PATHFINDER REPORT

451 Research® | Advisory

Metadata

Meeting the New Challenges of Unstructured Data Management

JUNE 2017

COMMISSIONED BY

HITACHI
Inspire the Next

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.

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
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
(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

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 life. The traditional model of 'save everything just in case' will become impractical and costly, and stress storage resources, the IT staff and user patience. Stockpiling exabytes of data doesn't make sense when you consider the simple fact that not all data is created equal. Companies end up paying to store dark data from myriad sources without actually knowing whether it is mission-critical, useless or even toxic, and it doesn't take 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 implementing these tools 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 provides marginal long-term value. Visionary solutions will include the tools to identify, categorize and search stored data, and help automate and control its movement and lifecycle.
- **Capturing useful metadata is a major challenge** – the best time to collect metadata is when data is created, 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 contains a common and extensible set of basic 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 AWS, Microsoft, IBM and Google vary substantially in their feature sets, metadata environments 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 organizations have been collecting. This identifying information, or metadata, remains with the data throughout its entire life 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.

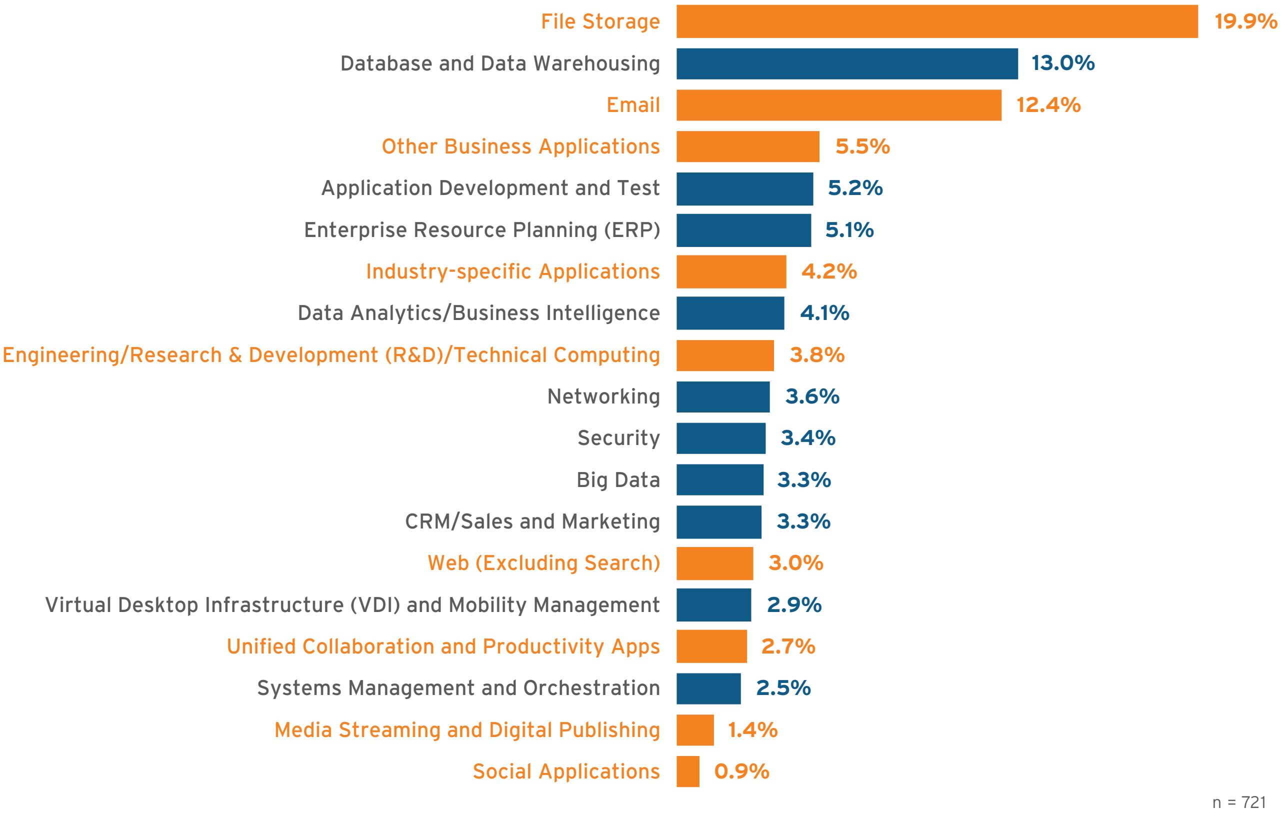
New Data Uses Drive a New Management Model

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 Internet of Things sensor data growth – can make the task of classification that much harder, especially when it comes to creating that metadata after the fact. Without metadata, unstructured data files easily accumulate to become hidden blobs of data that consume extremely expensive enterprise storage space while providing 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, numerous cloud storage vendors have emerged with technologies that combine virtually limitless storage repositories along with a varied set of metadata-enrichment capabilities that organizations need to understand and consider during the product-selection process.

At the beginning of 2016, 451 Research fielded a Voice of the Enterprise storage poll (see Figure 1) that delved into the nature of the data that enterprise customers were protecting within their storage systems. Highlighted in orange are the applications that are commonly associated with unstructured data. This poll shows that 53.8% of enterprise data may be made up of unstructured data, and some estimates rank that even higher. Clearly, unstructured data has become the majority of enterprise storage, and there is every indication that this percentage will increase. The poll also shows that dealing with unstructured data should become a mission-critical initiative, and we believe that this ultimately requires the collection of far better information about the unstructured data that companies choose to protect. Organizations can't control what they do not know.

Figure 1: A breakout of enterprise data distribution based on application/workload

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

Why Metadata Matters

File systems 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 and modification access dates; and a few checkbox 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, so is hidden and underutilized. The right tools could extract this contextual information and add it to the metadata that is key in establishing the contents or value of the file itself. This would allow 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 makes 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.

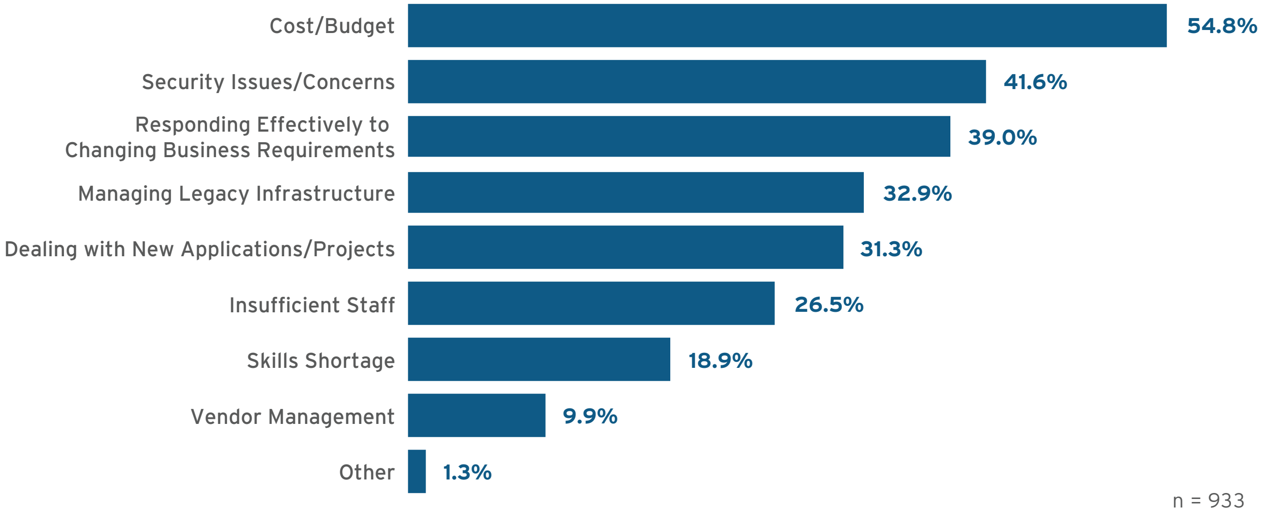
Imagine, if you will, the power of being able to automate the granular management of your unstructured data based on a privacy ranking, based on what countries it cannot leave (or enter), based on a warning that it contains proprietary information that your company must protect, or sorted by device, individual, department, division, branch or any combination of metadata you choose to collect. Then think of legal ramifications such as data sovereignty, e-discovery and the upcoming General Data Protection Regulation, which will take effect in 2018, 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 deriving 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 multiple storage platforms.

UNSTRUCTURED DATA IN THE HYBRID CLOUD

Classic primary storage wasn't 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. 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, but many companies have adopted a 'save everything' approach, and the safest place has traditionally been a well-protected SAN environment. Now, 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; however, 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 the organization's needs and is extensible to public cloud services, rather than the other way around.

Figure 2: Cost, security and responsiveness rule enterprise IT

Q15. What are the top three IT pain points in your organization?



Source: 451 Research, Voice of the Enterprise: Storage, Budgets and Outlook 2016

In Figure 2, we asked VoTE respondents to list the key challenges they face in their IT environment, and the top three issues were cost, security and responsiveness. This was no real surprise; these have been common IT issues for decades, and although the adoption of public cloud services can affect the cost/budget aspect of the formula, it's extremely difficult to pin down the actual costs of cloud storage because charges are incurred based on data access, as well as capacity at rest – access charges that don't exist for on-premises cloud 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 metadata-generation or data-recognition capabilities, much less ones that focus on individual customer needs.

Summary

In spite of the fact that object storage and its rich metadata capabilities have been known for decades, 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 is becoming the platform of choice for a growing number of tier one applications. In addition, the leaders in software-defined storage (SDS) platforms are now combining object's flexibility to utilize any form of storage medium with the platform's intelligence to support highly granular data management capabilities based on metadata. Buyers should strongly consider SDS vendors that focus on leveraging customer-facing metadata to unlock the power of object storage to provide smart alternatives to the 'save everything' approach.

Unstructured data management really depends on the creation and curation of quality information – metadata – about the data itself. It provides the tools for organization, automation, policy management and visibility that simply don't exist without metadata, 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. A well-designed metadata environment will ultimately give customers a whole new dimension for data management that allows a storage customer 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 storage hardware and extends to public cloud.
- **Examine how greater visibility into unstructured data can increase data value and protection** – 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 particular 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 leveraging 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. It 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 they offer metadata extraction from well-known file types?
  - Do they offer indexing or search capabilities? APIs?
  - Is metadata utilized to drive any management policies? (tier/retain/delete)

With the exception of some specific applications such as medical imaging, library services, e-discovery and earth sciences, the IT industry as a whole 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 have even greater financial, business and legal ramifications as data extends further 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.

HITACHI
Inspire the Next

In the era of digital everything, IT and the business are intricately connected and mutually dependent. End-users expect to be able to work anywhere, at any time, on any device. DevOps teams require creativity and look for open application programming interface (API) feature sets. IT must deliver true agility to support business demands while reducing risk. Leadership wants to derive insights from all types of data for a competitive edge and intelligent customer engagement.

This complex challenge is compounded by the need to rely on existing infrastructure while trying to deliver new services and capabilities. The Hitachi Content Platform (HCP) portfolio is a complete solution for bridging bi-modal challenges and next-generation data management. Built upon proven Hitachi-innovated object storage architecture, the HCP portfolio helps you reduce risk, accelerate productivity and increase profitability.

Data, particularly unstructured data, needs to be searchable, visible, pliable and secure — not mired down by limitations, silos and shadow IT. To emphasize a key perspective of this paper, Hitachi understands that metadata-based storage and indexing are the keys to unlocking the potential of unstructured data – Hitachi engineers recognized the importance of metadata when designing HCP more than 10 years ago. While virtually all object storage vendors offer some form of metadata tagging today, only Hitachi had the foresight to make metadata indexable and searchable from the outset.

To accomplish this, HCP software integrates two powerful open source database packages, Postgres and SOLR/Lucene. Not separate, but integrated databases that help you classify billions of stored objects and drive automated policies for proper placement, proper access and discovery. Together, these tools offer a different approach for managing unstructured data sprawl and associated storage infrastructure needs.

Metadata enrichment is achieved via structured XML annotations; entries include cast information to facilitate intelligent queries and better search results. Users can search on either operational or custom fields through a web-based search console or through APIs embedded directly into applications.

For example, in the case of e-discovery, metadata is used to quickly identify and refine the list of relevant documents, set legal hold, and make electronic copies to satisfy compliance obligations. In the case of regulatory compliance, metadata can be used to drive data governance policies.

Successful information management at petabyte scale requires not only solid object repository technologies, but scalable and elastic tools to build content intelligence. This continues to be a strategic R&D focus area for Hitachi and is continually integrated into the HCP content portfolio.

The HCP portfolio includes solutions that automate data management, movement and access via policies based on metadata and content intelligence. In addition to a strong metadata strategy, the portfolio can deliver adaptive cloud tiering that orchestrates data mobility to, from and between public and private clouds, streamlining mobile collaboration, end-user data protection, and enterprise file sync and share, while ensuring secure, continuous data. The HCP portfolio can also provide low-cost elastic, backup-free file services beyond the data center.

To learn more about how Hitachi Content Platform can help you bridge traditional and emerging technologies, visit *https://www.hds.com/en-us/products-solutions/storage/content-platform.html*