Lab Validation Brief

EMC Isilon Scale-Out Data Lake Foundation

Essential Capabilities for Building Big Data Infrastructure

By Ashish Nadkarni, IDC Storage Team

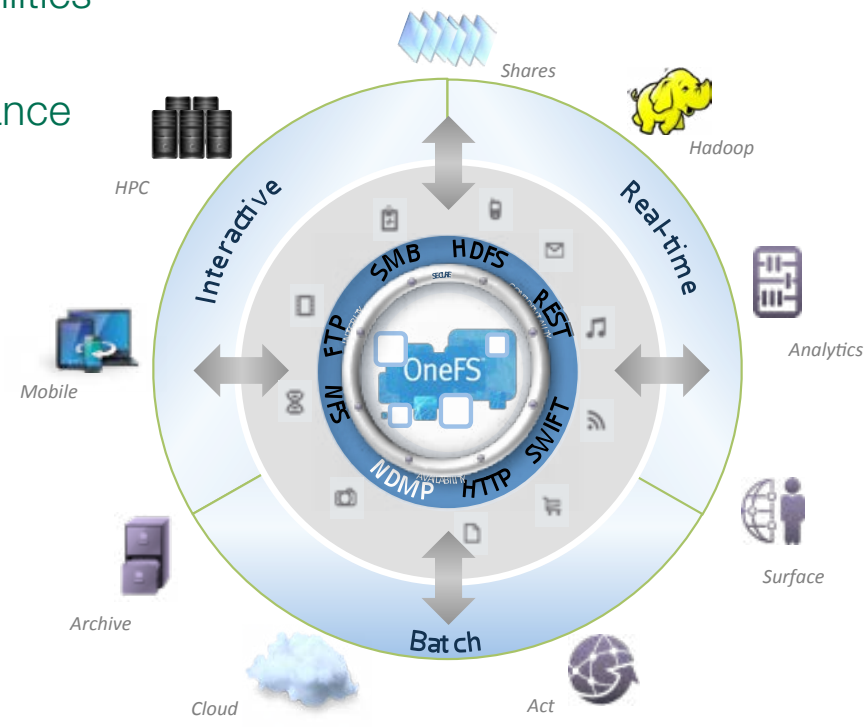
Sponsored by EMC Isilon | March 2016

IDC
Analyze the Future

Lab Validation Brief Executive Summary

IDC validated key features/functions of EMC Isilon in big data workflows

1. Multi-Protocol Capabilities
2. Availability
3. Security and Compliance
4. Simplified Operations



IDC Opinion

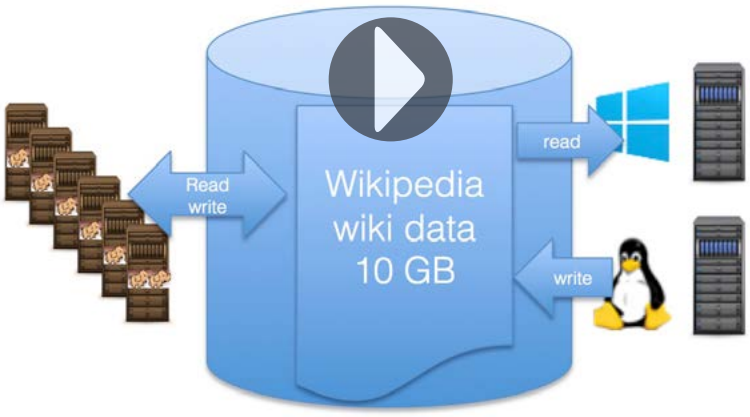
IDC believes that EMC Isilon is indeed an easy to operate, highly scalable and efficient Enterprise Data Lake Platform (EDLP)*. IDC validated that a shared storage model based on the Data Lake can in fact provide enterprise-grade service-levels while performing better than dedicated commodity off-the-shelf (COTS) storage for Hadoop workloads.

**The EMC Isilon Scale-out Data Lake is an Enterprise Data Lake Platform (EDLP) based on the OneFS distributed file system.*

Validated: Concurrent Ingest via NFS, SMB and HDFS

Feature/Validation Summary

The EMC Isilon Scale-out Data Lake is an ideal platform for multi-protocol ingest of data. This is a crucial function in Big Data environments, in which it is necessary to quickly and reliably ingest data into the Data Lake using protocols closest to the workload generating the data. With OneFS it is possible to ingest data via NFSv3, NFSv4, SMB2.0, SMB3.0 as well as via HDFS. This makes the platform very friendly for complex Big Data workflows.



IDC validated the robust multi-protocol ingest capabilities of the OneFS platform. This validation included the ability of the platform to preserve user access permissions across protocols.

Validation Process

For this validation, the EMC Isilon Scale-out Data Lake was configured for access via NFSv3, SMB3.0 and HDFS from the Hadoop Cluster. HDFS and NFS access was setup via the master node of the Hadoop DAS cluster. A large file download was simulated (Wikipedia wiki data, 10GB). The file was left compressed. The file was accessed and analyzed continuously via HDFS, while it was being copied via NFS to the EMC Isilon Scale-out Data Lake. It was also accessed via SMB3.0 while it was being read and written via HDFS and NFS, respectively.

Notes:

- Isilon OneFS uses DNS zone delegation and then subsequently uses DNS round robin for balancing incoming connections.
- No such inherent functionality is available in HDFS, which results in job failure should the data node go down. Once the node is marked as failed, HDFS calls bypass the failed node.

Key Finding: Multi-Protocol Ingest Capabilities

Robust multi-protocol ingest capabilities make it easier to build a Big Data workflow in a Data Lake built on the EMC Isilon Scale-out Data Lake platform

Capability	Why does it matter?
Ingest via NFS (v3, v4), SMB (2.0, 3.0) and HDFS	Simultaneous multiprotocol read/write access from various users, local and directory-based, allows concurrent handling of Big Data workflows

IDC Inference

Businesses will find it easy to build out workflows using the EMC Isilon Scale-out Data Lake because:

- It enables the use of existing and known file protocol mechanisms (instead of Hadoop-specific mechanisms that require specific application-level modifications).
- It's performance optimization capabilities make it an ideal platform for enterprise-wide data storage/analytics with a centralized storage repository.
- The use of native protocols allow in-place analytics (eliminate migrations), make data workflows faster and allow businesses to gain faster insights.

Validated: HDFS Performance in the Data Lake

Feature/Validation Summary

The EMC Isilon Scale-out Data Lake offers excellent read and write performance for Hadoop clusters accessing HDFS via OneFS vis-à-vis those accessing HDFS via local (internal) storage.

	TeraGen		TeraSort		TeraValidate	
	total sec	**MB/sec**	**total sec**	**MB/sec**	**total sec**	**MB/sec**
Hadoop Data Lake Cluster	594.652	1,681.66	1558.159	641.78	353.112	2,831.96
Hadoop DAS Cluster	1652.761	605.05	2405.953	415.64	547.05	1,827.99

IDC validated the performance profiles of both DAS and EDLP using well known Hadoop benchmarking jobs supplied with the Hadoop distribution.

Validation Process

For this validation, three standard benchmarking tests were used: TeraGen, TeraSort and TeraValidate. TeraGen benchmarks sequential write performance. TeraSort provides a good benchmark for mixed read/write tests. TeraValidate benchmarks read performance. Results of each script are summarized in the table above. The Hadoop Data Lake Cluster accessed the EMC Isilon Scale-out Data Lake via the HDFS via the API, whereas the Hadoop DAS Cluster accessed the HDFS locally. Exact same values were used for parameters passed to the 'Tera' jobs on both Hadoop Clusters respectively. In this configuration, the EMC Isilon Scale-out Data Lake is nearly 3x faster for writes, and over 1.5x faster for reads/writes and reads. Similarly, in spite of network access, 10GbE links provide significant improvement in bandwidth (MB/sec) for reads and writes.

Notes:

- The EMC Isilon Scale-out Data Lake was configured with SSD-based caching, whereas the Hadoop DAS Cluster was configured with 10k RPM SAS drives.
- IDC believes that with internal SSDs, the Hadoop DAS cluster may show a marked improvement in performance.

Validated: NFS Performance During Multi-Protocol Ingest

Feature/Validation Summary

SmartFlash L3 caching capabilities of OneFS lend themselves well to the preservation of protocol performance (read/write performance) during the multi-protocol ingest process in the EMC Isilon Scale-out Data Lake. This is a crucial function in Big Data environments, because analytics workloads cannot be paused during data ingest, nor can the ingest and analytics processes be serialized. Furthermore read/write performance is important in environments where the analytics workloads are operating on files that are constantly being updated.

Hadoop NFS read/write performance

	NFS Write		NFS Read	
	total sec	MB/sec	total sec	MB/sec
Hadoop Data Lake Cluster	34.4362	290.39	10.3208	968.92
Hadoop DAS Cluster	145.816	68.58	381.847	26.19

IDC validated that NFS performance of the EDLP is significantly faster than a Hadoop DAS cluster because of optimizations on the OneFS platform, including a native NFS daemon and L3 caching.

Validation Process

For this validation, the UNIX 'dd' command was used to write a set of blocks into a file and then subsequently read those blocks. (This test used the UNIX 'dd' command to sequentially write and read a file containing 10 GB worth of zeros.) These tests were performed on NFS mounted file systems from the EMC Isilon Scale-out Data Lake and the Hadoop DAS cluster. The results of these tests are summarized above. The EMC Isilon Scale-out Data Lake provides 4.2x faster write performance and 37x faster read performance.

Notes:

- An NFS Gateway was installed on one of the data nodes of the Hadoop DAS Cluster and mounted on the master node. Native NFS functionality in Hadoop is provided by way of a Java-based user process, and not via the system NFS daemons.
- NFS read performance in Isilon is greatly improved because of read hits (flash-based L3 caching), whereas the Java NFS implementation has limited caching capabilities.

Key Finding: Multi-Protocol Workload Performance

An Enterprise Data Lake platform should provide vastly improved Hadoop workload performance over a standard DAS configuration

Capability	Why does it matter?
HDFS read and write performance in a shared Data Lake platform	Excellent performance of Hadoop clusters attached to a shared Data Lake means marked improvement in Map/Reduce operations, which result in more efficient Big Data workflows. Enterprises benefit from quicker completion of tasks, thereby minimizing latency between subsequent operations
NFS performance during multi-protocol ingest	Performance improves because SmartFlash L3 caching means that read/write performance is increased during multi-protocol ingest

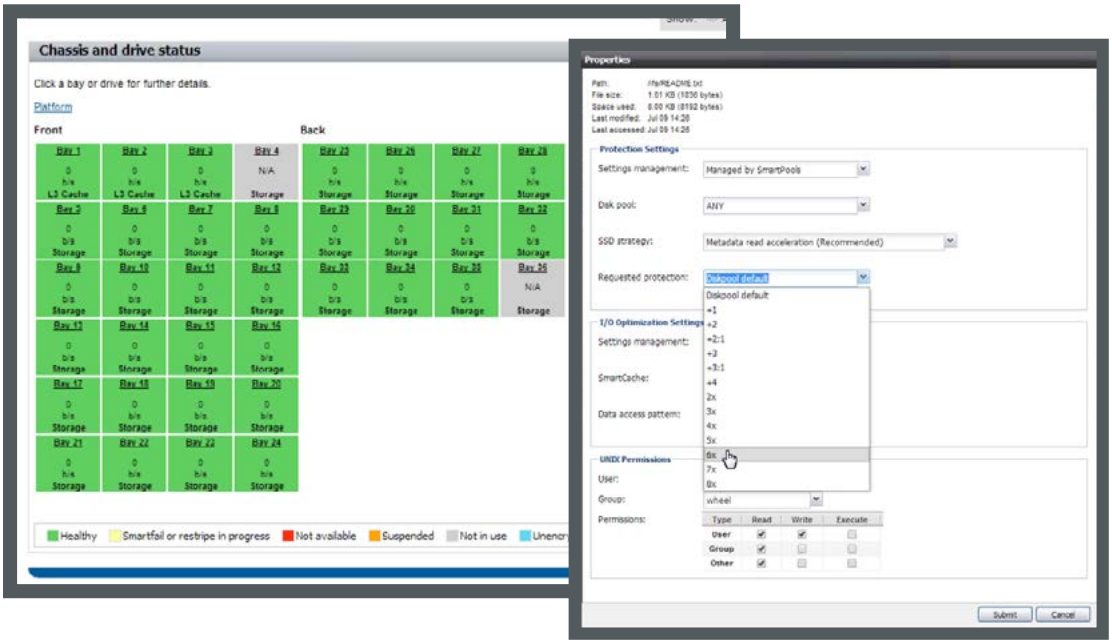
IDC Inference

Caching and SSD-based tiering capabilities of the EMC Isilon Scale-out Data Lake make it a suitable performance-optimized platform for shared mixed-profile Hadoop environments. However, it also provides the economics of capacity-optimized storage, thereby eliminating the need to archive post-processed data to another tier. The ability to handle concurrent Hadoop data streams make the EMC Isilon Scale-out Data Lake well suited for virtualized Hadoop workloads.

Validated: Isilon OneFS High Availability (Recovery from Disk-Level Failure)

Feature/Validation Summary

The EMC Isilon Scale-out Data Lake is designed to withstand one or more simultaneous component failures without preventing the cluster from serving data. It features a distributed RAID (Reed Solomon encoding, or mirroring as needed). When there is a component failure, such as a disk failure, OneFS only recovers file data that is compromised by the failure as opposed to the entire volume. Furthermore, because metadata and inodes are also protected by node-level mirroring in addition to being distributed across all the nodes in the cluster, a disk-level failure seldom causes performance degradation.



IDC validated that a disk failure on a single node has no noticeable impact on the cluster. Furthermore, the operation of replacing the drive is a seamless process and has little administrative overhead, no different than enterprise disk storage systems. This is in contrast to DAS in which the process of replacing a drive is rather involved and time consuming.

IDC Analyze the Future

Continued on next page

Validated: Isilon OneFS High Availability (Recovery from Disk-level Failure)

Continued from previous page

Validation Process

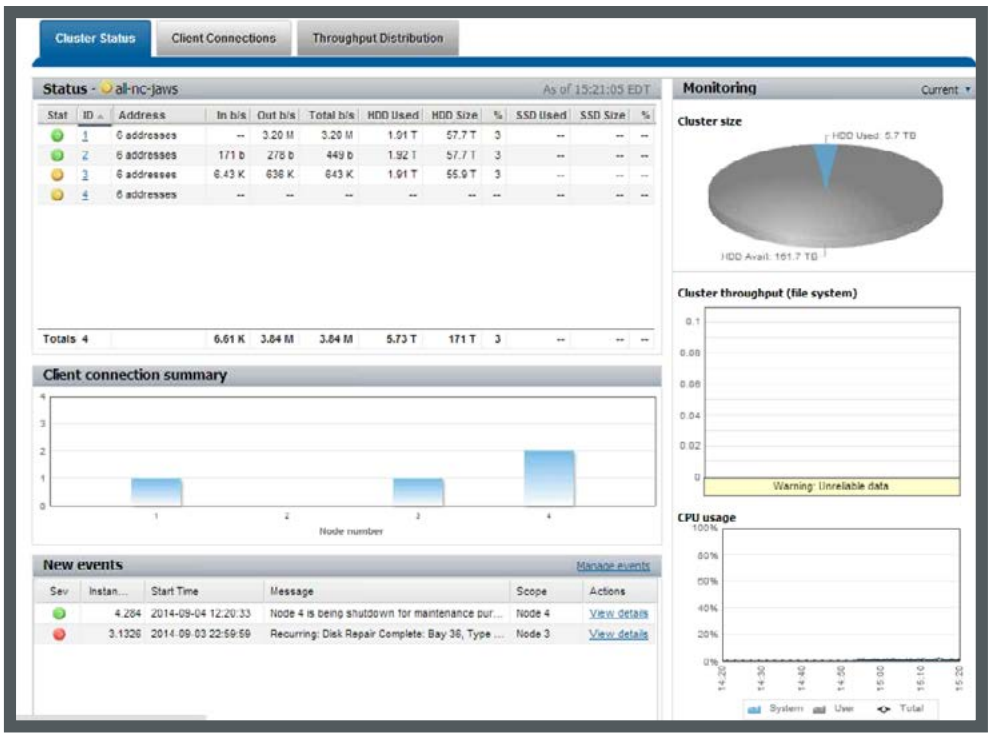
For this validation, disk failure was simulated using the 'smartfail' utility on the EMC Isilon Scaleout Data Lake. The 'smartfail' process ran until all data on the 'failed' drive was safely migrated to other drives in the cluster. When the 'smartfail' action completed, the drive status changed to 'REPLACE', at which point it was safe to remove the drive. This is a similar situation to a failed drive, or a physically displaced drive, except that in both these cases, the system automatically checks and balances under replicated blocks to other spare disks on the same node in the cluster. Similarly disk failure on the Hadoop DAS cluster was simulated using standard UNIX commands (which included forced unmounting of the file system setup on this drive on one of the worker nodes).

On the Isilon OneFS, data redundancy can be set at the file or file type level, directory level or at the disk pool level (default is +2.1 – i.e., any two drives can fail or a node can fail). The administrator specifies a cluster's Protection Level (i.e., the number of simultaneous failures of disks and/or nodes the cluster can tolerate before data is lost). OneFS responds to this setting by striping data appropriately. In the event of hardware failure, or the administrator changing the Protection Level, the FlexProtect job runs stripe rebuilds as needed. (The priority of the FlexProtect job can also be changed as needed.) Choosing a low Protection Level increases available capacity while simultaneously increasing the risk of data loss.

Validated: Isilon OneFS Lake High Availability (Recovery from Node-Level Failure)

Feature/Validation Summary

Data protection on the Isilon OneFS is not just limited to intra-node component failures (disk failures as an example), but also node-level failures. By default, OneFS can recover from a single-node failure without any performance degradation. However, this parameter is configurable to make the Data Lake more resilient to node failures.



IDC validated that a single-node failure has no noticeable impact on the cluster. Furthermore, the operation of removing a node from the cluster and adding it back to the cluster is a seamless process. Again, this was compared to the the process on removing a node and adding it back on the Hadoop DAS cluster, with the net result that the latter operation is far more cumbersome than in the Data Lake.

Continued on next page

Validated: Isilon OneFS Lake High Availability (Recovery from Node-Level Failure)

Continued from previous page

Validation Process

For this validation, node failure was simulated by simply powering off one of the healthy nodes of the EMC Isilon Scale-out Data Lake. When the node was shutdown, IDC observed that the IP-address configured on that node failed over to another node. IDC further observed that by default when a node is shutdown, OneFS does not start the process of replicating/balancing the cluster, with the assumption that an offline node is a temporary maintenance situation. However, the process of cluster rebuild (i.e., the process of replicating/balancing the cluster) was simulated by initiating the 'smartfail' of the node. A similar failure was simulated on the Hadoop DAS Cluster by powering off one of the nodes. Compared with the EMC Isilon Scale-out Data Lake, the Hadoop Cluster took a longer time (10 minutes before the data node was declared dead). During this time, the cluster continued to send jobs to the failed node, but returned errors. After the node was marked 'dead', the Hadoop cluster began the process of rebuilding/rebalancing the cluster automatically.

During both simulations, the Terasort job was run on the Hadoop Data Lake and Hadoop DAS clusters during and after the node failure (Teragen was used earlier to populate the data set for sorting). Results were noted and are summarized in the table below.

	Terasort (MB/sec)	% of baseline
Baseline (4 Isilon nodes)	642	100%
During smartfail of 1 node	429	67%
After smartfail of 1 node	507	79%
Baseline (6 DAS nodes)	416	100%
During rebalance from 1 node failure	123	29%
After rebalance from 1 node failure	356	86%

Key Finding: High Availability

Policy-based high availability capabilities are a must for enterprise adoption of Data Lakes

Capability	Key benefits	Why does it matter?
Recovery from disk(s) failure	No disruption of normal operations during intra-node component failures	Increased operational resiliency of the EMC Isilon scale-out Data Lake
Recovery from node-level failure	No disruption of normal operations during a single node failure. Limited degradation in performance during rebuild process	Built-in protection allows seconds to recovery, while preserving data integrity, ingest and access

IDC Inference

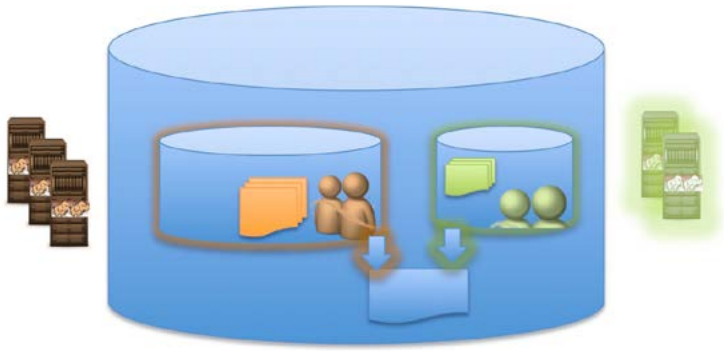
The EMC Isilon Scale out Data Lake provides robust data availability and protection, which is in line with most enterprise storage platforms. Furthermore, component and node-level failures do not cause a noticeable drop in performance, especially during and after the rebuild process. The process of recovering from such failures is also seamless and resource-friendly. This is in stark contrast to the limitations and overhead posed by a standard Hadoop cluster built with commodity components.

Note: IDC did not validate the site-level protection and resiliency capabilities of the EMC Isilon Scale-out Data Lake, but acknowledges that such capabilities are essential to a Data Lake.

Validated: Access Zones and Access Control Lists

Feature/Validation Summary

Access zones provide a method to logically partition cluster access and allocate resources to self-contained units. They are a crucial part of the multi-tenancy capabilities of the Isilon OneFS. They provide logical isolation and a mechanism that partitions the OneFS cluster into multiple authentication and access zones.



IDC validated that access zones do indeed provide no-crossover isolation between two separate Hadoop clusters – with different (local) authentication domains and data sets. IDC also validated sharing of data across access zones.

Validation Process

For this validation, the Isilon Cluster was configured with two additional access zones (in addition to the system zone). Each access zone was configured with two separate user lists (with non-conflicting UIDs/GIDs) and data sets that were accessible via HDFS. Two independent Hadoop clusters were then configured to access each zone respectively. IDC also validated the sharing of data between access zone by the creation of a soft link (UNIX symbolic link) from one access zone to another.

Notes:

- Validation was performed using local authentication, however access zones also work with other mechanisms like NIS, LDAP and Active Directory.
- Access zones provide isolation at the user level, therefore it is important to have unique UIDs/GIDs for each user in each cluster that needs to be isolated. The system zone can also be used to isolate users. However for consistency and symmetry, separate access zones were created for each isolation domain.
- The only legitimate mechanism to share data between two access zones is via symbolic links – which provide a referential path to a shared target directory in another access zone, from the initiating access zone.
- File permissions are important to ensure that users from a different access zone have the appropriate read, write and execute access.

Key Finding: Secure Multi-Tenancy

EMC Isilon Scale-out Data Lake provides a secure multi-tenant environment for multiple Big Data workloads (Hadoop). The shared storage model provides fine-grained control and sharing capabilities

Capability	Specific features	Why does it matter?
Access zones and access control lists	Independent Hadoop clusters can access different resources on same Isilon Cluster (partitioned users and data sets)	Logical separation of data on a shared/unified Data Lake
	Selectively share data between two or more access zones, based on referential links and file/ directory permissions	One data set shared between two or more clusters

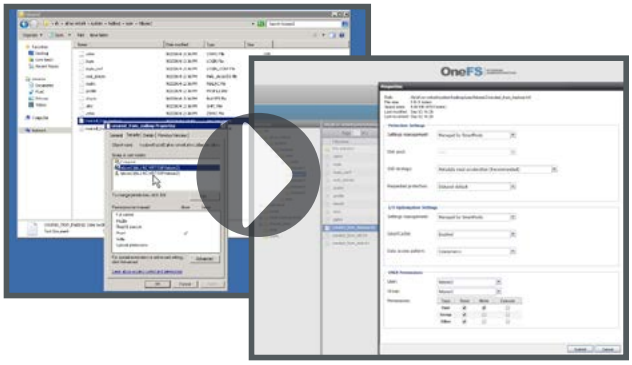
IDC Inference

The EMC Isilon Scale-out Data Lake provides deep, unified and scalable storage for Big Data workloads. It also provides an efficient mechanism for optimizing the number of data copies, by allowing multiple Big Data workloads (Hadoop clusters) to operate on the same data set while providing isolation capabilities to restrict access between logically separate data sets and users.

Validated: User-Level Authentication and Authorization

Feature/Validation Summary

The EMC Isilon Scale-out Data Lake provides multiple local and directory-based authentication and authorization schemes. A core component of secure multi-tenancy is the ability to provide a secure authentication and authorization mechanism for local and directory-based users and groups.



IDC validated that the EMC Isilon Scale-out Data Lake provides federated user-level authentication and authorization. User-level permissions are preserved across protocols, which include NFS, SMB and HDFS.

Validation Process

This validation is an extension of the secure multi-tenancy validation process outlined previously. For this validation, four separate users and groups were created in each of the Access Zones. UID and GIDs were matched on the EMC Isilon Scale-out Data Lake, for NFS to work properly (SMB and HDFS use a username/password combination). Files were then created from user account in NFS, and then accessed via the same and different user accounts via SMB and HDFS. The steps were repeated across various protocols for read and write access.

Notes:

- For a large installation, a directory service like LDAP and Active Directory is recommended to avoid potential UID/GID conflict, and provide centralized security and identity management.
- Local SMB users are able to modify file permissions, but cannot change authenticated users because of the inability to do a user look-up in a directory.
- For additional security, Kerberos can be enabled.
- Isilon supports authentication providers like NIS, Active Directory and LDAP along with local accounts.

Validated – Smart Lock (Sec 17a/4 Compliance)

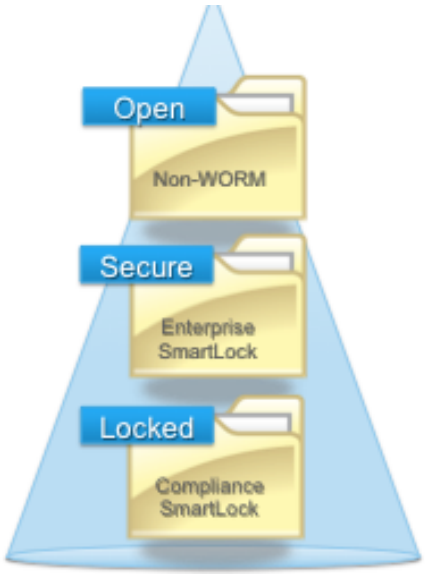
Feature/Validation Summary

SmartLock is a key security feature of the OneFS platform. It is specifically designed for deploying secure and compliant (SEC 17a/4) Enterprise Data Lake platforms. SmartLock has degrees of security Enterprise (Secure) and Compliance (Locked). It provides two operational components.

- The ability to restrict admin access cluster-wide
- The ability to control WORM (write once, read many) attributes on files and directories

The SmartLock Compliance mode is a cluster-wide setting, and when this mode is enabled the root user account is locked (this can be setup during initial configuration). All tasks are performed only by a special user account called 'compadmin' (short for compliance administrator), where commands are logged for auditing. This user can only run a preset list of commands as the privileged user that are assumed to be compliant.

As far as the WORM capabilities are concerned, they can be set in two modes on a per-resource level. The 'enterprise mode' WORM flag provides a per-directory setting of WORM attributes, but allows 'compadmin' to delete files before the retention period expires. When the 'compliance mode' WORM flag is set, files and directories cannot be deleted before the retention period expires. WORM modes are compliant across protocols, and cannot be bypassed. Files can be enrolled via any protocol, or locally on the OneFS cluster nodes.



IDC validated both Enterprise and Compliance modes of the OneFS platform. IDC validated the restrictions placed on the compadmin user when the cluster-wide compliance mode was set. Furthermore IDC validated both enterprise and compliance WORM modes for Hadoop workloads.

Continued on next page

Validated – Smart Lock (Sec 17a/4 Compliance)

Continued from previous page

Validation Process

For this validation, a virtual instance of the OneFS cluster was used (see notes). During the initialization process, the compliance option was selected thereby forcing the use of 'compadmin' for performing all approved tasks (via the sudoers files). The second step in the process was to set a compliance date – a one-time operation that enables a hardware-based compliance clock on the cluster nodes. This was followed by the creation of a user account that would be used to validate read-only permissions. The next step was to create a directory and set WORM attributes on it (using the 'worm domain' option, 'compliance' flag enabled and 'default-retention' flag set to one day). For the SmartLock compliance test, IDC validated that the file was locked by removing write permissions from all users. However, setting the the UNIX 'access time' (access time, before removing write permissions) allows for an explicit write lock expiration time which defaults to 1 day in this case. During this time it was readable (but not writeable) via NFS or HDFS.

Notes:

- SmartLock flags can be set at the directory level. Each directory has its own unique set of SmartLock permissions.
- Retention of files can be set in three ways. Files can get locked if it has not been accessed for a certain period of time. Any user with rights to change permissions can specifically remove all the write permissions. Finally, the user can also set the access time (UNIX atime) on the file to manually trigger the WORM flag.
- Default retention period: Admin can specify how long the file remains locked. After retention period expires, file can only be deleted – permissions cannot be changed.
- Disabling SmartLock once it has been enabled requires that the cluster be reformatted.

Key Finding: Security and Compliance

Federated security is an essential attribute of an Enterprise Data Lake Platform with the ability to maintain confidentiality and integrity of data irrespective of the protocols used

Capability	Specific features	Why does it matter?
Identity-based permissions	Federate access via a "one user, one identity" across multiple protocols (SMB, HDFS and NFS)	Multi-protocol ACLs that work with HDFS, NFS and SMB
	Logical separation, permission-based separation	Maintaining file-level security/data integrity in compliant environments
Smart Lock (Sec 17/a4 Compliance)	Restrict admin/privileged user in a compliant environment	Limiting the number of commands that can be run by a privileged user in a compliant environment
	Set WORM attributes on files and directories, so they cannot be deleted and/or updated by Hadoop users	Maintaining WORM file-level security and data Integrity in compliant environments

IDC Inference

The EMC Isilon Scale-out Data Lake provides a federated security fabric across the entire Data Lake. It brings enterprise-grade governance, regulatory and compliance (GRC) capabilities to Big Data environments.

Validated: Storage Pools

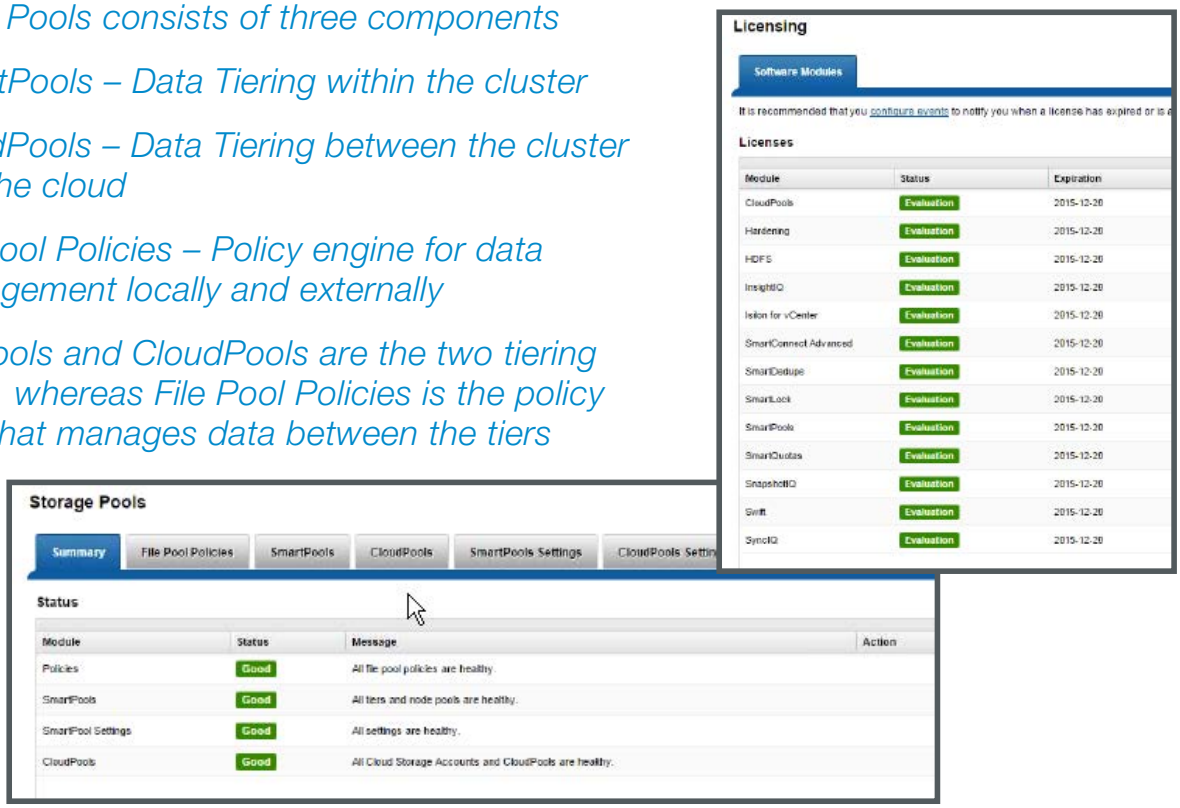
Feature/Validation Summary

Isilon OneFS allows the ability to manage data within the cluster, and extend this to the cloud (with OneFS 8.0). Known as Storage Pools, this capability allows administrators to apply common file policies across the cluster locally and extend them to the cloud.

Storage Pools consists of three components

1. *SmartPools – Data Tiering within the cluster*
2. *CloudPools – Data Tiering between the cluster and the cloud*
3. *File Pool Policies – Policy engine for data management locally and externally*

SmartPools and CloudPools are the two tiering engines, whereas File Pool Policies is the policy engine that manages data between the tiers



Notes:

- No additional hardware is required for Cloud Pools – it is a licensed feature
- CloudPools and SmartPools use the same policy engine
- Files placed in the cloud are stubbed locally so appear as online files to the user
- SmartPools is a requisite license for CloudPools

Validated – Storage Pools

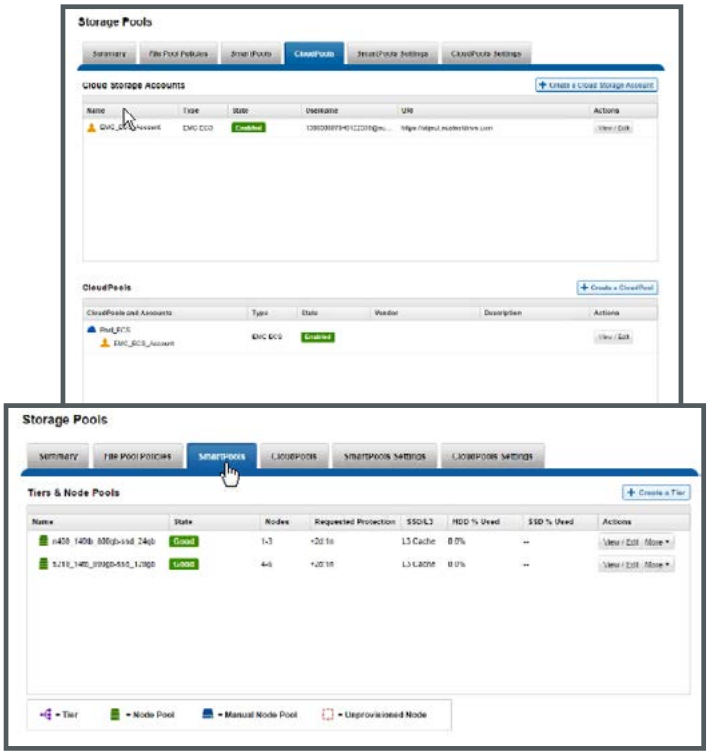
Step 1/2 – SmartPools and CloudPools (Policy-based data placement within Isilon cluster, and the cloud)

Feature/Validation Summary

SmartPools simplifies data management within the cluster by allowing files to be moved between "performance optimized" and "capacity optimized" cluster nodes. This greatly simplifies management overhead, as movement is automated via policies.

CloudPools is the ability to tier the Isilon cluster to a public cloud, private cloud (via an Object API) or to a remote Isilon Cluster. CloudPools supports EMC ECS, Amazon S3, Microsoft Azure and a remote Isilon cluster (Remote Access Node).

IDC validated the implementation of SmartPools in OneFS. IDC also validated the process to create CloudPools. Steps in the process are (1) Creation of a cloud user account, and (2) Choosing a logical container (creation of a "Cloud Pool").



Notes:

- CloudPools and SmartPools use the same policy engine
- Files placed in the cloud are stubbed locally so appear as online files to the user
- In case of public cloud, it is essential that the customer have some kind of relationship with a public cloud IaaS provider like Amazon or Microsoft
- Files placed anywhere in the cluster appear local and online files to the user, as they are part of the same global namespace

Validated – Storage Pools

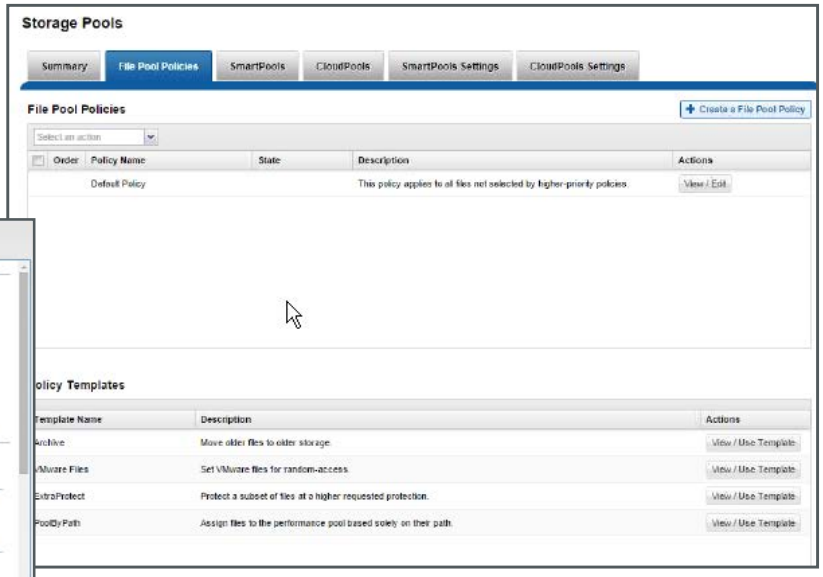
Step 2/2 – File Pool Policies

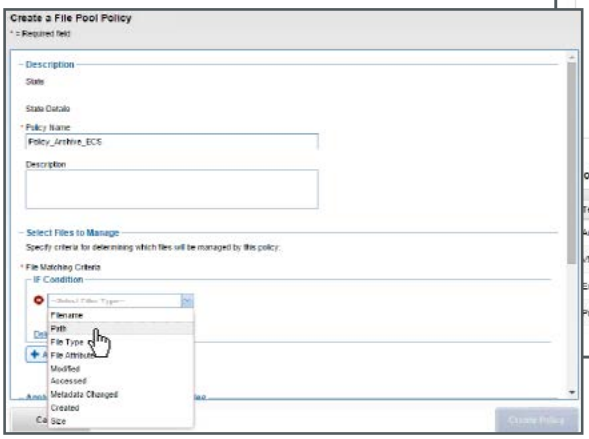
Validation Process

SmartPools, this capability allows administrators to apply common file policies (File Pool Policies) for data movement within the cluster (via SmartPools) and between the cluster and the cloud (via CloudPools).

File Pool Policies for data movement within the cluster (via SmartPools) and between the cluster and the cloud (via CloudPools).

IDC validated the creation of File Pool policies – the policy engine that allows data tiering between different cluster configurations, and the cluster and the cloud.





Notes:

- File Pool Policies are executed the same way regardless of whether they are SmartPool policies or CloudPool policies
- File Pool Policies can be applied on a schedule, or manually via the command line (either by running the policy, or archiving a single file)
- When the file is tiered to the cloud it is stubbed, whereas when it is tiered within the cluster it is relocated within the namespace
- Tiering to and from the cloud can be controlled granularly

Key Finding: Simplified Data Management

Simplified Data Management within the cluster and to/from the cloud is an essential functional characteristic of an Enterprise Data Lake Platform

Capability	Specific features	Why does it matter?
SmartPools	Data management tiering within the cluster	Essential for tiering between performance optimized and capacity optimized cluster nodes
CloudPools	Data management tiering between the cluster and the cloud	Essential for implementing a hybrid cloud, and placing archive data on a low-cost (cloud) tier
File Pool Policies	Policy engine for data management locally (within the cluster) and between the cluster and cloud	Essential for automating data movement within the cluster, and the cloud.

IDC Inference

The EMC Isilon Scale-out Data Lake provides a federated data tiering scheme across the entire Data Lake. With this feature IT administrators can right size the infrastructure by automating data placement on the right tier.

Validation Test Bed

IDC performed the validation at EMC's labs in North Carolina. The storage test bed consisted of Isilon Clusters and Hadoop Clusters. The table below provides a summary of the test environment.

Function	Components	Configuration	Validation specifics (if any)
Enterprise Data Lake Platform (EDLP)	4-node Isilon X410 cluster	Each 4U X410 node configured with dual Intel Xeon CPUs, 64GB RAM, 57.7 TB raw (Total cluster size is 231TB), 3.2TB SSD, 2x1GbE and 2x 10GbE SFP	For testing purposes two onboard 10GbE NICs were used. Two access zones were mapped to two subnet pools to provide IP-based isolation
Hadoop Data Lake Compute only Cluster	7-node Hadoop cluster	1 master-node and 6 worker-nodes (Each worker node configured with 16 Xeon 2.8GHz CPUs – 32 logical CPUs, 64GB RAM, 8 10K RPM 300GB HDDs) Cloudera Hadoop distribution (CDH5)	Name nodes setup to access an NFS data store. YARN used for benchmark testing
Hadoop DAS Cluster	7-node Hadoop cluster	1 master-node and 6 worker-nodes (Each worker node configured with 16 Xeon 2.8GHz CPUs – 32 logical CPUs, 64GB RAM, 8 10K RPM 300GB HDDs) Cloudera Hadoop distribution (CDH5)	Name nodes setup to access an NFS data store. YARN used for benchmark testing
SMB access	Windows 2008 R2 Server	Configured with the same user name as the one used by Hadoop for access zone –€common for both tests	
Script Server	Linux	Configured with the same user name as the one used by Hadoop for access zone –€common for both tests	

Notes:

- All Hadoop compute nodes were virtualized on a vSphere cluster.
- One virtual machine per physical server (vSphere host) to minimize any side effects of virtualization.
- All Hadoop nodes could access internal disk resources in the physical server (vSphere host) or access the Data Lake via 10GbE network connection.
- VMware Big Data Extensions were used for Hadoop nodes, VMware Big Data Extensions is an automated Big Data provisioning and management solution. It lets administrators deploy and centrally manage Hadoop and HBase clusters.
- Master node roles: HDFS Balancer, HDFS Secondary Name Node, Hive Gateway, Hive MetaStore Server, YARN Resource Manager, Zookeeper server, Hive Server2.
- Isilon cluster setup with its own DNS zone, delegated from the master running on a Windows server.

Essential Guidance: Advice for Buyers

A Data Lake should be a part of every Big Data workflow in the enterprise. By consolidating storage for multiple workloads onto a single shared storage platform, buyers can reduce costs and complexity in their environment, and make their Big Data efficient, agile and scalable. Furthermore, a data lake should not only cater to the Hadoop workload performance needs but also the needs of other workloads that use it as a reliable enterprise class store.

IDC believes that EDLPs should be a core part of enterprise storage infrastructure strategy. As businesses learn to collate data from various sources and convert it into consumable nuggets of information for their various organizational units, they will no doubt be compelled to establish enterprise-wide Data Lakes — upon which various workloads can concurrently operate. Such Data Lakes will enable existing workloads, as well as be future proof to seamlessly support new applications and workloads.

IDC concludes that EMC Isilon possesses the necessary attributes such as multi-protocol access, availability and security to provide the foundations to build an enterprise-grade Big Data Lake for most big data Hadoop workloads.

IDC Validation Methodology

This Lab Validation Brief provides a summary of an extensive validation process performed by IDC in collaboration with the supplier's teams. IDC relied on the supplier's equipment, facilities and their configuration to perform this validation. All of the tests were conducted during the presence of one or more IDC Analysts.

This Brief is meant to provide a quick set of inferences and insights for IT professionals and business decision makers seeking to perform further due diligence on the capabilities of the product and/or services that have been validated in this Brief. However, the goal of this Brief is not to supply detailed hands-on test plans and validation jobs. It is not meant to replace the evaluation process that most businesses will conduct before making any decision to purchase the product and/or services.

It is for this reason that this Brief is not designed to be an all-inclusive document on all the capabilities of the product, but rather as a concise document that highlights features/functions of products, their relative performance with respect to a traditional environment and the value these features bring to businesses looking to solving certain problems for Hadoop workloads.

Finally, even though this Brief is a sponsored document, it is not meant to be an IDC endorsement of the product, service or the sponsoring supplier. IDC's opinions are its own and not influenced by the production of this document.