DELL EMC

PROTECTING BIG DATA

Data Protection Solutions for Hadoop and the Business Data Lake

ABSTRACT

Big data analytics in the enterprise is rapidly maturing, driving transformative business outcomes for their companies. At the same time big data applications often enter production without a robust data protection strategy. Hadoop is the leading big data framework. Enterprises are asking for a comprehensive data protection story for Hadoop, from backup to disaster recovery. This paper discusses how Dell EMC's Data Domain Boost for Enterprise Applications, part of the Dell EMC Data Protection Suite Family, provides industry-first true backup application functionality for Hadoop, offering the Hadoop admin a set of CLI commands to perform their own backup and recovery to Data Domain protection storage. It also touches on Isilon, Networker and Elastic Cloud Storage (ECS) data protection options to backup other data lake components.

May, 2017

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Big data analytics in the enterprise is rapidly maturing, driving transformative business outcomes for companies. It is a matter of time before big data analytics drives business decisions for enterprises, thus becoming the new mission critical application. Today, as big data use cases rapidly mature, they often enter production without a robust data protection strategy. Homegrown backup approaches leveraging snapshots and replication reach their limits in the face of enterprise grade reliability, availability and serviceability expectations that are the norm for other workloads. Hadoop is the leading big data framework, and the lack of true backup and disaster recovery for it is not lost on companies using those solutions to protect other applications in their environment. Enterprises are asking for a comprehensive data protection story for Hadoop, covering backup and disaster recovery.

Dell EMC® has stepped up to the challenge with the DD Boost for Enterprise Applications® which is part of the Dell EMC Data Protection Suite Family®, to deliver a purpose built backup application for Hadoop. Now Hadoop admins can backup and recover their Hadoop data while using native UIs to and from Data Domain® systems, Dell EMCs market leading protection storage.

THE CHALLENGE

Hadoop natively lacks a true point-in-time backup capability. Although it does offer snapshots and replication capabilities, these are not sufficiently resilient to software errors, data corruption or human error. The lack of enterprise ready backup and disaster recovery for Hadoop is a big inhibitor for Hadoop adoption in enterprises. As big data applications become mainstream, the business risk of downtime or data loss becomes significant. Therefore, enterprises want their Hadoop data to be protected to similar SLAs as mainstream IT workloads.

SOLUTION OVERVIEW

Dell EMC is providing an effective data protection strategy to address the challenges associated with Hadoop and other big data environments. This paper discusses DD Boost for Enterprise Applications, which is licensed as part of the Dell EMC Data Protection Suite Family, and several other Dell EMC Business Data Lake protection solution options that include:

- Hadoop systems built with Cloudera and Hortonworks distributions and managed through command line tools and native management UIs like Cloudera Manager and Hortonworks Ambari.
- Supporting Hadoop clusters built using local DAS storage, or shared storage systems such as Dell EMC Isilon®.
- Using Hadoop-native constructs and integration into the Hadoop file system.
- Other Dell EMC products (e.g. NetWorker®, Isilon, Data Domain protection storage, and Elastic Cloud Storage® (ECS)) to backup Hadoop and other big data frameworks.

INTRODUCTION

The purpose of this white paper is to provide background information on why data lake, and in particular Hadoop protection is becoming critically important, and to describe the various Dell EMC protection solutions for it. This paper will help customers achieve higher degrees of business value and operational efficiency with their data lake implementation and big data frameworks.

AUDIENCE

This white paper is intended for IT & Hadoop Administrators, systems engineers, partners and members of the Dell EMC and partner professional services community who are looking to better understand and implement Dell EMC Business Data Lake protection solutions.

BACKGROUND

WHAT IS A DATA LAKE?

In simple terms, a data lake is a single centralized repository, collecting data from a wide range of sources, in turn feeding many analytical applications. Data lakes are comprised of a mix of structured, semi-structured, and unstructured data. Various analytics applications consume data in this “lake”, gaining efficiency through reusability and consistency of data. Data lakes evolved from

Enterprise Data Warehouses (EDW), but unlike EDW, data lakes don't require an upfront schema. This makes them capable of supporting new analytics frameworks like Hadoop, NoSQL databases, etc., which can analyze these new data sources. This flexibility allows customers to easily add and leverage many other data sources, enabling them to make better business decisions based on their data.

Data lakes aggregate a variety of data sources, from traditional enterprise applications, to new sources of semi-structured and unstructured data as illustrated in Figure 1 below.

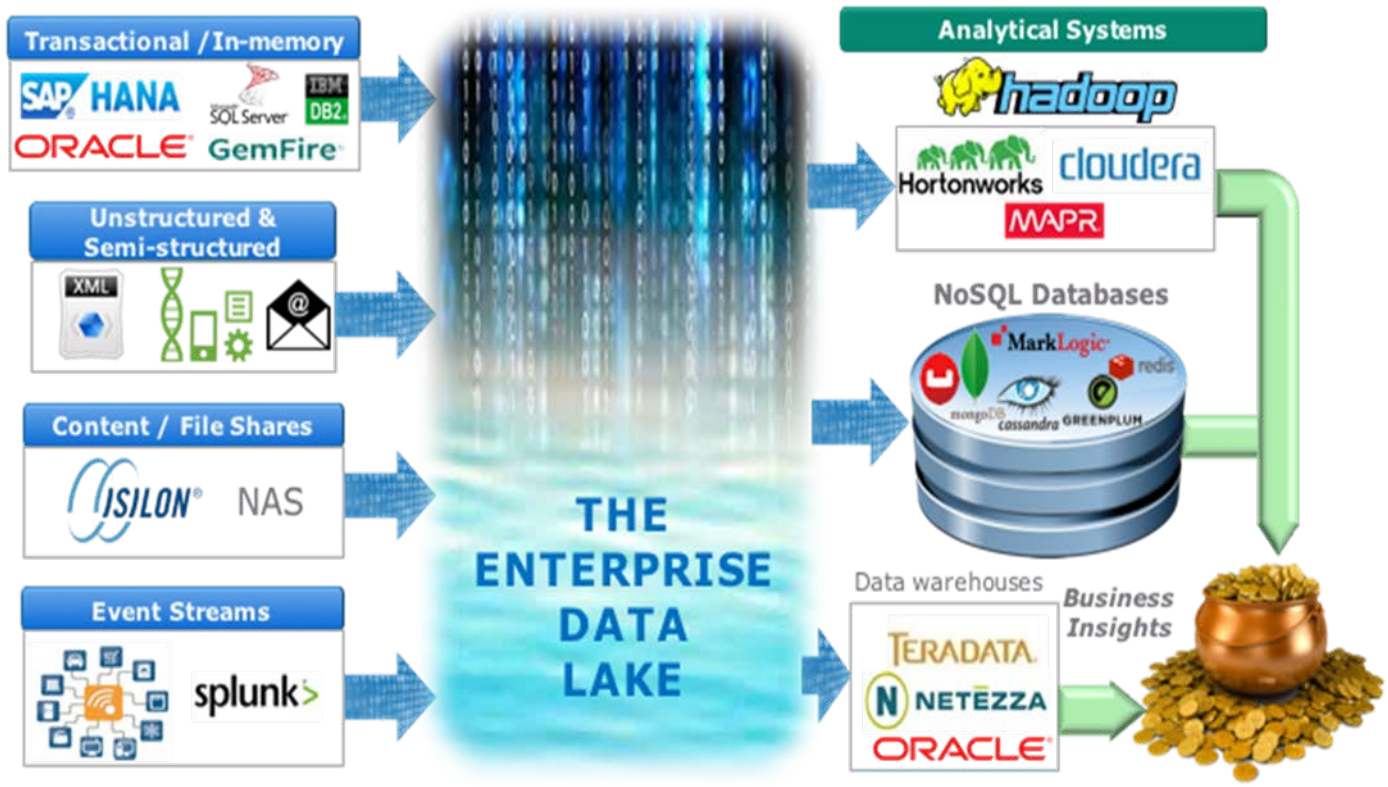


Figure 1: An Enterprise data lake

HADOOP IS THE FOUNDATION OF MOST DATA LAKES

Hadoop is an open source data platform for managing large volumes of data, from a variety of data sources, at scale and with speed. Managed by the Apache Software Foundation, Hadoop initially saw rapid adoption by large web companies. With the emergence of commercially supported distributions from companies like Cloudera and Hortonworks, Hadoop is now undergoing rapid adoption throughout enterprises as well.

Hadoop excels at distributed processing of large data sets, across clusters of commodity servers. It is especially good at processing and analyzing massive amounts of incoming unstructured and semi-structured data, in addition to traditional structured data sources. These qualities have fueled the popularity of Hadoop as an analytics platform. Market studies claim roughly 60% of big data systems involve Hadoop, making it the single most popular big data platform.

Many big data systems also involve databases for semi- and unstructured data, in many cases feeding off of data in a Hadoop system (the data lake), and optionally feeding results back into the data lake. Most data lake implementations therefore revolve around Hadoop.

CLOUDERA ENTERPRISE

Cloudera Enterprise (CDH), illustrated in Figure 2 below, packages Apache Hadoop with a number of other open source projects, and is one of the popular commercial Hadoop distributions used by enterprises. Cluster monitoring, management and operations are performed from the Cloudera Manager UI.

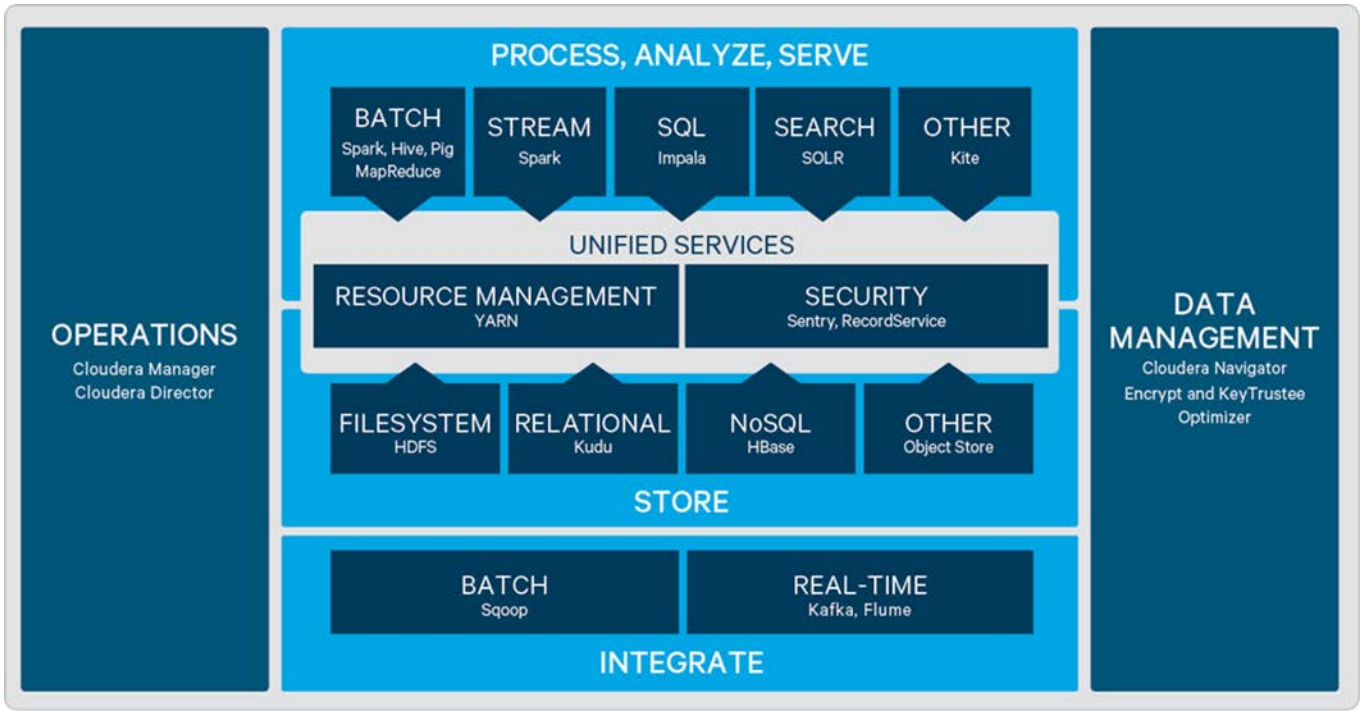


Figure 2: Cloudera Enterprise

HORTONWORKS DATA PLATFORM

Hortonworks analytics platform illustrated in Figure 3 below is based on Apache Hadoop is also popular among enterprises. It packages Apache Hadoop components for a wide range of analytical systems – batch, streaming, and real-time. Monitoring, management and operations are performed from the Hortonworks Ambari UI. Hortonworks also focuses on adding security features to Hadoop.

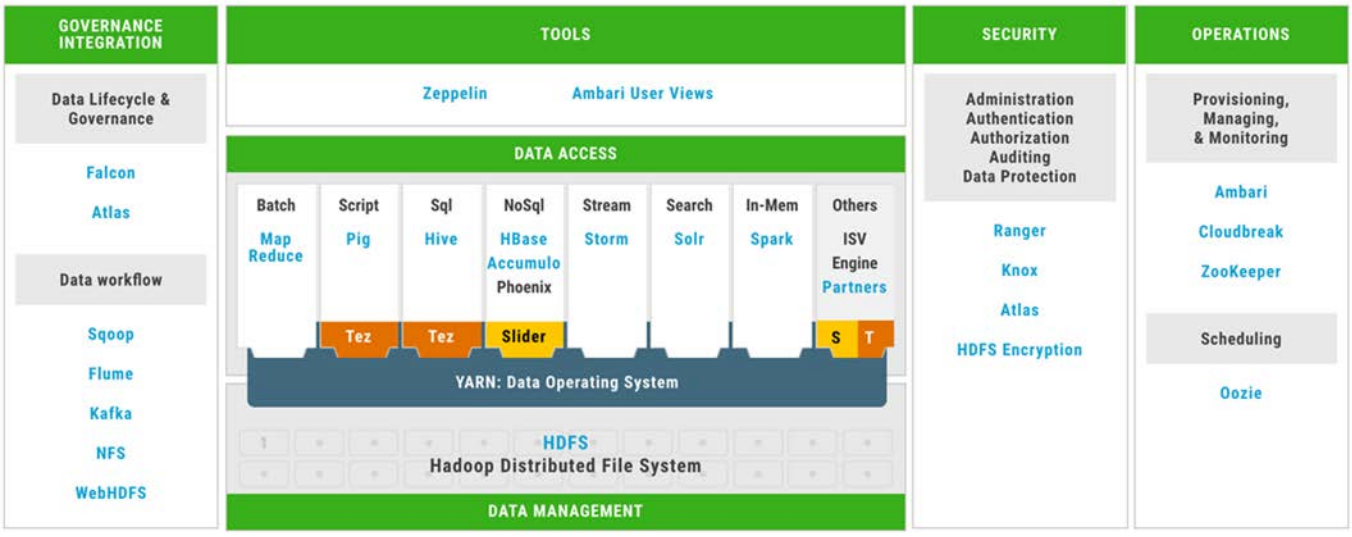


Figure 3: Hortonworks Data Platform (HDP)

COMMON HADOOP DEPLOYMENT MODELS

Regardless of distribution, there are three common ways in which Hadoop can be deployed which are illustrated in Figure 4 below.

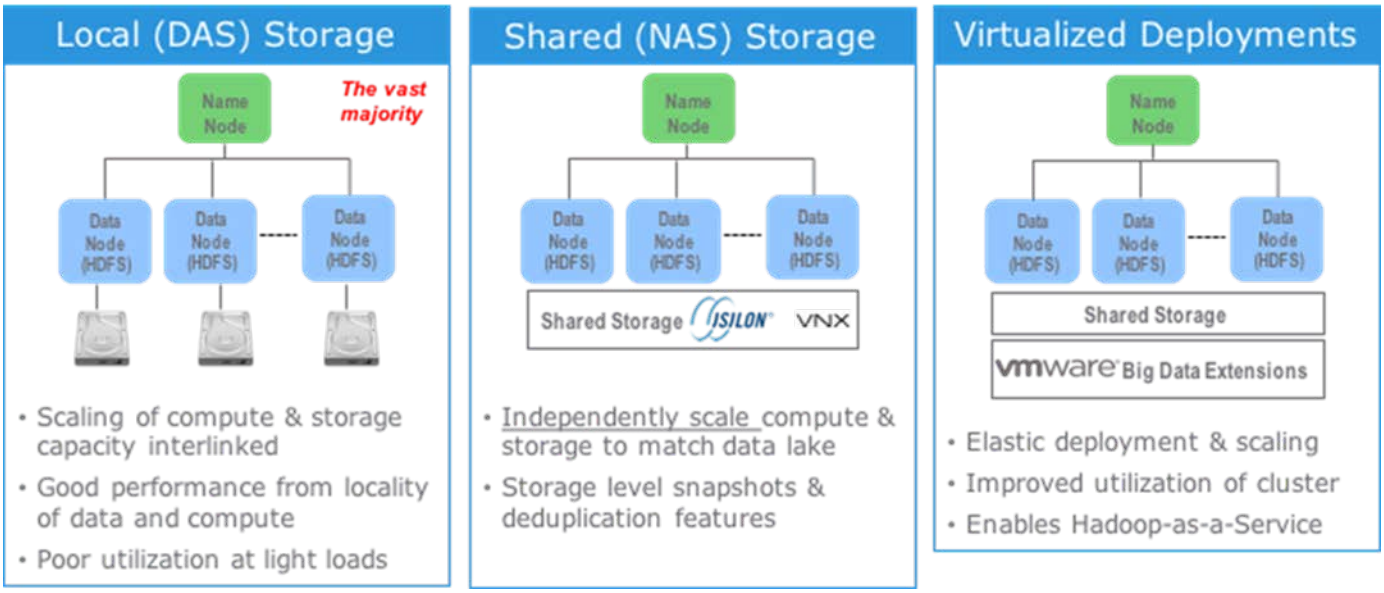


Figure 4: Common Hadoop deployment models

DATA PROTECTION IS THE BARRIER TO ENTERPRISE ADOPTION OF BIG DATA

With big data analytics on course to become the next mission-critical enterprise application, enterprises are demanding a robust level of backup, recovery and disaster recovery solutions for their big data systems, in particular Hadoop. However, Hadoop today natively lacks a true point-in-time backup capability. Although it does offer snapshots and replication capabilities, these are not sufficiently resilient to software errors, data corruption or human error. At small scale, in experimental systems snapshots and replication can suffice as a backup and DR strategy.

When enterprises speak of taking applications into production, they have specific expectations in the reliability, uptime, and recoverability of the application. This is where snapshots and replication as a data protection strategy no longer suffice. The following are some of the pros and cons of these approaches:

	Advantages	**Disadvantages**
Snapshots	• Quick recovery • Protects against human error • Built-into Hadoop	• Storage-hungry – consume costlier primary storage • Tedious to manage at scale • Live on primary storage (same failure domain)
Replication	• Copy on another system • Built-into Hadoop	• Requires like-for-like technologies • Not a defense against software bugs, human error, or data corruption • Not a point-in-time copy

The fundamental objectives of a backup strategy are:

1. To create a true point-in-time copy of the original data on another distinct system
2. So that recovery can be performed back to a known-good point-in-time

As you can see, Hadoop's built-in primitives cannot be called a true backup story. Enterprises using backup products for their other IT applications realize this, and are asking for a true backup capability for Hadoop. With the increasing importance of, and reliance on analytics for business decision making, the cost of downtime or data loss can be significant. Hence the need for a backup and recovery capability for Hadoop.

BACKUP & RECOVERY OF HADOOP WITH DELL EMC DATA PROTECTION SOLUTIONS

UNIQUE CONSIDERATIONS IN BACKING UP HADOOP

There are crucial differences between Hadoop and how traditional enterprise systems are architected:

- Hadoop is architected to be a scale-out system, built on clusters of commodity servers and storage, tolerant of failures in individual components. It is architected to process data in parallel across many servers or "nodes".
- The Hadoop File System (HDFS) is also distributed in nature. Files stored in HDFS are broken into blocks, which are dispersed among the nodes in the system.
- Hadoop clusters are most commonly deployed on server-local disks (also called DAS storage). Shared (NFS) storage systems such as Dell EMC Isilon are another way to deploy Hadoop, facilitated by Isilon's native integration with HDFS.
- HDFS offers high availability by replicating each block across multiple nodes (typically 3 times) for redundancy.
- Hadoop systems are operated and managed by dedicated administrators. Therefore, backup and recovery of Hadoop will most likely be the responsibility of Hadoop admins, not the backup or storage admins.

Therefore, backing up HDFS requires the backup application to be integrated into HDFS and the cluster's management node (also called the name node). Data volumes in HDFS can be large, requiring parallel data transfer to keep backup windows to a reasonable size.

DELL EMC DATA DOMAIN PROTECTION STORAGE HIGH LEVEL OVERVIEW

Dell EMC Data Domain protection storage systems deliver industry-leading speed and efficiency with throughput up to 68 TB/hour enabling more backups to complete sooner and reducing pressure on backup windows. Data Domain systems leverage variable-length deduplication to minimize disk requirements and ensure data lands on disk already deduplicated. This reduces backup and archive storage requirements by an average of 10 to 30x, making disk a cost-effective alternative to tape. Data on disk is available online and onsite for longer retention periods and restores and retrievals become fast and reliable. This efficiency enables Data Domain systems to protect up to 150 PB of logical capacity for backup and archive data on a single system.

Data Domain Boost (DD Boost) is a feature that improves backup performance by up to 50%, reduces bandwidth consumption by up to 99%, improves backup success through automatic link aggregation and path failover, and provides other benefits compared to backing up over NFS.

Data Domain systems are designed as the storage of last resort – built to ensure you can reliably recover your data with confidence. The Data Domain Data Invulnerability Architecture is built into the Data Domain Operating System (DD OS) to provide the industry's best defense against data integrity issues. For additional information on Data Domain systems please refer to the Dell EMC Data Domain Data Sheet, The Business Value of Data Domain Boost, and the Dell EMC Data Domain Data Invulnerability Architecture white paper.

DELL EMC DD BOOST FOR ENTERPRISE APPLICATIONS

DD Boost for Enterprise Applications is available as a component of the Dell EMC Data Protection Suite Family. DD Boost for Enterprise Apps provides true point-in-time backup and recovery of data to Dell EMC Data Domain protection storage via the DD Boost protocol and utilizes application agents when integrating with applications: Microsoft app agent, Database app agent and Hadoop app agent. The Hadoop app agent is used when protecting big data workloads. Leveraging the storage efficiency and reliability of Data Domain systems with the network-efficient DD Boost protocol, DD Boost for Enterprise Apps offers the Hadoop admin a set of CLI commands to perform their own backup and recovery.

The technical highlights of the DD Boost for Enterprise Apps and Data Domain based backup solution for Hadoop environments are:

- True point-in-time backup and recovery of Hadoop data to a Data Domain system.
- Integrated into native management UIs - Cloudera Manager and Hortonworks Ambari .
- HDFS integration transparently works through the 3-way storage redundancy to backup one consistent copy of the data.

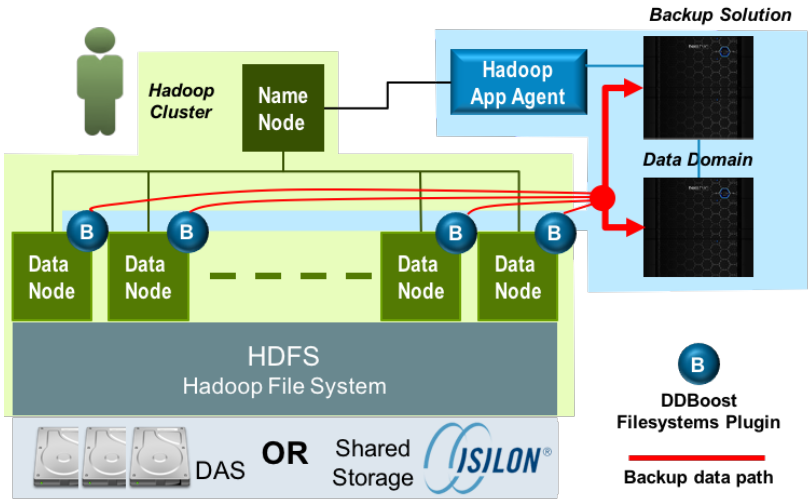
- Uses standard Hadoop constructs (e.g. MapReduce, distcp) to spawn distributed DD Boost agents to parallelize data transfer to a Data Domain system.
- Bandwidth-efficiency of DD Boost, sends only unique data over the network.
- Management and tooling simplicity. No need to deploy or manage individual DD Boost agents. The Hadoop admin performs backup & recovery from the Hadoop cluster management console.
- DD Boost for Enterprise Apps offers a set of Linux commands for backup, recovery, file search, retention etc. Every Hadoop administrator can readily use these commands and incorporate them into other workflows.
- Backup operations can be further scheduled and automated by Oozie.
- Audit log of configuration changes

The following table illustrates the salient points about the DD Boost for Enterprise Applications for Hadoop data protection:

Backup sources	HDFS directories and HBase tables Cloudera Manager and Hortonworks Ambari Backup policies can be associated between these sources and a target Data Domain system.
Backup Target(s)	One or more Data Domain systems, with DD Boost licenses
Storage configurations supported	Server-local direct-attached storage (DAS), and Shared (NAS) storage systems (e.g. Dell EMC Isilon)
User Interface	Linux command-line application
Distributions supported	Cloudera Enterprise 5.4 and above Hortonworks Data Platform 2.2 and above
Scheduling & Automation	None. Scheduling can be performed using Oozie or cron

DD Boost for Enterprise Apps requires minimal configuration and installs only on the Name Node of the Hadoop Cluster. It is tightly integrated into the Hadoop file system, and leverages Hadoop's scale-out distributed processing architecture to parallelize data transfer from Hadoop to a Data Domain system. DD Boost provides a network-efficient data transfer with client-side deduplication, and Data Domain provides storage-efficiency through deduplication and compression. Together this makes it the most efficient method of moving large amounts of data from a Hadoop cluster to a Data Domain target system. Internally standard Hadoop constructs like Distributed File Copy & HDFS/HBase snapshots are leveraged to accomplish tasks.

Figure 5 below shows a Hadoop cluster with DD Boost for Enterprise Apps being deployed.



OPERATIONALIZING HADOOP BACKUPS WITH DD BOOST FOR ENTEPRISE APPS

The High Level provisioning and setup workflow:

- Install the application on the Name Node of the Hadoop cluster.
- Provision one or more Data Domain systems (the backup targets). Existing Data Domain systems backing up other workloads can also receive backups from DD Boost for Enterprise Apps, provided they are running DD OS 6.0 and up.
- The target Data Domain systems must have one or more storage units provisioned, to receive backup streams from DD Boost for Enterprise Apps.
- Kerberos authentication (if enabled in the Hadoop cluster) should be enabled at this time.

The High Level backup & restore operations workflows:

- An HDFS directory or HBase table to be backed up is associated with a target Data Domain system, and storage unit where the backups will be stored. This is the backup provisioning step.
- Retention time characteristics are also specified when backups are provisioned.
- Optionally, you can also specify a secondary (or remote) Data Domain system that will receive replicated backups from the primary backup Data Domain system.
- Once provisioned, the backup command will backup the selected HDFS directory or HBase table to the provisioned Data Domain system and storage unit. The backup process uses HDFS snapshots in the course of its work, so ensure snapshots are enabled for the directories or tables to be backed up.
- When a restore is needed, the backup catalog on the target Data Domain system can be listed with the backup listing commands to select the restore point. The restore command is used to restore the HDFS directory or HBase table of interest back into HDFS.
- Due to the NDFS integration, backups are actually run as DistCp jobs from HDFS to the Data Domain system, and restore jobs are DistCp jobs in the reverse direction. The important distinction is that this process is transparently run in the background without the user having to manage any part of this process.
- Backups and restores internally leverage DD Boost, with its unique benefits of de-duplication, dynamic interface group, and TLS encryption.

The command set for Hadoop protection includes:

Data Domain Configuration	• Add/Remove Data Domain system • Browse configuration • Test Data Domain system connectivity
Backup Provisioning	• Associate a source HDFS directory or HBase table with a Data Domain target • Set retention time • Set secondary (offsite) Data Domain target system
Backup	• Backup HDFS directory / HBase table • Backup master configuration
Restore	• Restore HDFS directory /HBase table • Restore one subdir from backup • Restore master configuration
File Search	• Search backup for subdirectory/file

	• Search backup using regular expression
Retention Management	• Change absolute retention • Change relative retention
Expiration	• Expiration of old backups
Backup Listing	• List single backup • List backups by [range / date filter] • List configuration
Deletion	• Delete single backup • Delete backups [range/date filter]
Kerberos	• Kerberos authentication enable / disable

BENEFITS OF DD BOOST FOR ENTERPRISE APPS FOR HADOOP

DD Boost for Enterprise Applications provides a true backup and recovery solution for Hadoop data. Backup and recovery is managed by Hadoop Administrators from the clusters management tools, using Hadoop-native tools. This is an application-centric backup workflow, and is therefore an easy learning curve. The overall experience is that of using a backup application, instead of managing snapshots or managing replication through custom scripting.

The deep integration with HDFS allows standard Hadoop constructs and tools like MapReduce, Yarn and distcp to be used to backup and recover data to/from Data Domain.

DD Boost provides a network-efficient data transfer with client-side deduplication, and Data Domain provides storage-efficiency through deduplication and compression. Together this makes it the most efficient method of moving large amounts of data from a Hadoop cluster to a Data Domain target system. On the other hand, the user is not responsible for managing either Boost agents or NFS mounts individually, improving usability. Managing agents individually per node is not a scalable strategy for scale-out systems such as Hadoop.

OTHER DELL EMC DATA LAKE PROTECTION SOLUTIONS

OTHER DELL EMC DATA LAKE PROTECTION SOLUTIONS OVERVIEW

Dell EMC offers several other solution options for Business Data Lake protection: Isilon snapshots managed by Dell EMC NetWorker Snapshot Management for shared storage deployments, and Dell EMC Elastic Cloud Storage. These solutions are illustrated in Figure 6 below and explained in more detail throughout the rest of this paper.

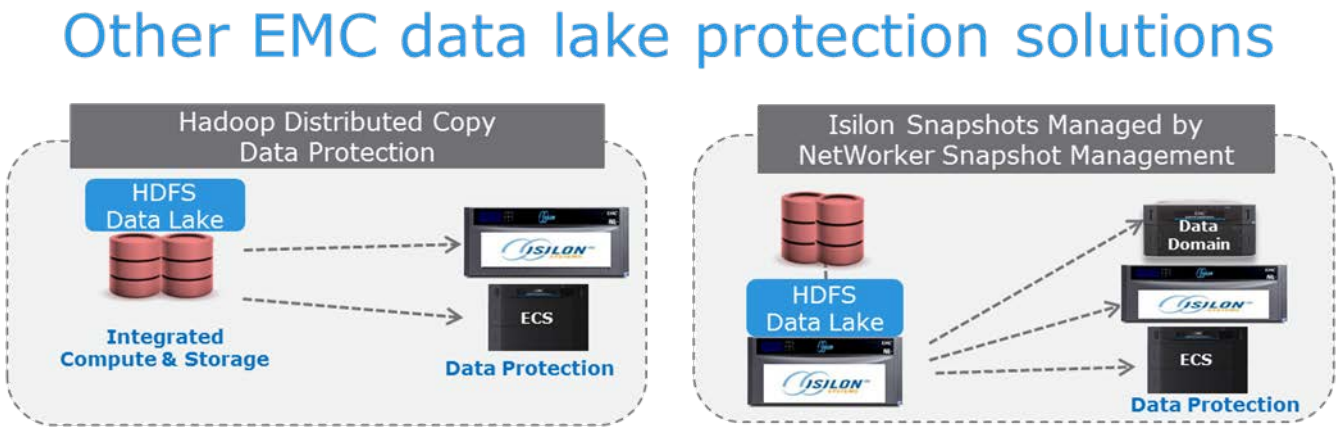


Figure 6: Other Dell EMC Business Data Lake protection solutions

DELL EMC TARGET STORAGE OPTIONS

As further described in the following paragraphs, the Dell EMC Business Data Lake protection solutions illustrated in Figure 6 can leverage Dell EMC Data Domain, Dell EMC Isilon, or Dell EMC Elastic Cloud Storage (ECS) as target storage depending on a number of factors including, accessibility, storage efficiency, and capacity needs. Data Domain systems are ideal for workloads that deduplicate well (databases, files, etc.) and provide storage savings through industry leading variable-length deduplication and compression. Isilon is a good fit for data sets that don't deduplicate well (video, voice, etc.) and provides efficient, cost-effective storage from a single system. ECS is a good fit for object workloads at Cloud (Exabytes) scale.

DELL EMC ISILON SCALE-OUT NAS STORAGE HIGH LEVEL OVERVIEW

Dell EMC Isilon scale-out storage solutions are designed for enterprises that want to manage their data, not their storage. Isilon storage systems are powerful yet simple to install, manage, and scale to virtually any size. And, unlike traditional enterprise storage, Isilon solutions stay simple no matter how much storage capacity is added, how much performance is required, or how business needs change in the future. Isilon challenges enterprises to think differently about their storage, because when they do, they'll recognize there's a better, simpler way – with Dell EMC Isilon.

Through the winning combination of the groundbreaking Isilon OneFS operating system, high-performance industry-standard hardware, and powerful data and storage management software, Isilon provides a complete portfolio of innovative storage solutions that drive business value for customers by optimizing mission-critical applications, workflows, and processes. Isilon storage enables enterprise and research organizations worldwide to manage large and rapidly growing amounts of data in a highly scalable, easy-to-manage, and cost effective way. Every Isilon solution is designed to accelerate workflow productivity and reduce capital and operational expenditures, while seamlessly scaling storage in lockstep with the growth of mission-critical data. For additional information on Isilon please refer to the Dell EMC Isilon Data Sheet.

DELL EMC ELASTIC CLOUD STORAGE (ECS) HIGH LEVEL OVERVIEW

Customers are continually looking for more efficient architectures to manage today's hyperscale growth. Powered by Dell EMC ViPR®, the new Elastic Cloud Storage (ECSTM) Appliance provides a complete hyperscale storage infrastructure designed to meet the requirements of modern applications. Regardless of the size of your organization, the ECS Appliance lets you deliver competitive cloud storage services and grow effortlessly. The ECS Appliance brings the cost profile, simplicity and scale of public cloud services to anyone – with the trust, reliability and support you expect from Dell EMC. The ECS Appliance helps:

- Data Scientists accelerate big data initiatives
- Cloud Providers deliver competitive Cloud Storage services at scale
- Enterprises and software developers to accelerate development

The ECS Appliance makes Hyperscale storage and cloud economics viable for any size business by combining the power of ViPR on a low-cost, high density, scale out commodity hardware platform. The ECS Appliance is available in multiple form factors that can be deployed and expanded incrementally, so each customer can choose the right size for their immediate needs and projected growth. Customers can now optimize their solution based on their application and access needs – giving them the flexibility and control they want. For additional information on Elastic Cloud Storage please refer to the Dell EMC ECS Data Sheet.

HADOOP DISTRIBUTED COPY DATA PROTECTION TO ECS

This section provides more detail about leveraging the native Distributed Copy (DistCp) utility built into HDFS (Hadoop File System) to backup & restore data from an integrated compute & storage data lake to an on premise Elastic Cloud Storage Appliance.

The choice of using ECS as the target storage for this solution will typically be made by customers based on a consideration of 3 primary factors:

1. Do you already know that your data would not gain significant storage savings from the variable-length deduplication & compression that Data Domain systems would provide?
2. Do you require the hyperscale that ECS provides? (Exabytes)
3. Do you require Object/HDFS accessibility?

DistCp (distributed copy) is a standard tool that comes with all Hadoop distributions and versions that can be used to copy entire Hadoop directories. DistCp runs as a MapReduce job to perform file copies in parallel, fully utilizing your systems if desired. There is also an option to limit the bandwidth to control the impact on other tasks.

This solution can be used in 2 different ways.

1. One approach takes an HDFS snapshot from the Hadoop application and then moves the snapshot using DistCp to the target storage.

2. The second approach uses DistCp directly to the target storage. The advantage of the first approach is that the application is freed up after the snapshot finishes.

In this data lake protection scenario, the Hadoop Administrator uses DistCp to perform full backups using NFS over Ethernet to an on premise ECS Appliance.

The standard method to restore a DistCp backup from ECS to a traditional Hadoop infrastructure is to run DistCp in the reverse direction. This is done simply by swapping the source and target paths. You can perform partial or full restores and restores can be directed to the original location or an alternate location.

Customers have the option of leveraging ECS replication to a separate ECS Appliance installed at a second site for additional disaster recovery protection. DistCp restores could then be performed from the second site ECS Appliance for disaster recovery.

BENEFITS OF USING HADOOP DISTRIBUTED COPY DATA PROTECTION TO ECS

Customers will realize very important benefits from Distributed Copy data lake protection to Elastic Cloud Storage. First and most importantly, this Business Data Lake protection solution provides enterprise-grade data protection for Hadoop from data loss or corruption. This solution also gives the Hadoop Administrator direct visibility and control over their data lake protection.

The ECS Appliance makes hyperscale storage and cloud economics viable for any size business by combining the power of ViPR on a low-cost, high density, scale out commodity hardware platform. The ECS Appliance can be deployed and expanded incrementally, so you can choose the right size for your immediate needs and your projected growth. ECS allows you to optimize your data lake protection solution based on your applications, storage requirements, and access needs – giving you the flexibility and control that you want.

If a customer already uses Elastic Cloud Storage for other needs then the same processes and expertise can be leveraged for data lake protection.

OVERVIEW OF ISILON SNAPSHOTS MANAGED BY NETWORKER SNAPSHOT MANAGEMENT

Isilon snapshots managed by NetWorker Snapshot Management, illustrated on the right in Figure 6, applies to data lake deployments where the compute and storage are separated and the HDFS layer is running on the shared storage. Because you are using shared storage, customers can leverage all the data management capabilities that are built into that storage layer. This means customers can leverage Isilon snapshot functionality managed by NetWorker and can also do rollovers to Data Domain protection storage. A rollover refers to performing a backup of a snapshot to a secondary protection storage device via NDMP. This is typically done when longer term retention of data is a requirement.

ISILON SNAPSHOTS MANAGED BY NETWORKER SNAPSHOT MANAGEMENT TO DATA DOMAIN

This section provides more detail about leveraging EMC NetWorker Snapshot Management for data lake protection in deployments where the compute and storage are separated and the HDFS layer is running on Isilon storage. Because you are using shared Isilon storage, you can leverage all Isilon data management capabilities that are built into the storage layer. In this data lake protection scenario, NetWorker manages Isilon snapshots which are then rolled over to an on premise Data Domain storage system.

The choice of using Data Domain systems as the target protection storage for this solution will typically be made by customers based on a consideration of 3 primary factors:

1. Will your data benefit from Data Domain variable-length deduplication & compression storage benefits?

2. Does Data Domain storage scalability meet your needs? (Terabytes)

3. Does NFS meet your accessibility requirements?

The NetWorker Administrator can define a single policy to automate the data protection process including initiating a snapshot on the data lake Isilon system and then a executing a rollover of that Isilon snapshot using NDMP Tape Server over Ethernet to an on premise Data Domain system. The Data Domain system will ingest the snapshot data and perform variable-length deduplication and compression.

NetWorker maintains catalogs for all backups, snapshots, and clones which makes restores for this data lake protection solution simple and straightforward. NetWorker can also manage snapshot retention. To perform a restore, the NetWorker Administrator can simply and quickly restore from the initial snapshot, or can select one of the NDMP backup savesets that has been rolled over to the Data Domain system and then restore it back to the primary Isilon system using NDMP over Ethernet. Restoring from the snapshot offers the benefit of a much quicker RTO, while recovery from the backup on a Data Domain provides quick access to longer RPOs. NetWorker can perform full or partial restores and restores can be directed to the original location or an alternate location on the same device.

Customers have the option of leveraging NetWorker controlled replication to a separate Data Domain system installed at a second site for additional disaster recovery protection. NetWorker restores could then be performed from the second site Data Domain system for disaster recovery.

BENEFITS OF USING NETWORKER MANAGED ISILON SNAPSHOTS TO DATA DOMAIN

Customers will realize very important benefits from NetWorker management of Isilon snapshots for data lake protection to a Data Domain system. First and most importantly, this Business Data Lake protection solution provides enterprise-grade data protection for Hadoop from data loss or corruption and provides superior RTOs.

NetWorker Snapshot Management simplifies the data protection process by automating both the array snapshots & the rollovers to Data Domain. This data protection solution provides multiple recovery options including recovery from the initial snapshot and from rollover savesets on Data Domain protection storage.

Data Domain's Data Invulnerability Architecture provides the best-in-class data protection ensuring that data from your data lake can be recovered when needed and the data can be trusted. Data Domain systems provide storage efficiency through variable-length deduplication and compression typically reducing storage requirements by 10-30x. Data Domain systems are also very fast, capable of ingesting data up to 68 TB/hour minimizing the time it takes to complete data lake protection backups. If customer already uses NetWorker or Data Domain systems for other data protection needs then the same processes and expertise can be leveraged for data lake protection. And finally, NetWorker can be leveraged to manage bandwidth efficient Data Domain replication to a Data Domain system at a second site for optional disaster recovery.

ISILON SNAPSHOTS MANAGED BY NETWORKER SNAPSHOT MANAGEMENT TO ISILON

This section provides more detail about leveraging Dell EMC NetWorker Snapshot Management for data lake protection in deployments where the compute and storage are separated and the HDFS layer is running on Isilon storage. Because you are using shared Isilon storage, you can leverage all Isilon data management capabilities that are built into the storage layer. In this data lake protection scenario, NetWorker manages Isilon snapshots which are then replicated to a second on premise Isilon storage system.

The choice of using Isilon snap and replicate protection for this solution will typically be made by customers based on a consideration of 4 primary factors:

1. Do you already know that your data would not gain significant storage savings from the variable-length deduplication & compression that Data Domain systems would provide?

2. Is it feasible to protect the amount of data that needs to be protected within the allotted backup windows?

3. Does Isilon storage scalability meet your needs? (Petabytes)

4. Does your organization have NFS/SMB (CIFS)/HDFS accessibility requirements?

The NetWorker Administrator can define a single policy to automate the data protection process including initiating a snapshot on the data lake Isilon system and automatically control the replication of that Isilon snapshot using Isilon SyncIQ to a second on premise Isilon system. The second Isilon system will store a copy of the snapshot data that has been replicated over by NetWorker and Isilon SyncIQ.

NetWorker maintains catalogs for all backups, snapshots, and clones which makes restores for this data lake protection solution simple and straightforward. NetWorker can also manage snapshot retention. To perform a restore, the NetWorker Administrator can simply restore from the initial snapshot, or can select one of the snapshots that have been replicated to the target Isilon system and then restore it back to the primary Isilon system. NetWorker can perform full or partial restores and restores can be directed to the original location or an alternate location on the same device.

In a Remote Replication scenario, NetWorker can additionally orchestrate and manage NDMP rollover to a Data Domain system or other backup target at the remote site, completely offloading backup from the production Isilon system. This allows for weekly or quarterly backups of larger datasets without impacting daily production.

BENEFITS OF USING NETWORKER MANAGED ISILON SNAPSHOTS TO ISILON

Customers will realize very important benefits from NetWorker management of Isilon snapshots for data lake protection to Isilon storage. First and most importantly, this Business Data Lake protection solution provides enterprise-grade data protection for Hadoop from data loss or corruption and provides superior RTOs.

NetWorker Snapshot Management simplifies the data protection process by automating both the initial snapshots & the replication process to a secondary Isilon. This data protection solution provides multiple recovery options including recovery from the initial snapshot on the source Isilon system and from replicated snapshots on the second Isilon system. In addition, the ability to rollover to a Data Domain system enables longer term retention and greater protection from data corruption and disaster. The snapshot, replicate, and rollover process can all be controlled by a single policy.

Isilon is an ideal platform for Hadoop and other Big Data applications. It uses erasure coding to protect data with greater than 80% storage efficiency, in contrast to traditional HDFS with 33% storage efficiency. Isilon has several classes of node types. This allows different Isilon tiers to be optimized for particular workloads.

If customer already uses Isilon or NetWorker for other needs then the same processes and expertise can be leveraged for this data lake protection solution. NetWorker Snapshot Management is an integrated feature in NetWorker utilizing common workflows and user interface for both snapshots and backup. And finally, NetWorker can be leveraged to manage Isilon replication to another Isilon system at a second site for optional disaster recovery.

ISILON SNAPSHOTS MANAGED BY NETWORKER SNAPSHOT MANAGEMENT TO ECS

This section provides more detail about leveraging Dell EMC NetWorker Snapshot Management for data lake protection in deployments where the compute and storage are separated and the HDFS layer is running on Isilon storage. Because you are using shared Isilon storage, you can leverage all Isilon data management capabilities that are built into the storage layer. In this data lake protection scenario, NetWorker manages Isilon snapshots which are then rolled over to an on premise Elastic Cloud Storage (ECS) Appliance.

The choice of using ECS as the target storage for this solution will typically be made by customers based on a consideration of 3 primary factors:

1. Do you already know that your data would not gain significant storage savings from the variable-length deduplication & compression that Data Domain systems would provide?

2. Do you require the hyperscale that ECS provides? (Exabytes)

3. Do you require Object/HDFS accessibility?

The NetWorker Administrator can define a single policy to automate the data protection process including initiating a snapshot on the data lake Isilon system and then executing a rollover of that Isilon snapshot using ECS APIs over Ethernet to a second on premise ECS Appliance.

NetWorker maintains catalogs for all backups, snapshots, and clones which makes restores for this data lake protection solution simple and straightforward. NetWorker can also manage snapshot retention. To perform a restore, the NetWorker Administrator can simply restore from the initial snapshot, or can select one of the savesets that has been rolled over to the ECS system and then restore it back to the primary Isilon system using ECS APIs over Ethernet. NetWorker can perform full or partial restores and restores can be directed to the original location or an alternate location on the same device.

Customers have the option of leveraging NetWorker controlled replication to a separate ECS Appliance installed at a second site for additional disaster recovery protection. NetWorker restores could then be performed from the second site ECS Appliance for disaster recovery.

BENEFITS OF USING NETWORKER MANAGED ISILON SNAPSHOTS TO ECS

Customers will realize very important benefits from NetWorker management of Isilon snapshots for data lake protection to Elastic Cloud Storage solution. First and most importantly, this Business Data Lake protection solution provides enterprise-grade data protection for Hadoop from data loss or corruption and provides superior RTOs.

NetWorker Snapshot Management simplifies the data protection process by automating both the initial snapshots & the rollovers to ECS. This data protection solution provides multiple recovery options including recovery from the initial snapshot and from rollover savesets on ECS storage.

The ECS Appliance makes hyperscale storage and cloud economics viable for any size business by combining the power of ViPR on a low-cost, high density, scale out commodity hardware platform. The ECS Appliance can be deployed and expanded incrementally, so you can choose the right size for your immediate needs and your projected growth. ECS allows you to optimize your data lake protection solution based on your applications, storage requirements, and access needs – giving you the flexibility and control that you want.

If customer already uses NetWorker or Elastic Cloud Storage for other needs then the same processes and expertise can be leveraged for data lake protection.

CUSTOMER BENEFITS

As stated previously, all of the Business Data Lake protection solutions presented in this paper provide much needed enterprise-grade data protection for Hadoop from data loss or corruption. Dell EMC gives customers choice in selecting the best data lake protection solution depending on the size of their data lake, their data types, their accessibility requirements, and their existing storage & data protection expertise.

The Business Data Lake protection solution options described in this paper that leverage Data Domain systems as the protection storage target provide additional benefits that are unique to Data Domain. Data Domain's Data Invulnerability Architecture provides the ultimate in data protection ensuring that data from your data lake can be recovered when needed and the data can be trusted. Data Domain systems provide storage efficiency through variable-length deduplication and compression typically reducing storage requirements by 10-30x. Data Domain systems are also very fast, capable of ingesting data at up to 68 TB/hour minimizing the time it takes to complete data lake protection backups. If customer already uses Data Domain for other data protection needs then the same processes and expertise can be leveraged to protect your data lake.

DD Boost for Enterprise Applications, which is part of Dell EMCs Data Protection Suite Family, provides Hadoop data protection. Hadoop customers further benefit from Data Domain using the power of DD Boost with backup performance that is superior to NFS, reduced bandwidth requirements, and improved load balancing and reliability. In doing so, DD Boost for Enterprise Apps offers a superior user experience by integrating into the Hadoop cluster management, the Hadoop filesystem, and by leveraging Hadoop-native constructs.

The Business Data lake protection solution options described in this paper that leverage Isilon systems as the storage target provide their own additional set of unique customer benefits. Isilon uses erasure coding to protect data with greater than 80% storage efficiency, in contrast to traditional HDFS with only 33% storage efficiency. Isilon has several classes of node types which allow different Isilon tiers to be optimized for particular workloads. If your organization already uses Isilon or for other needs then the same processes and expertise can be leveraged for these data lake protection solution options.

The Business Data Lake protection solution options described in this paper that leverage Elastic Cloud Storage (ECS) as the storage target provide scalability and accessibility advantages. The ECS Appliance makes hyperscale storage and cloud economics viable for any size business by combining the power of ViPR on a low-cost, high density, scale out commodity hardware platform. ECS allows you to optimize your data lake protection solution based on your applications, storage requirements, and access needs – giving you the flexibility and control that you want. And finally, if your organization already uses Elastic Cloud Storage for other needs then the same processes and expertise can be leveraged for data lake protection.

The Business Data Lake protection solutions described in this paper which leverage NetWorker provide a number of additional advantages regardless of the storage option used. The NetWorker administrator can define data protection policies that will automate all the snapshot and rollover activities making day to day operations simple and effective. NetWorker also provides control over retention of backups, snapshots, and rollovers minimizing manual retention effort. And the NetWorker solution options include the ability to recover from Isilon snapshots in addition to the rollover savesets providing superior RTOs and maximum flexibility.

CONCLUSION

This paper has stated that big data use cases have matured, has provided a definition for what is a data lake, and explained why customers are now demanding serious enterprise-grade data lake protection solutions. As a thought leader in big data solutions, Dell EMC has presented in this paper a data protection strategy and multiple data protection solution options to protect Hadoop and other data lakes. Dell EMC gives customers choice for which solution approach and which target storage option best meets their scalability & accessibility needs and can leverage any existing in-house storage or data protection expertise that may already exist.

For more information on Dell EMC big data, Hadoop, and Business Data Lake solutions, please check out our Big Data solutions page on Dell.com and these additional resources:

Dell EMC Data Domain Operating System Data Sheet

Dell EMC Isilon Scale-Out Storage Product Family Data Sheet

Dell EMC ECS Appliance, powered by ViPR Data Sheet

Dell EMC Data Domain Data Invulnerability Architecture white paper

Dell EMC NetWorker Data Sheet

The Business Value of Data Domain Boost