White Paper

EMC ISILON SCALE-OUT NAS FOR IN-PLACE HADOOP DATA ANALYTICS

Abstract

This white paper shows that storing data in EMC Isilon scale-out network-attached storage optimizes data management for Hadoop analytics. Separating data from HDFS clients and storing it in an Isilon cluster provides scalability, efficiency, and workflow flexibility.

November 2013

Part Number H12532

Table of Contents

Executive summary

EMC® Isilon® transforms data analytics by coupling a key tool of data science with the natural home of Big Data—scale-out network-attached storage (NAS). By its very nature, unstructured data flows into large storage systems, often over Server Message Block (SMB) and network file system (NFS). Meanwhile, data scientists and other data analysts are increasingly turning to Hadoop to analyze unstructured data.

By allowing Hadoop clients direct access through Hadoop Distributed File System (HDFS) to the data that is stored in an Isilon cluster with NFS, HTTP, FTP, or SMB, an EMC Isilon cluster pairs the standard tool of data science with a highly scalable storage system. Combining Hadoop with scale-out NAS supports a fundamental shift taking place in advanced organizations: Businesses are trying to analyze their data and extract value from it. Isilon fosters data analytics without placing a heavy reliance on Hadoop application developers and without ingesting data into HDFS.

An EMC Isilon cluster delivers value by separating data from compute. With an EMC Isilon cluster, you can store data on an enterprise storage platform with your existing workflows and standard protocols, including SMB, HTTP, FTP, Representational State Transfer (REST), and NFS. Regardless of whether you store the data with SMB or NFS, however, you can analyze it with a Hadoop compute grid through HDFS. There is no need to set up a separate HDFS and then move data to it with tedious HDFS copy commands or specialized Hadoop connectors. Instead of struggling to take your data to Hadoop, you can bring Hadoop to your data.

An Isilon cluster simplifies data management while cost-effectively maximizing the value of data. Although high-performance computing with Hadoop has traditionally stored data locally in compute clients' HDFS deployment, the following use cases make a compelling case for coupling the MapReduce programming model with Isilon scale-out NAS:

- Store data in a POSIX-compliant file system with SMB and NFS workflows and then access it through HDFS for MapReduce
- Scale storage independently of compute as your data sets grow
- Protect data more reliably and efficiently instead of replicating it
- Eliminate HDFS copy operations to ingest data and Hadoop FileSystem (FS) commands to manage data
- Implement NameNode redundancy
- Manage data with enterprise storage features such as deduplication and snapshots

This white paper shows that storing data in an Isilon scale-out NAS cluster instead of HDFS clients optimizes the management of Big Data for Hadoop analytics. An Isilon cluster provides MapReduce clients with file system scalability, storage efficiency, and workflow flexibility.

Scale-out storage and Big Data analysis

It's well known that the volume, velocity, and variety of unstructured data and digital information are exploding. The growth is putting companies of all sizes under pressure to harness their data. Traditional storage systems such as storage-attached networks, direct-attached storage, and scale-up network-attached storage are not up to the challenge—at least not in a highly scalable, efficient way. These traditional storage systems are inefficient because they require the use of many disparate volumes, defragmentation, and volume management—all of which require excess overhead. The overall efficiency rate of traditional storage systems that use logical unit numbers (LUNs) or volumes is about 55 percent to 65 percent. As a result, traditional storage systems, although dominant because of low barriers to entry and mature but inefficient RAID technologies, are marred by a high rate of capital expenditure and high operating expenses.

The data deluge is pressing many companies to try different approaches. "While not all companies, regardless of size, are in industries experiencing exponential data growth, the scale of content growth across all companies is significant. If you haven't done so already, your organization will need to consider new approaches for dealing with data growth and active access to archive data in the very near future," Richard L. Villars writes in an IDC white paper.[1]

Rapid data growth demands scale-out storage. The IDC paper argues that the pressures of Big Data and the limitations of file servers and scale-up storage devices require "the deployment of new classes of storage solutions (scale-out storage solutions) that are optimized for rapid data ingest, efficient storage management, and reliable access."

The convergence of stored data and data analysis

Mining stored data is quickly becoming as important as managing it. As data piles up, often in storage silos, businesses are turning to analyzing it with the hope of extracting value from it. Businesses want to find patterns in their data to predict behavior, create better products, innovate faster, increase revenue, or cut costs. "In fact, with the right mindset, data can be cleverly reused to become a fountain of innovation and new services. The data can reveal secrets to those with the humility, the willingness, and the tools to listen," Viktor Mayer-Schonberger and Kenneth Cukier write in their 2013 book "Big Data: A Revolution That Will Transform How We Live, Work, and Think." One of these tools is Hadoop, and its use is on the rise.

[1] Villars, Richard L. "Managing Data Growth and Monetizing Information Value: The Role of Scale-Out Storage Solutions in the Expanding Universe of Digital Information and Big Data," IDC white paper, July 2012.

Data analysis is on a course to converge with data storage. "In the very near future, the management, organization, and continuous mining of large content pools will become an equally important task for many data center administrators. The greater use of robust 'Big Data' solutions in ongoing business processes is where these two developments will intersect,"[2] Richard L. Villars and Benjamin Woo write in an IDC white paper on competing in the new world of Big Data.

At present, however, businesses that see value in their data often hire engineers to build stand-alone grids of client computers that can run Hadoop to analyze data with MapReduce, a processing paradigm for data-intensive computational analysis. But before engineers can analyze data, they must move it from one or more storage systems to the machines running Hadoop. Importing the data into the clients can be a complicated, time-consuming task that involves running HDFS copy operations or using specialized Hadoop connectors. And it's an operation that tends to be repeated: As the data expands or changes, it must be loaded in the machines in the compute grid again.

In addition, some data-intensive workloads must process data in another system after extracting it from a storage system but before loading it into HDFS, which creates an Extract, Transform, Load (ETL) workflow. Finally, after porting the data to HDFS and analyzing it, the results must be exported to another system. It is a slow, expensive method of preparing data for analysis.

"Enabling a distributed, scale-out architecture, particularly as it relates to data storage, protection, and access, is the foundational building block to being able to evolve organizations from application-focused to data/information-focused enterprises."
—IDC white paper on managing data growth and monetizing information value[2]

Separating data from compute

EMC Isilon scale-out NAS fosters the convergence of data analytics with stored data. As you work to extract value from stored data, you can use an Isilon cluster's HDFS implementation to point your data analytics tools at the storage system. Instead of requiring application developers to move data to the compute grid, you can take the compute function to where the data usually already resides—in storage.

The convergence of stored data and data analysis helps streamline the entire analytics workflow. The convergence eliminates the need to extract the data from a storage system and load it into a traditional Hadoop deployment. The convergence also eliminates the need to export the data after it is analyzed. Streamlining the analytics workflow cost-effectively speeds the transition to a data-focused enterprise: You not only increase the ease and flexibility with which you can analyze data but also reduce your capital expenditures and operating expenses.

[2] Villars, Richard L., and Benjamin Woo. "Managing Data Growth and Monetizing Information Value: Competing in the Expanding Universe of Digital Information and Big Data," IDC white paper, October 2011.

Scale-out storage for Big Data

The EMC Isilon scale-out platform combines modular hardware with unified software to provide a storage foundation for in-place data analysis. Isilon scale-out NAS is a fully distributed system that consists of modular hardware nodes arranged in a cluster. The distributed EMC Isilon OneFS® operating system combines the nodes' memory, I/O, CPUs, and disks into a cohesive storage unit that presents a global namespace as a single file system.

The nodes work together as peers in a shared-nothing hardware architecture with no single point of failure. Every node adds capacity, performance, and resiliency to the cluster, and each node acts as a Hadoop NameNode and DataNode. The NameNode daemon is a distributed process that runs on all nodes in the cluster.

As nodes are added, the file system expands dynamically and redistributes data, eliminating the work of partitioning disks and creating volumes. The result is a highly efficient and resilient storage architecture that brings all the advantages of an enterprise scale-out NAS system to storing data for analysis.

To handle Big Data, an Isilon cluster scales multidimensionally, optimizes data protection, supports existing workflows with standard network protocols like SMB and NFS, and manages data intelligently.

An Isilon cluster scales out multidimensionally. For Hadoop, scalability applies to more than just hardware. Scalability must also address throughput, file volume, and RAM. Each Isilon node adds memory, capacity, 10 GbE network connections, and increasingly efficient data protection. An Isilon cluster can scale nondisruptively while your MapReduce jobs run.

Unlike traditional storage, Hadoop's ratio of CPU, RAM, and disk space depends on the workload—factors that make it difficult to size a Hadoop cluster before you have had a chance to measure your MapReduce workload. In addition, the data you are analyzing with Hadoop is probably growing daily, and possibly growing at an increasing rate, making up-front sizing decisions problematic.

Isilon scale-out NAS lends itself perfectly to this scenario: It lets you increase CPUs, RAM, and disk space by adding nodes to dynamically match storage capacity and performance with the demands of a dynamic Hadoop workload.

An Isilon cluster optimizes data protection. Hadoop presupposes that your data is an asset, and an Isilon cluster includes enterprise features that protect your data.

The OneFS operating system more efficiently and reliably protects data than HDFS. By default, the HDFS protocol replicates a block of data three times. OneFS stripes the data across the cluster and protects the data with forward error correction (FEC) codes, which consume less space than replication and provide better protection.

An Isilon cluster also includes enterprise features to back up your data and provide high availability. For example, in managing your DataNode data, a good approach to take with a traditional Hadoop system is to back up your data to another system—an operation that must be performed with brute force by using a tool like Hadoop RCP.

OneFS availability features include clones, Network Data Management Protocol (NDMP) backups, synchronization, automated cluster replication and failover, snapshots, file system journal, virtual hot spare, antivirus, IntegrityScan, dynamic

sector repair, and accelerated drive rebuilds. For complete information about the data availability features of OneFS, see the white paper titled "Isilon: Data Availability & Protection."

An Isilon cluster supports your workflows. Instead of running HDFS copy operations to move your data to Hadoop clients in your compute grid, you can continue to store data using existing workflows. An EMC Isilon cluster provides multiprotocol data access with SMB, NFS, HTTP, REST, and FTP as well as HDFS.

Why is supporting existing workflows so key to advancing data analytics within your organization? Because it empowers a business to analyze its data without a heavy dependence on IT. Instead of relying on Hadoop application developers to painstakingly deploy HDFS and move data to it, business personnel can use their existing workflows—especially SMB and NFS—to collect and manage the data that they want to analyze. The implications of this shift not only help prepare a business now for the future of data-driven analysis, but also open up data analysis to the core business.

For example, if your technical support staff has collected log data from a year's worth of support cases and stored the log files on a file server, support managers can analyze the data in place to identify patterns of recurring problems to help predict when a problem might arise and address it before it affects a system.

A 2012 article in "MITSloan Management Review" sums up this shift: "Advanced organizations are moving analytics from IT into their core business and operational functions. As big data evolves, a new information ecosystem is also evolving, a network that is continuously sharing information, optimizing decisions, communicating results and generating new insights for businesses."[3] With HDFS access to its POSIX file system, an Isilon cluster lays the foundation for such an ecosystem.

An Isilon cluster lets you manage data intelligently. OneFS includes storage pools, deduplication, automated tiering, quotas, high-performing SSDs, capacity-optimized HDDs, and monitoring with EMC Isilon InsightIQ®.

Deduplication, for example, decreases the space required to store data. The post-process deduplication of OneFS eliminates duplicate blocks of identical data stored on disk and replaces the blocks with pointers to shadow stores. With post-process deduplication, data is analyzed for identical blocks after it is stored on disk, not during write operations, so that the deduplication process does not affect the performance of file operations such as writing data or modifying data.

For security, OneFS can authenticate HDFS connections with the Kerberos protocol. EMC Isilon SmartLock® can protect sensitive data from malicious, accidental, or premature alteration or deletion and help meet compliance with SEC 17a-4 regulations.

For more information about the enterprise features of OneFS, see the white paper titled "Hadoop on EMC Isilon Scale-Out NAS."

[3] Davenport, Thomas H., Paul Barth, and Randy Bean. "How 'Big Data' Is Different," MITSloan Management Review. Fall 2012.

A use case from the life Sciences

A real-world use case illustrates how an Isilon cluster dynamically scales to adapt to a changing workload. Although organizations working in the life sciences have been early adopters of Hadoop, which is ideally suited to analyze genomics data, scientists face exponential growth in data.

An article titled "Will Computers Crash Genomics" points to the exponential growth of the total capacity of the genomics sequencing market. In 2010, the capacity was about 200 petabytes but was growing to about 1 exabyte by late 2012.[4]

The growth is drowning out storage technologies that cannot scale rapidly and efficiently (see "Hadoop in the Life Sciences: An Introduction," EMC Isilon).

Life science workflows require a high-performance computing (HPC) infrastructure to process and analyze data to determine the variations in the genome, and the workflows require a proper scale of storage to retain the data.

With next-generation genome sequencing workflows generating up to 2 terabytes of data per run per week per sequencer, not including raw images, scale-out storage that integrates easily with HPC is a requirement.

EMC Isilon has provided the scale-out storage for nearly all the next-generation DNA sequencing workflows that exist today.

In the life sciences, EMC Isilon scale-out NAS serves more than 300 customers. The EMC Isilon storage platform has a life sciences installed base of more than 65 petabytes.

How Hadoop works with Isilon scale-out NAS

An Isilon cluster separates data from compute. As Hadoop clients run MapReduce jobs, the clients access the data stored on an Isilon cluster over HDFS. OneFS becomes the native HDFS deployment for MapReduce clients.

OneFS implements the server-side operations of the HDFS protocol on every node, and each node functions as both a NameNode and a DataNode. An Isilon node, however, does not act as a job tracker or a task tracker; those functions remain the purview of Hadoop clients. OneFS contains no concept of a Secondary NameNode: Because every Isilon node functions as a NameNode, the function of the Secondary NameNode—checking the internal NameNode transaction log—is unnecessary.

The cluster load balances HDFS connections across all the nodes in the cluster. Because OneFS stripes Hadoop data across the cluster and protects it with parity blocks at the file level, any node can simultaneously serve DataNode traffic as well as NameNode requests for file blocks.

A virtual racking feature mimics data locality. For example, you can create a virtual rack of nodes to assign compute clients to the nodes that are closest to a client's network switch, if doing so is necessary to work with your network topology or optimize performance.

Even though Hadoop clients connect to the cluster over HDFS as they run MapReduce jobs, the data can be stored on the cluster and managed through other common application-layer network protocols, including SMB, HTTP, FTP, REST, and NFS. You

[4] Pennisi, Elizabeth. *Science 331 No. 6018* (February 2011): 666-668.

can, for instance, load the data with NFS, analyze it through HDFS, and then export or share it with SMB.

There is no need to store data over HDFS with time-consuming copy operations or to manage the data with cumbersome Hadoop FS commands. Instead, you can manage the files with standard Linux commands, such as `chmod`, `chown`, `ls`, and `cp`. Managing your Hadoop files with familiar Linux commands saves time and makes data management easy.

Supported distributions

An EMC Isilon cluster is platform agnostic for compute, and there is no vendor lock-in: You can run most of the common Hadoop distributions with an Isilon cluster, including Apache Hadoop, Hortonworks Data Platform, Cloudera, and Pivotal HD. After you activate an Isilon Hadoop license, the cluster tries to automatically detect a client's Hadoop distribution.

Clients running different Hadoop distributions or versions can connect to the cluster simultaneously. For example, you can point both Cloudera and Pivotal HD at the same data on your Isilon cluster and run MapReduce jobs from both distributions at the same time.

An EMC Isilon cluster running OneFS 7.0.2.2 or later works with the following Hadoop distributions and projects. Earlier versions of OneFS work with many of these distributions and projects, too. For more information, contact an EMC Isilon representative.

- Apache Hadoop 0.20.203
- Apache Hadoop 0.20.205
- Apache Hadoop 1.0.0-1.0.3
- Apache Hadoop 1.2.1
- Apache Hadoop 2.0.x
- Cloudera CDH3u2
- Cloudera CDH3u3
- Cloudera CDH3u4
- Cloudera CDH3u5
- Cloudera CDH4.2
- Cloudera Manager CDH4
- Greenplum® HD 1.1
- Greenplum HD 1.2
- Hortonworks Data Platform/Apache 1.0.3
- Pivotal HD 1.0.1
- HAWQ 1.1.0.1
- Apache HBase
- Apache Hive

- Apache Pig

Availability

The Isilon architecture provides a resilient foundation for data availability and data protection. In its 2013 report titled "Critical Capabilities for Scale-Out File System Storage," Gartner rated EMC Isilon highest among storage vendors for resiliency—a platform's capabilities for provisioning a high level of system availability and uptime.[5]

For availability, an Isilon cluster includes the following features:

- No single point of failure
- Unparalleled levels of data protection
- Tolerance for multi-failure scenarios
- A fully distributed single file system
- Proactive failure detection
- Fast drive rebuilds
- Flexible, efficient data protection
- A fully journaled file system
- High transient availability
- NameNode redundancy
- DataNode load balancing

For more information on the availability features of OneFS for Hadoop, see the white paper titled "Hadoop on EMC Isilon Scale-Out NAS."

Hardware

An Isilon cluster is built on a highly redundant architecture governed by the hardware premise of shared nothing. The cluster's fundamental building blocks are platform nodes. As a rack-mountable appliance, a node includes the following components in a 2U or 4U rack-mountable chassis: memory, CPUs, RAM, non-volatile RAM (NVRAM), network interfaces, InfiniBand adapters, disk controllers, and storage media. The redundant InfiniBand adapters provide the distributed system bus that connects all the nodes. Each node houses a battery-backed file system journal. NVRAM is grouped to protect write operations from power failures.

With Hadoop, RAM matters. Hadoop jobs generally consist of many sequential reader threads. An Isilon cluster's large level-two cache, which is what most of a node's DRAM is used for, supports MapReduce jobs with sequential reader threads.

Network

Client computers can access any node in the cluster through dual 1 GbE or dual 10 GbE network connections. Client connections are, by default, distributed across the cluster with round-robin load balancing. On the network side, Isilon's logical network interface (LNI) framework provides a robust, dynamic abstraction for easily combining

[5] "Critical Capabilities for Scale-Out File System Storage," Gartner, Inc. Jan. 24, 2013.

and managing different interfaces for network resilience. Multiple network interfaces can be trunked together with Link Aggregation Control Protocol (LACP) and LAGG to aggregate bandwidth. An EMC Isilon SmartConnect™ license adds additional network resilience with IP address pools that support multiple DNS zones in a subnet as well as IP failover.

File system

The cluster's highly extensible file system provides mirrored volumes for the root and /var volumes that are stored on flash drives. For additional resilience, OneFS saves last known good boot partitions.

Data protection overview

OneFS takes a more efficient approach to data protection than HDFS. By default, the HDFS protocol replicates a block of data three times to protect it and to make it highly available. Instead of replicating the data, OneFS stripes the data across the cluster over its internal InfiniBand network and protects the data with forward error correction (FEC) codes.

FEC is a highly efficient method of reliably protecting data. FEC encodes a file's data in a distributed set of symbols, adding space-efficient redundancy. With only a part of the symbol set, OneFS can recover the original file data. In a cluster with five or more nodes, FEC delivers as much as 80 percent efficiency. As you add nodes to a cluster, data protection becomes increasingly efficient.

Striping data with FEC codes consumes much less storage space than replicating data three times—as much as 2.5 times fewer drives. Striping data lets a Hadoop client connecting to any node take advantage of the entire cluster's performance to read or write data.

Efficient data protection

The difference in the efficiency with which OneFS and a traditional HDFS deployment protect data is dramatic. To support an effective capacity goal of 4 petabyes, the difference between traditional HDFS and OneFS can be seen in the table that follows.

File system	Capacity goal	Consumption with overhead	Description
HDFS	4 PB	12 PB of disk space	4 PB plus 3 copies of each block
OneFS	4 PB	5 PB of disk space	4 PB plus FEC protection

With the Isilon data protection scheme, more than 80 percent of an Isilon cluster's capacity can be utilized, bringing efficiency to a data analytics workflow. In contrast to HDFS, which uses triple replication for every block, the efficiency of Isilon data protection optimizes return on investment (ROI) and total cost of ownership (TCO).

For example, if an enterprise wanted to store 4 PB of Hadoop data, it would typically need to purchase more than 12 PB of raw disk capacity in a traditional Hadoop cluster using a default of 3x mirroring to store data in it. Storing the same 4 PB of Hadoop data with data protection on OneFS, however, would only require 5 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.

If you set the replication level from an HDFS client, OneFS ignores it and instead uses the protection level that you set for the directory or the file pool that contains your Hadoop data.

By default, OneFS optimizes striping for concurrent access. With Hadoop, however, the dominant data access pattern might be streaming. You can set OneFS to lay out data for streaming access patterns to increase sequential read performance for MapReduce jobs. To better handle streaming access, OneFS stripes data across more drives. Streaming is most effective on directories or subpools serving large files or handling large compute jobs.

NameNode redundancy

Every Isilon node acts as a NameNode and a DataNode. Because every node runs the OneFS HDFS service, every node can simultaneously serve NameNode requests for file blocks and DataNode traffic. A cluster thus inherently provides NameNode redundancy as long as you follow the standard Isilon practice of setting your clients to connect to the cluster's SmartConnect zone's DNS entry. The result: There is no single point of failure.

SmartConnect distributes NameNode sessions with round-robin routing. Here's how it works: When a Hadoop client first tries to connect to a NameNode, OneFS routes the traffic to a node, which serves as the client's NameNode. The client's subsequent NameNode requests go the same node. When a second Hadoop client connects to the cluster's SmartConnect DNS entry, OneFS balances the traffic by default with round-robin, and routes the connection to a different node than the one used by the previous client. In this way, OneFS evenly distributes NameNode connections across the cluster to significantly improve the performance of read/write intensive traffic like that of TeraSort.

If a node that a client is using as a NameNode goes down, SmartConnect moves the IP address of the connection to another node, which then becomes the node that services NameNode traffic for the Hadoop clients that had been connected to the down node. Although reassigning the NameNode IP address to another node might temporarily disrupt the in-flight connections, the MapReduce job will continue. There may be tasks that need to be restarted from a checkpoint, however.

DataNode load balancing

OneFS load balances connections across DataNodes with round-robin routing. When a compute client submits a request to an Isilon node, the node acts like a NameNode and responds dynamically with a DataNode.

Architecture

Isilon storage architecture

The OneFS HDFS service relies on the cluster's scale-out architecture and distributed file system. Each node adds resources to the cluster. Because each node contains globally coherent RAM, as a cluster becomes larger, it serves data for MapReduce jobs

more quickly. A compute client can connect to any node to access its data for MapReduce. OneFS distributes Hadoop client connections among all the nodes in the cluster.

When you add a node to a cluster, you increase the cluster's aggregate disk, cache, CPU, RAM, and network capacity. OneFS groups RAM into a single coherent cache so that a data request on a node benefits from data that is cached anywhere. NVRAM is grouped to write data with high throughput and to protect HDFS write operations from power failures. As the cluster expands, spindles and CPU combine to increase throughput, capacity, and I/Os per second. Meanwhile, the Hadoop file system expands dynamically and redistributes content, which eliminates the need to add storage to compute clients. You can add Isilon nodes to the cluster without interrupting MapReduce jobs.

As a result, a scale-out Isilon cluster radically simplifies storage management for Hadoop data. You can manage more storage with fewer people. More importantly, storing your data on an Isilon cluster liberates your Hadoop application developers from managing storage, and shifts the work to storage administrators. Storing data on an Isilon cluster leaves developers free to focus on what they do best: developing applications to analyze data.

In contrast to storing data on Hadoop clients, storing data on an Isilon cluster saves on power, cooling, and the other costs that are associated with storage. The rack space as well as power needed to run a 12 PB traditional Hadoop cluster using direct attached storage can be significantly more than what is needed to run a 5 PB Isilon cluster that can support the same storage requirements.

The EMC Isilon OneFS file system can scale to more than 20 PB in a single file system and a single global namespace today. It can also scale to 85 GB/s concurrent throughput at that capacity. See the SPECsfs2008 benchmarking results (www.spec.org) for more information on how OneFS can scale linearly to meet the capacity and performance requirements of a Hadoop workflow.

Rack awareness

OneFS can contain a virtual rack of nodes to assign a pool of Hadoop compute clients to a pool of DataNodes that are closest to the clients' main network switch. The virtual rack also ensures that OneFS does not route compute clients to unreachable DataNodes in networks that lack full connectivity.

A virtual rack mimics data locality. The rack associates the clients' IP addresses with a pool of DataNodes so that when a client connects to the cluster, OneFS assigns the connection to one of the DataNodes in the pool. In this way, a virtual rack can, for example, route a client's connection through its optimal network switch. This improves performance by reducing read throughput latency while minimizing traffic through a top-of-rack switch.

More specifically, the virtual rack simulates two DataNodes in the rack and a third DataNode in a different rack to optimize the arrangement of DataNodes for your network switch topology. When a NameNode returns three IP addresses to a Hadoop client's request to read data, the NameNode selects the DataNode. The OneFS HDFS daemon checks which client is connecting and then returns the IP addresses for the

DataNode and the secondary DataNode from one rack and an IP address for the third DataNode from another rack.

A Hadoop client, such as a Pivotal Data Computing Appliance, can use a virtual rack to connect to a node even when a networking switch fails. If, for example, a client connects to two networking switches, a main switch and a top-of-rack switch, a virtual rack ensures that the client can connect to a DataNode even if the top-of-rack switch fails. In such a case, the client's connection comes into the Isilon cluster through the main switch.

A Hadoop client connects over HDFS to the DataNodes with interfaces that are assigned to the pool. After you add a pool with EMC Isilon SmartPools®, you can change the IP address allocation for clients that connect to the cluster. For more information, see the OneFS Command Reference or the OneFS Administration Guide.

The HDFS architecture of OneFS

OneFS implements the server-side operations of the HDFS protocol on every node. The architecture employs a single thread pool with a daemon—named isi_hdfs_d—that allocates a thread to each HDFS connection to handle RPC calls to NameNodes and read/write requests to DataNodes.

A NameNode resides on every node in the cluster. An HDFS client connects to a NameNode to query or modify metadata in the OneFS file system. The metadata includes the logical location of data for the file stream—that is, the address of the DataNode on which a block resides. An HDFS client can modify the metadata through the NameNode's RPC interface. OneFS protects HDFS metadata at the same protection level as HDFS file data. In fact, OneFS handles all the metadata. You do not need to worry about managing the data or backing it up.

With Isilon, the NameNode daemon translates HDFS semantics and data layout into OneFS semantics and file layout. For example, the NameNode translates a file's path, offset, and LEN into lists of block IDs and generation stamps. The NameNode also translates a client's relative path request into a LIN and then returns to the client the address of a DataNode and the location of a block.

A DataNode stores blocks of files. More specifically, the DataNode maps blocks to block data. With HDFS, a block is an inode-offset pair that refers to a part of a file. With OneFS, you can set the size of an HDFS block to optimize performance. A Hadoop client can connect to a DataNode to read or write a block, but the client may not write a block twice or delete a block. To transfer a block to a DataNode, the HDFS client encapsulates the block in packets and sends them over a TCP/IP connection. The Isilon HDFS daemon performs zero-copy system calls to read and write blocks to the file system. On OneFS, the DataNode reads packets from and writes packets to disk.

To manage writes, OneFS implements the same write semantics as the Apache implementation of HDFS: Files are append only and may be written to by only one client at a time. Concurrent writes are permitted only to different files. As with the Apache implementation, OneFS permits one lock per file and provides a mechanism for releasing the locks or leases of expired clients.

Conclusion

An EMC Isilon cluster optimizes the storage of Big Data for data analysis. Combining Hadoop clients with Isilon scale-out NAS and the OneFS implementation of HDFS delivers the following solutions:

- Store your analytics data with existing workflows and protocols like NFS, HTTP, and SMB instead of spending time importing and exporting data with HDFS
- Protect data efficiently, reliably, and cost-effectively with forward error correction instead of triple replication
- Manage data with such enterprise features as snapshots, deduplication, clones, and replication
- Receive NameNode redundancy with a distributed NameNode daemon that eliminates a single point of failure
- Support HDFS 1.0 and 2.0 simultaneously without migrating data or modifying metadata
- Run multiple Hadoop distributions—including Cloudera, Pivotal HD, Apache Hadoop, and Hortonworks Data Platform—against the same dataset at the same time
- Implement security for HDFS clients with Kerberos and address compliance requirements with write once, read many (WORM) protection for Hadoop data
- Scale storage independently of compute to handle expanding datasets

By scaling multidimensionally to handle the exponential growth of Big Data, an EMC Isilon cluster pairs with Hadoop to provide the best of both worlds: Data analytics and enterprise scale-out storage. The combination helps you adapt to fluid storage requirements, nondisruptively add capacity and performance in cost-effective increments, reduce storage overhead, and exploit your data through in-place analytics.

About EMC

EMC Corporation is a global leader in enabling businesses and service providers to transform their operations and deliver IT as a service. Fundamental to this transformation is cloud computing. Through innovative products and services, EMC accelerates the journey to cloud computing, helping IT departments to store, manage, protect and analyze their most valuable asset—information—in a more agile, trusted and cost-efficient way. Additional information about EMC can be found at www.EMC.com.