451 Research
Now a Part of
S&P Global Market Intelligence

PATHFINDER REPORT

Hybrid File Storage

Meeting the New Challenges of Unstructured Data Management

COMMISSIONED BY

COHESITY

JUNE 2020

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 THE AUTHOR

STEVEN HILL

SENIOR ANALYST, APPLIED INFRASTRUCTURE AND STORAGE TECHNOLOGIES

Steven Hill is a Senior Analyst of Applied Infrastructure and Storage Technologies. He covers the latest generation of hyperconverged systems, hybrid storage, business continuity, disaster recovery, and unstructured data management, governance and archiving technologies for enterprise customers.

Executive Summary

In the classic world of information services (IT), most of the challenges of enterprise-class storage were addressed on-premises via large, purpose-built storage area network (SAN) systems for primary block applications and network-attached storage (NAS) systems that focused on shared file-based storage. These hardware-based, scale-up, on-premises options remain a common model for enterprise storage, but software-defined storage (SDS) has evolved to offer an alternative path for enterprise-class storage that combines high-performance commodity server and storage hardware with next-generation storage software based on a distributed, multi-node, cloud-friendly model.

The cloud delivery model has changed everything, and business IT is now shifting toward a hybrid cloud model that allows customers to seamlessly adopt a combination of on- and off-premises compute and storage resources based on the best combination of cost, scalability, performance, availability and developmental flexibility. At the same time, the makeup of data itself has been changing, and there is a distinct trend toward file-based, unstructured data over traditional database information. While both forms of data can be business-critical, unstructured data presents a different set of challenges, especially in an environment where hybrid applications and their data increasingly operate outside the traditional datacenter firewall. In this paper, we look at the evolving challenges of unstructured data in the context of hybrid cloud, and some of the new opportunities made available by a cloud-native approach to data management.

Key Findings

- Business data continues to grow at an unprecedented pace. The problem of unchecked data growth has been a constant refrain for decades, and we expect it to grow by about 25% in 2020.
- Storage budgets aren't keeping up. Our polling has shown that average storage spending is expected to increase by roughly 10% annually.
- Unstructured data is making up an increasing majority of business data growth. File-based data in the form of documents, sensor data, email and social media interactions, as well as audio, video, images and other highly dense media is growing much faster than structured database information.
- Storage-based challenges often top the list of IT pain points. Managing data growth while ensuring data availability, protection and security remain top IT priorities.
- Data is being kept longer for legal, compliance and ongoing analytics purposes. While it varies by vertical market, it's increasingly common to see companies archiving data for at least five to seven years, with several use cases extending well into multiple decades.
- Data from Internet of Things (IoT) initiatives has the potential to dramatically increase the unstructured data management problem. IoT may become the largest contributor to unstructured data management problems through billions of new data sources generating zettabytes of file-based data.

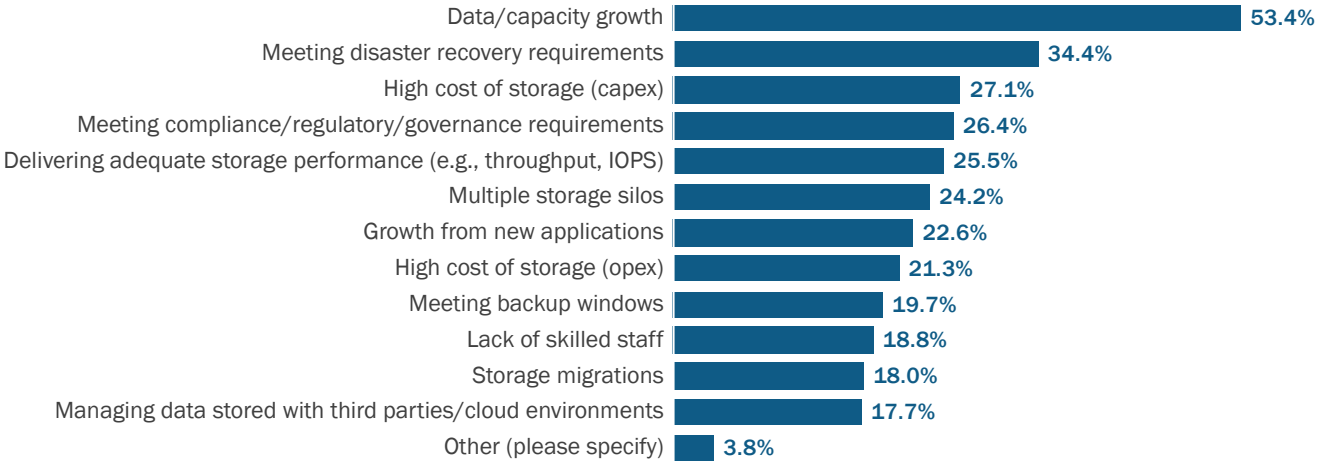
Tracking the Next Generation of Data Storage

Working to Fix the Same Old Problems

The IT industry is in love with data. To a certain extent it always has been, but it's also been a dysfunctional relationship because of the cost and challenges of storing and maintaining the growing amount of data that is needed to feed our insatiable desire for more data. It's an endless loop that's increasingly fed with new forms of unstructured data, as well as with data that's being kept for much longer periods. To illustrate the point, Figure 1 shows the most recent responses from 451's Voice of the Enterprise (VotE) 2020 Storage polling.

Figure 1: Top enterprise storage pain points

Source: 451 Research's Voice of the Enterprise: Storage, Budgets and Outlook 2020 (sample size: 451 respondents).
Q: What are your organization's top pain points from a storage perspective? (Please select all that apply.)



The ranking of these storage-based pain points has remained relatively consistent for years, and though they represent different challenges within the context of data storage, almost every point is directly affected by the overall unchecked growth of data. These problems remain despite capacity-reduction technologies such as deduplication and data compression that have been commonly available for more than a decade. So far, it's simply been easier to throw capacity at the data-growth problem rather than deal with the challenges of evolving storage from a relatively blind repository to a more intelligent and automatable system based on an awareness of the information within that data.

Unfortunately, most file-based data offers little or no visibility or context into what it contains beyond a creation date and filename, both of which are of little use in determining the data's business value. This is far less of a problem for structured database information that has already been normalized and internally cross-referenced; today, unstructured data represents one of the greatest enterprise storage challenges.

Filesystems are far from new. There are literally dozens that were created to meet the need for operating systems to provide a logical abstraction to keep individual application data separate on block storage, and the common tree-like structure of most filesystems provides a method to sort data in such a way that humans can remember where it resides on a large disk system. The centralized NAS model expanded the capabilities so that file-based data could be shared across multiple systems and users, and it provided a way to lock files against concurrent write access that could corrupt data. With few exceptions, traditional NAS systems offer an on-premises, device-specific approach that focuses on controlling and managing data residing on a physical system, but this approach isn't applicable in a hybrid environment where data may need to span multiple storage hosts outside the datacenter.

The Difficulties of Getting a Handle on Data Growth

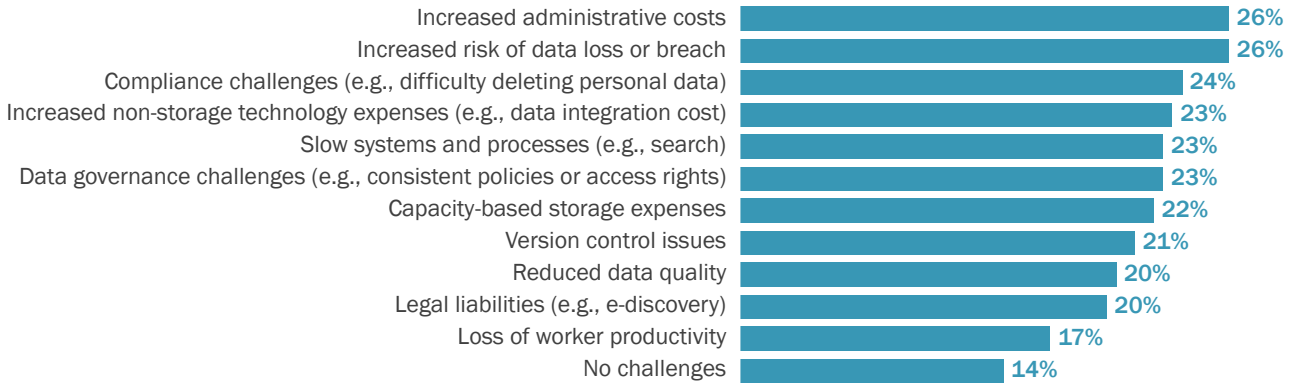
Part of the challenge of data growth comes from the need to keep multiple copies of data as part of a good data protection policy. The 3-2-1 data protection rule states that you should have at least three copies of your data, on two different platforms, with at least one copy kept physically off-site, but in 451 Research's 2020 Voice of the Enterprise polling on Storage, Data Management & Disaster Recovery, we found that while 25% of respondents said they keep the recommended three copies of their data, the average number of copies was closer to five, and there were a surprising number of respondents that said they maintain 10 or more copies of their data.

There are any number of good reasons to keep multiple copies of important business data – enhanced data protection, regional distribution, application development and ongoing analytics, for example. But there's also a downside to having too many copies, and the polling data in Figure 2 shows some of the real-world ramifications of maintaining excessive copies of data. While this reflects the impact from all forms of data, we contend that these problems are magnified by the lack of a cohesive, content-aware model for addressing the short-, medium- and long-term management of unstructured data that's based on its business value rather than when it was created or last accessed.

Figure 2: Risks and drawbacks of maintaining excessive data copies

Source: 451 Research's Voice of the Enterprise: Storage, Data Management & Disaster Recovery 2020 (sample size: 197 respondents)

Q: Which of the following data-related challenges, if any, is your organization experiencing by having multiple copies of its business data?



Unstructured data in the form of documents, sensor data, email and social media interactions, as well as audio, video, images and other highly dense media are a common part of business data today, and the growth of unstructured data is rapidly exceeding the growth of traditional database data in most vertical markets. While database information is typically ASCII data that's highly compressible, many common forms of unstructured media files such as high-definition video, medical imaging and scientific data often reach gigabytes in size and are based on file formats that are already compressed as much as possible.

As mentioned earlier, perhaps the largest challenge of managing unstructured data lies in establishing an 'identity' for unstructured data. This is where traditional filesystems fail; they don't have the ability to attach custom metadata to file data that could then be used as criteria for establishing policies automating granular security, data protection, access control and lifecycle management based on an understanding of a file's contents and other relevant factors. Unfortunately, most unstructured data goes 'dark' once it leaves the direct control of its creator, and the trend toward data decentralization via hybrid-cloud-based services makes it even more difficult to maintain visibility and control over unstructured data that's already hard to manage. The data growth problem is at or nearing crisis levels for many companies, and the future only promises more data with which to contend. It's likely that most companies still handle unstructured data the same way they did 20 years ago, and it's time to bring unstructured data management kicking and screaming into the 21st century.

Hybrid File Storage: Teaching an Old Dog New Software-Defined Tricks

Since the introduction of dedicated, enterprise-grade NAS systems in the mid-90s, the ongoing challenge has been to provide shared file storage that offers comparable capacity, performance, security and reliability to that of block-based SAN systems. While SAN and NAS share the same underlying challenge of building a large, logical pool of storage capacity from hundreds or thousands of physical disks, NAS systems need additional computing power to support the abstraction of a POSIX-compliant file system.

Filesystems have been the norm for local storage on all computing platforms for decades, but providing commonly shareable file storage for a large number of client machines was a relatively difficult challenge in the days of single-processor systems. Most of the first high-performance enterprise NAS vendors focused on a scale-up approach, usually based on a combination of purpose-built hardware that connected many drives to a single or paired set of storage controllers. When the capacity of a scale-up system reached its peak, the system required a major upgrade to more powerful NAS controllers in order to increase both capacity and performance, but that's no longer the case.

Today's hardware environment is radically different than that of a decade ago, and the average server today offers dozens of dual-threaded CPU cores, gigabytes of system memory, multi-gigabit networking and a wide choice of high-performance solid-state disk (SSD) and high-capacity hard disk drive (HDD) devices. This wealth of performance has enabled the industry-wide movement toward software-defined storage, which supports a scale-out approach offering flexible storage resources supplied by multiple network-connected, server-based nodes. This proven, distributed-node model offers linear scalability for both compute and storage capacity, and individual SDS storage nodes can even be optimized for desired capacity and performance by adjusting the matrix of CPU, DRAM, SSD and HDD devices, as well as the latest generation of high-capacity, persistent memory products that bridge storage and system memory.

While scale-out, node-based SDS raised the bar for storage simplicity and economy, perhaps its biggest feature is behind the scenes. The distributed, software-based storage abstraction offers a new model for scalability and data reliability that leverages the combined resources of clustered server nodes to present a dynamic pool of flexible storage capacity. This type of abstraction was originally introduced as the distributed-node model for object storage over 20 years ago, and given the luxury of relatively massive hardware resources, most SDS vendors embrace variations of this distributed-storage theme as a basis for a scale-out storage model that's functional for both on-premises infrastructure and cloud-based storage offerings. The question remains: how does all this affect the challenges of unstructured data?

Part of the beauty of many SDS platforms lies in the remarkable flexibility their software abstraction layer offers for providing a variety of front-end storage options that are no longer defined by the underlying hardware. A growing number of SDS platforms support a flexible combination of block, file and object storage services, with some even offering customized front-end APIs for Hadoop HDFS and others. Perhaps more importantly, an object-capable software abstraction layer offers nearly unlimited scalability, erasure-coding-based data resilience, plus the major value-add of enhanced metadata capabilities. We believe an SDS model that combines file-based application capabilities (e.g., via SMB or NFS) and object capabilities (e.g., via S3 or Swift) offers the best of both worlds.

A merged, software-defined approach to unstructured data storage that combines file and object capabilities presents a storage model that meets the needs of legacy, file-based applications while providing a framework for leveraging the advanced management, scaling and ubiquity of flexible object storage to enhance long-term data management. Perhaps more importantly, a combined file/object model is ideally suited to the hybrid cloud, where a unified approach to data visibility is needed to provide common, policy-based management for data regardless of its physical location.

What Does the Future Hold for Unstructured Data?

Unstructured data is already taking a lead role in traditional business IT, but perhaps one of the largest potential contributors to all forms of next-generation data growth will be the Internet of Things. IoT is an initiative throughout the IT industry that aims to harvest and capitalize on a broad range of new data gathered from a rapidly expanding number of systems and devices. IoT has the potential to touch nearly every vertical market, and there are already a large number of IoT-centric vendors developing products for transportation, manufacturing, sciences, medicine and retail industries, to mention but a few.

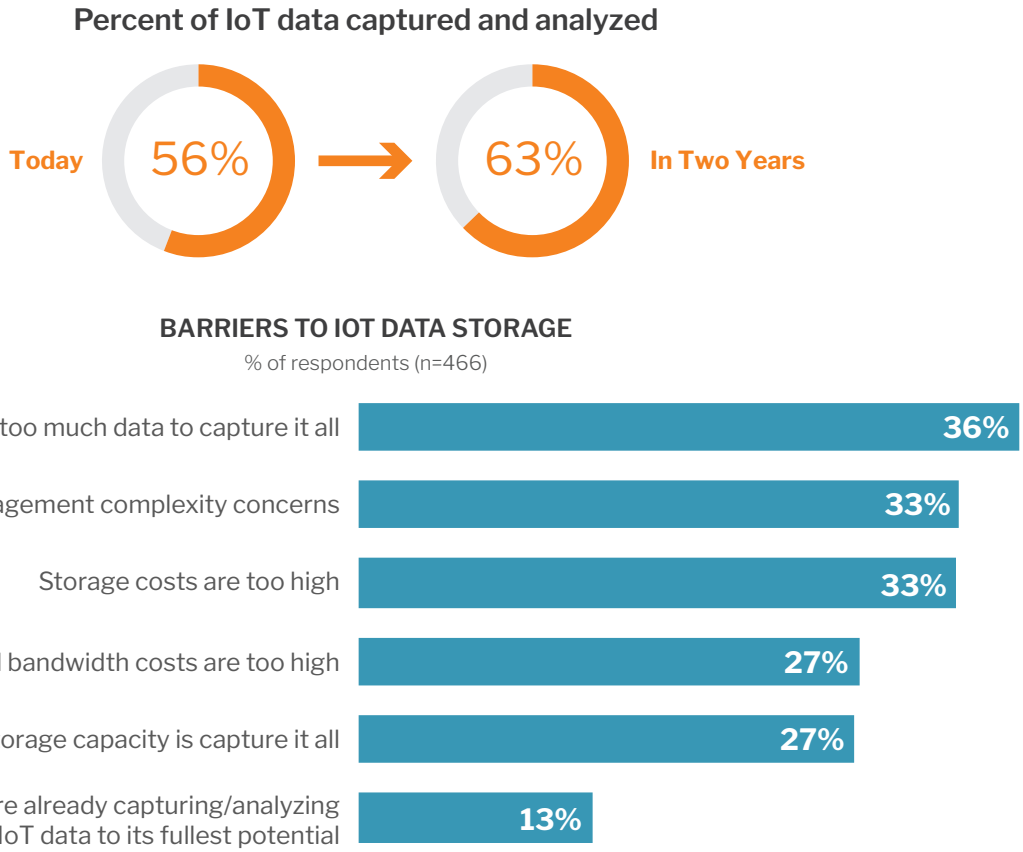
Some of these data sources already existed and are simply being unified under the IoT banner, but there will also be a significant increase in data from new sources, many of which will likely be generating some form of unstructured data. Given the machine-generated nature of many IoT data sources, it will be even more critical to develop a model for identifying and contextualizing this file-based information as it's being created, a problem that could likely be addressed by adding metadata generation as part of the initial data creation and storage process. In this case, IoT applications would be well served by a hybrid SDS platform that offers a combination of file and object capabilities and would open the door for a workflow that could possibly go directly to object and then remain accessible through either file-based or object APIs. To get a feel for the scale of the IoT data challenge, see the 451 VoTE-based polling responses in Figure 3 below, which indicate an annual increase in the amount of IoT data being captured and processed, but the responses also reveal some key challenges presented by the growth of IoT-generated data.

Figure 3: The IoT data storage challenge

Source: 451 Research's Voice of the Enterprise: Internet of Things, Workloads & Key Projects 2019

Q: What percentage of your organization's total IoT data ends up being captured and analyzed? (n=497) What percentage do you expect in two years? (n=494)

Q: Which, if any, of the following are barriers to your organization's ability to capture and analyze IoT data to its fullest potential? (n=466)

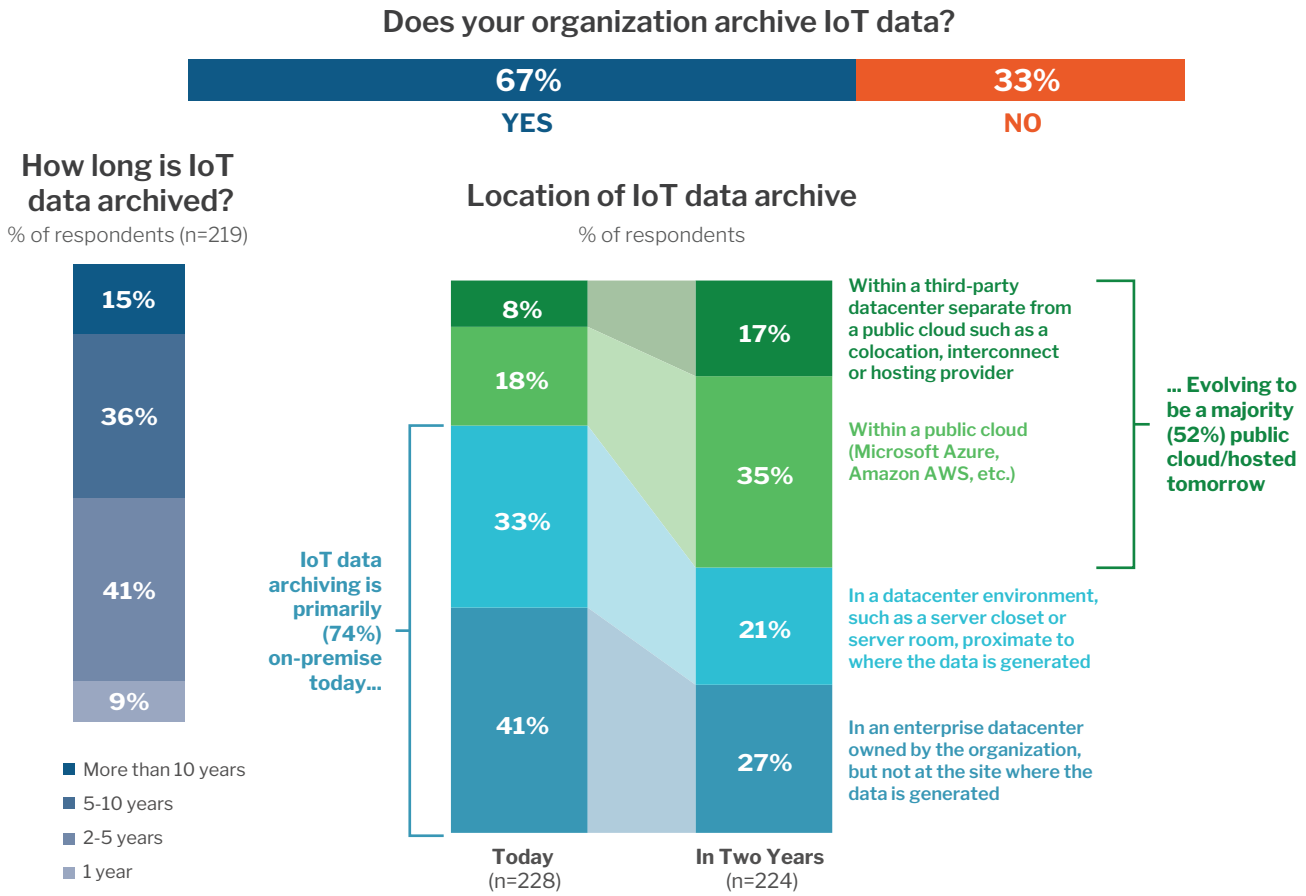


Many of the challenges from Figure 3 are similar to those of traditional storage, but with the added difficulty involved in supporting an application environment that's rapidly evolving. There is little agreement on where the IoT 'edge' actually is, and there are a lot of variables in the IoT formula in terms of data type, quantity, local processing power and network bandwidth necessary for utilizing IoT data at the location where it's most efficient, cost-effective or needed to provide real-time artificial intelligence/machine learning capabilities.

Of course, much depends on the specific nature of the IoT use case, and aside from the near-term data challenges, there will also be a growing need to archive IoT data for a broad range of purposes such as R&D, performance optimization, trend analysis, and even liability and IP protection. Figure 4 shows that the majority of IoT-focused respondents currently archive IoT data, that most will be archiving for between two to five years, and that there will be a growing trend toward utilizing flexible hybrid cloud storage as part of their long-term plan for supporting ongoing IoT initiatives.

Figure 4: IoT data archiving outlook, on-premises and in the cloud

Source: 451 Research's Voice of the Enterprise: Internet of Things, Workloads & Key Projects 2019
Q: Does your organization archive IoT data; if so, for how long does your organization archive IoT data?
Q: What is the primary location for archiving IoT data for the long term today? In two years?



The potential large-scale adoption of IoT throughout the IT industry is just another reason to reevaluate and modernize the traditional storage model for unstructured data. We are firm believers in the future of hybrid technology, but the transition to hybrid requires both a philosophical and a technological shift toward a more intelligent storage environment based on an awareness of the origin, contents and value of unstructured data. Next-generation storage presents a different set of challenges than either compute or networking in that it requires an additional level of consideration for ensuring security, protection and governance that continues for months or years after the data was created. The granular control and visibility offered by intelligent, next-generation hybrid storage systems that provide unified file and object capabilities may be the best option for addressing the evolving needs of business data management through a combination of metadata-based policies and automation.

Conclusions

Crisis…What Crisis?

We've made the observation in the past that the IT world is headed toward an 'unstructured data crisis,' and although it sounds like hyperbole, we suspect that worldwide, we have already reached the point where we're no longer capable of effectively managing all of our data (assuming that we ever were). Recent 451 VoTE polling shows that on average, companies are experiencing 20-25% annual data growth per year, with some vertical markets reaching upwards of an 80% annual growth rate. To get an idea of scale, this can be referenced against estimates that worldwide storage already contains roughly 50-75 zettabytes (one zettabyte = one billion terabytes) of data. These are numbers that are hard to fathom now, and the future looks prime for even more data growth as business finds more ways to parlay data from initiatives like IoT into useful insights and ongoing value.

The IT industry has undergone a radical transformation over the past two decades. Infrastructure options based on multi-core CPUs, gigabytes of memory and petabytes of storage are now the norm, and the combination of server virtualization and the cloud delivery model is freeing business IT from the physical limitations of the traditional datacenter. With an SDS-based platform that combines file and object services, many of the pieces are already in place for hybrid storage modernization. The next step is to focus on defining and gathering relevant metadata that can be used as criteria for automating granular data security, visibility, governance and lifecycle management of unstructured data no matter where it resides while seamlessly supporting legacy business applications well into the future.

Customer Recommendations

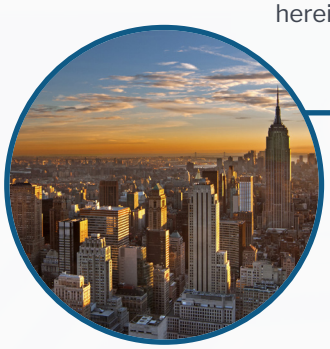
- Understand the key requirements of your applications. Hybrid or not, many of the same rules still apply when it comes to determining the right storage platform for hosting a given application. While performance remains a key consideration, next-generation hybrid storage may also provide integrated, granular services such as data protection, security, tiered archiving and data lifecycle management that can be tied to a common and universal set of policies.
- Plan for increasing storage decentralization. Part of the value proposition of hybrid storage lies in the ability to place data where it makes the most technological and business sense. A hybrid cloud storage environment should offer the visibility to manage data and provide a unified set of storage services regardless of physical location or infrastructure.
- Optimize the placement of data. As data grows, it becomes increasingly difficult and expensive to move. A metadata-based model for data management can reduce the amount of data that needs to be moved, automate tiering based on a flexible set of options, and manage the number of copies of data that may need to exist on multiple storage platforms for production needs.
- Embrace automation to mitigate complexity. Automation is a huge benefit of the hybrid cloud model and can be used to simplify storage management, enable BC/DR capabilities, as well as support deterministic workflows that allow unstructured data to move through a predictable path of actions throughout its lifecycle.

451 Research®
Now a Part of

S&P Global Market Intelligence

About 451 Research

451 Research is a leading information technology research and advisory company focusing on technology innovation and market disruption. More than 100 analysts and consultants provide essential insight to more than 1,000 client organizations globally through a combination of syndicated research and data, advisory and go-to-market services, and live events. Founded in 2000, 451 Research is a part of S&P Global Market Intelligence.



NEW YORK
55 Water Street
New York, NY 10041
+1 212 505 3030

SAN FRANCISCO
One California Street,
31st Floor
San Francisco, CA 94111
+1 212 505 3030

LONDON
20 Canada Square
Canary Wharf
London E14 5LH, UK
+44 (0) 203 929 5700

BOSTON
75-101 Federal Street
Boston, MA 02110
+1 617 598 7200