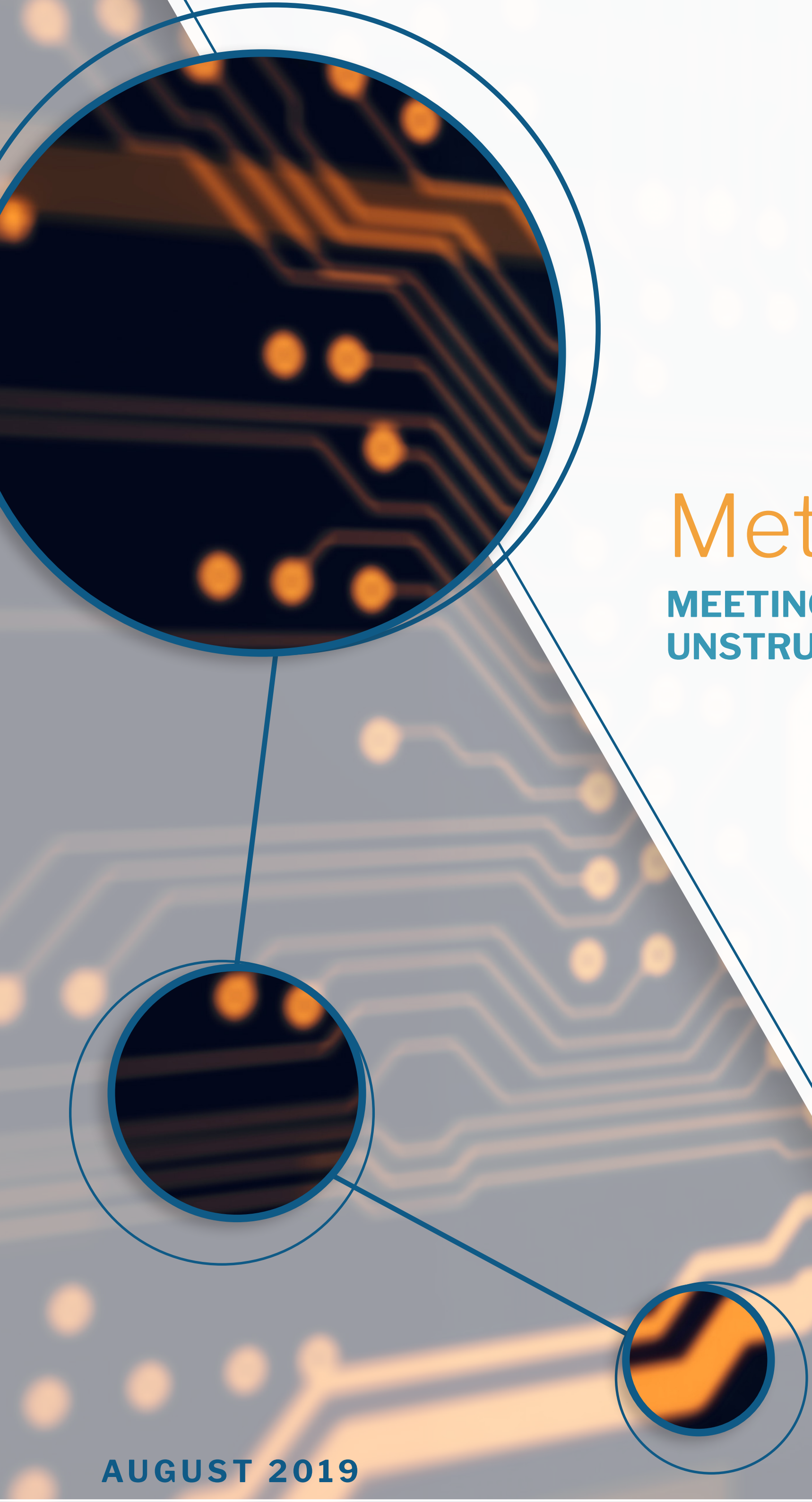
451 Research | PATHFINDER REPORT



Metadata

MEETING THE GROWING CHALLENGES OF UNSTRUCTURED DATA MANAGEMENT

COMMISSIONED BY

HITACHI
Inspire the Next

AUGUST 2019

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 at 451 Research. He covers the latest generation of software-defined systems, hybrid cloud storage, unstructured data management and business continuity/disaster recovery solutions for enterprise customers.

Executive Summary

IT managers and business stakeholders alike are charged with the challenge of squeezing every bit of value possible out of their IT dollars, which means keeping infrastructure costs in line as well as making the best possible long-term use of their business data. Cloud-based services, both private and public, offer a highly flexible combination of scalability and use-based pricing, but making the best use of a hybrid environment that encompasses both public and private services is no slam dunk, and there are still a number of applications that aren't particularly well-suited to exist outside the corporate firewall.

While tackling infrastructure costs is a good first-order strategy for managing data growth, even bigger challenges and value lie in making the best use of business data throughout its entire lifecycle. The traditional model of 'save everything just in case' will become impractical and costly, as well as stress storage resources, the IT staff and user patience. Organizations that are stockpiling exabytes of data aren't considering the simple fact that not all data is created equal. Companies end up paying to store dark data from myriads of sources without knowing whether it's mission-critical, useless or even toxic, and it doesn't take very long for unchecked data sprawl to drive storage costs out of control. We believe the answer lies in next-generation storage systems that leverage the power of metadata to better identify, utilize, protect and control unstructured business data, but it takes foresight and planning.

Key Findings:

- **Unstructured data is becoming the new mission-critical data.** Documents and digital media files represent a substantial percentage of the business data being generated today, and the need to protect unstructured data can be directly tied to the overwhelming growth of data storage costs.
- **Legal and industry compliance issues are driving the need for data awareness.** Healthcare, financial services and other heavily regulated sectors require ready access to all forms of data – unstructured or not – and availability can mean the difference between financial success and failure, or perhaps even life and death.
- **Metadata-based storage and indexing is the key to long-term unstructured data management.** Metadata 'sticky notes' with no indexing provide marginal long-term value. Visionary solutions will offer the tools to identify, categorize and search stored data, and help automate and control its movement throughout its lifecycle.
- **Capturing useful metadata is a major challenge.** The best time to collect metadata is at the time of data creation, but there is no common mechanism at the OS/storage level for metadata creation. Until metadata gathering is enforced at data creation, the big challenge will be generating metadata after the fact, which requires systems with search and cataloging abilities that can address text, sensor and digital media files.
- **Best practices for business metadata generation are poorly defined.** Unstructured data storage should contain a common and extensible set of basic metadata fields that enable policy-based management regardless of where the data physically resides or what business environment it serves.

- **All object-based cloud storage platforms are not alike.** Both private and public cloud services from vendors such as Amazon, Microsoft, IBM and Google vary substantially in their feature sets, metadata environment and performance tiers, making movement between providers a challenge. Hybrid cloud customers should have the flexibility to utilize all hybrid cloud storage options based on the combination of cost, performance, resilience and availability that best suits business needs.

The problem of unchecked data growth has been a constant refrain for decades, and the IT industry has focused on addressing that challenge by making storage larger, cheaper and faster, but we believe the key to dealing with data growth lies in making storage smarter. The next generation of intelligent storage will need much more information about the data itself than we've been collecting. This identifying information, or metadata, remains with the data throughout its lifecycle and provides the hooks necessary to classify the data's contents, establish context and enable highly granular data management automation and lifecycle controls that are missing in most traditional storage architectures.

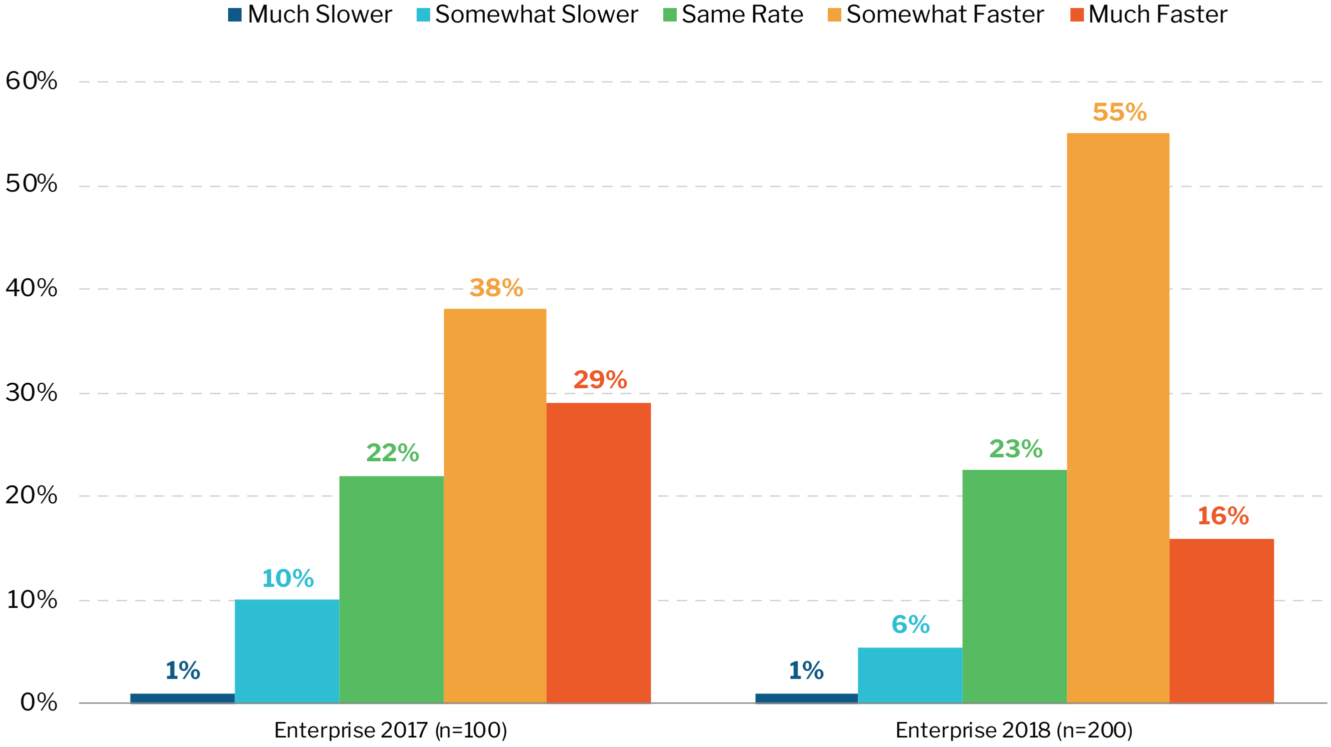
New Data Uses Drive a New Management Model

The term 'dark data' is an accurate representation of the challenges companies face when dealing with most unstructured data. Text-based information is substantially easier to classify and index, but the increasing use of audio, video and digital images – not to mention the massive potential of IoT sensor data growth – can make the task of classification that much harder, especially when it comes to creating that metadata after the fact. And without metadata, unstructured data files easily accumulate to become hidden blobs of data that consume extremely expensive enterprise storage space with little ongoing value. This means that dark, unstructured data eventually falls deeper into the traditional backup oubliette, where it becomes more about raw warehousing than effective management. To address this challenge at a massive scale, multiple storage vendors have emerged with object technology that combines virtually limitless storage repositories along with a varied set of metadata enrichment capabilities that can serve as a model for automated and granular data management based on a wide range of business criteria.

From 2017 to 2018, 451 Research polled enterprise customers to track the growth of unstructured data compared with the rest of their enterprise data. The chart in Figure 1 shows that 72% of the enterprises polled in 2018 said they are experiencing unstructured data growth that is increasing faster than traditional data, and there's every indication that this percentage will continue to increase. It also follows that dealing with unstructured data management is becoming a more business-critical concern, especially given the growing importance of unstructured data as a shared resource for business applications, as critical evidence for e-discovery-based litigation, and as a historical data archive to support ongoing analytics and deep learning initiatives. We believe that this ultimately requires the collection of object-based metadata as a tool for identifying, contextualizing and governing unstructured data to make it useful. To put it simply, companies can't control what they do not know or cannot see.

Figure 1: Unstructured data growth, 2017-2018

Source: 451 Research/Western Digital Object Storage Poll 2018
Q: What are your organization's top pain points from a storage perspective?



Why Metadata Matters

Filesystems typically only collect a few minor data points about the information they hold: a user-designated filename; a three-letter extension for a loose association with an application; creation, modification and access dates; and a few check-box system-based attributes. Some applications such as office suites and media-creation apps, as well as endpoint devices such as digital cameras and smartphones generate additional metadata, but that information is imbedded as part of the file and thus hidden and underutilized. With the right tools to extract this contextual information, it can be added as metadata that is key in establishing the contents or value of the file itself, allowing the separation of business-related information from irrelevant or potentially toxic data. Furthermore, metadata information defines the criteria needed to establish long-term tiering, retention, deletion, security and access policies. Unfortunately, there is no industry-wide metadata model at this point that establishes universal, business-related fields that cover these specific issues, as well as other practical fields such as ownership, nation of origin, privacy, HIPAA control, litigation hold and perhaps a handful of other tags that specifically address business needs. This lack of standardization can make it even more important to establish a relationship with vendors that understand the challenges of business data and can help define a metadata framework that works best over the long run.

Imagine, if you will, the power of being able to automate the granular management of your unstructured data based on a privacy ranking; based on what countries it cannot leave (or enter); based on a warning that it contains proprietary information that your company must protect; or sorted by device, individual, department, division, branch or any combination of metadata you may choose to collect. Then think of legal ramifications such as data sovereignty, e-discovery or the General Data Protection Regulation (GDPR), which became law for customers in the EU in 2018 – especially in the context of data that will increasingly be placed in hybrid cloud scenarios. On-premises or off, cloud-based object storage is a perfect fit for highly accessible unstructured data storage, but strangely enough, there is little commonality between public cloud providers when it comes to customer-facing metadata. While virtually all vendors offer some form of metadata tagging, few offer advanced enrichment, indexing, search or metadata-driven disposition capabilities. Successfully integrating these technologies will become the foundation for realizing the amazing potential enabled by metadata, and making the right choices from the start can be critical when undertaking a more functional, active-archive-based approach to managing unstructured data. It also makes the creation of an accurate and useful metadata strategy that addresses a company's long-term business needs of the utmost importance when architecting a long-term unstructured data management strategy that can span several storage platforms.

Unstructured Data in the Hybrid Cloud

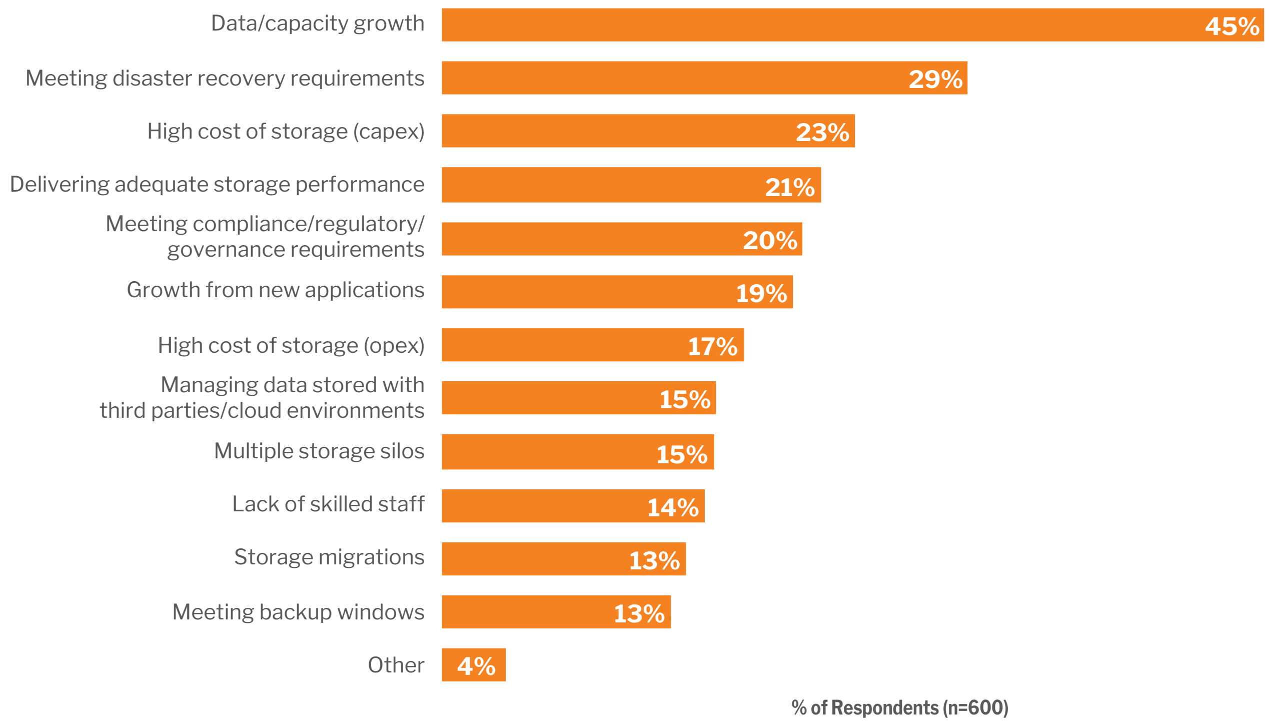
Classic primary storage wasn't really designed to deal with the growth of unstructured data. Large SAN architectures were developed to serve the performance and protection of large systems, and they still manage that admirably. But the unstructured data that's rapidly consuming very expensive primary storage doesn't have the same performance needs and access patterns as systems of record. Many companies have adopted a 'save everything' approach, and the safest place has traditionally been a well-protected SAN environment. Today, a hybrid cloud environment offers a more flexible alternative with a combination of common accessibility and unprecedented scalability for unstructured data growth, as well as the benefits of metadata-based management automation and policy-based management, but not all cloud object storage platforms are identical. Public cloud services offer a variety of metadata capabilities, but there's no standardization between vendor platforms, making it difficult to move between cloud providers, much less craft a metadata environment that best suits a company's individual business needs. There is a strong case to be made for building an on-premises unstructured data environment that matches your needs and is extensible to public cloud services, rather than the other way around.

We regularly ask our Voice of the Enterprise (VotE) respondents to list the key challenges they face in their storage environment, and the top three issues have always been data growth, disaster recovery and cost (capex). This is no real surprise; these have been common IT issues for decades, and though the adoption of public cloud object services can affect the cost/budget aspect of the formula, it's difficult to pin down the actual costs of cloud storage because raw capacity is only the beginning. Additional charges can be incurred based on data access patterns, along with egress penalties – charges that don't exist for on-premises object storage. Private cloud also alleviates the security concerns over data in the public cloud, although a well-designed cloud platform for unstructured data that extends to the cloud can address that problem. In addition, most public cloud offerings don't offer any standardized metadata generation and data recognition capabilities, much less ones that focus on individual customer needs.

In the 2018 VotE Storage chart in Figure 2 below, we've highlighted in orange some of the key areas where we believe content-aware object storage can alleviate the problems associated with enterprise storage. The traditional, file-based model for data storage lacks the information needed to make intelligent storage decisions, and adding even a small amount of metadata can provide content-awareness and serve as criteria for common, granular, policy-based management that can span any hybrid storage venue.

Figure 2: Storage challenges addressable via object-based metadata

Source: 451 Research, Voice of the Enterprise: Storage, Budgets and Outlook 2018
Q: What are your organization's top pain points from a storage perspective?



While automated data governance offers solid benefits for compliance and storage reduction, there is potentially even greater value in the further classification of data through analytics, as well as in the use of metadata as criteria for the inclusion of data in ongoing artificial intelligence/ machine learning and deep learning applications. The Internet of Things has the potential to dramatically increase the amount of unstructured data that companies must face, and metadata offers a model for indexing and contextualizing data generated by this growing number of new data sources. From an analytical standpoint, matching the appropriate data to analytics initiatives improves both efficiency and accuracy, and many AI/ML/DL applications depend on reliable training data that aligns with the business intent of the analytics process, which makes enhanced data identification a win-win scenario.

Summary

Object storage and its rich metadata capabilities have been known for decades, but the IT industry is only now starting to understand the opportunities offered by metadata to help manage and harness the unbridled growth of unstructured data. Part of the problem lies in the fact that object storage has a reputation for being old and slow when the reality is that – with all the processing power, memory, flash storage and high-speed networking – object storage today can accommodate a growing number of tier 1 applications. In addition, many of the leaders in software-defined storage platforms are now combining object's flexibility to use any form of storage medium with the platform's intelligence to support highly granular data management capabilities based on metadata. Buyers should strongly consider storage vendors that focus on leveraging customer-facing metadata to unlock the power of object storage to provide smart alternatives to the 'save everything' approach.

In the past, saving everything was simply the easiest way to prevent the potential loss of important data, but it may also be one of primary causes of the data growth problem because of the continued storage of redundant, outdated or trivial information. But storage growth is only a part of the challenge for next-generation storage, and privacy-based initiatives such as the GDPR and the upcoming California Consumer Protection Act (which is due to become active in January 2020) add a completely new set of expectations for the appropriate governance and protection of personally identifiable information.

These and future initiatives not only add new requirements for security and protection, but they also include the need to be able to identify, produce and even delete consumer data on demand. Traditional storage isn't even remotely capable of such granular distinctions, and added to this is the continuing decentralization of storage through the adoption of hybrid infrastructure. Intelligent and automated policy-based data governance is no longer a 'nice to have' capability, and we're rapidly reaching an environment where improperly managing data is more than an inconvenience; it could have severe business and economic ramifications.

The future of unstructured data management really depends on the creation and curation of quality information about the data itself. Metadata provides the tools for organization, automation, policy management and visibility that simply don't exist without it, and it is the key to managing data growth over the long run. The storage industry is in the early stages of metadata adoption, making the need for reliable metadata the next major challenge for storage customers and vendors alike. While the options in the public cloud for metadata extraction are limited, a few companies are thinking past the basic framework of object storage to address the full potential of metadata as a management tool. Data should be classified by its business value rather than by date alone, and a well-designed metadata environment will ultimately give business a whole new dimension for data management that supports content-aware automation as well as the flexibility to extend to a variety of cloud platforms based on the best combination of cost, performance, security and business needs.

Recommendations

- **Quantify the impact of unstructured data management on your current storage environment.** Many of the capacity-growth challenges companies are facing in costly primary storage can be addressed by a more flexible model for unstructured data that leverages less costly and more efficient on-premises storage infrastructure that extends to public cloud.
- **Examine how greater visibility into unstructured data can increase data value and protections.** Many companies are looking to better leverage the information that's locked away in dark data through analytics, and the problem becomes even more critical when data suddenly becomes evidence in a legal dispute. Understanding the contents of data is the only way to establish its relative importance and treat it accordingly.
- **Start thinking about metadata that's relevant to your business.** A metadata environment doesn't have to be complex to be effective. The only limit to the flexibility of metadata-based management is our inventiveness, and it's a useful exercise to envision what data fields are most important to your environment, and how policies based on those fields can simplify data management.
- **Understand the true cost, security and management issues of on and off-premises hybrid cloud storage.** The low numbers quoted for public cloud storage mask several costs that vary substantially based on access patterns. A hybrid cloud offering that starts with an on-premises approach that's extensible to multiple public cloud offerings allows customers to formulate an unstructured metadata environment that can exceed the capabilities of the public cloud while still being able to leverage its cost and scalability as an off-premises alternative tier.
- **Evaluate a vendor's ability to generate quality metadata for new and existing data.** The only thing worse than no metadata is bad metadata, and the lack of industry standards and enforceable policies for new business metadata means that extraction will need to be done after the fact for the foreseeable future. It's important to find a vendor that can guide you through establishing a metadata framework that works for your business needs and can generate useful, trustworthy metadata, and can be the key to unlocking a flexible, long-term unstructured data management strategy. Ask these questions of potential vendors:
  - What sort of metadata is added to files?
  - Do you offer metadata extraction from well-known file types?
  - Do you offer indexing or search capabilities? APIs?
  - Is metadata used to drive any management policies? (tier/retain/delete)
  - Is that metadata-based management capable of spanning other clouds/platforms?

Except for some specific applications such as medical imaging, library services, e-discovery and earth sciences, the IT industry hasn't focused on dealing with the problems of building metadata for unstructured data management as it relates to business. It's a growing challenge that will only continue to get larger, and it will have even greater financial, business and legal ramifications as data extends beyond the corporate firewall. Knowing the contents of your unstructured data is the first step to enabling a vast array of management tools that will allow you to take control of unstructured business information on your own terms.

About Hitachi Vantara

Data is your greatest asset – if you know how to use it. It reveals your path to innovation and outcomes that matter for business and society. Hitachi Vantara combines 100 years of OT and 60 years of IT experience to help data-driven leaders unlock the value in their data. Our unique Stairway to Value model uses artificial intelligence and machine learning to deliver tangible benefits driven by your data. We help you store, enrich, activate and monetize your data to improve customer experiences, create new revenue streams and lower costs. We listen. We understand. We work with you.

Learn more at *hitachivantara.com*.

Content provided by

HITACHI
Inspire the Next

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 and headquartered in New York, 451 Research is a division of The 451 Group.

The information contained herein has been obtained from sources believed to be reliable. 451 Research disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although 451 Research may discuss legal issues related to the information technology business, 451 Research does not provide legal advice or services and their research should not be construed or used as such.

451 Research shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice.

NEW YORK
Chrysler Building
405 Lexington Avenue,
9th Floor
New York, NY 10174
+1 212 505 3030

SAN FRANCISCO
505 Montgomery Street,
Suite 1052
San Francisco, CA 94111
+1 212 505 3030

LONDON
Paxton House
30, Artillery Lane
London, E1 7LS, UK
+44 (0) 203 929 5700

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