Technical Report

Big Data Analytics Data to Artificial Intelligence

Data Mover for Artificial Intelligence

Karthikeyan Nagalingham, NetApp

November 2018 | TR-4732

Abstract

This document describes how to move Big Data Analytics data to Artificial Intelligence (AI). AI processes NFS data through NFS exports. A customer has their AI data in Big Data Analytics platform specifically in Hadoop Distributed File System (HDFS), Blob, and S3 storage. We move the Big Data Analytics Platform data through Hadoop and NetApp® In-Place Analytics Module (NIPAM). This document also provides the business benefits of moving data from Big Data to AI.

NetApp®

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

This document describes how to move the Big Data Analytics data to AI. AI process the NFS data via nfs exports, the customer has their AI data in Big Data Analytics platform specifically in HDFS, Blob, and S3 storage. This paper provides guidelines to move the Big Data Analytics Platform data through Hadoop and NIPAM. We will also provide the business benefits of moving data from Big Data to AI.

2 Concepts and Components

2.1 Big Data Analytics Storage

Big Data Analytics is the major storage provider for HDFS. Sometimes a customer uses the Hadoop-compatible File System (HCFS) such as Windows Azure Blob Storage, MapR-FS, and S3 object storage.

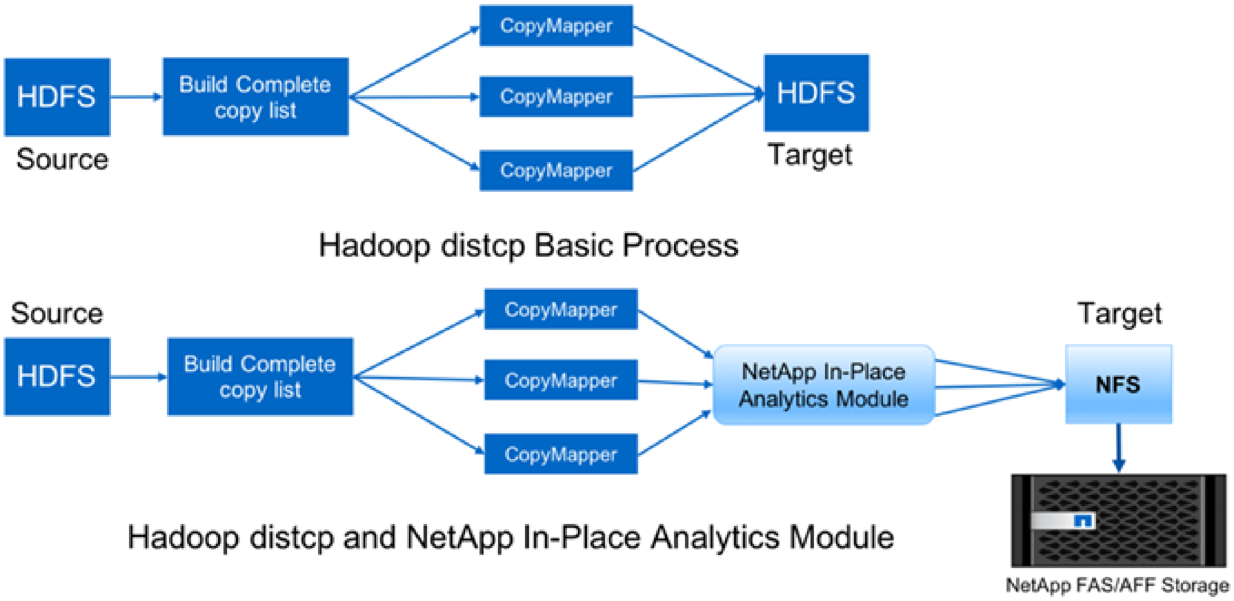
2.2 NetApp In-Place Analytics Module

NIPAM act as a driver for the Hadoop cluster to access NFS data. It has four components: connection pool, NFS InputStream, file handle cache, and NFS OutputStream. For more information, see TR-4382: NetApp In-Place Analytics Module.

2.3 Hadoop Distributed Copy

Hadoop Distributed Copy (DistCp) is a distributed copy tool used for large inter- and intra-cluster coping tasks. This tool uses MapReduce to distribute data, error handling, and reporting. It expands the list of files and directories and inputs them to map tasks to copy the data from the source list. Figure 1 shows the DistCp operation in HDFS and non-HDFS.

Figure 1) DistCp operation in HDFS and non-HDFS (such as NFS).



Hadoop DistCp moves data between the two HDFS systems without using an additional driver. NetApp provides the driver for non-HDFS systems. For an NFS destination, NIPAM provides the driver to copy data that Hadoop DistCp uses to communicate with NFS destination when copy the data.

2.4 NetApp Cloud Volume Service

NetApp Cloud Volume Service (CVS) is a cloud native file service with extreme performance. This service helps customers accelerate their time-to-market business by instantly spinning up and down the resources, leveraging NetApp features to improve productivity and reducing staff downtime. CVS is the right alternative for disaster recovery and back up to cloud by reducing the overall data center footprint and consuming less native public cloud storage.

2.5 NetApp XCP Migration Tool

NetApp XCP Migration Tool is a client software that enables fast and reliable any-to-NetApp and NetApp-to-NetApp data migration. This tool is designed to copy a large amount of unstructured NAS data from any NAS system to a NetApp storage controller. XCP Migration Tool leverages a multicore, multichannel I/O streaming engine that can process many requests in parallel, such as data migration, file or directory listings, and space reportings.

2.6 Cloud Sync

Cloud sync is a hybrid data replication software-as-a-service that transfers and synchronizes NFS, S3, and CIFS data seamlessly and securely between on-the-premises or cloud storage. This software is used for data migration, archiving, collaboration, analytics, and more. After data is transferred, cloud sync continuously syncs the data between the source and destination. Going forward, it will transfer the delta. It also secures the data with your own network in the cloud or on the premises. This software is based on the pay-as-you-go model, which provides a cost-effective solution and provides monitoring and reporting capabilities for your data transfer.

3 Customer Challenges

Customers might face the following challenges when trying to access data from Big Data Analytics for AI operations:

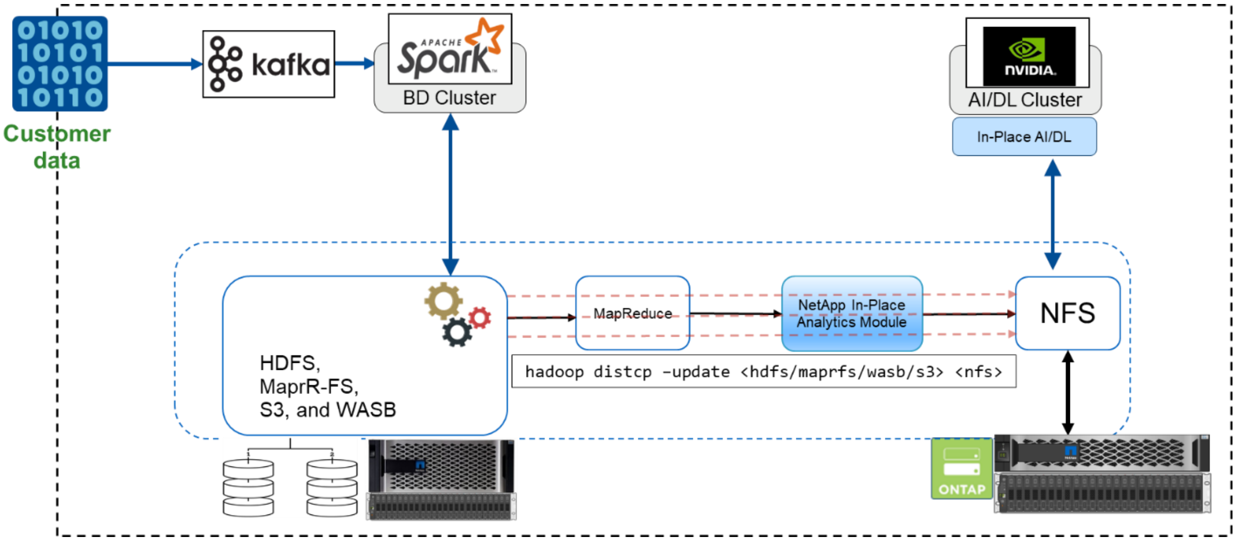
- The customer data is in a data lake repository. The data lake can contain different types of data such as structured, unstructured, semi-structured, logs, and machine-to-machine data. All these types of data are required to process in AI.
- AI is not compatible with Hadoop file systems. AI architecture is not able to directly access the HDFS and HCFS data, which must be moved to an AI understandable file system (NFS).
- Moving data lake data to AI. The amount of data in the data lake is very large. A customer must have an efficient, high-throughput, and cost-effective way to move data to AI.
- Syncing data. If a customer wants to sync data between the Big Data platform and AI, sometimes the data processed through AI can be used with Big Data for analytical processing.

4 Data Mover

In a Big Data cluster, data storage in HDFS or HCFS, such as MapR-FS, Windows Azure Storage Blob (WASB), S3, and Google file system. NetApp performed testing with HDFS, MapR-FS, and S3 as the source to copy data to NetApp ONTAP® NFS export with the help of NIPAMe by using the `hadoop distcp` command from source.

Figure 2 illustrates the typical data movement from a Spark cluster running with HDFS storage to a NetApp ONTAP NFS volume so that NVIDIA can process AI operations.

Figure 2) Hadoop DistCp basic workflow—data mover.



```
hadoop distcp -update <hdfs/maprfs/wasb/s3> <nfs>
```

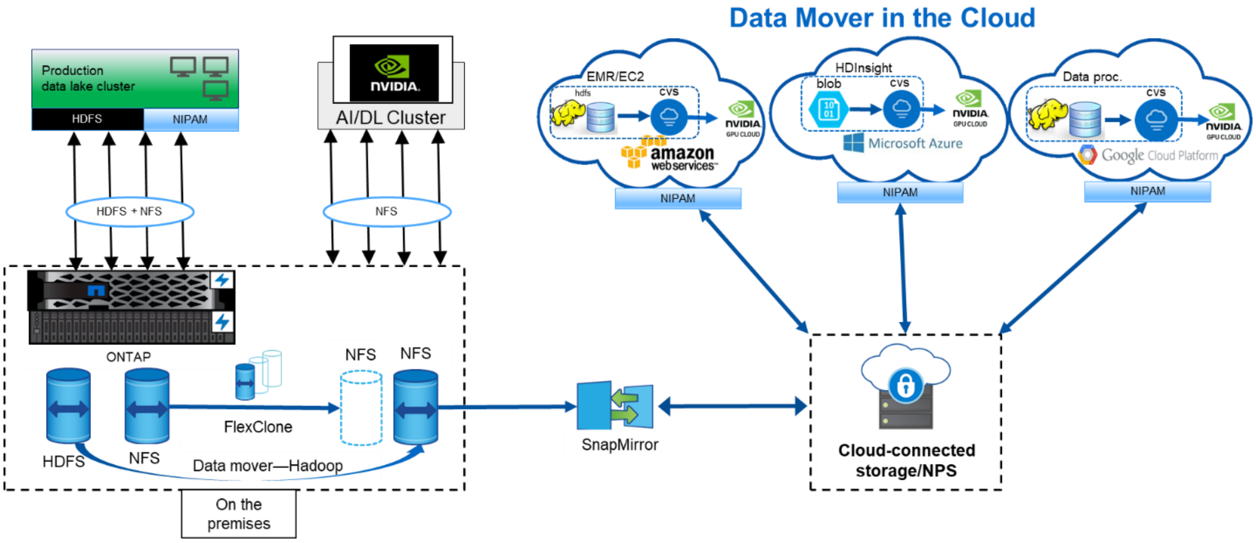
The hadoop distcp command uses the MapReduce program to copy the data. NIPAM works with MapReduce to act as a driver for the Hadoop cluster when copying data. NIPAM can distribute a load across multiple network interfaces for a single export. This process maximizes the network throughput by distributing the data across multiple network interfaces when you copy the data from HDFS or HCFS to NFS.

5 Data Mover Solution for AI

The data mover solution for AI is based on customers' needs to process Hadoop data from AI operations. NetApp moves data from HDFS to NFS by using the NIPAM. In one use case, the customer needed to move data to NFS on the premises and another customer needed to move data from WASB to cloud volume services in order to process the data from the General Processor Unit (GPU) in the cloud.

Figure 3 illustrates the data mover solution details.

Figure 3) Data mover solution for AI on the premises and in the cloud.



The steps required to build the data mover solution include:

1. ONTAP SAN provides HDFS and NAS provides NFS volume through NIPAM to production data lake cluster.
2. The customer has their data in HDFS and NFS. The NFS data can be production data from other application which is used for Big Data Analytics and AI.
3. The NetApp FlexClone® technology creates the clone of the production NFS volume and provisioned to AI cluster on the premises.
4. Data from HDFS SAN LUN is copied into NFS volume by using NIPAM and the `hadoop distcp` command. The NIPAM leverages the bandwidth of multiple network interfaces to transfer data. This process reduces the data copy time and more data can be transferred.
5. Both NFS volumes are provisioned to the AI cluster for AI operations.
6. To process the on-the-premises NFS data with GPU in cloud, the NFS volumes are mirrored to NetApp Private Storage (NPS) by using NetApp SnapMirror® technology and mounted to cloud service providers for GPUs.
7. The customer wants to process data in EC2/EMR or HDInsight or DataProc services in GPUs such as NVIDIA from cloud service providers. Hadoop data mover moves the data from Hadoop services to cloud volume services by using NIPAM and the `hadoop distcp` command.
8. The CVS data provisioned to AI through NFS protocol.

Data that is processed through AI can be sent to on the premises for Big Data Analytics as well as the NVIDIA cluster through NIPAM, SnapMirror technology, and NPS.

In this scenario, the customer has a large file count data in the NAS system on a remote location that is required for AI processing on the NetApp storage controller on the premises. In this scenario, it's better to use the XCP Migration Tool to migrate the data at a faster speed.

The hybrid use case customer can use cloud sync to migrate on-the-premises data from NFS, CIFS, S3 data to cloud and vice versa for AI processing by using GPUs such as a NVIDIA cluster. Both cloud sync and the XCP Migration Tool are used for the non-HDFS data migration to NFS.

6 Business Benefits

Moving data from Big Data Analytics to AI provides the following benefits:

- Ability to extract data from different Hadoop file systems into a unified NFS storage system
- A Hadoop-integrated and automated way to transfer data
- Reduction in the cost of library development for moving data from Hadoop file systems
- Maximum performance by aggregated throughput of multiple network interfaces from a single source of data by using NIPAM
- Scheduled and on-demand methods to transfer data
- Storage efficiency and enterprise management capability on the unified NFS data by using ONTAP data management software
- Zero cost for data movement because of using Hadoop way of transfer

Where to Find Additional Information

To learn more about the information that is described in this document, review the following documents and/or websites:

- NetApp In-Place Analytics Module Best Practices
  https://www.netapp.com/us/media/tr-4382.pdf

- NetApp FlexGroup Volume Best Practices and Implementation Guide
https://www.netapp.com/us/media/tr-4571.pdf
- NetApp Product Documentation
https://www.netapp.com/us/documentation/index.aspx

Version History

Version	Date	Document Version History
Version 1.0	November 2018	Initial release.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

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.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, 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).

Trademark Information

TR-4732-1118

NetApp®