PATHFINDER REPORT

451 Research | Advisory

Next-Generation Data Management in a Multi-Cloud World

APRIL 2017

COMMISSIONED BY:

SwiftStack

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
New York, NY 10018
+1 212 505 3030

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

LONDON
Paxton House
30, Artillery Lane
London, E1 7LS, UK
+44 207 426 1050

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

Executive Summary

For the last few decades, the storage industry has focused on decreasing the cost per terabyte while increasing capacity and performance, but little has been done to make the enterprise use of storage any smarter. File systems such as CIFS/SMB and NFS were designed to provide an easily understandable, shared model that emulated the simplicity of a filing cabinet, with multiple drawers and folders for storing data. This may have worked in the past, but in today's extended IT environment, traditional file systems lack the metadata capabilities necessary to provide deeper context and enable programmatic management of unstructured data. With storage growth expected to continue unabated for the foreseeable future, businesses really need a better model for classifying, organizing and utilizing data that spans multiple clouds to provide common policy enforcement and ensure long-term information value.

Key Findings

The metadata capabilities of object storage offer detailed information about data content and ownership. This supports the creation of comprehensive, business-specific policies for managing data access, worldwide governance, data protection and legal compliance of unstructured data that can span multiple geographical locations and public cloud providers.

Most cloud storage providers – such as AWS, Microsoft, IBM, Rackspace and Google – have adopted object as the primary storage platform for unstructured data, making object storage the common denominator for hybrid cloud applications.

The increasing use of unstructured data in the form of text-based and log information, machine and sensor data, and rich media files is changing the definition of mission-critical data, and the traditional 'save everything' approach to data storage is counter-productive in terms of the growing cost of storage, not to mention the potential liability for maintaining outdated or toxic data.

There are significant differences between object storage platforms that involve their native user interfaces, as well as the specific nature of the metadata being collected. These differences can make it difficult to seamlessly migrate data from one object storage platform to another, so it's important to consider cross-cloud platform compatibility as part of the evaluation process.

Industry-wide recognition of the value of metadata as an organizational model is very low at present, in part because legacy storage platforms didn't need or support its creation.

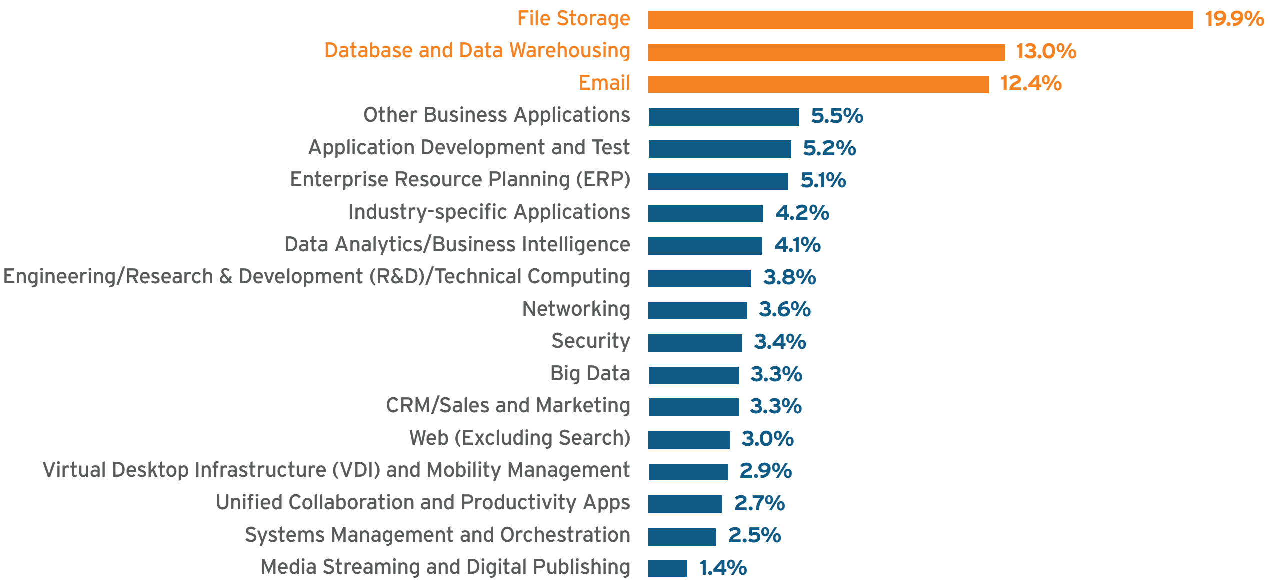
Unstructured Data is the New Mission-Critical Data

For most enterprise IT environments, primary storage was built to protect the database information that constitutes a substantial amount of highly transactional business data. However, the preservation of files and other unstructured data is often piled on as an afterthought because IT was forced to adopt the policy of saving all business-related data due to lack of better information about the contents. Unstructured data such as documents, media files and other user-created content is mostly made up of separate entities with little or no context other than a nondescript file name and extension, plus a handful of generic attributes established by the file system upon which it was created.

Today, unstructured data appears to be rapidly becoming the new mission-critical data. This premise is supported by data from a Q1 2016 Voice of the Enterprise Storage poll (see Figure 1), which shows that unstructured data growth in the form of file storage has surpassed that of traditional database data. What this poll doesn't show is the likelihood that the data from several of the other categories shown could also be added to the list of unstructured data. In addition, the growing interest in IoT technology portends even greater growth of sensor and log data generated from manufacturing, retail, industrial, transportation, consumer activity and other sources where metadata will greatly enhance the value of the raw data.

Figure 1: Storage Capacity Growth by Application

Q. Approximately how is your organization's total storage capacity, including primary and backup/archive storage distributed across the following applications/workloads



Source: 451 Research, Voice of the Enterprise: Storage, Q1 2016

This growth reflects a number of changes to the business computing environment in general, with an increasing amount of extremely important data now being generated in the form of file-based work products and communications. Managing this unstructured data growth is especially troublesome because existing storage technology doesn't collect enough useful information about the contents of the files themselves. This makes it very difficult for IT administrators to gauge the relative importance of unstructured, file-based data, much less identify and eliminate toxic data such as copyrighted materials.

This problem can be substantially reduced for unstructured data with the addition of metadata that establishes content and context. This has been addressed by object storage technology for decades, but only recently has the IT industry started to understand the remarkable flexibility and management capabilities of what we term 'unified object.' These newest iterations of object storage use metadata as the organizing principle for data management in the cloud, and massively extend the scalability of all forms of storage media, as well as provide a rich-metadata framework that enables programmatic data organization and management across all tiers of storage. As storage growth continues unabated, there is a real need for metadata to manage that data growth long term, but the path to universal, rich-metadata adoption will be a challenge.

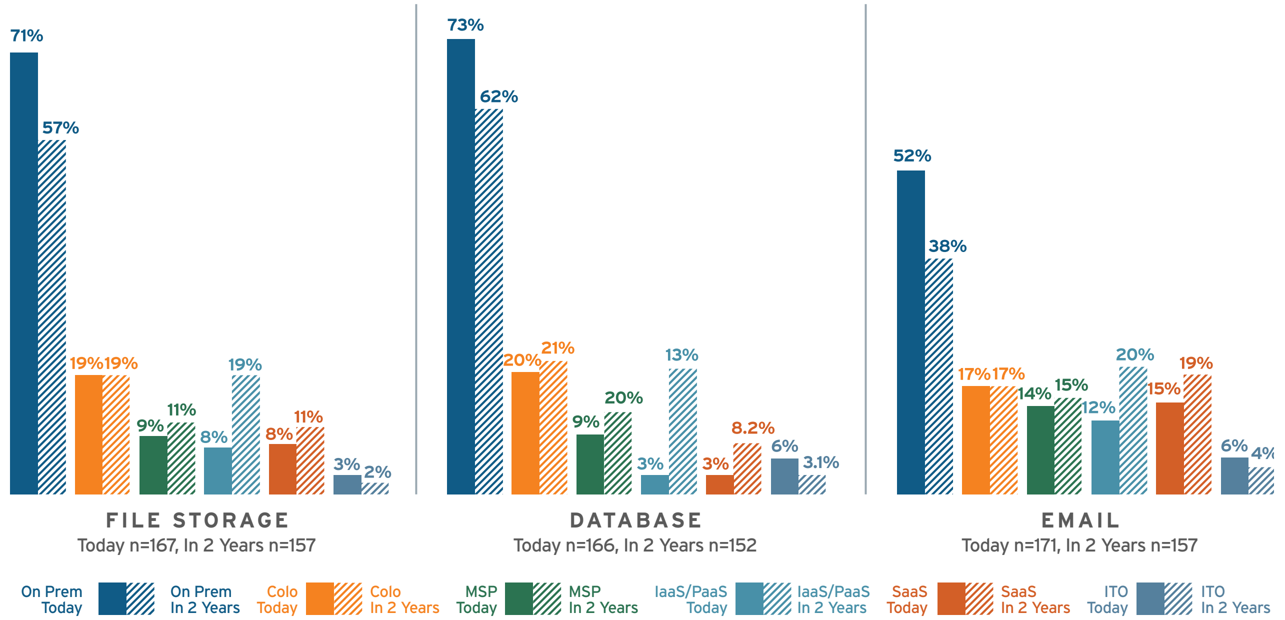
Part of the challenge lies in the collection of useful metadata for existing data. This is a relatively simple process for text-based data, but there are trillions of unstructured media files in existence today that have no, little or even questionable metadata, and the challenge of building that metadata can be a substantial undertaking. Fortunately, the number of private and cloud-based options for metadata classification of legacy data is slowly growing, and we are only beginning to recognize the value and importance of rich metadata as a tool for better organization and greater visibility across storage tiers.

Unstructured Data in the Cloud

The rise of cloud-based compute and storage services has revolutionized the IT industry more than nearly any other advancement in business computing. The cloud storage consumption model is appealing because it offers features such as on-demand availability, usage-based billing, virtually limitless scalability and very attractive pricing. Currently, most cloud storage is stratified into two major use cases: flexible, high-performance block/file primary storage for cloud-based production workloads, and multi-tiered object stores for secondary storage applications with more flexible performance needs.

Figure 2: Two-Year Data-Movement Estimates

Q. Which deployment locations does your organization use for primary storage for the following workloads, today and in 2 years?



Source: 451 Research, Voice of the Enterprise: Storage, Q1 2016

The estimates for data movement illustrated in Figure 2 show a substantial shift in storage positioning for a number of enterprise applications, and as dependence on these systems and features grows, cross-platform data management and visibility become more complex, not less. Granted, the cloud delivery model is designed to mask that complexity from the end user; it doesn't actually eliminate it. There are substantial differences between public cloud storage vendors in the form of dissimilar storage classes (tiers), APIs and metadata frameworks that must be mitigated to seamlessly move data from one provider to another. A truly flexible hybrid cloud strategy should offer the freedom to move between services, as well as provide unified and flexible policy-based management across multi-cloud platforms.

Public cloud is only part of the cloud storage equation; a hybrid cloud offering that seeks to align private cloud and public cloud capabilities offers what may be the best combination of privacy, security, cost and performance for many enterprise customers. The challenge there comes in choosing a private cloud storage platform that also offers the greatest flexibility in making public cloud storage choices.

New Options for On-Premises Storage

Today's enterprise storage customer has several platforms to choose from when it comes to public and private cloud storage options. In classic datacenter storage, there were usually only two:

Storage area network (SAN)	A dedicated, scale-up storage platform that combines the capacity of scores of hard disks to deliver a dynamic pool of tier one, block-level storage that is delivered as logical units, which are mountable as bootable block devices similar to a directly attached disk or RAID system.
Network-attached storage (NAS)	A dedicated, scale-up storage platform that combines the capacity capabilities of a SAN with an embedded and optimized file system, such as SMB and NFS, and delivers shareable volumes of file-based storage to client servers.

Now, there are two new midrange enterprise storage platforms that focus on delivering a combination of block, file and object capabilities, as well as options for features such as Hadoop Distributed File System for analytics. These offer an interesting alternative to monolithic storage systems.

Software-defined storage (SDS)	A scale-out dedicated platform designed to provide distributed storage based on a combination of storage software, commodity server hardware and onboard disk storage. These modular systems are based on x86 nodes that are then clustered in a scale-out model to provide dynamic pools of block, file, object and other storage services to other systems.
Hyperconverged infrastructure (HCI)	Scale-out compute platforms that combine commodity server hardware and onboard disk storage with SDS software and a virtualization stack to provide modular application hosting capabilities and the integrated storage services necessary to support them. .

The key difference between SDS and HCI is the inclusion of production workloads as part of the HCI model. As a whole, the enabling technology for both SDS and HCI is the underlying storage software platform that provides both the front-end application connectivity and the back-end distributed data environment that enables data protection and system resiliency that spans all the drives, nodes and clusters of the entire storage system.

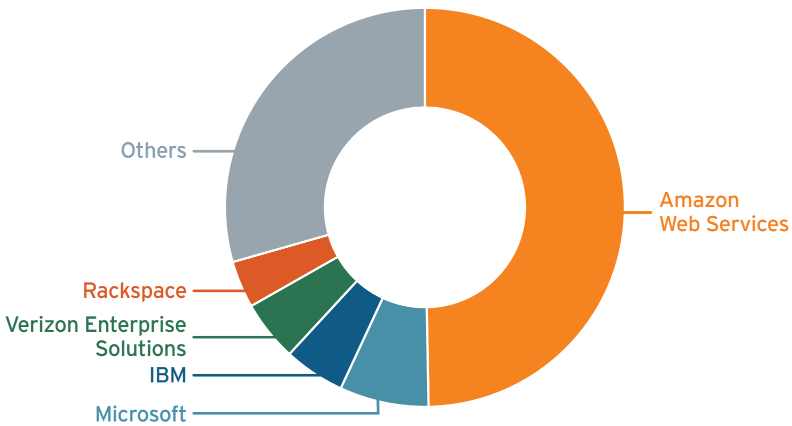
There are several things that make both SDS and HCI very appealing for secondary, unstructured data applications. Aside from the modular scalability and attractive pricing of the commodity hardware, most SDS platforms are based on a highly flexible 'unified object' model that provides the intelligent abstraction layer and metadata capabilities necessary to support next-generation data management. This flexibility is ideally suited for private cloud applications, as well as for storage applications that go beyond the capabilities of traditional file and block storage systems.

Ultimately, modular SDS systems give customers the opportunity to begin an on-premises transition to metadata-based unstructured data management on a much smaller scale, yet with the ability to scale out both performance and capacity as necessary. HCI systems can offer similar capabilities, but adding virtualized workloads to the equation introduces a performance variable that can be difficult to anticipate. Regardless, the availability of metadata capabilities is a major positive for these two platforms.

Taking Storage into the Clouds

As mentioned earlier, there are a number of public cloud storage offerings that can provide block, file and object storage, but there are also several variables among object-based cloud storage platforms that make the movement between clouds a challenge. These differences can necessitate some form of export/ingest function to move data between object platforms, and there's a lack of a unified metadata model for business data that's easily extensible, and no common model for policy management for data that spans multiple clouds. AWS is the largest provider of public cloud services by far (see Figure 3), and the Amazon S3 object API has become the de facto standard. Although it only offers limited user-definable metadata, it still represents a substantial improvement over traditional file systems.

Figure 3: A Rough Assessment of Cloud Market Share



Source: 451 Research Market Monitor: Cloud Computing

AWS's substantial lead in the public cloud market suggests that compatibility with the S3 API should be a key consideration when choosing a private cloud storage option, but as mentioned earlier, part of the remarkable flexibility of object storage lies in its inherent ability to support multiple object APIs, as well as read capabilities for traditional file systems such as NFS and SMB. The OpenStack Swift object API is another desirable interface for hybrid cloud applications because it offers enhanced metadata capabilities, and it serves as a baseline, open source object framework that many storage vendors have agreed to support. Object storage vendors may also offer a native API that could offer advanced features unavailable in either S3 or OpenStack, so the decision to use a given API should be based on developing a hybrid cloud strategy that utilizes the best combination of application support, management capabilities, policy control and cross-platform flexibility.

Metadata: The Great Enabler

We've mentioned the power of metadata many times in this document, but it's important to understand that the wonderful flexibility of metadata is also its biggest challenge because – just as with any database environment – the customization options of user-defined metadata fields can be practically limitless. There are actually dozens of existing metadata standards that were developed to address the specific needs of applications such as library services, life sciences, global positioning, education, multimedia management and government services, but these standards are too specific for a general business application.

What's missing is a baseline metadata set that addresses some of the most common issues of data management and protection, regardless of business application. Even a simple, industry-wide metadata model based on common business-related fields such as security level, data author, access rules, nation of origin, retention level, privacy requirements, HIPAA control, litigation hold and perhaps a handful of other practical tags would go a long way in helping customers address challenges of managing unstructured data across all storage tiers. This standardization would by no means restrict additional customization based on customer needs or specific industry compliance requirements, but customers would benefit from a reasonable baseline metadata set as a starting point, and that core framework of customer metadata should be commonly extensible across all metadata-based storage platforms.

The IT industry is at the earliest stages of exploring metadata-based data management, and the growing challenges of managing storage that spans multiple clouds will eventually become a driving force in the greater adoption and use of rich metadata. Today, most operating systems and applications are primarily dependent on file-based storage systems that don't enable and enforce the collection of metadata at the time of data creation, but we believe that this will begin to change as application developers, IT administrators and business stakeholders come to realize the critical need for metadata to more effectively organize and manage the inevitable growth of unstructured business data.

Ultimately, the rich metadata capabilities of object storage enable customers to move, mine and manage unstructured data – no matter where it travels across the cloud.

Conclusions

Many of the storage challenges that enterprises face will continue to be caused by the growth of unstructured data, and enterprise administrators will be challenged to reduce storage management costs and make more efficient use of storage resources. In addition, business stakeholders are coming to realize the potential of extracting information from legacy data, much of which now goes dark as it moves through traditional tiers of storage.

The metadata capabilities of object storage have the ability to unlock the information within unstructured data and provide a powerful set of tools to leverage its content and manage its growth. A metadata-rich object storage environment enables seamless and extremely granular data manipulation and protection capabilities – behind the scenes – while providing the flexibility of front-end data-delivery capabilities that can match practically any application.

There are substantial benefits to be reaped from the combination of multi-tiered public cloud storage and private cloud, but tapping into this hybrid storage model requires a new approach to storage management. New hybrid cloud storage customers should assess the opportunities offered by both on-premises and off-premises cloud, and make choices that are based on the best combination of application support, management capabilities and policy control, as well as the ability to easily move data between multiple cloud storage platforms.