Enterprise Strategy Group™

BRIEF

JULY 2025

Starfish Storage's Metadata-driven Approach for Unstructured Data Management Strikes a Chord in the AI Era

Simon Robinson, Principal Analyst

Abstract: Data is the lifeblood of any modern organization, but the sheer volume of unstructured data at many organizations threatens to overwhelm, pushing up costs, introducing risks, and limiting their ability to innovate at a time when the importance of leveraging digital data has never been stronger. These are challenges that Starfish Storage has spent the last decade addressing, with a comprehensive set of federated data management capabilities that help many of the world's largest organizations control, optimize, and fully leverage their critical unstructured data assets.

Analysis

In today's digital world it's a given that data is the lifeblood of the modern organization. It follows that organizations are looking to effectively harness their data to set them apart. But alongside this sits another, more awkward truism: organizations are drowning in data, and this threatens to undermine their ability to effectively leverage it.

A particular challenge organizations are facing is around unstructured data, which continues to balloon in both volume and variety as organizations generate massive volumes of documents, rich media content, images, logs and other data types to support all aspects of their business. According to research by Enterprise Strategy Group, the majority of respondents reported having more than a petabyte of data stored on their corporate servers and storage systems, with many storing a good deal more than that. The same study found that unstructured data accounts for over half of total data in most organizations.[1]

This presents many organizations with multiple challenges at the data management level (see Figure 1). For file data in particular, top issues include performing data migrations effectively, lacking automation, meeting compliance and governance requirements, and managing data across a disparate environment. Additionally, 63% of respondents said their organizations regularly encounter issues with visibility across all of their data.[2]

Managing unstructured data is complex for many reasons. It's particularly challenging considering that it is often locked behind user permissions, buried in directories multiple levels deep, and subject to constant revisions. Another issue is that there are typically multiple "owners" of the data (e.g., the application and line-of-business teams, creatives, and others who chiefly create and use the content; the governance and stewardship teams that are responsible for ensuring data complies with regulations, laws, and other compliance mandates; and the IT infrastructure teams that are responsible for physically storing, protecting, and securing the data). Storage managers usually have no connection with the content value creators, making it difficult for them to make decisions around moving, deleting, and archiving data.

This federated ownership structure requires clear rules around usage, management, retention, protection, and so on. Unfortunately for many organizations, and often despite their best intentions, the sheer volume and pace of growth of unstructured data means this is often difficult to implement in practice.

[1] Source: Enterprise Strategy Group Research Report, *Achieving Cyber and Data Resilience*, September 2024.
[2] Source: Enterprise Strategy Group Research Report, *Navigating the Cloud and AI Revolution: The State of Enterprise Storage and HCI*, March 2024.

Another issue is data longevity of data. Data is typically at its most useful when it is first created, but it is often retained for years on corporate storage systems or in public clouds despite being seldom or even never accessed again. Though this can seem like a trivial concern given the falling cost of storage, when aggregated across millions or even hundreds of millions of files, this can represent a significant waste of storage capacity, plus the space and power required to store it. The advent of AI is also creating a new middle ground where data needs to be both archived and easily retrievable in order to meet the dual needs of cost control and availability for AI.

This data can also be risky, especially if it contains sensitive, secret, or classified information. Unstructured data can be so easy to create, move, and copy, but in an IT environment where security threats are ever-present and growing in sophistication every day, organizations that don't fully understand their aggregate data environments are increasingly vulnerable. Providing traditional backup for all of a company's unstructured data can prove to be prohibitively expensive, yet IT managers typically have very little visibility into what data is valuable and what is redundant or no longer needed. The result is organizations often overpaying for backup or living with additional risk of data loss.

These challenges are further compounded by data entropy, where unstructured data is scattered across a broad range of storage locations, both on premises, in the public cloud, and across myriad edge locations. The fragmentation of data across multiple storage types (NAS, object, cloud, etc.) and media types (flash, HDD, tape, etc.) and across different (often incompatible) vendor solutions compounds the problem, and gaining overall control of data across the entire environment—and managed through a single pane of glass—can seem all but impossible.

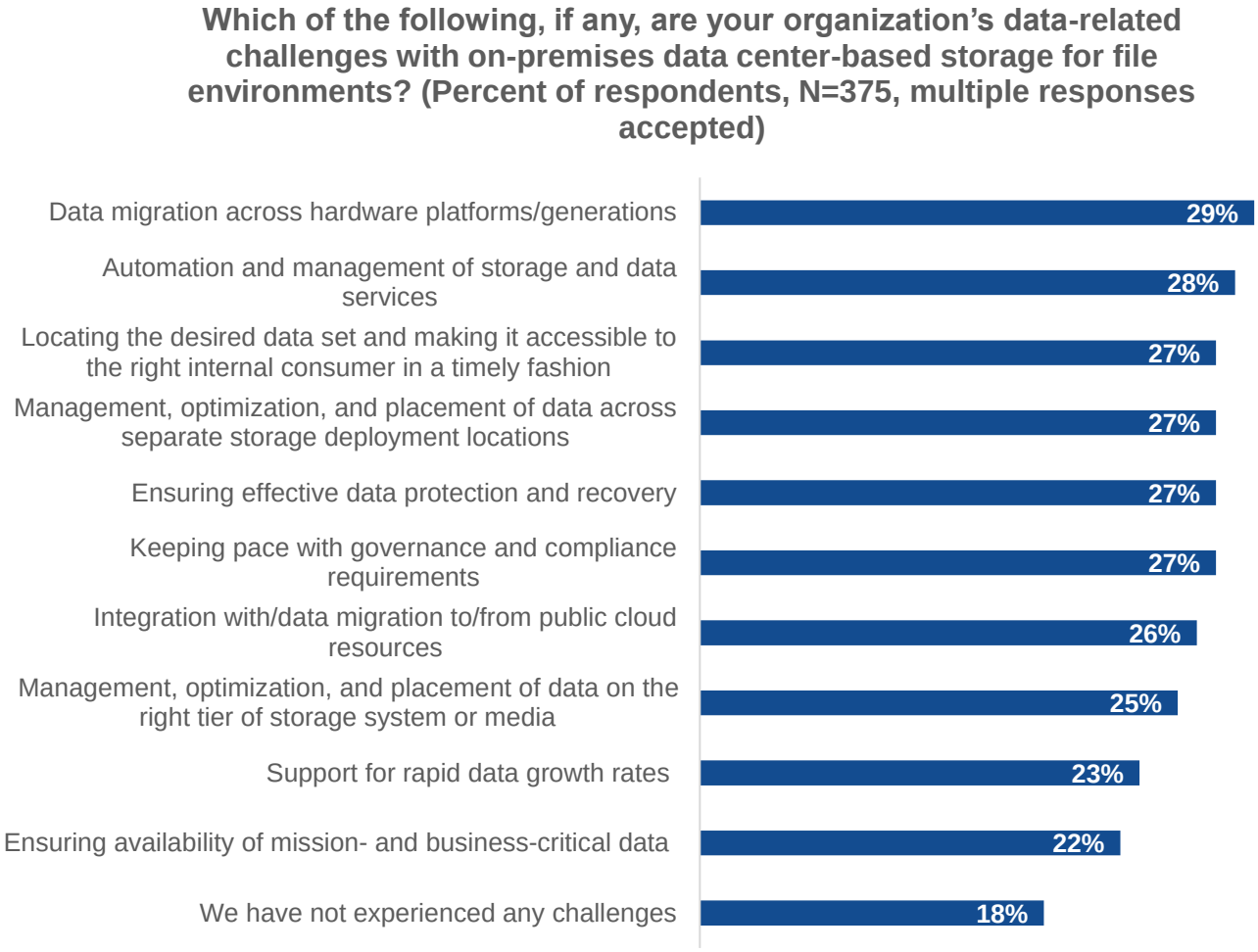
Though these challenges with unstructured data management are not new, they are certainly not going away. Furthermore, organizations increasingly recognize the growing imperative to manage and control their unstructured data holistically.

Of course, the surging levels of interest around AI have a large role to play here. Good AI requires good data—unstructured data, in particular—and lots of it. The challenge in this is that organizations typically do not have a good understanding of where their good data lives, and even if they do, they often struggle to get it into their AI models in a timely fashion. According to Enterprise Strategy Group research, data management and/or data quality issues was the second most frequently cited challenge that organizations reported experiencing when implementing AI, behind only costs.[3] Hence, a good AI strategy must begin with a solid data strategy.

For all these reasons, organizations across the board are starting to prioritize unstructured data management. For many, this has been a long time coming, but, at the same time, it is a problem that many cannot afford to put off any longer.

[3] Source: Enterprise Strategy Group Research Report, *Navigating Build-versus-buy Dynamics for Enterprise-ready AI*, January 2025.

Figure 1. Data-related Challenges With File Storage Environments



Source: Enterprise Strategy Group, now part of Omdia

Starfish Offers a Single Pane of Glass for Federated Unstructured Data Management

Starfish Storage is one of those companies that deserves a "best kept secret" award. The Massachusetts-based company, which recently celebrated its first decade of operations, has developed a rich set of unstructured data management capabilities that are in use by a number of large global organizations. The company helps to solve some of their most demanding data-related challenges. Its relatively small stature belies the critical role it plays for its customers; Starfish's software manages more than exabyte of capacity over its client base, which includes eight of the world's top ten pharmaceutical firms, six Ivy League universities, the U.S. Department of Energy's supercomputing sites, and leading corporations across all major industry segments, including semiconductor, gas and oil, fintech, automotive, healthcare, and media and entertainment.

Starfish's offerings combine a file system metadata catalog—the Starfish Unstructured Data Catalog (UDC)—with a parallelized data mover and batch processor—the Starfish Automation Engine. The UDC creates an index across all of an organization's file storage devices that associate metadata with files and directories and enables the business to understand how file contents relate to projects, intellectual property, workflows, and cost centers even when spread across multiple storage devices.

Using the catalog, data is discovered and reported upon, before being moved using the automation engine. The software is able to handle a wide variety of use cases at scale, including traditional data management use cases such as archiving, migration, and data movement, as well as more sophisticated use cases the company believes helps it stand out from the crowd and enable customers to manage their data more intelligently than is possible elsewhere. Such use cases include data protection; cloud bursting; cost accounting; data disposition; redundant, obsolete, trivial (ROT) cleanup; and AI/ML workflows.

A key part of the Starfish proposition is that it empowers content creators and users to take control of their storage management across a wide range of use cases through a feature called Zones. Starfish Zones helps organizations solve the problem of linking data storage to data value by enabling owners, creators, and researchers to tag data for archive, deletion, movement, consolidation, etc. and helping to bring together storage management and data management.

A good example of Starfish's capabilities including Zones is from Harvard University. Harvard's Research Computing organization within its Faculty of Arts and Sciences (FAS-RC)—the largest computing environment at Harvard—was facing such runaway storage growth that it lost the ability to account for storage consumption and implement chargeback with the granularity it required. Starfish's software helped provide users with fine-grained visibility into their storage usage, such that they were able to delete nearly a petabyte of data within a few weeks of deployment; within two years, Harvard's FAS-RC had generated $1.5 million in chargeback revenue and is now responsible for tens of petabytes of primary storage being freed up or retired.

The software is storage vendor agnostic and, hence, is compatible with a range of HPC file systems, enterprise NAS, object stores, cloud services. and archiving technologies.

With this solid foundation in place, and buoyed by growing market demand for comprehensive unstructured data management capabilities, Starfish Storage's opportunity to expand further is evident. If the company can continue to build on this momentum with smart product development and a go-to-market and partnership model that can improve its visibility and awareness, it will likely no longer be a "best kept secret" but rather a widely recognized market leader.

About Enterprise Strategy Group
Enterprise Strategy Group, now part of Omdia, provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com