PANZURA

SPECIAL REPORT

5 Things Most CIOs Don’t Know About Their Unstructured Data

The CIO’s Guide to Uncovering the Hidden Dangers and Untapped Potential of Unstructured Data.

CIOs are often laser-focused on structured data — the neatly organized rows and columns of databases that power core business processes. But unstructured data, which includes everything from email messages, Word documents and PDFs to images and videos, is growing exponentially and holds untapped potential.

Unstructured data contains thoughts, opinions, patterns, symptoms, outcomes, observations, sentiments, and more. It contains the context and intelligence that would allow organizations to identify patterns that could transform outcomes, if only they could access it.

Within healthcare for example, scanned doctors’ notes and patient discharge summaries can hold the key to identifying the most effective treatment regime for patients, based on their unique characteristics.

For financial institutions, identifying unusual patterns in communication or transactions within documents attached to more structured communications can help detect and prevent fraudulent activities.

Within manufacturing, the ability to assess photo or video footage to spot defects or identify process optimization can improve output and efficiency.

Generative AI has dramatically boosted organizational ability to parse and query large volumes of data, yet for many CIOs, accessing their unstructured data’s potential remains stubbornly elusive.

Those who succeed can unlock insights and understanding that propel their organization towards innovation and ultimately, competitive advantage. Here are five critical things most CIOs don’t know about their unstructured data, with deeper insights into their implications.

TABLE OF CONTENTS

ONE

Unstructured Data Is More Vulnerable Than They Realize

Unstructured data makes up an estimated 80-90% of enterprise data. Despite its prevalence, unstructured data often lacks the governance and security frameworks that structured data enjoys. This disparity arises because unstructured data is harder to manage; it's scattered across silos, lacks standardized formats, and is constantly growing.

Many CIOs fail to realize the full extent of their unstructured data footprint, let alone its vulnerabilities — and it's been their downfall. Without a clear understanding of what unstructured data exists, where it's stored, and who has access to it, organizations are flying blind in the face of mounting risks.

Understanding and Addressing the Causes of Vulnerability

Fragmented Storage Environments
Unstructured data is often spread across on-premises servers, cloud services, employee devices, and legacy systems. This fragmentation makes it difficult to enforce consistent security policies, increasing the risk of data breaches and unauthorized access.

Lack of Visibility
Many organizations lack the ability to see or understand all of the data within their estate, even if they have tools capable of monitoring and classifying unstructured data effectively. Sensitive information such as personally identifiable information (PII) or intellectual property can reside in unprotected files, leaving the organization exposed to regulatory fines and intellectual property theft.

Inadequate Detection and Recovery Capabilities
Ransomware attacks thrive on exploiting unstructured data because it's often inadequately backed up or poorly indexed. Attacks can run below the radar for significant periods of time, resulting in the exfiltration and subsequent exposure of sensitive and critical information. When critical files are encrypted or destroyed, organizations can face massive downtime and data loss.

Regulatory Compliance Risks
Data protection regulations like GDPR, CCPA, and HIPAA require stringent safeguards for sensitive data. Unstructured data, however, often evades compliance checks due to its dispersed and perceived unclassified nature. This puts organizations at risk of severe penalties.

Insider Threats
Unstructured data stored in shared drives or email systems is particularly vulnerable to insider threats. Employees with unnecessary access to sensitive files can inadvertently or maliciously expose the organization to risk.

Real-World Consequences

Numerous organizations have fallen victim to their lack of effective unstructured data management and it has threatened their survival:

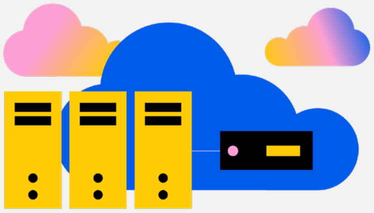
Data Breaches
Unprotected files containing sensitive customer or employee data have been exfiltrated, leading to financial and reputational damage.

Regulatory Fines
Companies have faced multi-million-dollar penalties for failing to secure sensitive unstructured data in compliance with GDPR or HIPAA.

Ransomware Attacks
Businesses have been brought to a standstill when attackers encrypted their unstructured data, forcing them to pay ransoms or rebuild from scratch.

TWO

Dispersed Storage Drives Up Costs, Reduces Efficiency, and Creates Data Growth



Unstructured data often ends up scattered across multiple storage environments, including local file systems, shared network drives, cloud services, shadow IT, and legacy on-premises systems. This dispersed storage leads to inefficiencies that go beyond wasted storage space. Redundant copies of files, inconsistencies in data access, and increased maintenance costs, as well as potential compliance risks (GDPR, etc) are all symptoms of a fragmented storage strategy.

When unstructured data is a byproduct of virtually every interaction and that data must be stored, its volume will naturally grow over time. However, the way that many organizations manage their data is directly contributing to an accelerated rate of growth. In part, this is driven by the sheer volume of data CIOs are already managing. It can seem faster and easier to simply purchase more storage than to find ways to bring data under control.

Understanding and Addressing the Causes of Unstructured Data Growth

According to various industry analysts, unstructured data grows within most organizations between 40% to 60% every year. This places enormous strain on budgets as well as on operational teams tasked with managing and protecting it.

Data is Generated Everywhere

Users, processes, and workloads generate a torrent of information every day. Moving large quantities of data can take a significant amount of time, so the default position is often to store data as close to its source as possible. This saves on bandwidth and network utilization but contributes to data silos.

File Synchronization Between Sites is Problematic

The closer a user or process is to data, the better file performance they experience. However, most organizations have a workforce that is at least somewhat distributed, while needing to access and work with files that have been generated in another location. Slow performance leads to file copies and versions, which substantially contribute to the growth of data volumes.

Data Resilience Requires Replication

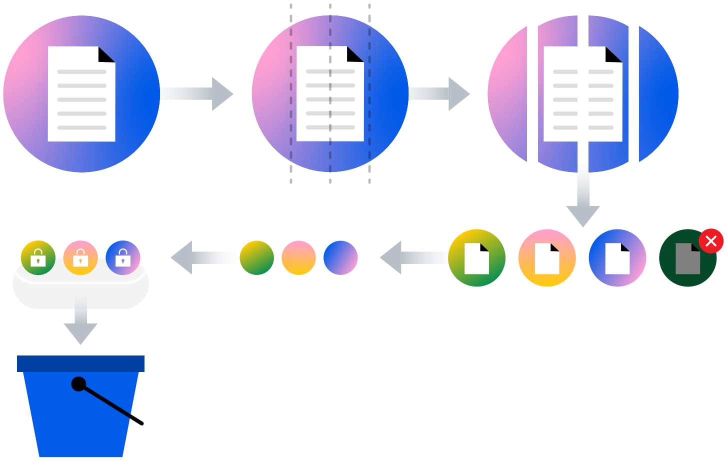
Legacy approaches to backup and disaster recovery exacerbate data growth, as organizations replicate datasets aiming to ensure they can recover in the event of data damage. In addition to file replication between sites, which already contains a significant amount of duplicated and therefore redundant data, entire datasets are backed up and then backed up again to achieve an acceptable level of data resilience.

Storage Providers Have Little Incentive to Optimize Data Volumes
The role of most storage is to provide a secure data repository that is accessible and sufficiently resilient. However, most storage is priced by capacity and storage providers do not inherently optimize data to consume less volume. Deduplication and compression is the role of data management and done well, is an intelligent approach designed to optimize storage consumption, access, and resilience.

The financial and operational impact of dispersed storage and traditional back ups can be significant. CIOs may find themselves paying for redundant storage, contributing to data corruption through versioning across sites, struggling with data recovery during outages, or facing delays in accessing critical files.

In contrast, consolidating and centralizing unstructured data using a hybrid cloud file services platform can eliminate redundancies, streamline access, and reduce overall costs.

Panzura CloudFS, for example, consolidates unstructured data into a unified, scalable hybrid cloud file services platform utilizing cloud or on-premises object storage. The platform is underpinned by a global file system that, by deduplicating data across the enterprise and providing a single source of truth, enables organizations to reduce storage expenses while improving data availability and management.

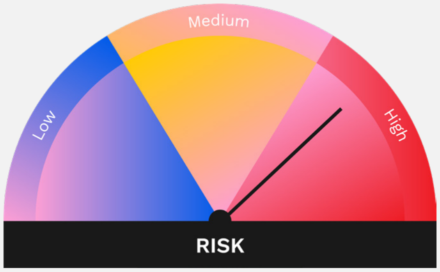


On average, organizations realize a 35% reduction in unstructured data volume, with some achieving up to 80% reduction.

Centralized storage also supports hybrid and multi-cloud strategies, allowing enterprises to adapt to evolving business needs without sacrificing control over their data.

THREE

Compliance Risks Are Hidden in Unstructured Data



The regulatory landscape is becoming increasingly complex, with frameworks like GDPR, CCPA, CMMC, and HIPAA imposing strict requirements on how organizations handle data.

While structured data is often managed with robust compliance tools, unstructured data presents a much greater challenge. Sensitive information, such as personally identifiable information (PII), health records, or financial details, is often buried in unstructured formats like emails, PDFs, and multimedia files.

The lack of visibility CIOs have across their data estate may place them in significant jeopardy. Even organizations with robust data governance frameworks setting out clear policies and procedures for managing unstructured data, including data retention, access control, and disposal can be caught by surprise through lack of awareness of data that has fallen out of compliance.

Moreover, timely response to regulatory requests or legal discovery require the ability to quickly locate, analyze, and package relevant data. Yet many organizations struggle to even see, effectively classify, and manage unstructured data, which makes responding to these types of requests both difficult and time consuming.

Storage, Security, and Compliance Cannot Work in Silos

Today’s regulatory environment means that organizations cannot simply handle data appropriately. Compliance means they must be able to prove that all relevant data has been treated in accordance with frequently changing regulations.

Organizations must have mechanisms in place to detect, remediate, and report breaches both from internal and external sources. This goes beyond the implementation of data discovery and classification tools, or the creation of governance policies.

Instead, CIOs must ensure they have the ability to effectively monitor user activity and alert on anomalies that could indicate insiders are inappropriately accessing or copying data.

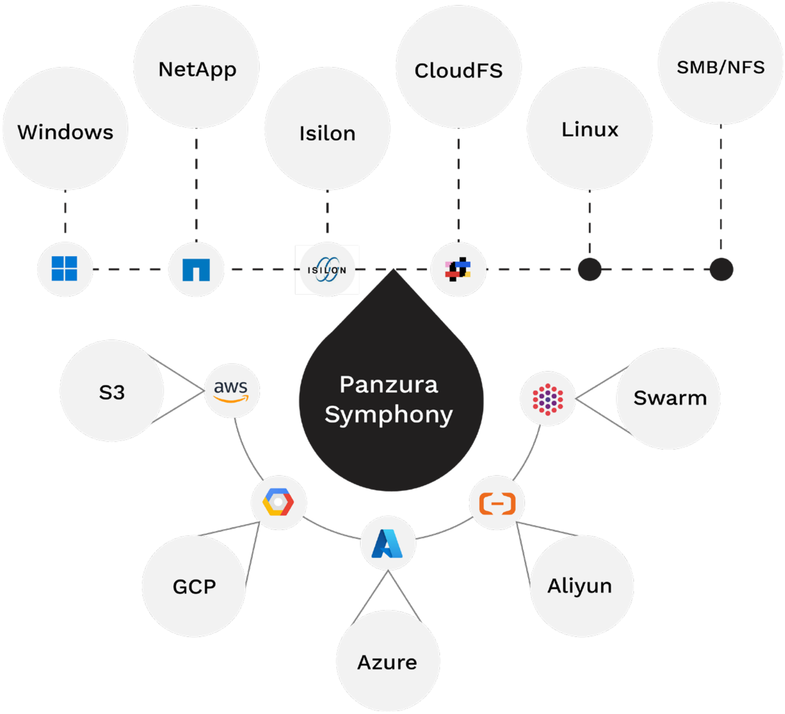
Additionally, they must be able to detect and shut down ransomware activity within their unstructured data in near real time, to avoid becoming yet another CIO whose organization failed to detect ransomware for months, allowing data to be damaged and exfiltrated, and electing to pay a ransom because they were otherwise unable to respond and recover.

Consolidating unstructured data into a unified platform like Panzura CloudFS simplifies compliance by enabling consistent policy application across all data. Advanced classification tools can also help identify sensitive information and enforce automated controls, reducing the likelihood of accidental exposure or non-compliance.

Panzura Symphony, a hybrid cloud data services platform, can scan any source storage metadata to find and flag sensitive information, improper access, orphaned files, and myriad other governance and compliance related issues. This flag acts as a trigger to custom Symphony automations — policies and tasks — which can be set up by either the governing body (required actions) or the company (voluntary actions).

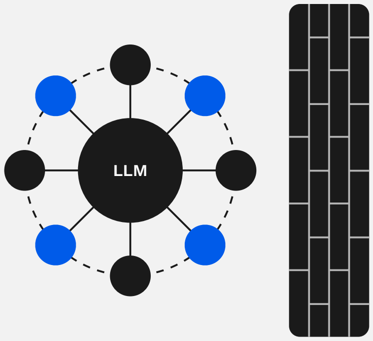
This means that your organization can stay compliant at all times, with no human interaction required. Moreover, all actions may be stored in a database for external auditing.

More information on how Panzura Symphony can help with compliance can be found within this Solution Brief.



FOUR

Training LLMs on Unstructured Data is Complex and Costly



Large language models (LLMs) like GPT have become transformative tools for businesses, powering capabilities like natural language processing, sentiment analysis, and advanced decision-making. However, training these models requires vast amounts of data, much of which is unstructured. The size and complexity of unstructured data make it challenging to prepare for machine learning workflows.

Centralizing unstructured data on platforms like Panzura CloudFS can address some of these challenges. By consolidating data and ensuring it is readily accessible, enterprises can reduce the time and cost associated with preprocessing. Additionally, centralized storage ensures that training datasets are consistent and up-to-date, improving the quality of the resulting models. CloudFS nodes can be deployed locally to the LLM training region to minimize egress fees, for example.

The implications of using unstructured data for LLM training extend beyond efficiency. Rich datasets, including text, images, and video, can provide models with nuanced context that drives better insights. Organizations that effectively leverage this data for AI training can gain a competitive advantage in automation, customer insights, and operational efficiency.

Panzura Symphony ensures efficient LLM data access by delivering only relevant subsets of data, reducing computational overhead. It enforces strict security measures, mitigating risks associated with unauthorized access or breaches. Additionally, Symphony's ability to maintain compliance simplifies regulatory adherence and protects sensitive information.

The scalability of Symphony also ensures that it can handle large datasets and complex queries, making it an indispensable tool for organizations looking to maximize the value of their LLM investments.

Integrating Panzura Symphony with LLM workflows provides a secure, efficient, and compliant method for accessing and processing unstructured data. By leveraging metadata, the data broker optimizes data access while ensuring that the LLM interacts only with relevant and permissible content. This approach not only enhances the performance of LLMs and saves cloud resources (and money) but also protects organizations from data misuse and compliance risks, making it a foundational component of modern data management strategies.

More information on leveraging Symphony for LLM training can be found within this Solution Brief.

FIVE

Unstructured Data Requires a Different Approach to Management

Traditional data management tools and strategies were designed with structured data in mind. Unstructured data, with its variety of formats and lack of predefined schemas, requires a fundamentally different approach. CIOs must adopt tools and technologies specifically designed to handle the unique challenges of unstructured data.

Panzura CloudFS and Panzura Symphony offer a comprehensive suite of features tailored to the unique needs of unstructured data management. These include:

Active and Passive Ransomware Protection

Panzura CloudFS's immutable data architecture ensures that unstructured data is protected from ransomware attacks. By creating unchangeable snapshots and enabling quick restoration of compromised files with a sub-minute recovery point objective (RPO), organizations can mitigate the impact of ransomware without losing critical data.

Panzura Detect and Rescue parses and assesses file operations within CloudFS in near real time, looking for anomalies, such as ransomware, malware and other threats. It also automatically intervenes; if suspicious activity is detected across multiple files in a short period, the system will interdict it by shutting off write access for the affected user or users. This prevents further damage or data modification while the security team investigates the issue.

Audit Logs and Alerting

Compliance and governance require detailed visibility into data access and usage. Panzura provides robust audit trails that track who accessed what data and when, enabling organizations to meet regulatory requirements and identify potential security vulnerabilities.

Managing alerts across multiple systems can quickly become overwhelming, and introducing additional alerts from Data Services might seem like it could add to the noise. That's why Panzura Data Services seamlessly integrates with SaaS SIEMs like Rapid7, providing centralized, streamlined alert management for improved efficiency and clarity.

However, no alert system is perfect from the start. Regularly reviewing and refining alert configurations is essential to reduce false positives and ensure new threats are accounted for. For example, adjusting thresholds or incorporating feedback from past incidents can significantly improve the accuracy of your alerts.

Data Classification and Search

Advanced tools help identify sensitive information within unstructured data, ensuring compliance with data protection regulations. The ability to search and classify data also aids in eDiscovery and other legal processes.

Panzura Data Services is an extension to the Panzura CloudFS hybrid cloud file services platform, providing a single, unified view and management of unstructured data within CloudFS. Data Services ingests both metadata and audit log data from CloudFS, to enable visibility and observability over the global file system and related infrastructure, and offer lightning-fast file search, audit, alerting, recovery and analysis across files in CloudFS.

Data Analytics with Panzura Symphony

Symphony includes tools for analyzing unstructured data. Uncover data usage patterns and trends, identify security risks, increase operational efficiency, and automate regulatory compliance across your entire unstructured data estate.

Aggregating data across multiple file systems and object stores, Symphony enables:

- Understanding of stored data, associated costs, and ownership.
- Identification of cost optimization opportunities based on data volatility, temperature, file type, file size, ownership, and metadata tags.
- Granular control at the folder, share, entire file system, or bucket level.
- Streamlined data placement, transformation, and restructuring to support AI workflows.
- Support for DevOps teams through webhooks, APIs, post-run actions, and relational database integration.

How to Take Action

To address these challenges and unlock the potential of unstructured data, CIOs should consider the following steps:

01 **Invest in Unified Storage Solutions:** Adopt a hybrid cloud file services platform like Panzura CloudFS to centralize and protect unstructured data, reduce costs, and improve access.

02 **Enhance Compliance Frameworks:** Use automated classification and data loss prevention tools like Panzura Symphony to identify sensitive information and enforce regulatory policies consistently.

03 **Streamline AI and LLM Training:** Simplify training workflows by consolidating unstructured data and reducing duplication, enabling faster and more cost-effective analytics.

04 **Break Down Silos:** Implement unified data management platforms to consolidate data across the organization and eliminate inefficiencies.

05 **Adopt a Strategic Vision:** Align unstructured data management with overall business objectives, focusing on cost savings, compliance, and innovation.

Panzura empowers today’s digital-first organizations to do impossible things with file data, making them more agile, efficient, and productive. They trust Panzura to help them consolidate dispersed data, see and manage data in and out of the cloud, make it more cyber-resilient and AI-ready, and ensure it is available to people and processes where and when it’s needed.

Discover how Panzura can fuel your success at **panzura.com.**