CTERA RESEARCH REPORT

The Cold Data Storage Report

The Hidden Budget, Security, and AI Readiness Strain Associated with Cold and ROT Data

Executive Summary

Most organizations have a strategy for buying storage. But what about a strategy for how to identify and manage cold data that sits in costly primary storage?

CTERA's analysis of 856 enterprise file storage environments finds that the overwhelming majority of files, 97.5%, have not been modified in over a year. Just under 10% of stored data is accessed in any given 90-day period. The active working set, the data users and applications regularly use and modify, represents only 4.4% of total storage capacity.

Primary storage is built to serve that 4.4%. The rest of the data is cold, or considered redundant, obsolete, or trivial (ROT), and is typically forgotten as it ages. Cold data costs the same to store, protect, and manage as active data. It also expands the ransomware attack surface, and creates governance liabilities when retention policies are absent and permissions are set too broadly. Additionally, as organizations invest in AI, that aging data degrades the quality of AI-powered results.

Organizations can address these challenges by gaining deeper visibility across their entire data estate. Only then can they identify which files are active and which are cold, and make informed decisions about what to move, at what priority, and how to manage the process at scale as data volumes grow.

Methodology

This report draws on CTERA's analysis of file-share discovery scans conducted across enterprise NAS environments in 2026. The dataset comprises 6,608,449,256 files across 16.1 PB of total storage, sourced from 354 distinct NAS source devices. Each scan captured file access patterns, modification dates, file types, and storage capacity distribution. Organizations represented span financial services, government, manufacturing, professional services, legal, and media sectors. Organizations ranged from mid-size businesses to those with more than 250,000 employees.

97.5%

of files have not been modified in the last year.

9.9%

of total data is accessed in any given 90-day period.

4.4%

of total storage capacity was modified by users and applications in the last 90 days.

Primary Storage Became an Archive

Primary storage is designed as fast, network-facing infrastructure for the files people open, edit, and share every day. In practice, it now holds something very different.

Eighty-three percent of total stored capacity lives in files that have not been modified in over one year. Eighty-eight percent of all files see no access in over a year. The active working set, the amount of data modified in the last 90 days, accounts for just 4.4% of total capacity.

IT professionals have a term for this accumulated inactive data: ROT, which stands for Redundant, Obsolete, or Trivial. It describes files that exist in duplicate, no longer reflect current operations, or serve no clear business purpose. Organizations accumulate ROT data due to a range of factors including:

- Employees changing teams or leaving the company
- Duplicate or multiple versions of the same file
- Email attachment downloads
- Automated system behaviors such as system snapshots
- Teams completing or abandoning a project

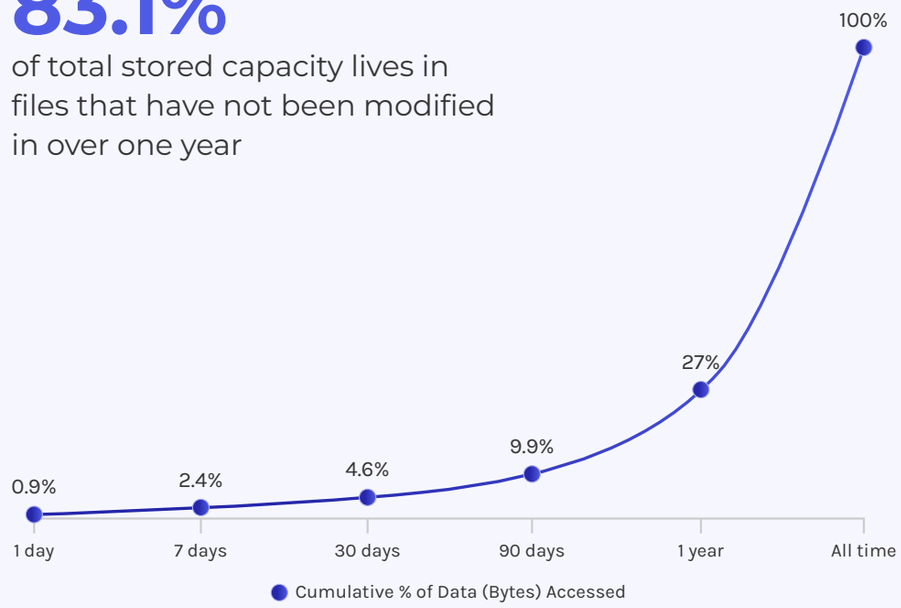
ROT data is typically framed as a storage capacity problem, but it creates a burden that extends well beyond capacity, impacting the entire organization. It increases the storage bill, widens the attack surface, adds governance complexities, and hinders AI investments.

The Data Cooling Curve

This chart shows the percentage of total storage capacity that was accessed in each time window, cumulatively. It indicates how much storage and data an archiving policy would affect. A 90-day threshold vs. a 1-year threshold produce very different scale of potential data to archive. IT leaders will need to balance potential cost-savings benefits vs. operational overhead of servicing requests to restore data for access.

83.1%

of total stored capacity lives in files that have not been modified in over one year



The Storage Bill Is Only One Part of Broader Data Costs

Storage capacity is one financial charge. But it is rarely the largest.

A file in primary storage generates more than one cost. That same file also lives in snapshots, backup systems, disaster recovery replicas, security scanning queues, and search indexes. The same logical terabyte generates several terabytes of physical storage consumption across these layers, along with the operational overhead of managing and eventually migrating all of it.

These charges appear across multiple budget lines, which is part of why the total cost of inactive data stays invisible. No single report consolidates them, so no single owner has visibility into the full number or the authority to act against it.

Consider what happens to a single file in primary storage. It lives in snapshots taken daily or weekly. It is copied into backup systems and replicated to a disaster recovery environment. Security tools scan it. Search systems index it. When migration eventually comes, every one of those copies must be accounted for and moved. The storage bill reflects one copy. The infrastructure carries many more.

For organizations carrying large bodies of inactive data, this is not a rounding error. It is a compounding structural cost that grows with data volumes.

Three reasons old data creates hidden costs:

1. Every file in primary storage also lives in snapshots, backups, and disaster recovery replicas.
2. Security tools scan and index every file, active or not.
3. Future migration projects must account for all data, not just the active subset.

Inactive Data Carries Active Risk

Inactive files feel low risk. They are old, rarely opened, and unlikely to attract attention. The assumption is that because they sit untouched, they also sit safely.

That assumption does not hold up.

Ransomware attacks and data exfiltration do not limit themselves to just active files. An attacker who gains access through a compromised credential can reach old files as readily as new ones. Both are accessible through the same network-facing shares.

An organization's attack surface depends not on how much data is active, but on how much data is accessible. Cold and ROT data should only be accessible to users and systems with a documented, current business need. Enforcing that boundary using the principle of least privilege directly reduces the attack surface.

Three things that compound when inactive data stays in primary storage

1. **Infrastructure costs accumulate across backup, disaster recovery, and security layers.**
2. **Lack of an audit trail and retention enforcement complicates compliance enforcement.**
3. **The attack surface expands, exposing dormant data to ransomware and other vulnerabilities.**

97.5%

of files receive no writes, making them perfect candidates for long-term storage governance policies with a WORM-immutability policy.

Data that is not expected to be modified should be protected accordingly. When no defined retention lifecycle exists, removing write permissions prevents unauthorized changes without requiring a formal policy. When a retention period is defined, applying a WORM policy enforces that lifecycle automatically, ensuring archived content cannot be modified, encrypted, or deleted by unauthorized parties.

Primary storage policies are built around active data with known owners. When data ages and ownership disappears, those policies remain on paper but stop being enforced in practice. That is the governance gap.

Without a governed archive workflow, there is no documented retention policy, no enforceable chain of custody, and no record that can be produced on demand. When a discovery request or regulatory audit arrives, that absence is itself a liability, independent of what the data contains.

Cold and ROT data sit in primary storage with no governance infrastructure around it. There is no defined retention period, no access restriction, and no documented record of what it contains or how long it should stay. These files share directory space and permissions with active business records. Anyone who can access current operations can reach old, ungoverned files just as easily.

Data Temperature Definitions

Tier	Modification Age (mtime)	Access Age (atime)
Hot (Active)	< 30 days	< 30 days
Warm	30–90 days	30–90 days
Cool	91 days – 1 year	91 days – 1 year
Cold	> 1 year	> 1 year

A **warm tier** keeps data online, immediately accessible, off primary storage, at lower cost. It suits data that must remain available to users, search systems, and AI retrieval pipelines without a restoration step.

Cool archive accommodates data that sees access infrequently but with predictable recovery requirements, typically within hours. It fits compliance records, older project archives, and collections with defined but occasional access needs.

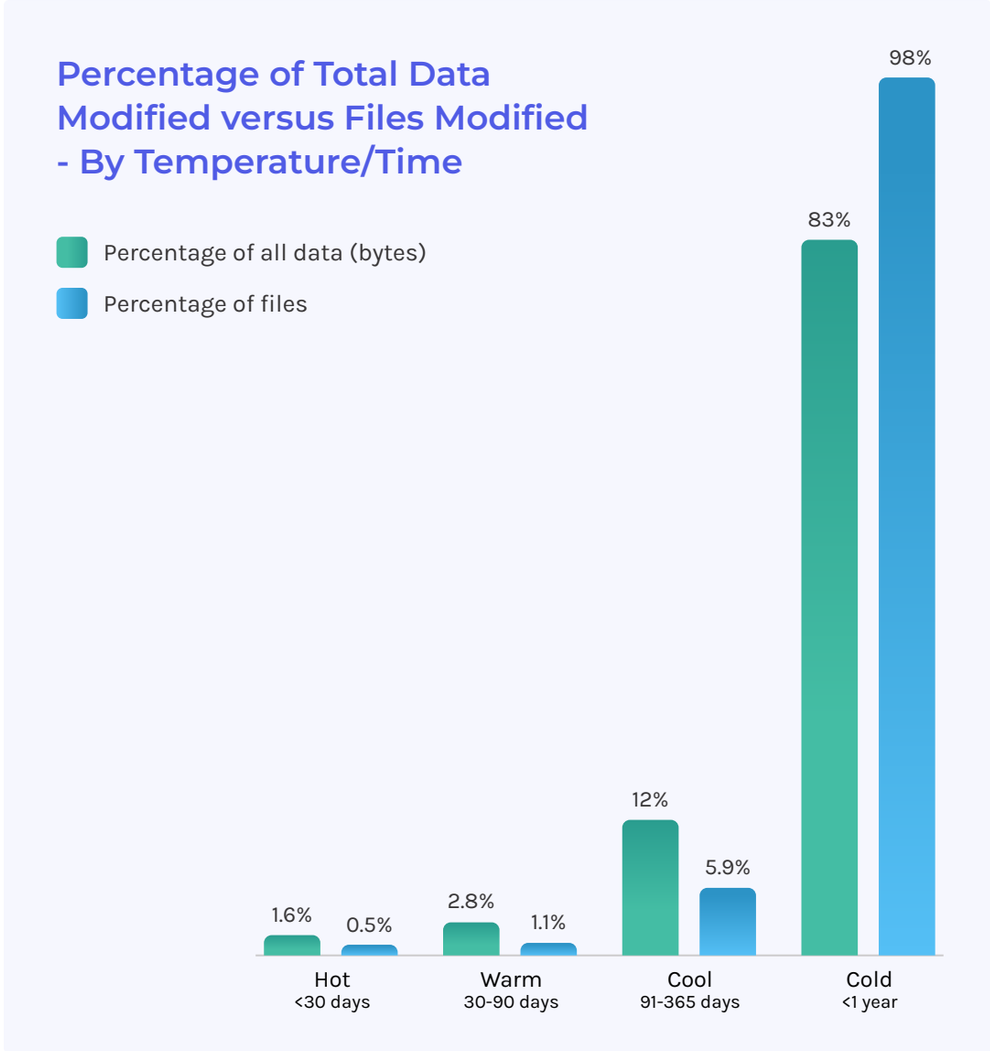
Cold data that is expected to remain untouched for extended periods is an excellent candidate for cold archiving. Retrieval should be measured in hours to days.

Cold & ROT Data Undermine Your AI Investments

Old data creates big and potentially negative impact on AI projects.

Enterprise AI systems rely on accurate data to provide useful and valid information back to the user. When those AI systems access old data, including data with superseded policies, closed project files, and years of accumulated records that are no longer relevant to current operations, the quality of those AI responses degrade. An AI-powered solution treats a five-year-old document and a current one with equal authority. It surfaces both with equal confidence.

Data quality has typically been considered a governance issue. Now it's impacting business outcomes across every department, as organizations make more investments in AI-assisted workflows, which rely on accurate and up-to-date data. The same data estate that inflates storage costs and expands the attack surface actively degrades the reliability of AI-generated answers across every function that touches it.



A Framework for Informed Decisions

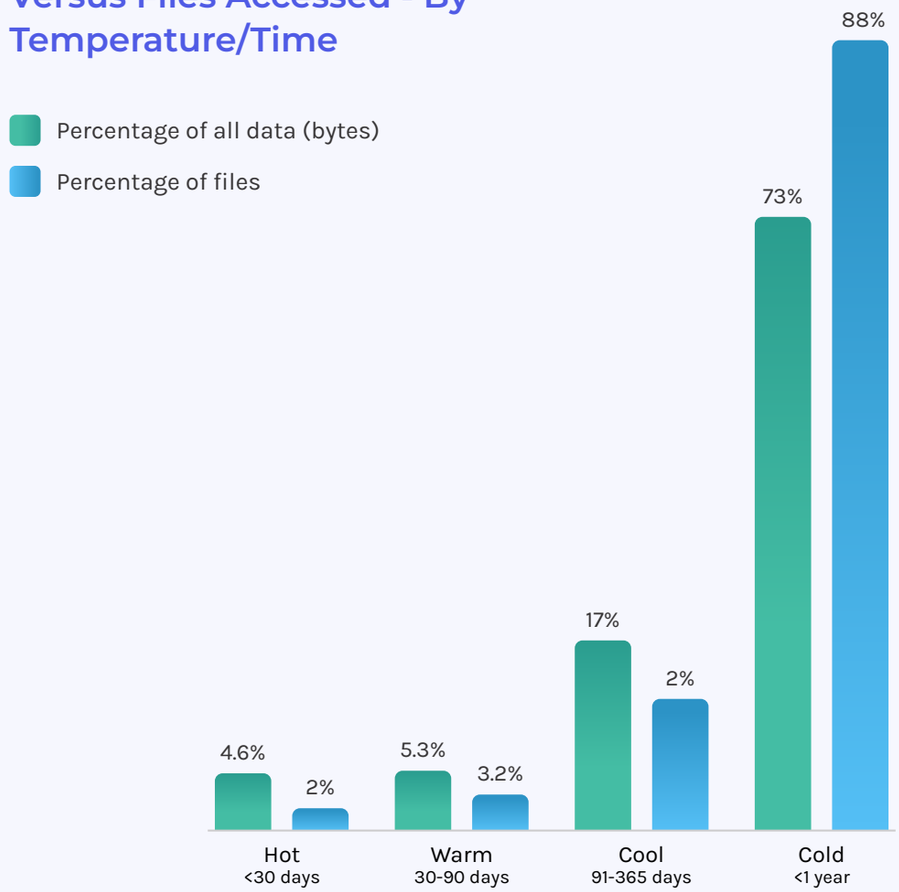
For most IT leaders, the case for action is clear. The challenge that often leads to stalled decision-making is identification. Determining which data should move to an archive, at what priority, and under what governance conditions takes more resources than most teams can deploy at scale. It is the reason most data lifecycle management programs stall before the team archives a single file.

Data Lifecycle Management and the Archive Spectrum

Data has a lifecycle, and managing it well means taking deliberate action at each stage rather than letting files accumulate unchecked. What begins as an active working file eventually becomes infrequently accessed, then cold, then a candidate for long-term retention or deletion. The archive spectrum is how organizations act on that progression.

Archiving is a range of decisions depending on the data and its use. Cold data requires different access and retrieval requirements depending on its use. Additionally, matching data to the right tier before moving it is critical to reducing operational friction and avoiding cost surprises later.

Percentage of All Data Accessed Versus Files Accessed - By Temperature/Time



Where to Start

A practical program does not require a complete governance overhaul first. Follow these steps for an organized process.

- **Map the estate.**
  Use storage telemetry, audit logs, and AI-assisted sampling to build a high-level catalog of major data collections.
- **Identify concentration points.**
  Prioritize repositories with high cost, weak ownership, and large amounts of inactive data.
- **Separate active and inactive content.**
  Move low-activity data out of primary user and application paths.
- **Apply archive controls.**
  Enforce retention, narrow permissions, define recovery workflows, and implement audit logging.
- **Improve the catalog continuously.**
  Learn from restores, searches, owner feedback, and AI usage patterns over time.

Moving data to less expensive storage without a governance layer is not archiving, as it lacks retention enforcement, access controls, audit logging, and defined recovery workflows.

Three Decisions Every IT Leader Should Make for Each Major Data Category

For every significant body of data, these three questions require separate answers:

1. Does this data need to be retained?
2. Is primary storage the best location?
3. Should the data appear in search and AI retrieval results?

Treating all data as a single keep-or-delete decision is where most retention programs fail. Data that must be retained does not necessarily belong in primary storage. Data that is searchable does not necessarily belong in an AI retrieval index.

AI Brings Clarity and Purpose-Driven Action

The organizations beginning to close this gap have stopped relying on manual processes to identify archiving targets.

Manual identification has always been the core barrier. Administrators review files by hand, build and maintain scripts, and sample repositories that grow faster than any team can keep pace with. Data lifecycle programs often fail because the identification step is too time-consuming to execute at scale.

AI-guided discovery creates a more reliable starting point. When AI surfaces archive candidates automatically, by access date, file type, size, and usage pattern, the question for IT teams shifts from “what should we archive?” to “do we agree with what the system found?” This creates a scalable process.

The combination of AI-guided identification with a governed execution layer, one that enforces retention policies, logs every operation, and maintains a chain of custody, is the foundation of a data lifecycle program that runs at the speed data accumulates.

The Solution: AI-Guided Data Lifecycle Management

The CTERA Data Archiving Solution surfaces which files are candidates for archiving, identifying cold data to move to affordable storage tiers, while enforcing retention policies and logging every action.

For organizations managing large volumes of cold and ROT data, the solution addresses the core challenges directly: reducing the compounding cost of inactive data in primary storage, limiting the security and compliance exposure that ungoverned files create, and improving the quality of data available to AI systems.

CTERA Data Archiving Solution Delivers:

- AI-guided identification of archive candidates by age and usage pattern.
- Governed, auditable archive and restore workflow with full operator logging.
- Configurable retention modes for administrative flexibility and strict compliance enforcement.
- WORM immutability via CTERA Vault for regulated environments.
- Reduced primary storage costs by moving inactive data to lower-cost cloud or on-premises object storage.

Assess your storage and data estate.

CTERA offers a free data assessments to provide a clear picture of cold data volumes, cost implications, and archive opportunities. Learn more at ctera.com.

TALK WITH US

ABOUT CTERA

CTERA is a global leader in the integrated data intelligence market enabling enterprises to lower their total cost of data ownership, maintain business continuity, and optimize workforce productivity by turning data into actionable insights to empower smarter decisions and drive sustainable business growth. With a foundation built on security, scale, and seamless integration, the CTERA Intelligent Data Platform enables organizations to align their data management strategies to continuously deliver against today's business needs and tomorrow's vision.

CTERA.COM info@ctera.com USA: (917) 768-7193 Intl: +972-3-679-9000

EB-AAI-0726