PATHFINDER REPORT

Advisory

Managing the Data Explosion Challenge with Open Source Storage

MAY 2018

COMMISSIONED BY

About this paper

A Pathfinder paper navigates decision-makers through the issues surrounding a specific technology or business case, explores the business value of adoption, and recommends the range of considerations and concrete next steps in the decision-making process.

About 451 Research

451 Research is a preeminent information technology research and advisory company. With a core focus on technology innovation and market disruption, we provide essential insight for leaders of the digital economy. More than 100 analysts and consultants deliver that insight via syndicated research, advisory services and live events to over 1,000 client organizations in North America, Europe and around the world. Founded in 2000 and headquartered in New York, 451 Research is a division of The 451 Group.

NEW YORK
1411 Broadway,
Suite 3200
New York NY 10018
+1 212 505 3030

SAN FRANCISCO
140 Geary Street,
9th Floor
San Francisco, CA 94108
+1 415 989 1555

LONDON
Paxton House
(Ground floor)
30, Artillery Lane
London, E1 7LS, UK
P +44 (0) 207 426 1050

BOSTON
75-101 Federal Street
5th Floor
Boston, MA 02110
Phone: +1 617.598.7200
Fax: +1 617.357.7495

Executive Summary

Business IT is facing storage growth that's exceeding even the highest estimates, and there's no sign of it slowing down anytime soon. Unstructured data in the form of audio, video, digital images and sensor data now makes up an increasingly large majority of business data and presents a new set of challenges that calls for a different approach to storage. Next-generation storage systems need to provide greater flexibility and choice, as well as the ability to better identify unstructured data in order to categorize, utilize and automate the management of it throughout its lifecycle.

Software-defined storage (SDS) is rapidly becoming the preferred solution for 'secondary storage' applications that don't have the performance requirements of highly transactional workloads. As a result, every major storage vendor now offers scale-out SDS options based on commodity hardware. This levels the playing field for open source SDS offerings that offer the same flexibility of closely integrated block, file and object capabilities, as well as a modular, scale-out approach to capacity expansion.

A key consideration for open source SDS adoption lies in choosing a partner capable of providing enterprise-class service and support for software integration and development, as well as guidance in the process of migration from legacy to next-generation, open source SDS storage.

Key Findings

- Scale-out SDS systems combine enterprise-class data protection and management with the improved economics of commodity server hardware, Ethernet connectivity and whatever mix of flash and disk best suits application needs.
- In the secondary storage space, unstructured data is growing at a greater rate than traditional database information, yet it shares many of the same critical concerns when it comes to data protection, availability and long-term management.
- Systems that offer close integration between Portable Operating System Interface (POSIX)-based file services and metadata-rich object storage capabilities provide the best balance between the need to support existing applications and a flexible environment for metadata-based automation and management that extends the value of data throughout its lifecycle.
- Next-generation, open source platforms such as Ceph offer a viable alternative to proprietary secondary storage options, and offer a competitive mix of unified storage management, data protection and scale-out capabilities covering block, file and object storage.
- Because secondary storage is the ultimate repository of all IT information, it's important to thoroughly consider critical issues such as data protection, security, scalability, performance, availability, management and the flexibility to fit seamlessly into a hybrid strategy that increasingly extends beyond the traditional datacenter.

The Changing Nature of Enterprise Storage

REDRAWING THE LINES FOR PRIMARY STORAGE

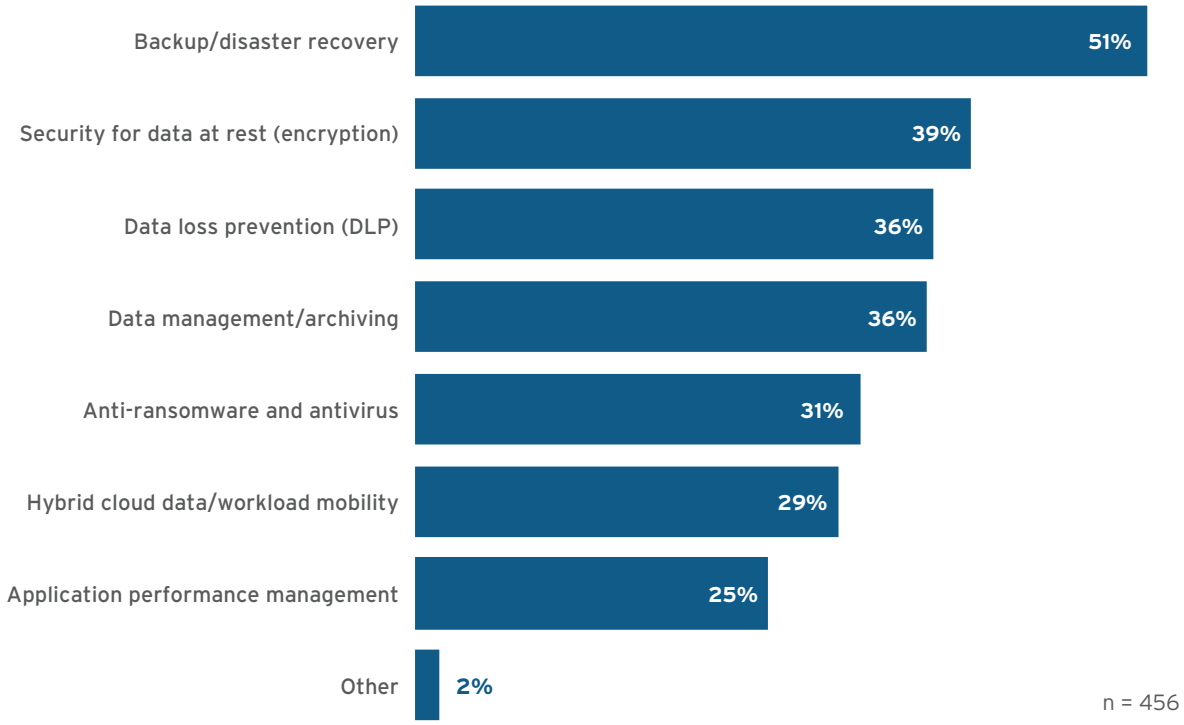
Enterprise storage as we know it today is just over 20 years old, and the monolithic storage area network (SAN) and network-attached storage (NAS) systems that made up the lion's share of enterprise storage have evolved based on the limitations of existing technology and the challenges of providing increasingly large amounts of well-protected, shared storage. SAN-based systems focused on providing high-performance block-level storage for large database and application server environments, and were typically connected via dedicated, Fibre Channel networks. SAN storage provided mountable block volumes called logical unit numbers, which offered a centrally managed alternative to onboard storage. Large-scale, enterprise NAS systems evolved in roughly the same time frame but, instead, specifically focused on providing shared, file-based services via Ethernet rather than block-level volumes.

The main difference between enterprise SAN and NAS was the file system because both still needed to create large, dynamic and protected pools of addressable storage from a massive number of individual disk drives. Of course, servers based on SAN storage could simply run local file services, and some NAS systems could provide block volumes as well, but the question ultimately came down to performance. SANs were considered optimal for performance-based workloads, and dedicated NAS systems were ideal for applications where optimized, shared file services were the primary consideration. We still believe that performance remains a priority when determining the best storage platform for a given use case, but

an increasing percentage of today's business applications focus more on the new challenges posed by unstructured data rather than database performance. The 451 Voice of the Enterprise (VotE) Storage survey data (Figure 1) shows that performance ranked at the bottom of customer expectations for integrated storage systems.

Figure 1: Services and features expected of integrated storage platforms

Q. Which types of products would you like to see integrated or bundled with your organization's current infrastructure platforms (CI, Servers, Storage)? Please select up to three.



Source: 451 Research Voice of the Enterprise: Storage, Budgets and Outlook 2017

Thanks to the substantial increases in the performance of CPU, memory, networking and storage devices, today's highest-performance storage systems are more than an order of magnitude faster than in the recent past, and we believe this substantially changes the entire formula of what workloads belong on primary vs. secondary storage. Highly transactional database applications remain the sweet spot for primary storage systems, and there is growing use of all-flash systems for ultra-high-performance analytics and technical computing applications that offer a new tier above primary storage to accelerate these new workloads. Fortunately, 'the rising tide lifts all boats,' so this increased performance translates to the next generation of secondary storage systems as well, making integrated, software-defined storage an attractive option to support a growing number of workloads, as well as providing a platform for delivering the new capabilities customers expect from an SDS system.

THE RISE OF SCALE-OUT STORAGE

Legacy SAN and NAS systems took a 'scale-up' approach based on proprietary, hardware-based storage controllers and the software necessary to combine thousands of small drives into a very large storage pool. This was no small task at the time and required highly optimized technology and specialized systems administrators to monitor, protect and deliver that pool across a wide variety of client systems. But the demand for storage capacity has been relentless, and scale-up systems eventually presented scalability problems because these controllers were only capable of handling a finite amount of capacity. When organizations reached that capacity limit, they either had to replace the controllers or add an entirely new system. Either option was a substantial undertaking.

Over the last decade, the scale-up approach to storage has evolved into a new approach known as 'scale-out,' which is based on clusters of individual, networked storage nodes that increase processing power and storage capacity on a linear basis. It also provides a model that offers data protection at the device, node and even cluster level. Although this can be done using proprietary node designs, most scale-out SDS systems today instead use commodity, x86-based servers as storage nodes. Server vendors have embraced this approach by designing storage-dense systems with flash and/or disk options that can substantially reduce the hardware costs for distributed storage, as well as accommodate the new, storage-specific features that enterprise customers expect.

Fortunately, the software tools for managing, monitoring and protecting SDS systems have improved along with the conveniences offered by commodity hardware, reducing many of the day-to-day challenges of storage administration and making it easier for storage customers at any skill level to build a flexible, scale-out SDS platform suitable for a broad variety of secondary storage applications. But the trend toward scale-out SDS for secondary storage is about more than new hardware and software options; it's about adapting to the changing nature of business data overall.

Today, business applications are increasingly dependent on file-based data such as documents, sensor data, images, video and other forms of media, and there's strong evidence that unstructured data growth is rapidly outpacing traditional database information across the IT industry. Perhaps more important, all this unstructured data presents serious challenges when it comes to identifying and categorizing the contents of billions of files. For these reasons, we believe that metadata-rich object storage offers a better model for long-term unstructured data management.

Object storage is fundamentally different from file systems because of the abstraction offered by the storage platform itself. When objects are 'ingested,' or initially added to the system, the data itself becomes immutable, and a metadata record is started that follows that data element for the rest of its life. This means the original data object can't be appended or modified, but elements within the metadata record can be updated. The metadata record of the object is fundamentally like any other database record, and the elements within that metadata record can be used by the storage platform itself to categorize, sort and define data management policies such as access rules, data protection, encryption and lifecycle management. It's this metadata-based model that affords object platforms practically unlimited scalability, as evidenced by the fact that there are now tens of trillions of objects residing on hundreds of exabytes of public-cloud-based object storage.

We believe that modern SDS platforms that tightly integrate the rich metadata capabilities of object storage with a POSIX-based file system offer the most flexible solution for addressing the challenges of legacy application support, data growth and hybrid cloud with the tools needed to identify and automate the management of this monsoon of unstructured data, both on-premises and in the cloud.

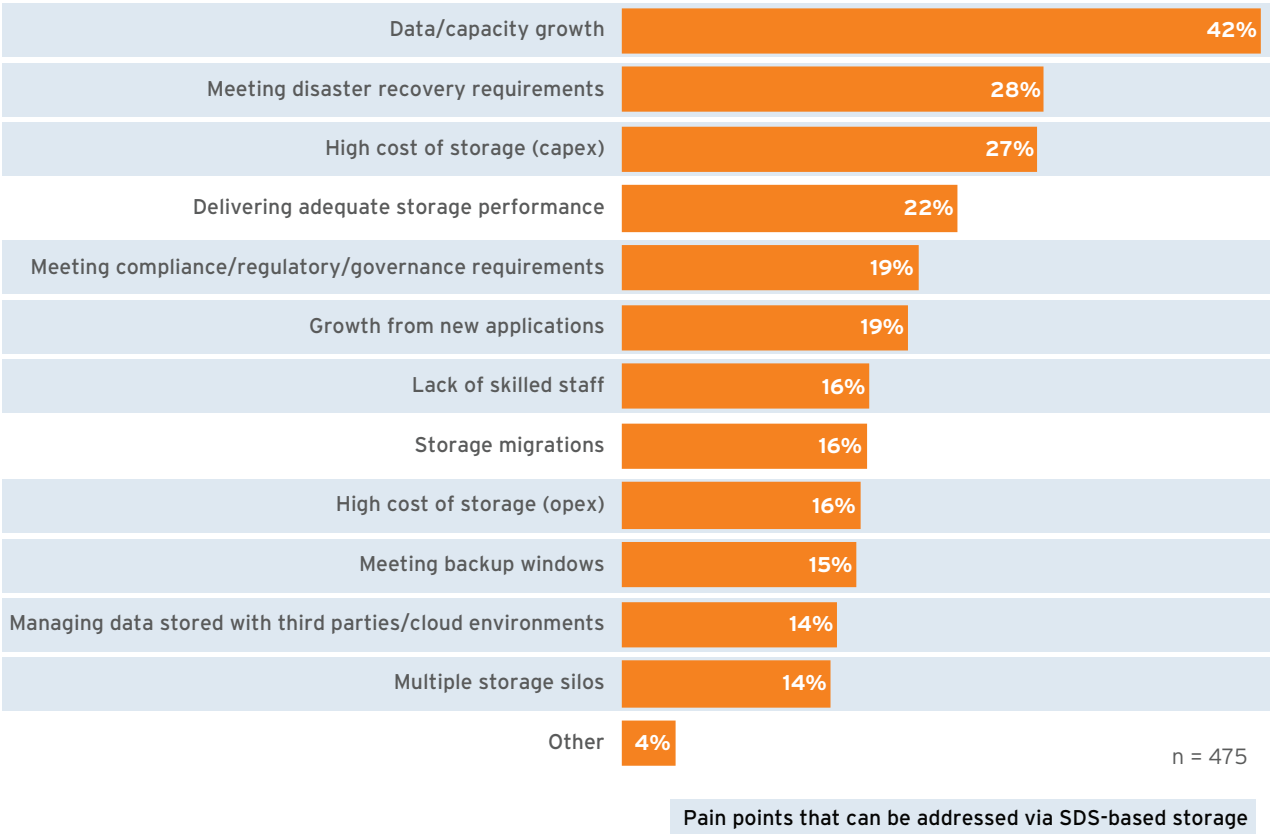
SDS – New Data Challenges Call for New Solutions

SIMPLIFYING STORAGE IN AN INCREASINGLY COMPLEX ENVIRONMENT

As we noted earlier, administering enterprise-class storage was a challenging and complex process that required highly specialized training to allocate resources, provide data protection, optimize performance and ensure system connectivity in a large storage infrastructure. Figure 2 below reflects many of these challenges, and highlights those we believe can be addressed by adopting an SDS platform that is based on commodity hardware and offers simplified management.

Figure 2: Top storage pain points that can be addressed by SDS

Q. What are your organization's top pain points from a storage perspective? Please select up to three.



Source: 451 Research Voice of the Enterprise: Storage, Budgets and Outlook 2017

The flexible capabilities and mid-level performance of most SDS platforms can make them an ideal choice for handling the storage needs of a wide variety of applications, especially when high performance isn't a critical requirement. With the convenience of modular scalability, the economy of commodity hardware, simplified management, and the ability to choose between a strong lineup of commercial and open source vendors, scale-out SDS is rapidly becoming the platform of choice for mid-range block, file and object services. This is especially true in the context of unstructured data, an area where we see a substantial need for improvement across the industry.

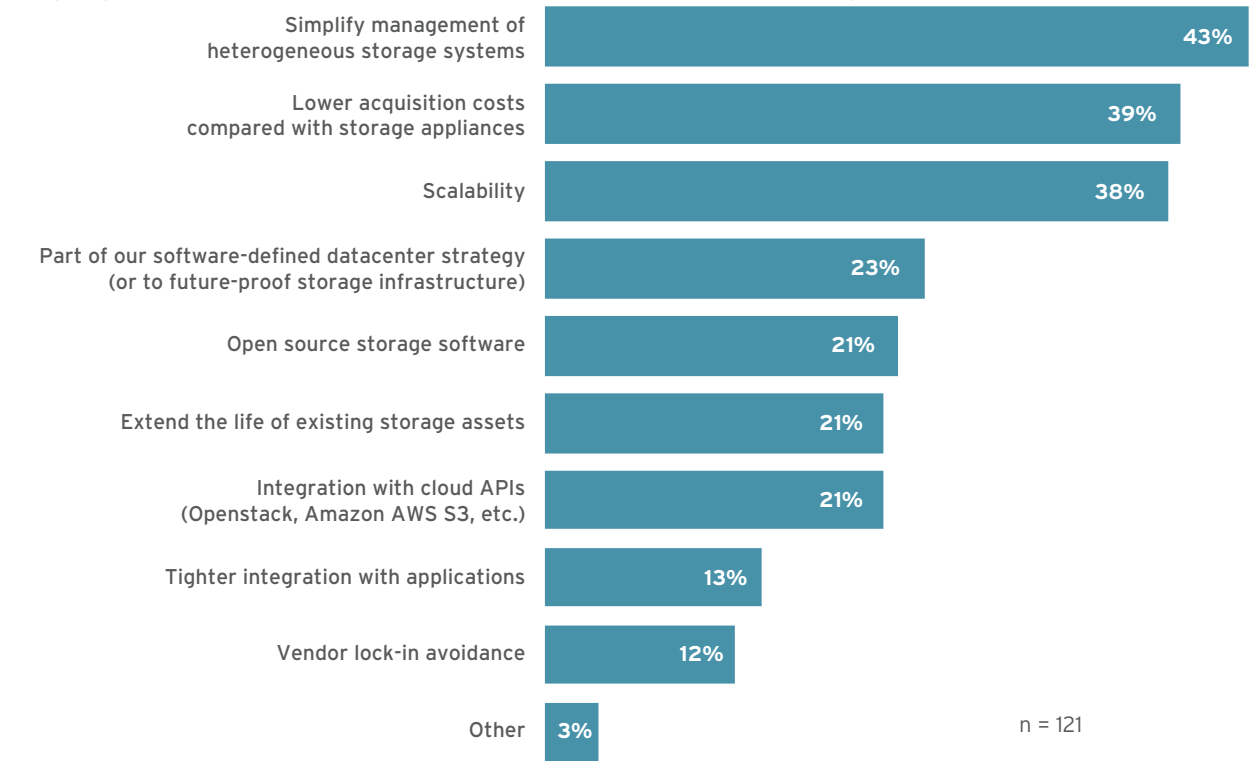
In the past, both business and IT personnel typically viewed unstructured data as a necessary but relatively minor annoyance that took up valuable space on costly primary storage, was relegated to an ad hoc mix of file servers spanning multiple locations, or worst of all, remained spread across hundreds or thousands of user end points. As a result, a large majority of today's unstructured data goes 'dark' as it moves through the traditional backup process. Historically, the cost and difficulty of adopting a cohesive, birth-to-death strategy for unstructured data simply wasn't worth the effort, but modern

businesses are becoming increasingly dependent on unstructured data, and there is a growing need to accurately classify, contextualize and triage that data to separate the data assets from useless or toxic data, as well as continue to extract value from the information that costs so much to store.

Rising storage costs consistently ranks at the top of business concerns across the industry, but data growth is only one part of a more complex equation. The greatest ongoing cost for IT technology usually lies in system support and management, and our latest VotE survey showed 'Simplify management of heterogeneous storage systems' at the top of the list of reasons for adopting scale-out SDS on commodity hardware.

Figure 3: Customer reasons for adopting SDS

Q. Why did your organization choose to use software-defined storage running on commodity x86 hardware? Please select up to three.



Source: 451 Research Voice of the Enterprise: Storage, Budgets and Outlook 2017

All of the options in Figure 3 actually contribute in some way to reducing the cost of storage, but the information in Figure 3 also reflects a customer shift away from proprietary storage toward more open storage systems for secondary storage applications. This data reinforces the growing interest in public cloud compatibility at the enterprise level and indicates an increased adoption of object storage in the form of on- and off-premises hybrid cloud storage.

Perhaps most interesting is that 'open source storage software' ranked in the middle of the results. Open source storage has been around for a long time, but it was more commonly found in large computing, academic and research environments rather than in enterprises. Although there are a variety of open source storage platforms, Ceph is currently the only open source SDS option that offers unified block, file and object storage capabilities suitable for enterprise-class applications, along with an integrated management platform and the availability of ongoing support options from key international Linux distributors.

There are also a substantial (and growing) number of commercial SDS offerings available from both traditional and emerging storage vendors that target secondary storage applications and embrace both file and object storage. Most, if not all, of these new offerings are available either as software or as pre-integrated storage appliances. In either case, we believe that the industry-wide availability of SDS reflects a growing desire for more flexible, open secondary storage options and validates the scale-out SDS model for an increasing number of enterprise applications.

Technology Considerations for SDS-based Secondary Storage Adoption

AS ALWAYS, PROCEED WITH CAUTION

Whether the underlying storage software is open source or commercial, adopting a modern SDS environment that will suit your long-term data challenges requires planning and forethought. The IT industry is still coming to grips with the problem of unchecked data growth, so the first step in a due-diligence process is understanding the nature and scope of your data problem, as well as the business challenges and types of applications that best serve those needs. Storage is, and should continue to be, one of the segments of an IT strategy where it pays to be risk-averse, but this doesn't mean that it's particularly dangerous to explore SDS-based, scale-out storage. It just takes a little common sense.

- **Data protection is still job one** – storage doesn't capture the imagination quite like other IT segments, but eventually, all things come from and return to the storage systems. Regardless of data type or source, it's universally critical to ensure that data protection is a top priority.
- **Not all commodity hardware is created equal** – for all their similarities, x86 servers are not all alike. One of the most interesting features of SDS is the ability to run on commonly available server technology, but when it comes to clustered, scale-out storage, it's usually advisable to narrow your hardware choices to pre-validated systems to ensure full compatibility and simplify management.
- **There's nothing like a strong partner** – simplicity in operation and management is the ultimate goal of SDS systems, but that simplicity is usually a result of up-front development to navigate the initial complexity of knitting together a system that serves so many purposes. There will be challenges to overcome, and it makes sense to choose a vendor that can offer enterprise-class technical support for the hardware, software and network integration work, as well as provide ongoing support and services throughout the system's life.
- **Head in the clouds, feet on the ground** – the hybrid approach to cloud adoption offers the greatest choice of price, features and services, so planning an SDS strategy that translates easily to public cloud storage offerings ensures greater flexibility when choosing between on-premises and off-premises options. Like it or not, Amazon's S3 API has become a de facto industry standard for object storage, so specifying S3 compatibility currently offers the greatest freedom of choice among competing public cloud storage offerings.
- **Try before you buy** – the modular approach of scale-out SDS substantially lowers the cost of entry for enterprise-class storage, making it relatively inexpensive and easy to test an open source SDS platform such as Ceph on a smaller scale, and then continue to grow the system as more applications arise. Although it's possible to run Ceph on repurposed hardware, it's generally advisable to use more modern and matched components to establish the actual performance parameters in production and leave the door open for expansion as the system proves itself.
- **SDS vs. HCI** – there is a lot of confusion between the terms software-defined-storage and hyperconverged infrastructure (HCI). Some vendors use HCI to describe any platform that's based on commodity server hardware and integrated storage, but we find it easier to draw the line based on tasking. In this case, dedicated, storage-specific nodes are most easily described as SDS, while nodes that run BOTH storage and virtualized production on the same hardware defines HCI. Regardless of the definition you choose, it's important to understand the impact that generalized virtual workloads mixed with SDS will have on the hardware and software cost of individual nodes, as well as the on the performance of a cluster as it scales.
- **Ground control to Major Tom** – network connectivity is one of the more commonly overlooked issues of SDS adoption. Traditional SANs were often based on a separate and dedicated Fibre Channel network that only carried storage traffic, while both SDS and HCI use shared Ethernet for networked storage and application IP traffic. Modern 10Gb Ethernet networking is proving more than sufficient to support a wide variety of mixed storage/application use cases, but it's also possible that sharing a common network could create resource contention between the largely east-west storage traffic, and the traditional north-south movement of application traffic as it scales.
- **Know thyself** – we believe that automation can be the key to managing the challenges of unstructured data growth, and that metadata-rich object storage provides the framework for granular data management, regardless of physical location. The next challenge lies in understanding the nature of your unstructured data and the key information that should be gathered as metadata to classify, protect and manage data over the long term.

- **Continue to seek wisdom** – content-based management is both old and new, and a growing number of vendors are dedicated to helping establish a model, mechanism and workflow for identifying unstructured data through metadata. Our survey shows that more than half of enterprises today look to their vendors for guidance on defining and collecting germane and functional metadata.

We believe that next-generation SDS technology offers benefits that will prove its value, especially through the granular data management automation that's enabled by better defining the contents and context of unstructured data. Like any other modernization initiative, dealing with the problem of the unstructured data explosion will take time and require a change in perspective to manage data based on the information it contains rather than blindly by its size or age.

This a difficult problem, made worse by the colossal amount of legacy data that's been piling up across the industry. Simply classifying unstructured data can be a daunting task, and establishing a reasonable return on investment for more intelligent data management can be difficult. However, from a business perspective, it may not be long before the risks of *not* managing data could far exceed the cost of fixing the problem. The GDPR legislation alone has penalties of up to €20m or 4% of a company's annual revenue, whichever is higher, and in 2016, the US Financial Industry Regulatory Authority fined 12 US firms $14.4m for deficiencies in preserving records in a format that prevents alteration. Understanding the contents of your unstructured data and establishing policies and automation accordingly is changing from a 'nice to have' capability to a necessity in order to protect your data from loss, corruption and exposure, as well as to address the growing legal requirements for identity protection.