THE DELL EMC ISILON SCALE-OUT DATA LAKE

ABSTRACT

This white paper provides an introduction to the Dell EMC Isilon scale-out data lake as the key enabler to store, manage, and protect unstructured data for traditional and emerging workloads. Business decision makers and architects can leverage the information provided here to make key strategy and implementation decisions for their storage infrastructure.

October 2016

The information in this publication is provided “as is.” DELL 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.

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

DELL EMC is now part of the Dell group of companies.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Data is growing rapidly as the numbers of people using digital devices interact with systems and networks grow across the globe. A Majority of this data growth is in the unstructured form—and as organizations are experiencing, contains valuable insights that could be used to improve business results. The technology to capture, store and analyze this data is maturing rapidly, as enterprises are looking for ways to effectively handle the data growth. By some industry estimates, over 80 percent of the new storage capacity deployed in organizations around the world will be for unstructured data.

Dell EMC® Isilon® scale-out network-attached storage (NAS) provides a simple, scalable, and efficient platform to store massive amounts of unstructured data and enable various applications to create a scalable and accessible data repository without the overhead associated with traditional storage systems. Isilon enables organizations to build a scale-out data lake where they can store their current data, and scale capacity, performance, or protection as their business data grows in the future. The scale-out data lake helps lower storage costs by efficient storage utilization, eliminate islands or silos of storage, and lower storage management costs of migration, security and protection.

Dell EMC Isilon OneFS®, the intelligence behind Isilon systems, is the industry's leading scale-out NAS operating system, which is known for its massive capacity, operational simplicity, extreme performance, and unmatched storage utilization. With proven enterprise-grade protection and security capabilities, Isilon is an ideal platform to meet a wide variety of storage needs.

AUDIENCE

This white paper is intended for business decision makers, IT managers, architects, and implementers. By leveraging the Isilon scale-out data lake—a key enabler for storing and managing massive quantities of unstructured data—enterprises can build their storage strategies and implement their infrastructure to maximize the return of their IT investments.

TERMINOLOGY

The acronyms used in this paper are summarized in the Table 1.

ACRONYM	DESCRIPTION
CIFS	Common Internet File System
DAS	direct-attached storage
HDFS	Hadoop Distributed File System
LAN	local area network
NAS	network-attached storage
NFS	network file system
SAN	storage area network
SMB	Server Message Block

Table 1. Acronyms used in this paper

OVERVIEW

According to an IDC "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things" study sponsored by Dell EMC, the digital universe will grow from 4.4 zettabytes (1 trillion gigabytes) in 2013 to 44 zettabytes in 2020, doubling in size every two years. Although enterprises create only 30 percent of this data (1.5 ZB in 2013), they come in contact with 85 percent (2.3 ZB in 2013) of it, and have some liability associated with the data. IDC further estimates that 22 percent of the data is a candidate for analysis and contains valuable information that organizations can use to make critical decisions. Figure 1 shows this trend for enterprise data.

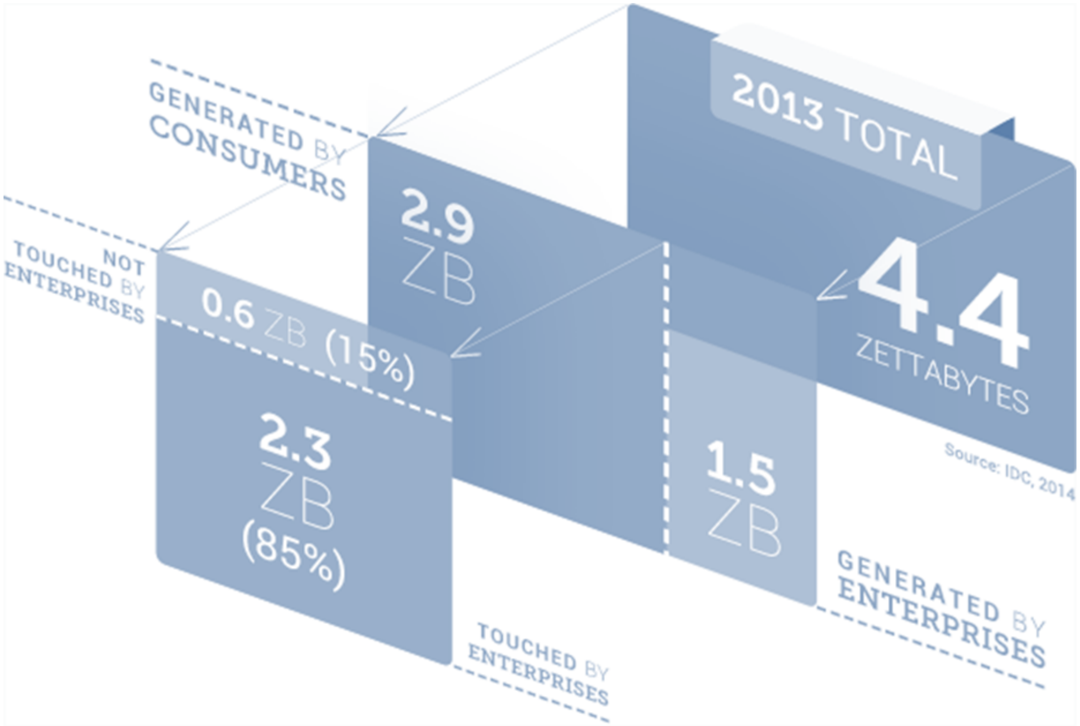

Figure 1. Enterprise data[1]

Not all of this data is stored by the enterprises in the datacenter for cost, liability or process reasons.

Isilon scale-out NAS provides key capabilities to build a data lake that physically de-couple compute from storage without affecting the seamless nature of the data flow outlined below. You can leverage a data lake to get the most value from your data. Data from a variety of sources can be converged using native protocols, protected, secured and exposed to analytics systems that drive value, organizations can plan, implement and manage loosely coupled but seamless storage and applications.

DATA FLOW

Organizational data typically follows a linear data flow starting with various sources both consumer and corporate; ingested into a store, analyzed and surfaced for actions that augment value creation for an organization; as shown in Figure 2. The five distinct stages are typically broken down into three broad categories of data, analytics and application for easy reference from a holistic point of view.

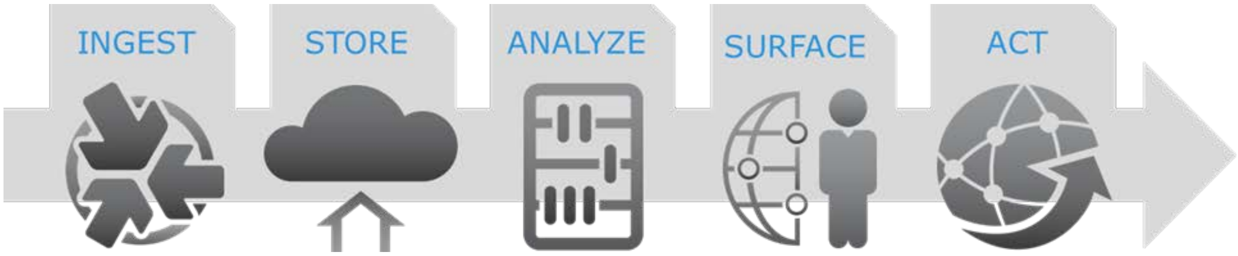

Figure 2. Data flow

In the following section we detail each of the stages in the data flow process as we understand the implications of each of these phases of the data flow. The choices at each of these stages have a profound impact in deriving value of the organizations data and ultimately the data store.

[1] Dell EMC Digital Universe Study—with Research and Analysis by IDC, http://www.emc.com/leadership/digital-universe/index.htm

INGEST

Data Ingestion is the process of obtaining and processing data for later use by storing in the most appropriate system for the application. An effective ingestion methodology validates the data, prioritizes the sources and with reasonable speed and efficiency commits data to storage. The velocity, volume and variety tax the capture speeds, throughput and efficiency of the ingest systems and therefore the store as discussed in the following section. The ingest process can act as the starting point of a storage silo if the organization planners do not look at the dataset holistically.

High velocity data- characterized by a continuous feed of data requires a specialized ingestion mechanism that typically captures the continuous stream and commits the records in batches of varying sizes to storage for further processing. Capture strategies come in two flavors, commit records to storage directly or use higher speed buffers as an intermediate step before storing to a more persistent downstream system. Examples of high velocity data include clickstreams, video surveillance feeds etc.

The volume of data is not only a factor of the velocity but also the size of the data set generated in batches of non-streaming origins. High volume data typically requires time to move from one point in the system to another depending on the network speed and places burdens on both the network and storage. Optimization for volume as in large files can affect velocity and vice versa. High definition video and audio are the typical examples of high volume data in the form of files- separate from streaming.

Variety places translation challenges at the data, file and protocol levels as disparate organic systems may need varying levels of translation for data to be of value for the downstream processes. A combination of CRM, LOB, social media, web and mobile information on customers is a good example of variety of data.

STORAGE

Storing data is typically dictated by the type of storage strategy namely block or file, the flow mechanism and the application. Over the years, storage has evolved into an optimization problem between storage costs and performance. As the volume and variety of data grows, the cost to reliably store, manage, secure, and protect it grows as well, particularly for data that is subject to compliance and regulatory mandates like personal identifiable information (PII), financial transactions, and medical records. Adding the availability component adds cost pressures at the expense of performance and vice versa.

The segregation started by the choice of ingestion system is further exacerbated by storage bringing about true silos or islands of information catering to various application requirements of real-time, interactive or batch processing. As silos permeate the IT infrastructure, hot spots arise in systems of heavier use, while capacity goes unused elsewhere. Downstream applications, compliance, security and data policies can contribute to create further silos within silos- giving rise to a very complex system.

ANALYSIS

Data analysis technologies load, process, surface, secure, and manage data in ways that enable organizations to mine value from their data. Traditional data analysis systems are expensive, and extending them beyond their critical purposes can place a heavy burden on IT resources and costs. Integrating data from disparate sources adds complexity and management overhead that can be a deterrent to most organizations looking to derive value from their data.

APPLICATION: SURFACE AND ACT

Post analysis, results and insights have to be surfaced for actions like e-discovery, post mortem analysis, business process improvements, decision making or a host of other applications. Traditional systems use traditional protocols and access mechanisms while new and emerging systems are redefining access requirements to data already stored within an organization. A system is not complete unless it caters to the range of requirements placed by traditional and next-generation workloads, systems and processes.

BUSINESS CHALLENGES

As organizations face large datasets, data and application growth, they are observing a tremendous increase in costs both CAPEX & OPEX; limitations in the ability to realize the value of the data; and protection and compliance lapses to name a few. These challenges can be attributed to a combination of factors throughout the data flow process but fall in the following broad categories:

SILOS OR ISLANDS OF DATA

The ingestion strategy employed for real-time, interactive or batch applications originate a silo that diverges as the data flows downstream. For example, if an organization choses to setup a combination of SAN- buffer stream of data and NAS - persistent storage for a real-time application like customer targeting; while elsewhere, CRM analysis happens on a combination of DAS and cloud. A third silo consisting of archived data uses a combination of DAS and NAS to run batch analytics- a typical scenario in organizations today.

In the first silo, the compliance requirement for handling of PII data can pressure admins to choose between applying policy to the entire silo or carve out a protected zone if the technology so supports. This adds complexity to the system without the inclusion of access control for financial data to meet SEC requirements; or HIPPA for medical data. If the design of the system does not support protections, organizations risk going out of compliance with large fines at a minimum and liability of data leaks causing massive lawsuits on the other extreme.

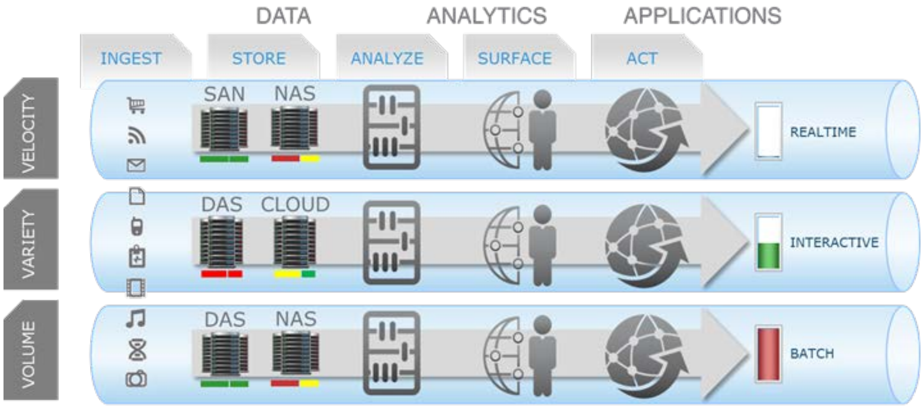

Figure 3: OneFS Job Engine Job Descriptions

Also, the silos of data adds cost pressures due to inefficiencies addressed in the following section, locking of insights within silos at the expense of business and management and protection inconsistencies. In many organizations, applying learnings from a batch processing system to a real-time system may involve lengthy change management processes- creating friction across departments or adding barriers to value generation.

INEFFICIENCIES ACROSS THE SYSTEM

If one dataset of say 10 terabytes is used for three different analyses by three systems, you will have a minimum of three copies requiring three times the storage capacity. If this dataset grows by a 30% a year the scaling requirement grows by 90% annually- demonstrating inefficiencies experienced by organizations at the most basic level. If one silo has lower utilization than the other, hotspots arise in the system while capacity goes unutilized elsewhere. Inefficient utilization of the budget can arise as a result of low value data residing on high cost, high performance storage.

Organizations are faced with management, datacenter footprint, power and cooling inefficiencies due to silos and hotspots in addition to reconfigurations, migrations and complicated maintenance activities.

SECURITY AND COMPLIANCE

Organizations with silos are faced with duplicate and inconsistent application of policies, security measures and governance policies, whereas organizations with shared infrastructure face access control violations. Working around these is typically time-consuming, painful and diverts resources away from standard procedures as issues arise. Securing data against leaks or destruction- both accidental and malicious; presents a separate set of challenges. These challenges are compounded in regulated sectors like finance, healthcare and government.

TIME TO INSIGHTS

More users than ever before are mobile or geographically dispersed with access to a larger dataset as a basic requirement to perform their duties effectively. Traditional systems operated within the confines of Corporate IT through a regulated and measured interconnections with the external world. With the growth of cloud, mobility and devices; the capabilities of traditional systems are being challenged in new ways increasing the time required to access, process and consume insights.

Figure 4. Enterprise storage challenges

Providing access while enforcing policies consistently across a wide range of approved and unapproved devices and platforms is a growing challenge for today's storage administrators. Traditional systems, silos and implementation strategies add latencies as protection and security layers are added around the application or sources of information.

DATA LAKE

The data lake represents a paradigm shift from the linear data flow model. As data and the insights gleaned from it increase in value, enterprise-wide consolidated storage is transformed into a hub around which the ingestion and consumption systems work (see Figure 4). This enables enterprises to bring analytics to data and avoid expensive costs of multiple systems, storage and time for ingestion and analysis.

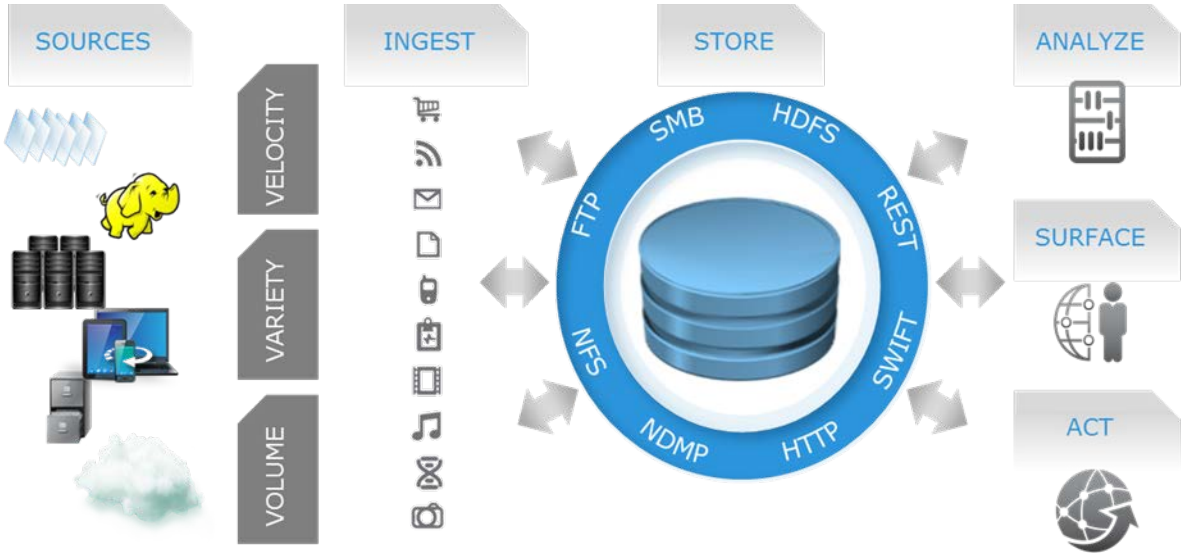

Figure 5. Scale out data lake

By eliminating a number of parallel linear data flows, enterprises can consolidate vast amounts of their data into a single store—a data lake—through a native and simple ingestion process. This data can be secured and analysis performed, insights surfaced, and actions taken in an iterative manner as the organization and technology matures. Enterprises can thus eliminate the cost of having silos or islands of information spread across their enterprises.

The scale-out data lake further enhances this paradigm by providing scaling capabilities in terms of capacity, performance, security, and protection. The key characteristics of a scale-out data lake are that it:

- Accepts data from a variety of sources like file shares, archives, web applications, devices, and the cloud, in both streaming and batch processes
- Enables access to this data for a variety of uses from conventional purposes to next-gen mobile, analytics, and cloud applications
- Secures and safeguards data with the appropriate level of protection from highly critical data like medical records, financial transactions, credit card data, and PII to website logs and temporary data that might not require any security
- Scales to meet the demands of future consolidation and growth as technology evolves and new possibilities emerge for applying data to gain competitive advantage in the marketplace
- Provides a tiering ability that enables organizations to manage their costs without setting up specialized infrastructures for cost optimization
- Maintains simplicity, even at the petabyte scale

ISILON SCALE-OUT DATA LAKE

Isilon enhances the data lake concept by enriching your storage with improved cost efficiencies, reduced risks, data protection, security, compliance & governance while enabling you to get to insights faster. You can reduce the risks of your big data project implementation, operational expenses and try out pilot projects on real business data before investing in a solution that meets your exact business needs. Isilon is based on a fully distributed architecture that consists of modular hardware nodes arranged in a cluster. As nodes are added, the file system expands dynamically scaling out capacity and performance without adding corresponding administrative overhead.

Multiple access methods

Isilon natively supports multiple protocols like SMB, NFS, File Transfer Protocol (FTP), and Hypertext Transfer Protocol (HTTP) for traditional workloads, and HDFS for emerging workloads like Hadoop analytics. By the very nature, a shared storage with multiple access methods eliminates storage silos bringing efficiencies and consistency to your IT infrastructure. This enables batch, real-time or interactive applications and systems to store and access data from one shared storage pool without the need for any migrations, loading, or conversion. Accessing data for read or write purposes are achieved at the protocol level. This implies that data can be created by any of the myriad systems, ingested into the data lake using a natively supported protocol such as SMB for Windows or Mac, and accessed or modified seamlessly using another protocol like NFS, FTP or HDFS.

Multiprotocol support enables the storage infrastructure to provide access to applications that leverage third platform protocols like HTTP or HDFS, which drive emerging workloads. Adding future protocol support, data access mechanisms, or interfaces can be easily achieved to scale the interoperability of the data lake. Without a data lake, interoperability would require an expensive and time-consuming sequence of operations on data across multiple silos, or even costly and inefficient data duplication.

Cost efficiency

You can achieve great cost efficiencies by investing in the storage capacity and capability required today- smaller than traditional or Hadoop requirements; scale in smaller steps proportionately to the data growth; tiering data according to value and; simplified management.

Isilon scale-out NAS can scale from 18 terabytes to over 50 petabytes in a single cluster so you do not have to overprovision your storage infrastructure. With over 80% utilization, you can keep your CAPEX in check with a smaller footprint. SmartDedupe provides the ability to reduce the physical data footprint by locating and sharing common elements across files with minimal impact to performance during writes or concurrent reads. You will observe a reduction in storage expansion costs, typically in the range of 30% with smaller storage capacity, power and cooling and of course less rack space requirement.

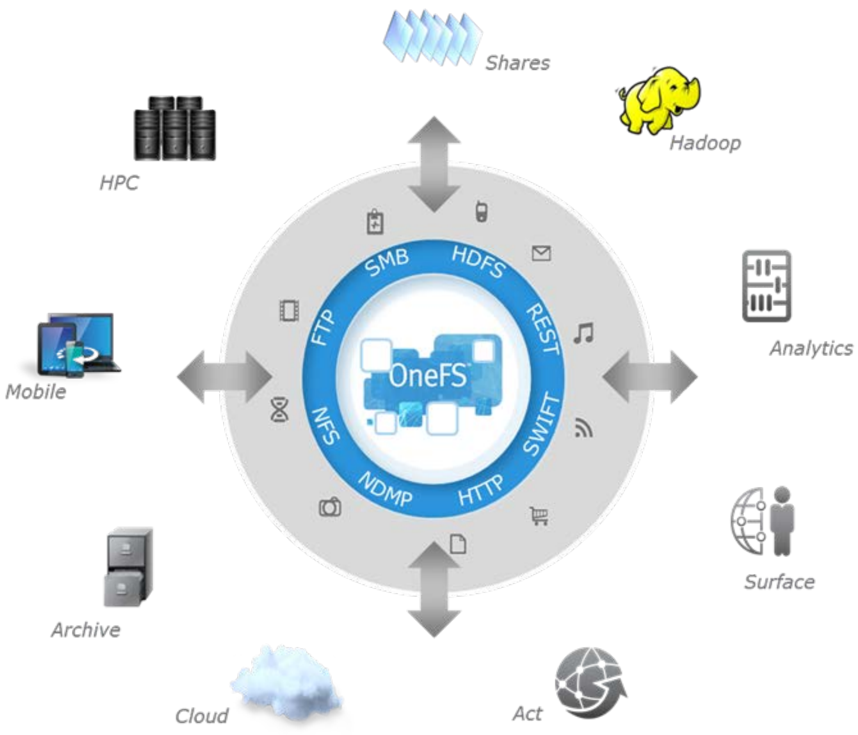

Figure 6. Isilon Scale-out data lake

Isilon enables you to tier the data lake to further drive cost efficiencies by optimizing performance, throughput, and density. Using policy-based tiering, you can reduce the cost of storing your data based on the inherent value and utility of the data. Tiering using Dell EMC Isilon S-Series, X-Series, NL-Series and HD-Series nodes is shown in Figure 7. High-value, readily needed data can be kept at a high-performance tier, whereas low-value, infrequently used data can be moved to a more cost-effective active archive without having any impact on your application or analytics infrastructure.

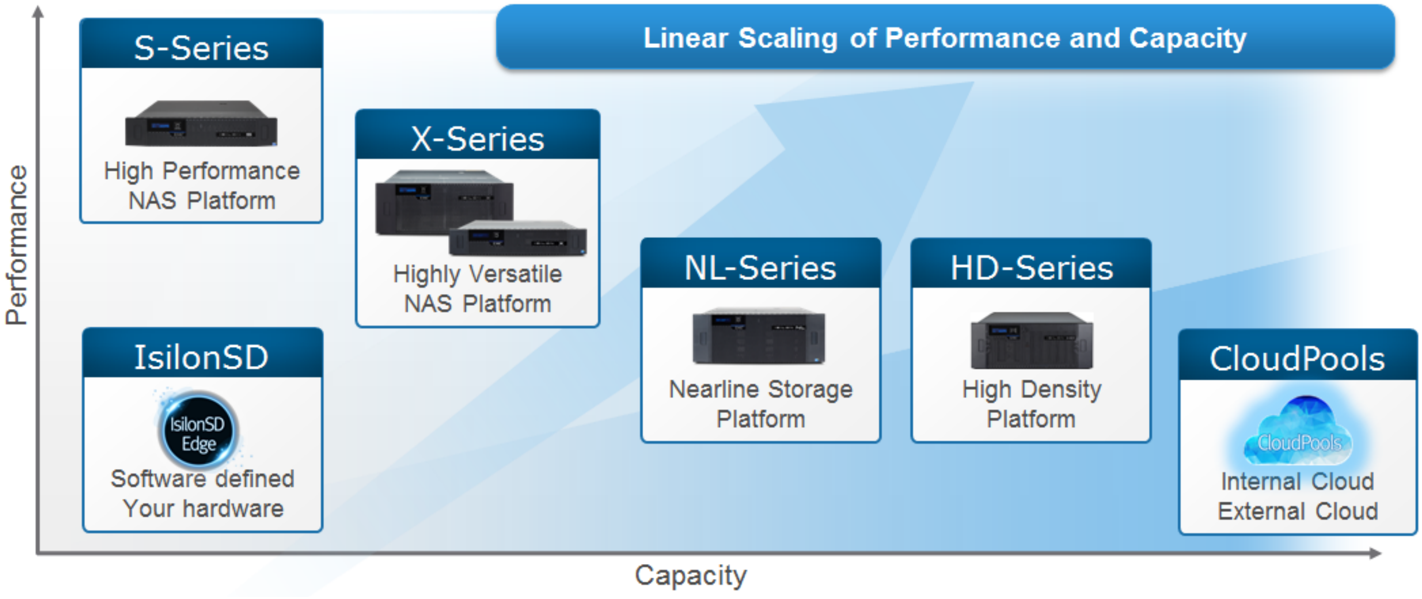
Linear Scaling of Performance and Capacity
S-Series
High Performance NAS Platform
X-Series
Highly Versatile NAS Platform
NL-Series
Nearline Storage Platform
HD-Series
High Density Platform
IsilonSD
IsilonSD Edge
Software defined Your hardware
CloudPools
CloudPools
Internal Cloud External Cloud
Performance
Capacity

Figure 7 . Tiering using Isilon storage nodes

Under the covers, an Isilon cluster provides automatic storage balancing and deduplication that not only enhances storage efficiency but also enables storage administrators to weather hardware outages better.

Reduce risks

Any large and growing data infrastructure risks can be classified as

- Ability to successfully implement an initial solution
- Ability and flexibility to scale solution as the environment changes
- Protect against current and future data loss
- Protect against data theft

These risks are further amplified as the dependency between the ingestion, storage and analytics system remains strong. With a loosely coupled system you will be able to build mitigation strategies. A data lake based on Isilon is able to provide mitigation for all these components. By decoupling storage from compute and using multiple access methods, you can deploy a storage system capable of ingesting the data you need for your solution effortlessly. By using protection and security features outlined in the following sections, you can ensure your data is safe and resilient to external forces.

Isilon is the only scale-out NAS to support multiple distributions of Hadoop through native HDFS implementation. This allows you the flexibility to analyze, surface and act on data using the best tool for your application and scale storage compute or both as your needs change.

Protect and secure data assets

Isilon storage systems are highly resilient and provide unmatched data protection and availability. Isilon uses the proven Reed-Solomon erasure encoding algorithm rather than RAID to provide a level of data protection that goes far beyond traditional storage systems. With N+1 protection, data is 100 percent available even if a single drive or a complete node fails, which is comparable to RAID 5 in conventional storage. You can also deploy Isilon for N+ 2 protections, which allows two components to fail within the system, similar to RAID 6; N+3 or N+ 4 protections, where three or four components can fail, keeping the data 100 percent available.

Isilon is able to recover from hardware failures faster than traditional systems as recovery entails rebuilding lost data as opposed to the entire disk. In a data Lake, this can have a huge impact on the operation as storage is shared across systems with varying levels of performance demands. A truly distributed storage like Isilon can orchestrate all the nodes to participate in the restoration or recovery of data from the outages on a dedicated backend infiniband network speeding up recovery without impacting front end performance.

As organizations view data as an asset, securing data for inherent value is even more important than meeting regulatory compliance and corporate governance requirements. Isilon helps organization address security needs by providing robust and flexible security features outlined below as described in the figure below.

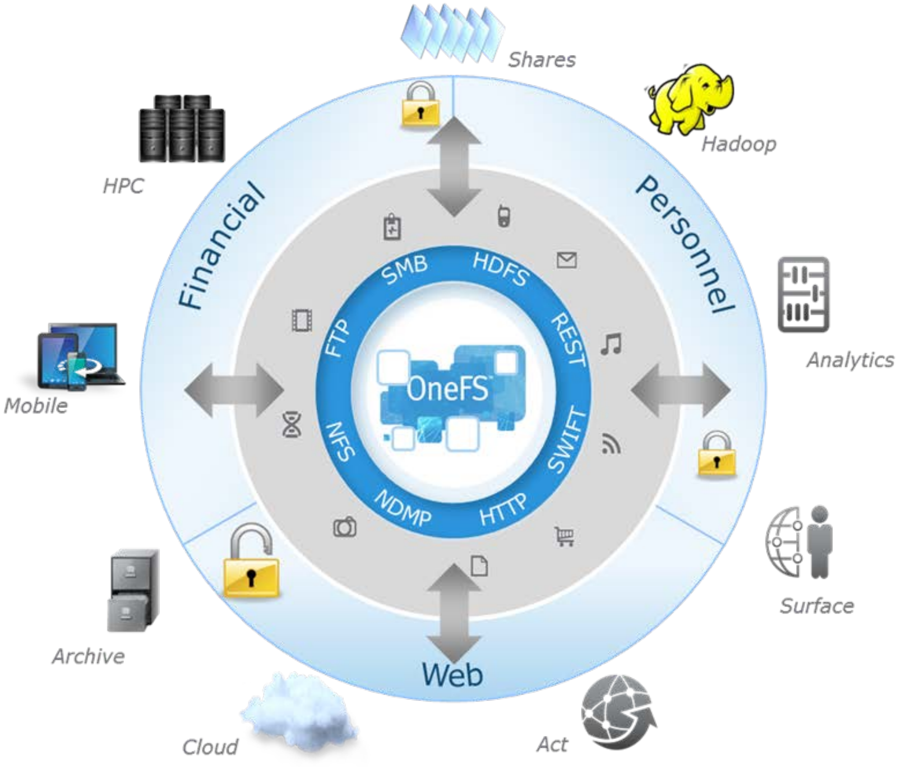

Figure 8. Authentication zones

- Secure role separation enables roles-based access control (RBAC), where a clear separation between storage administration and file system access is enforced.
- Authentication zones that serve as secure, isolated storage pools for insulating departments within the organization from access to data that they are not supposed to see. For example: legal, financial, PII and employee data can be seen by only authorized employees in the authentication zone.
- Write once-read many (WORM) protection is achieved by using SmartLock software, which prevents accidental or malicious alterations or deletion of data- a key requirement for governance and compliance.
- Data at rest encryptions through the use of Self-Encrypting drives enables organizations to ensure physical loss of hardware does not equate to data leak.

Faster time to insights

By utilizing a shared infrastructure, you can consolidate data from multiple islands into one single storage system based on Isilon. This is made possible through he multiple access mechanisms at the protocol level where data can be ingested into the Isilon store from a wide variety of sources and surfaced for use by others. This would eliminate the need for data migration or extraction-translation-loading (ETL) operations typical with any data analytics solution, saving you precious time and resources. You can then run Hadoop analytics with your dataset in-place. As depicted in the figure below. By using a large loosely coupled shared scalable storage on a typical dataset of around 100 terabytes, you can save over 24 hours of data moving time on a 10 GBps network; time that you can use to actually generate insights from your data.

Isilon is the only scale-out NAS that works with multiple distributions of Hadoop from a variety of vendors. This enables you to try out tools from all of these vendors at the same time if necessary to find the best solution to meet your business requirements.

The data lake is the key enabler for driving business value into customer environments through the multiprotocol, multi-access, tiered, single namespace, protected and scalable data repository. Leveraging the scale-out data lake, customers can consolidate multiple disparate islands of storage into a single cohesive and unified data repository that is easier to manage and more cost-effective.

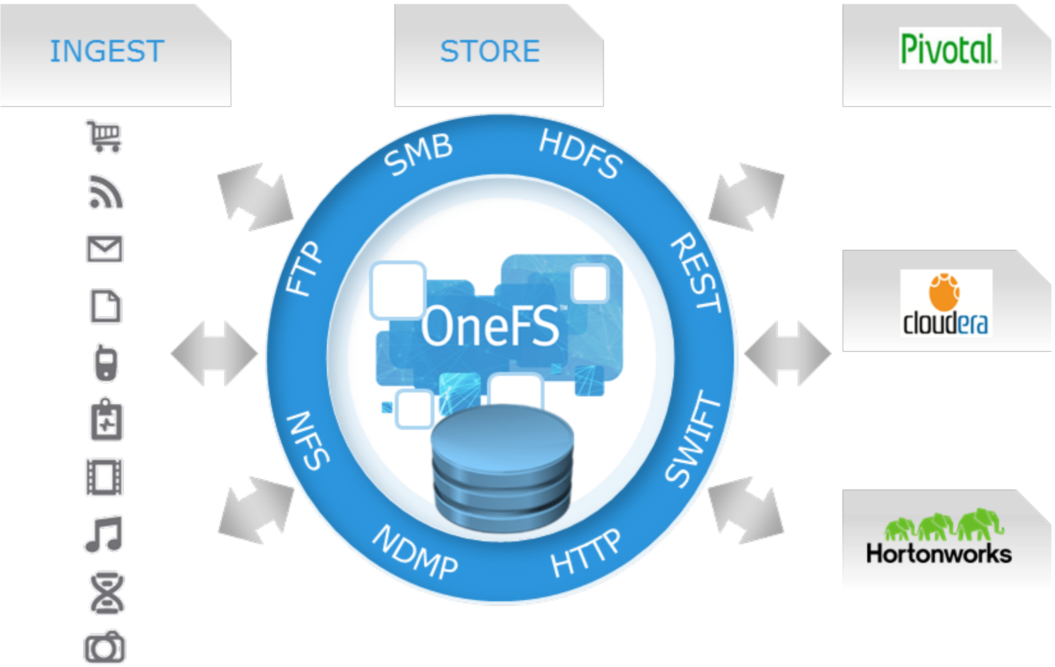

Figure 9. Faster insights

STORAGE AND DATA SERVICES

A scale-out data lake leverages the industry leading enterprise-grade storage and data services that extend the business value of your data. Isilon Infrastructure Software provides powerful storage management software that helps protect your data assets, control costs, and optimize the storage resources and system performance of your scale-out data lake.

The data management capabilities include Dell EMC Isilon InsightIQ®, SmartDedupe, SmartPools®, CloudPools, and SmartQuotas™, which together help you improve the ROI of your data lake infrastructure. For more information on these data management services, review the Isilon OneFS whitepaper available here.

The scale-out data lake is further strengthened by proven data protection capabilities provided by Dell EMC Isilon SmartConnect™, SmartLock®, SnapshotIQ™, and SyncIQ®. For more information on these data protection services, review the Isilon data availability whitepaper available here.

DATA LAKE 2.0 STRATEGY

With OneFS 8.0, Isilon introduced the data lake 2.0 strategy which extends the data lake from the core data center to address the needs of the enterprise edge with IsilonSD Edge and to the cloud with Isilon CloudPools software. IsilonSD Edge is a software defined storage version of OneFS that runs on your hardware on top of VMware ESX. It is best used for enterprises with a lot of branch or remote offices that need to simplify storage management and streamline backups by replicating the data from the edge to the core. IsilonSD Edge can scale up to 36 TB and can scale from a minimum of 3 nodes to a maximum of 6 nodes. It is also available as a “free and frictionless” download for non-production use.

CloudPools software helps enterprises optimize their storage by tiering cold or frozen data to your choice of public or private cloud providers. CloudPools is seamless to users and applications and can transparently tier the inactive data based on a flexible and powerful policy engine. As an extension to the SmartPools tiering, CloudPools can tier data to public cloud options like Microsoft Azure or Amazon AWS S3. Or, it can tier data to Dell EMC options including Virtustream, ECS or Isilon.

CONSUMPTION MODELS

Isilon provides a number of consumption models that enables you to choose a strategy that is best for your business. The simplest and most common strategy is an appliance that is a preinstalled combination of hardware and software. You can choose a converged infrastructure solution in conjunction with Dell EMC Vblock®. Or, you may choose a cloud-based utility infrastructure-as-a-service; in which you pay for the service based on usage. Or, you can use a software-defined storage offering of OneFS like IsilonSD Edge. The key is that you have choices in the storage purchasing model and flexibility to fit your business procurement needs.

CONCLUSION

Given that unstructured data will be doubling every two years, enterprises need higher efficiencies, architectural simplicity and more protection as they scale capacity and capabilities. A scale-out data lake provides key capabilities to eliminate silos of data; secure and protect information assets; support existing and next generation workloads while speeding time to insights. Starting with a scale-out data lake, organizations can

1. Invest in the infrastructure today to get started
2. Realize the value of data, store, process, and analyze it- in the most cost effective manner; and
3. Grow capabilities as needs grow in the future.

This enables organizations to store everything, analyze anything and build a solution with the best ROI. By de-coupling storage from analysis and application, organizations gain flexibility to choose between a larger number of strategies to deploy solutions without risking data loss, cost overruns and dataset leaks. The data lake based on Isilon offers organizations this along with a capability to simplify the IT infrastructure, tier, secure and protect data efficiently; and get to insights faster.

As a large dataset is the reason big data exists, organizations can start with the data. Understand the value locked in their large and growing datasets by running pilots; not worry about ingesting or surfacing and; pilot applications or emerging technologies- to make better informed strategic decisions.

REFERENCES

Dell EMC Digital Universe Study—with Research and Analysis by IDC (http://www.emc.com/leadership/digital-universe/index.htm)

Dell EMC Isilon OneFS Operating System http://www.emc.com/collateral/hardware/white-papers/h8202-isilon-onefs-wp.pdf

High Availability and Data Protection with Dell EMC Isilon scale-out NAS http://www.emc.com/collateral/hardware/white-papers/h10588-isilon-data-availability-protection-wp.pdf