times.

Customer's existing backup solution

The figure below shows the original Hadoop native backup solution.



The production data is protected to tape through the intermediate backup cluster:

- HDFS1 data is copied to HDFS2 by running the `hadoop distcp -update <hdfs1> <hdfs2>` command.
- The backup cluster acts as an NFS gateway, and the data is manually copied to tape through the Linux `cp` command through the tape library.

The benefits of the original Hadoop native backup solution include:

- The solution is based on Hadoop native commands, which saves the user from having to learn new procedures.
- The solution leverages industry-standard architecture and hardware.

The disadvantages of the original Hadoop native backup solution include:

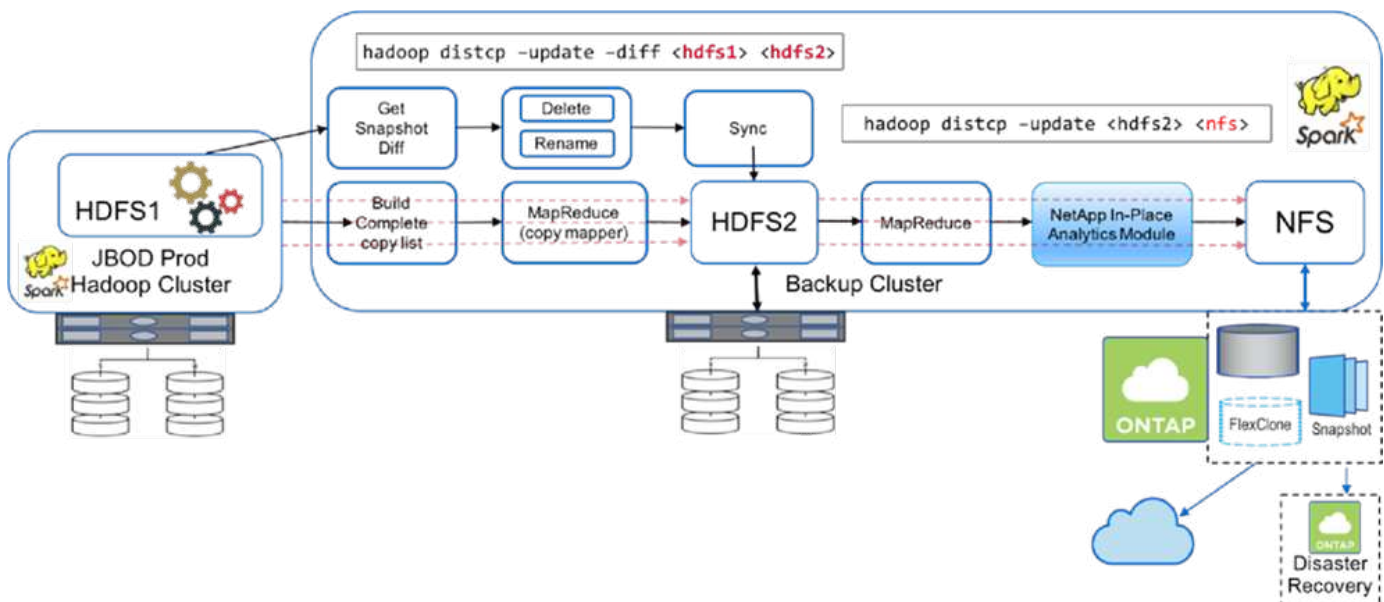
- The long backup window time exceeds 24 hours, which makes the production data vulnerable.
- Significant cluster performance degradation during backup times.
- Copying to tape is a manual process.
- The backup solution is expensive in terms of the hardware required and the human hours required for manual processes.

Backup solutions

Based on these challenges and requirements, and taking into consideration the existing backup system, three possible backup solutions were suggested. The following subsections describe each of these three different backup solutions, labeled solution A through solution C.

Solution A

Solution A adds the In-Place Analytics Module to the backup Hadoop cluster, which allows secondary backups to NetApp NFS storage systems, eliminating the tape requirement, as shown in the figure below.



The detailed tasks for solution A include:

- The production Hadoop cluster has the customer's analytics data in the HDFS that requires protection.
- The backup Hadoop cluster with HDFS acts as an intermediate location for the data. Just a bunch of disks (JBOD) provides the storage for HDFS in both the production and backup Hadoop clusters.
- Protect the Hadoop production data is protected from the production cluster HDFS to the backup cluster HDFS by running the `Hadoop distcp -update -diff <hdfs1> <hdfs2>` command.



The Hadoop snapshot is used to protect the data from production to the backup Hadoop cluster.

- The NetApp ONTAP storage controller provides an NFS exported volume, which is provisioned to the backup Hadoop cluster.
- By running the `Hadoop distcp` command leveraging MapReduce and multiple mappers, the analytics data is protected from the backup Hadoop cluster to NFS by using the In-Place Analytics Module.

After the data is stored in NFS on the NetApp storage system, NetApp Snapshot, SnapRestore, and FlexClone technologies are used to back up, restore, and duplicate the Hadoop data as needed.



Hadoop data can be protected to the cloud as well as disaster recovery locations by using SnapMirror technology.

The benefits of solution A include:

- Hadoop production data is protected from the backup cluster.
- HDFS data is protected through NFS enabling protection to cloud and disaster recovery locations.
- Improves performance by offloading backup operations to the backup cluster.
- Eliminates manual tape operations
- Allows for enterprise management functions through NetApp tools.
- Requires minimal changes to the existing environment.
- Is a cost-effective solution.

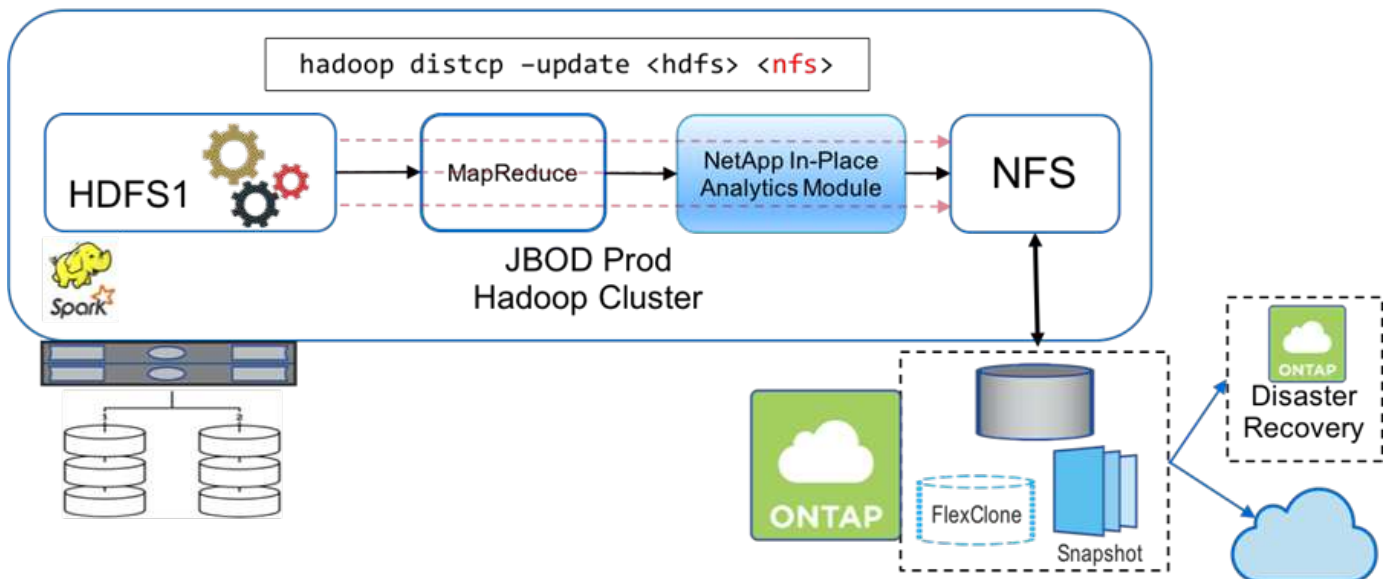
The disadvantage of this solution is that it requires a backup cluster and additional mappers to improve performance.

The customer recently deployed solution A due to its simplicity, cost, and overall performance.

In this solution, SAN disks from ONTAP can be used instead of JBOD. This option offloads the backup cluster storage load to ONTAP; however, the downside is that SAN fabric switches are required.

Solution B

Solution B adds the In-Place Analytics Module to the production Hadoop cluster, which eliminates the need for the backup Hadoop cluster, as shown in the figure below.



The detailed tasks for solution B include:

- The NetApp ONTAP storage controller provisions the NFS export to the production Hadoop cluster.

The Hadoop native `hadoop distcp` command protects the Hadoop data from the production cluster HDFS to NFS through the In-Place Analytics Module.

- After the data is stored in NFS on the NetApp storage system, Snapshot, SnapRestore, and FlexClone technologies are used to back up, restore, and duplicate the Hadoop data as needed.

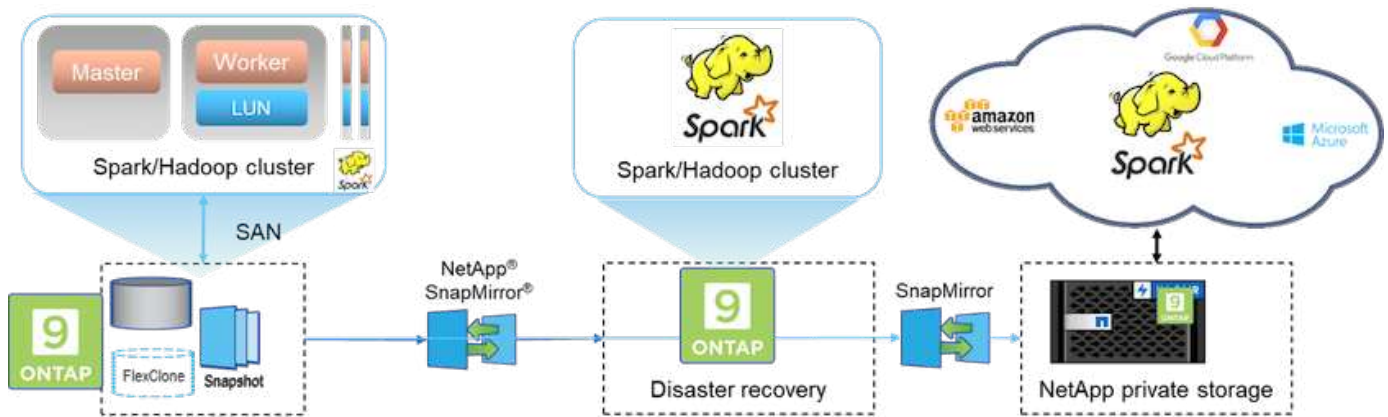
The benefits of solution B include:

- The production cluster is slightly modified for the backup solution, which simplifies implementation and reduces additional infrastructure cost.
- A backup cluster for the backup operation is not required.
- HDFS production data is protected in the conversion to NFS data.
- The solution allows for enterprise management functions through NetApp tools.

The disadvantage of this solution is that it's implemented in the production cluster, which can add additional administrator tasks in the production cluster.

Solution C

In solution C, the NetApp SAN volumes are directly provisioned to the Hadoop production cluster for HDFS storage, as shown in the figure below.



The detailed steps for solution C include:

- NetApp ONTAP SAN storage is provisioned at the production Hadoop cluster for HDFS data storage.
- NetApp Snapshot and SnapMirror technologies are used to back up the HDFS data from the production Hadoop cluster.
- There is no performance effect to production for the Hadoop/Spark cluster during the Snapshot copy backup process because the backup is at the storage layer.



Snapshot technology provides backups that complete in seconds regardless of the size of the data.

The benefits of solution C include:

- Space-efficient backup can be created by using Snapshot technology.
- Allows for enterprise management functions through NetApp tools.

Next: [Use case 2 - Backup and disaster recovery from the cloud to on-premises.](#)

Copyright information

Copyright © 2023 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at <http://www.netapp.com/TM> are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.