White Paper

HADOOP ON EMC ISILON SCALE-OUT NAS

Abstract

This white paper details the way EMC Isilon Scale-out NAS can be used to support a Hadoop data analytics workflow for an enterprise. It describes the core architectural components involved as well as highlights the benefits that an enterprise can leverage to gain reliable business insight quickly and efficiently while maintaining simplicity to meet the storage requirements of an evolving Big Data analytics workflow.

December 2012

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

The information in this publication is provided “as is.” EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Part Number H10528.1

Table of Contents

Introduction

Enterprises have been continuously dealing with storing and managing rapidly growing amounts of data, also known as Big Data. Though drive sizes have expanded to keep up with the compute capacity, the tools to analyze this Big Data and produce valuable insight has not kept up with this growth. Existing analytics architectures have proven to be too expensive and too slow. They have also been very challenging to maintain and manage.

Hadoop is an innovative open source Big Data analytics engine that is designed to minimize time to derive valuable insight from an enterprise's dataset. It is comprised of two major components; MapReduce and the Hadoop distributed file system (HDFS). MapReduce is the distributed task processing framework that runs jobs in parallel on multiple nodes to derive results faster from large datasets. HDFS is the distributed file system that a Hadoop compute farm uses to store all the input data that needs to be analyzed as well as any output produced by MapReduce jobs.

Hadoop is built on principles of scale-out and uses intelligent software running on a cluster of commodity hardware to quickly and cost effectively derive valuable insight. It is this distributed parallel task processing engine that makes Hadoop superb to analyze Big Data.

Enterprises have continued to rely on EMC's Isilon scale-out network attached storage (NAS) for various Big Data storage needs. OneFS is the operating system as well as the underlying distributed file system that runs on multiple nodes that form the EMC Isilon scale-out NAS. OneFS is designed to scale not just in terms of machines, but also in human terms—allowing large-scale systems to be managed with a fraction of the personnel required for traditional storage systems. OneFS eliminates complexity and incorporates self-healing and self-managing functionality that dramatically reduces the burden of storage management. OneFS also incorporates parallelism at a very deep level of the OS, such that every key system service is distributed across multiple units of hardware. This allows OneFS to scale in virtually every dimension as the infrastructure is expanded, ensuring that what works today, will continue to work as the dataset grows and workflows change.

This ability to be flexible and adapt to not only changing infrastructure and data capacity needs but also to adapt to evolving workflows with simplicity and ease makes EMC Isilon scale-out NAS an extremely attractive element of a Big Data storage and analytics workflow solution using Hadoop.

Hadoop Software Overview

Hadoop is an industry leading innovative open source Big Data analytics engine that is designed to minimize time to derive valuable insight from an enterprise's dataset. Below are the key components of Hadoop:

Hadoop MapReduce: the distributed task processing framework that runs jobs in parallel on large datasets across a cluster of compute nodes to derive results faster.

Hadoop Distributed File System (HDFS): the distributed file system that a Hadoop compute farm uses to store all the data that needs to be analyzed by Hadoop.

MapReduce as a computing paradigm was introduced by Google and Hadoop was written and donated to open source by Yahoo as an implementation of that paradigm.

Hadoop MapReduce

Hadoop MapReduce is a software framework for easily writing applications which process large amounts of data in-parallel on large clusters of commodity compute nodes.

The MapReduce framework consists of the following:

JobTracker: single master per cluster of nodes that schedules, monitors and manages jobs as well as its component tasks.

TaskTracker: one slave TaskTracker per cluster node that execute task components for a job as directed by the JobTracker.

A MapReduce job (query) comprises of multiple map tasks which are distributed and processed in a completely parallel manner across the cluster. The framework sorts the output of the maps, which are then used as input to the reduce tasks. Typically both the input and the output of the job are stored across the cluster of compute nodes using HDFS. The framework takes care of scheduling tasks, monitoring them and managing the re-execution of failed tasks.

Typically in a Hadoop cluster, the MapReduce compute nodes and the HDFS storage layer (HDFS) reside on the same set of nodes. This configuration allows the framework to effectively schedule tasks on the nodes where data is already present in order to avoid network bottlenecks involved with moving data within a cluster of nodes. This is how the compute layer derives key insight efficiently by aligning with data locality in the HDFS layer.

Hadoop is completely written in Java but MapReduce applications do not need to be. MapReduce applications can utilize the Hadoop Streaming interface to specify any executable to be the mapper or reducer for a particular job.

Hadoop Distributed Filesystem

HDFS is a block based file system that spans multiple nodes in a cluster and allows user data to be stored in files. It presents a traditional hierarchical file organization so that users or applications can manipulate (create, rename, move or remove) files and directories. It also presents a streaming interface that can be used to run any application of choice using the MapReduce framework. HDFS does not support setting hard or soft links and you cannot seek to particular blocks or overwrite files. HDFS requires programmatic access and so you cannot mount it as a file system. All HDFS communication is layered on top of the TCP/IP protocol.

Below are the key components for HDFS:

NameNode: single master metadata server that has in-memory maps of every file, file locations as well as all the blocks within the files and which DataNodes they reside on.

DataNode: one slave DataNode per cluster node that serves read/write requests as well as performs block creation, deletion and replication as directed by the NameNode.

HDFS is the storage layer where all the data resides before a MapReduce job can run on it. HDFS uses block mirroring to spread the data around in the Hadoop cluster for protection as well as data locality across multiple compute nodes. The default block size is 64 MB and the default replication factor is 3x.

Hadoop Distributions

The open source Apached Foundation maintains releases of Apache Hadoop at apache.org. All other distributions are derivatives of work that build upon or extend Apache Hadoop. Below is a list of common Hadoop distributions that are available today:

- Apache Hadoop
- Cloudera CDH3
- Greenplum HD
- Hortonworks Data Platform

The above list is not an exhaustive list of all the Hadoop distributions available today but a snapshot of popular choices. A detailed list of Hadoop distributions available today can be found at:
http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support

This is the software stack that customers run to analyze data with Hadoop.

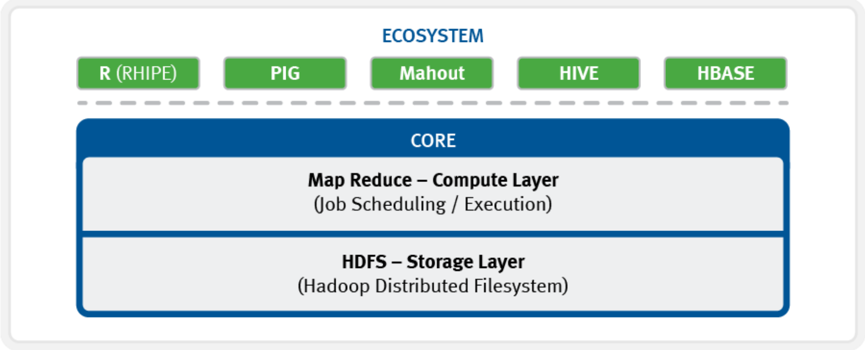
Hadoop Ecosystem

The following is the software stack that customers run to analyze data with Hadoop.

The ecosystem components are add-on components that sit on top of the Hadoop stack to provide additional features and benefits to the analytics workflows. Some popular choices in this area are:

- Hive: a SQL-like, ad hoc querying interface for data stored in HDFS
- HBase: a high-performance random read/writeable column oriented structured storage system that sits atop HDFS
- Pig: high level data flow language and execution framework for parallel computation
- Mahout: scalable machine learning algorithms using Hadoop
- R (RHIPE): divide and recombine statistical analysis for large complex data sets

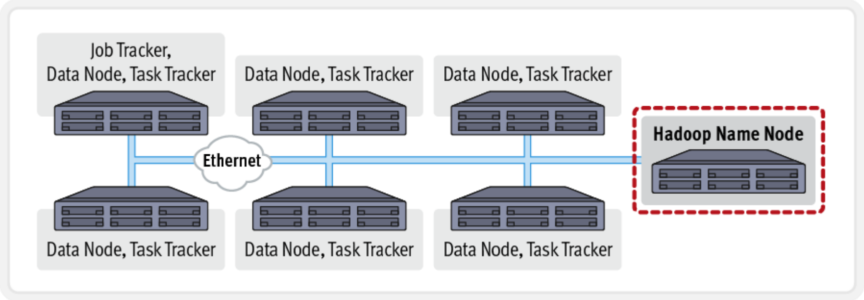
The above is not an exhaustive list of all Hadoop ecosystem components.


ECOSYSTEM
R (RHIPE)
PIG
Mahout
HIVE
HBASE
CORE
Map Reduce – Compute Layer
(Job Scheduling / Execution)
HDFS – Storage Layer
(Hadoop Distributed Filesystem)

All components of Hadoop

Hadoop Architecture

Below is what an architecture diagram would look like that shows all of the core Hadoop components that run on a Hadoop compute cluster.



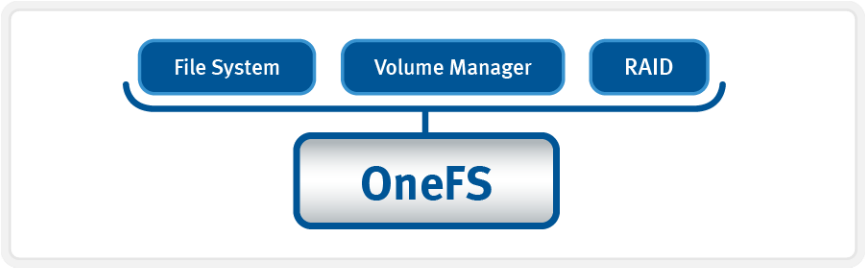
The general interaction that happens in this compute environment are:

1. Data must be ingested into the HDFS layer.
2. Computation or analysis occurs on the data using MapReduce.
3. Storing or exporting of results, either in HDFS or other infrastructure to accommodate the overall Hadoop workflow.

The above architecture also shows that the NameNode is a singleton in the environment and so if it has any issues, the entire Hadoop environment becomes unusable.

EMC Isilon OneFS Overview

OneFS combines the three layers of traditional storage architectures—filesystem, volume manager, and RAID—into one unified software layer, creating a single intelligent distributed filesystem that runs on an Isilon storage cluster.



OneFS combines filesystem, volume manager and protection into one single intelligent, distributed system

This is the core innovation that directly enables enterprises to successfully utilize the scale-out NAS in their environments today. It adheres to the key principles of scale-out; intelligent software, commodity hardware and distributed architecture. OneFS is not only the operating system but also the underlying filesystem that stores data in the Isilon Storage cluster.

Isilon Architecture

OneFS works exclusively with many Isilon scale-out storage nodes, referred to as a “cluster”. A single Isilon cluster consists of multiple “nodes”, which are constructed as rack-mountable enterprise appliances containing components such as memory, CPU, 1GB or 10GB networking, NVRAM, low latency Infiniband inter connects, disk controllers and storage media. Each node in the distributed cluster thus has processing capabilities as well as storage capabilities.

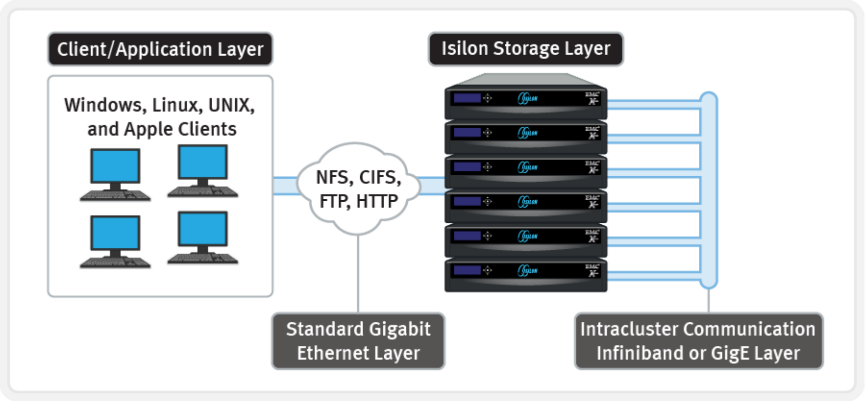
An Isilon cluster starts with as few as three nodes, and can currently scale up to 144 nodes. There are many different types of nodes, all of which can be incorporated into a single cluster where different nodes provide different ratios of capacity to throughput or IOPS.

OneFS is theoretically limitless in terms of the number of nodes that can be included in a single system. Each node added to a cluster increases aggregate disk, cache, CPU, and network capacity. OneFS leverages each of the hardware building blocks, so that the whole becomes greater than the sum of the parts. The RAM is grouped together into a single coherent cache, allowing I/O on any part of the cluster to benefit from data cached anywhere. NVRAM is grouped together to allow for high-throughput writes that are safe across power failures. Spindles and CPU are combined to increase throughput, capacity and IOPS as the cluster grows, for access to one file or for multiple files. A cluster’s storage capacity can range from a minimum of 18 terabytes (TB) to a maximum of 20 petabytes (PB) in a single filesystem.

EMC Isilon node types are segmented into several classes, based on their functionality:

- S-Series: IOPS-intensive applications
- X-Series: High-concurrency and throughput-driven workflows

- **NL-Series**: Near-primary accessibility, with near-tape value
- **Performance Accelerator**: Independent scaling for ultimate performance
- **Backup Accelerator Node**: High-speed and scalable backup and restore solution


Client/Application Layer
Windows, Linux, UNIX, and Apple Clients
Isilon Storage Layer
NFS, CIFS, FTP, HTTP
Standard Gigabit Ethernet Layer
Intracluster Communication Infiniband or GigE Layer

All components of OneFS at work in your environment

Above is the complete architecture; software, hardware and network connectivity all working together in your environment with your servers to provide a completely distributed single filesystem that can scale dynamically as workloads and capacity needs or throughput needs change.

OneFS Optional Software Modules

OneFS has add-on licensed software modules that can be enabled based on a customer's needs. The list below shows you a brief description of all the available modules and their functionality.

SnapshotIQ™ — Simple, scalable, and flexible snapshot based local data protection

SmartConnect™ — Policy-based data access, load balancing with failover for high availability

SmartQuotas™ — Data management with quotas and thin provisioning for clustered storage

SyncIQ™ — Fast and flexible file-based asynchronous data replication

SmartPools™ — Data management using different disk tiers and applying Information Lifecycle Management (ILM) policies based on file attributes

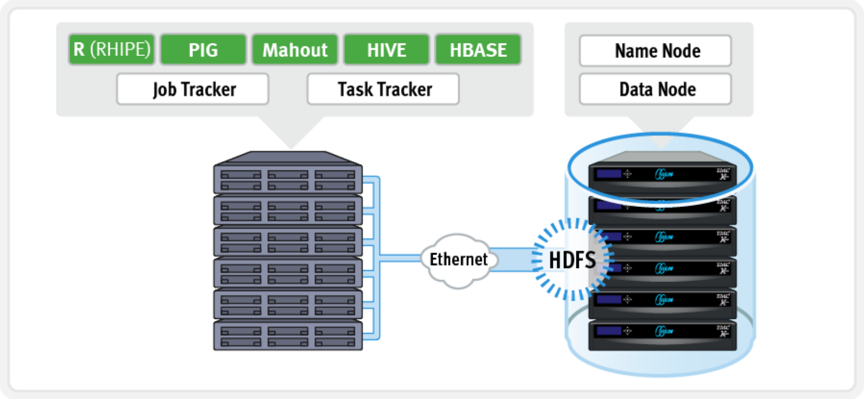
SmartLock™ — Store data in Enterprise WORM compliant format

InsightIQ™ — Powerful, yet simple analytics platform to identify trends, hot spots and key cluster statistics and information

Please refer to product documentation for details on all of the above software modules.

Hadoop On Isilon

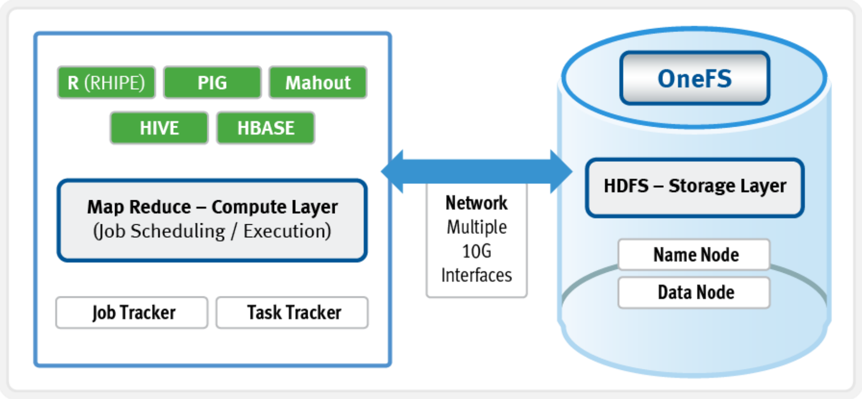
Since all of the communication in HDFS is layered on top of the TCP/IP protocol, Isilon has integrated the HDFS layer as an over-the-wire protocol for OneFS. This integration allows customers to leverage the scale-out NAS platform as a native part of their Hadoop architecture for both Hadoop core components as well as any ecosystem components. It also gives them the ability to leverage the simplicity, flexibility, reliability and efficiency of OneFS in their entire Hadoop workflow.



Isilon scale-out NAS (storage layer) connected to Hadoop (compute layer)

The above diagram shows what the architecture would look like when an Isilon scale-out NAS is integrated to a Hadoop compute cluster via an over-the-wire protocol (HDFS). This makes it possible for Isilon scale-out NAS to be a first class Hadoop citizen in an enterprise's Hadoop workflow. It also allows for separating two core components of a Hadoop workflow; the compute or the MapReduce layer as well as the Hadoop distributed file system (HDFS) or the storage layer. Since the network bandwidth available today has improved dramatically and because OneFS was built with distributed parallelism at its core, it makes sense for enterprise customers to leverage a shared scale-out storage model for their data analytics workflow using Hadoop.

Below is an in depth view that shows all of the core components as well ecosystem components of Hadoop at work with Isilon's scale-out NAS.



All Hadoop components working with Isilon scale-out NAS in a Hadoop environment

The EMC Isilon OneFS file system can scale to over 20 PB in a single file system and a single global namespace today. It can also scale to 100 GB/s concurrent throughput at that capacity. Please refer to the specsfs2008 benchmarking results (www.spec.org) for details on how OneFS can scale linearly up to the capacity and performance requirements of a Hadoop workflow.

The next few sections of this paper will detail the significant advantages of utilizing EMC Isilon scale-out NAS in a Hadoop workflow.

Simplicity

EMC Isilon OneFS removes all the complexities involved with managing racks of disk pools; provisioning file systems on them and applying appropriate data protection mechanisms to them so that an enterprise's data set is accessible at all times and is adequately protected against failures. OneFS brings simplicity to Hadoop so that an enterprise can focus on leveraging their data to derive business accelerating insight from it. This enables an enterprise to focus on using Hadoop to uncover key trends and identify opportunities that can help accelerate their business rather than spending time managing the storage infrastructure for their Hadoop ecosystem.

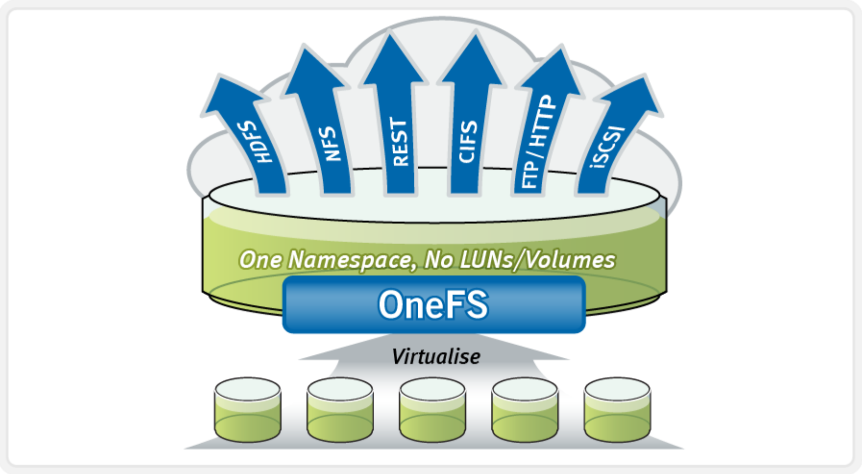
Scaling Isilon storage typically takes less than 60 seconds and can be accomplished with the press of a button while the cluster remains online. Adding an Isilon node to an existing cluster is as simple as powering it up after installing it in a rack and asking it to join the existing Isilon scale-out cluster. This “join” process will ensure that the additional capacity is available right away and that the correct OneFS version and configuration is running on the Isilon node being joined. Not only is the additional capacity available in 60 seconds but a background job runs to rebalance the current utilization evenly across the Isilon cluster to avoid hot spots of data. All of this capacity expansion happens as the scale-out storage remains online and is servicing MapReduce jobs without any impact.

In addition to supporting the HDFS protocol, OneFS also supports all of the following protocols:

- NFS

- CIFS/SMB
- FTP
- HTTP
- iSCSI
- REST

The Isilon HDFS implementation is a lightweight protocol layer between OneFS filesystem and HDFS clients. This means that files are stored in standard POSIX compatible filesystem on an Isilon cluster. This really makes it simple for an organization to utilize any of the above mentioned protocols to ingest data for their Hadoop workflow or export the Hadoop derived business critical insight to other components of the data analytics workflow. If the data is already stored on the EMC Isilon scale-out NAS, then the customer simply points their Hadoop compute farm at OneFS without having to perform a time and resource intensive load operation for their Hadoop workflow. OneFS allows enterprises to simply use the HDFS layer in their Hadoop environment as a true and proven filesystem.



Isilon scale-out NAS brings simplicity to Hadoop workflows

Efficiency

OneFS is designed to allow over 80% capacity utilization out of an Isilon scale-out cluster which makes it very efficient for Hadoop data analytics workflows. When compared to a traditional Hadoop architecture that typically uses 3X mirroring for every block that resides in the filesystem, OneFS is extremely capacity efficient and can provide an optimized ROI and TCO for the enterprise customers. For example, if an enterprise wanted to be able to store 12 PB of Hadoop data, they would typically need to purchase more than 36 PB of raw disk capacity in a traditional Hadoop cluster using a default of 3x mirroring to store data in it. Storing the same 12 PB of Hadoop data with data protection on OneFS, would, however, only require 15 PB of raw disk capacity in an Isilon cluster. This results in a significant CAPEX savings as well as a much simpler infrastructure environment to manage.

Along with the operational ease and simple management that Isilon brings to improve OPEX savings, there are other efficiencies that can be achieved in the environment as well. For example, Isilon nodes can get very dense from a capacity perspective. As a result, the rack space as well as power needed to run a 36 PB traditional Hadoop cluster using direct attached storage can be significantly more than a 15 PB Isilon cluster that can support the same data requirements. This advantage of the Isilon cluster results in additional cost savings.

Using the Isilon scale-out NAS as a shared storage layer for the Hadoop environment also allows customers to converge and minimize their Hadoop compute farm. By off-loading all the storage-related HDFS overhead to the Isilon scale-out NAS, Hadoop compute farms can be better utilized for performing more analysis jobs instead of managing local storage, providing protection on the data in the local storage as well as performing analysis on the data residing in the local storage. By alleviating the Hadoop compute farm from performing all of these HDFS related tasks, OneFS can thereby help to reduce the Hadoop compute farm footprint as well as potentially leveraging the existing Hadoop compute infrastructure for other tasks in the data analytics workflow. The entire data analytics workflow benefits from the efficiency of having the shared storage being accessible via other standard protocols for getting key Hadoop derived insight to other parts of the data analytics workflow. This converged storage approach helps streamline the entire data analytics workflow so that enterprises can realize significant CAPEX and OPEX savings.

Flexibility

In traditional Hadoop clusters using direct attached storage, the compute layer and the storage layer are tightly coupled and so you cannot expand one without the other. The drawback this leads to is that customers expand their Hadoop clusters because they need more storage capacity and not compute capacity. However as they expand, they have now added more network infrastructure as well as compute infrastructure. This proves to be very inefficient and inflexible in terms of overall usage.

De-coupling the Hadoop compute and storage layers allows an enterprise to have the flexibility to independently scale one (storage) vs. the other (compute) when they need to. This flexible, pay-as-you-grow architecture allows customers to buy only what they need and when they need it making the entire complex Hadoop environment more efficient. With this capability, organizations can start small and scale-out as they need, up to 100 GB/s of concurrent throughput on their Hadoop storage layer with OneFS.

A key advantage of using OneFS for Hadoop storage needs is that it is Apache Hadoop compliant. This gives organizations the flexibility to select the Hadoop distribution of their choice for use in their Hadoop data analytics workflow. EMC Isilon Scale-out NAS has been tested with the following Hadoop distribution:

- Apache Hadoop 0.20.203
- Apache Hadoop 0.20.205
- EMC Greenplum HD 1.1

Reliability

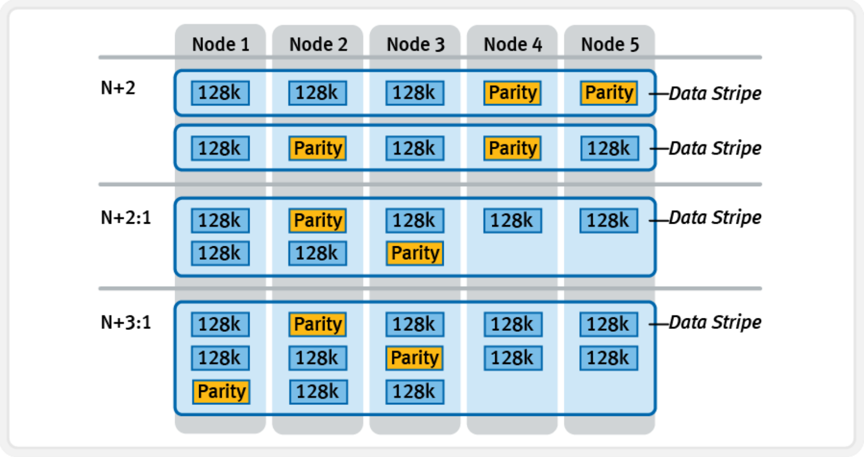
To get the most value of their Hadoop analytics investments, enterprises require a resilient Big Data storage infrastructure. Isilon Scale-out NAS and the OneFS operating system provides a highly reliable infrastructure with unmatched levels of data protection to safeguard data assets and deliver a highly available data environment.

In traditional Hadoop clusters, enterprises must rely on custom sub-block CRC checksums for providing data protection at the hardware layer along with mirroring technology at the HDFS layer so that there is some level of data redundancy. However, this becomes a very expensive proposition at scale.

Data protection for Isilon is implemented at the OneFS file system level and, as such, is not dependent on any hardware RAID controllers. This provides many benefits, including the ability to add new data protection schemes as market conditions or hardware attributes and characteristics evolve. Since protection is applied at the file-level, a OneFS software upgrade is all that’s required in order to make new protection and performance schemes available.

OneFS employs the widely proven Reed-Solomon erasure coding algorithm for its parity protection calculations. Protection is applied at the file-level, enabling the cluster to recover data quickly and efficiently. Inodes, directories and other metadata are protected at the same or higher level as the data blocks they reference. Since all data, metadata and forward error correction (FEC) blocks are striped across multiple nodes, there is no requirement for dedicated parity drives. This both guards against single points of failure and bottlenecks and allows file reconstruction to be a highly parallelized process. Today, OneFS provides N+1 through N+4 parity protection levels, providing protection against up to four simultaneous component failures respectively. A single failure can be as little as an individual disk or, at the other end of the spectrum, an entire node.

OneFS also supports several hybrid protection schemes. These include N+2:1 and N+3:1, which protect against two drive failures or one node failure, and three drive failures or one node failure, respectively. These protection schemes are particularly useful for high density node configurations, where each node contains up to thirty six, multi-terabyte SATA drives. Here, the probability of multiple drives failing far surpasses that of an entire node failure. In the unlikely event that multiple devices have simultaneously failed, such that the file is “beyond its protection level”, OneFS will re-protect everything possible and report errors on the individual files affected to the Isilon cluster’s logs.



OneFS Hybrid Parity Protection Schemes (N+M:x)

File System Journal

Every Isilon node is equipped with a dual-battery backed 512MB NVRAM card, which guards that node's file system journal. Each journal is used by OneFS as stable storage, and guards write transactions against sudden power loss or other catastrophic events. The journal protects the consistency of the file system and the battery charge lasts up to three days. Since each member node of an Isilon cluster contains an NVRAM controller, the entire OneFS file system is therefore fully journaled.

Proactive Node/Device Failure

OneFS will proactively remove, or SmartFail, any drive that reaches a particular threshold of detected ECC errors, and automatically reconstruct the data from that drive and locate it elsewhere on the cluster. Both SmartFail and the subsequent repair process are fully automated and hence require no administrator intervention. Because OneFS protects its data at the file-level, any inconsistencies or data loss is isolated to the unavailable or failing device - the rest of the file system remains intact and available.

Since OneFS is built upon a highly distributed architecture, it is able to leverage the CPU, memory and spindles from multiple nodes to reconstruct data from failed drives in a highly parallel and efficient manner. Because an Isilon storage system is not bound by the speed of any particular drive, OneFS is able to recover from drive failures extremely quickly and this efficiency grows relative to cluster size. As such, failed drive within an Isilon cluster will be rebuilt an order of magnitude faster than hardware RAID-based storage devices -- in minutes or hours as compared to many hours or days. Additionally, OneFS has no requirement for dedicated ‘hot-spare’ drives.

Isilon Data Integrity

ISI Data Integrity (IDI) is the OneFS process that protects file system structures against corruption via 32-bit CRC checksums. All Isilon blocks, both for file and metadata, utilize checksum verification. Metadata checksums are housed in the metadata blocks themselves, whereas file data checksums are stored as metadata, thereby providing referential integrity. All checksums are recomputed by the initiator, the node servicing a particular read, on every request.

In the event that the recomputed checksum does not match the stored checksum, OneFS will generate a system alert, log the event, retrieve and return the corresponding parity block to the client and attempt to repair the suspect data block automatically.

Protocol Checksums

In addition to blocks and metadata, OneFS also provides checksum verification for Remote Block Management (RBM) protocol data. RBM is a unicast, RPC-based protocol developed by Isilon for use over the back-end cluster interconnect. Checksums on the RBM protocol are in addition to the Infiniband hardware checksums provided at the network layer, and are used to detect and isolate machines with certain faulty hardware components and exhibiting other failure states.

Dynamic Sector Repair

OneFS includes a Dynamic Sector Repair (DSR) feature whereby bad disk sectors are fenced off and good data can be redirected by the file system to be rewritten elsewhere. When OneFS fails to read a block during normal operation, DSR is invoked to reconstruct the missing data and write it to either a different location on the drive or to another drive on the node. This is done to ensure that subsequent reads of the block do not fail. DSR is fully automated and completely transparent to the end-user. Disk sector errors & CRC mismatches use almost the same mechanism as the drive rebuild process.

Mediascan

MediaScan's role within OneFS is to check disk sectors and deploy the above DSR mechanism in order to force disk drives to fix any sector ECC errors they may encounter. Implemented as one of the part of OneFS, MediaScan is run automatically based on a predefined schedule. Designed as a low-impact, background process, MediaScan is fully distributed and can thereby leverage the benefits of Isilon's unique parallel architecture.

Integrity Scan

IntegrityScan, another component of OneFS, is responsible for examining the entire file system for inconsistencies. It does this by systematically reading every block and verifying its associated checksum. Unlike traditional 'fsck' style file system integrity checking tools, IntegrityScan is designed to run while the cluster is fully operational, thereby removing the need for any downtime. In the event that IntegrityScan detects a checksum mismatch, a system alert is generated and written to the syslog and OneFS automatically attempts to repair the suspect block.

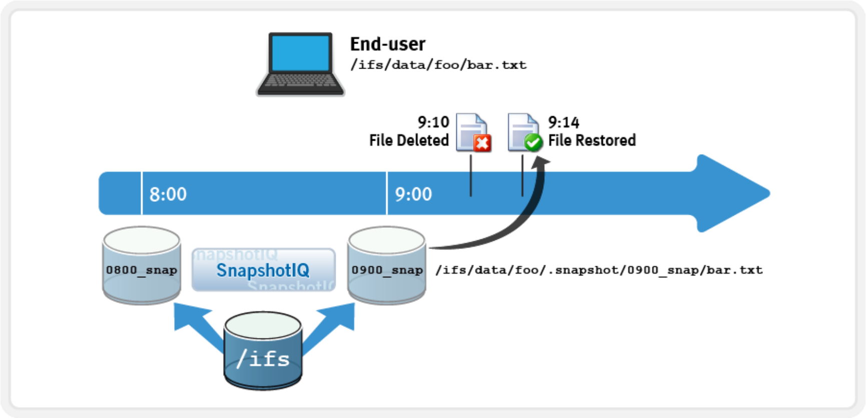
Data High Availability

In a traditional Hadoop cluster using direct attached storage, there is only a single NameNode that manages any filesystem namespace operation. However with Isilon storage, every Isilon node can reply to NameNode or DataNode requests. Each time a Hadoop compute client sends a request for a file to the Isilon OneFS scale-out NAS, the request is sent to a different Isilon node address. In a Hadoop environment, every Isilon node in the cluster is a NameNode and DataNode. This allows for load balancing of IO to occur from multiple mapper and reducer tasks that run on multiple Hadoop compute nodes to occur across multiple Isilon nodes. In this way OneFS also eliminates single points of failure that exist in a traditional Hadoop cluster and enables load balancing.

The Isilon SmartConnect module contributes to data high availability by supporting dynamic failover and failback for Hadoop compute clients. This ensures that when a node failure occurs, all in-flight reads and writes associated with a MapReduce job are handed off to another node in the Isilon cluster to finish its operation without any MapReduce job or task interruption. This is accomplished when using a SmartConnect Zone name as the configuration parameter in the Hadoop core-site.xml configuration file as well as dynamic IP addresses on the Isilon cluster. Refer to the Isilon User guide for details on SmartConnect Zone configuration. This capability provides load balancing as well as continuous data availability in the event of a failure when running Hadoop MapReduce jobs with Isilon OneFS providing the HDFS storage layer.

Business Continuity

OneFS has a robust mechanism to provide a highly reliable data backup strategy necessary for business continuity in an enterprise environment. Isilon's SnapshotIQ can take read-only, point-in-time copies of any directory or subdirectory within OneFS to serve as the fastest local backup. OneFS Snapshots are highly scalable and typically take less than one second to create. They create little performance overhead, regardless of the level of activity of the file system, the size of the file system, or the size of the directory being copied. Also, only the changed blocks of a file are stored when updating the snapshots, thereby ensuring highly-efficient snapshot storage utilization. User access to the available snapshots is via a /.snapshot hidden directory under each file system directory. Isilon SnapshotIQ can also create unlimited snapshots on a cluster. This provides a substantial benefit over the majority of other snapshot implementations because the snapshot intervals can be far more granular and hence offer improved RPO time frames.



User driven file recovery using SnapshotIQ

In addition to the benefits provided by SnapshotIQ in terms of user recovery of lost or corrupted files, it also offers a powerful way to perform backups while minimizing the impact on the file system. Initiating backups from snapshots affords several substantial benefits. The most significant of these is that the file system does not need to be quiesced, since the backup is taken directly from the read-only snapshot. This eliminates lock contention issues around open files and allows users full access to data throughout the duration of the backup job.

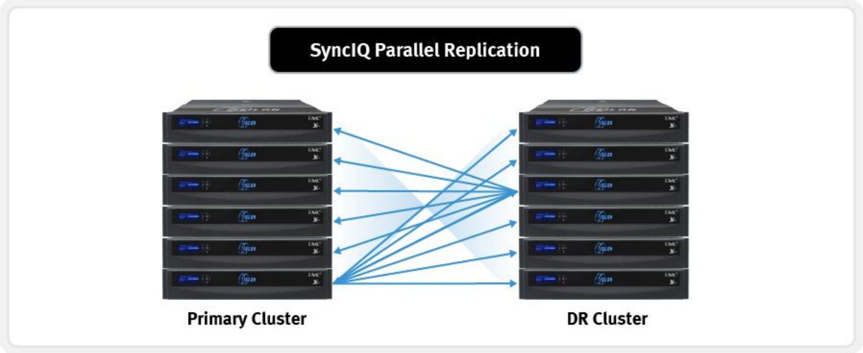
SnapshotIQ also automatically creates an alias which points to the latest version of each snapshot on the cluster, which facilitates the backup process by allowing the backup to always refer to that alias. Since a snapshot is by definition a point-in-time copy, by backing up from a snapshot, the consistency of the file system or sub-directory is maintained.

This process can be further streamlined by using the NDMP snapshot capability to create a snapshot as part of the NDMP backup job, then delete it upon successful completion of the backup. OneFS facilitates backup and restore functionality via its support of the ubiquitous Network Data Management Protocol (NDMP). NDMP is an open-standard protocol that provides interoperability with leading data-backup products. Isilon supports both NDMP versions 3 and 4. The OneFS NDMP module includes the following functionality:

- Full and incremental backups and restores using NDMP
- Direct access restore (DAR/DDAR), single-file restores, and three-way backups
- Restore-to-arbitrary systems
- Seamless integration with access control lists (ACLs), alternate data streams and resource forks
- Selective File Recovery
- Replicate then backup

While snapshots and NDMP provide an ideal solution for localized failure scenarios, when it comes to catastrophic failures or natural disasters, a second, geographically separate copy of a dataset is clearly beneficial.

The Isilon SyncIQ module delivers high-performance, asynchronous replication of data to address a broad range of recovery point objectives (RPO) and recovery time objectives (RTO). This enables customers to make an optimal tradeoff between infrastructure cost and potential for data loss if a disaster occurs. SyncIQ does not impose a hard limit on the size of a replicated file system so will scale linearly with an organization's data growth up into the multiple petabyte ranges.



Replicate Hadoop data using SyncIQ to local or geographically separate Isilon OneFS clusters

SyncIQ is easily optimized for either LAN or WAN connectivity to replicate over short or long distances, thereby providing protection from both site-specific and regional disasters. Additionally, SyncIQ utilizes a highly-parallel, policy-based replication architecture designed to leverage the performance and efficiency of clustered storage. As such, aggregate throughput scales with capacity and allows a consistent RPO over expanding data sets.

In summary, a number of shortcomings in traditional Hadoop storage implementations can be addressed by inserting EMC Isilon storage as part of the HDFS storage layer, as follows:

Traditional Hadoop Storage Implementation	EMC Isilon Storage Implementation
Dedicated Storage Infrastructure – One-off for Hadoop only	Scale-Out Storage Platform – Multiple applications & workflows
Single Point of Failure – Namenode	No Single Point of Failure – Distributed Namenode
Lacking Enterprise Data Protection – No Snapshots, replication, backup	End-to-End Data Protection – SnapshotIQ, SyncIQ, NDMP Backup
Poor Storage Efficiency – 3X mirroring	Industry-Leading Storage Efficiency – ›80% Storage Utilization

Fixed Scalability – Rigid compute to storage ratio	**Independent Scalability** – Add compute & storage separately
Manual Import/Export – No protocol support	**Multi-Protocol** – Industry standard protocols – NFS, CIFS, FTP, HTTP, HDFS

Conclusion

Hadoop is an innovative analytics engine that can significantly reduce the time and resources needed by an enterprise to derive valuable insight from their Big Data assets. As detailed in this paper, EMC Isilon scale-out NAS and the Isilon OneFS operating system can be an over the wire HDFS layer and thereby provide significant advantages. This integration allows organizations to utilize the scale-out NAS platform as a native part of their Hadoop architecture for both Hadoop core components as well as any ecosystem components. It also provides the ability to leverage the simplicity, flexibility, reliability and efficiency of OneFS in their entire Hadoop workflow. By treating HDFS as an over the wire protocol, organizations can readily deploy a Big Data analytics solution that combines any industry standard Apache Hadoop distribution with Isilon scale-out NAS storage systems to achieve a powerful, highly efficient and flexible Big Data storage and analytics ecosystem. This approach enables organizations to avoid the resource-intensive complexity of traditional Hadoop deployments that use direct-attached storage. With Isilon Scale-out NAS, organizations can provide a highly resilient storage infrastructure for their Hadoop environment that increases data protection and improves reliability while maintaining simplicity to meet the requirements of an evolving Big Data analytics workflow.

About Isilon

Isilon, a division of EMC, is the global leader in scale-out NAS. We deliver powerful yet simple solutions for enterprises that want to manage their data, not their storage. Isilon's products are simple to install, manage and scale, at any size. And, unlike traditional enterprise storage, Isilon stays simple no matter how much storage is added, how much performance is required or how business needs change in the future. We're challenging enterprises to think differently about their storage, because when they do, they'll recognize there's a better, simpler way. Learn what we mean at http://www.isilon.com.