Managing Archival of Unstructured Data

Challenges and Strategies of Cost-Efficient
Long-Term Data Preservation, Security, and Accessibility

Introduction

Data is being generated faster than ever, and the rate is accelerating exponentially. Businesses are reluctant to discard information they've developed, and they are often legally required to preserve the information they have collected.
The overwhelming majority of information being saved is unstructured. Unstructured data is far more challenging to manage, maintain, and access than information stored in traditional databases.

This paper discusses:

The Acceleration of Data Generation

Unstructured Data Definition

How to Optimize Long-Term Storage and Retrieval

The focus is on the largest data archives: those containing less-frequently accessed, longterm storage. We examine two superior media for archival data: Optical and LTO.

The Growth of Data Generation and Storage

International Data Corporation (www.idc.com) is a technology analysis company highly regarded for its Global DataSphere. The Global DataSphere measures the data created, captured, and replicated across the world. It also examines how data is stored.

According to IDC, data generation is increasing by roughly 20% every year. Unstructured data accounts for 92.8% of all data in 2024.

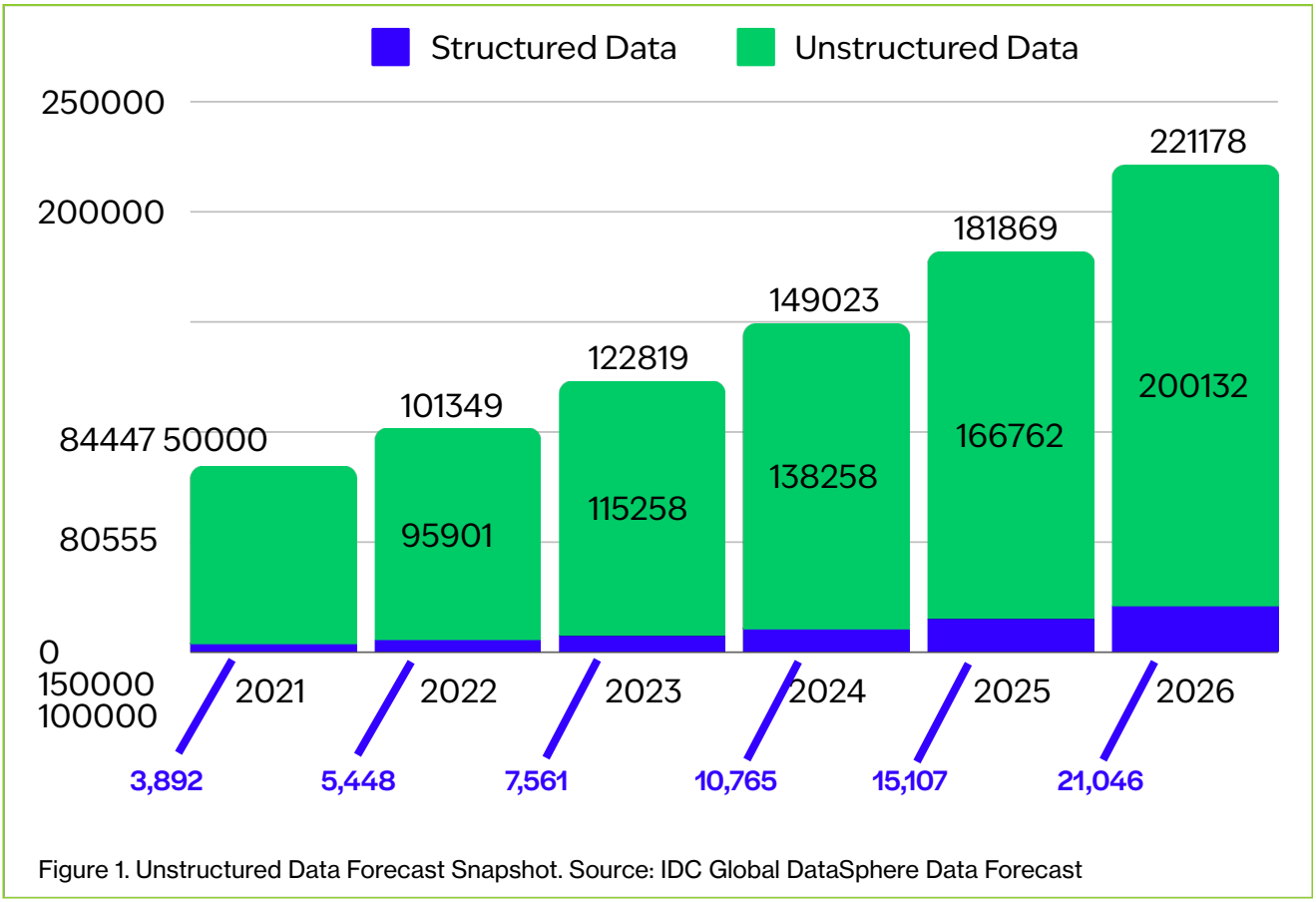


Figure 1. Unstructured Data Forecast Snapshot. Source: IDC Global DataSphere Data Forecast

The Growth of Data Generation and Storage

Table 1 breaks down the year-over-year growth of global data generation.

IDC also estimates that 80% of the information in organizational data archives is "cold data" which is infrequently accessed.

Table 1.
Increased Global Data Generation Year-Over-Year With Percentage of Unstructured Data to the Total

Year	Year-Over-Year Growth Structured	Year-Over-Year Growth Unstructured	% of Total Unstructured
2021			95.4%
2022	20.01%	19.05%	94.6%
2023	21.18%	20.18%	93.8%
2024	21.34%	19.96%	92.8%
2025	22.04%	20.62%	91.7%
2026	21.61%	20.01%	90.5%

Note: Adapted from IDC Global DataSphere Data Forecast.

Challenges of Unstructured Data

Structured data is carefully organized in an easily indexed format. Indexing by serial number or coded identifier makes locating and retrieving information simple.

Unstructured data lacks formatting to manage, index, catalog, search, store, or retrieve it effectively. This data includes email, documents, multimedia, and social media content which does not conform to traditional data structures. Without analysis and indexing before storage, data can only be searched by reading an entire archive – an extremely inefficient, often cost- and time-prohibitive proposition.

Many organizations use outdated storage or poorly maintained archival systems. Their data controls often perform poorly. They risk unauthorized data disclosure, are unable to use their archives effectively, and many of their archived records become corrupted, orphaned, or inaccessible.

Challenges of Unstructured Data

All storage – whether the data is structured or unstructured – must meet some common requirements. The data's integrity must be preserved, its security must be ensured, its accessibility must be maintained. It must also balance availability and cost.
Unstructured data introduces additional challenges:

Analysis during archival:

Efficient retrieval of unstructured data requires the data be analyzed and indexed as it is archived, not after it is already stored. This relies on the best profiling of data manageable.

Cost of analysis and metadata storage:

Analysis and storage of unstructured data must efficiently balance the computational processing and metadata storage cost with the value of that indexing at the time of retrieval.

Managing data with an intelligent tiered approach is a common way to meet these challenges. An additional option, also available in tiered storage, is to utilize object storage with its metadata and tagging options. Utilizing object storage will add virtually unlimited scalability.

Tiered Storage Planning

Tiered data storage is a new idea to many computer users, but it mirrors the way businesses, libraries, and archives have managed their stored information for centuries.
Files actively being worked on are kept close at hand, as in a desk drawer. Files being retrieved routinely for review might be kept in a filing cabinet in the office, while finished or large files are moved to an on-site file room, perhaps on a different floor. Files not active, but which must be kept indefinitely are indexed and moved to a warehouse that may be some distance away. These warehouses have come to be colloquially referred to as “cold storage.”

Large stores of digital information are handled in a similar way. The most active data is kept in the fastest local storage. As data is used less and less frequently, it is moved to slower, less expensive storage. These distinct categories of storage are called tiers.

Figure 2 shows a typical distribution of data in a four-tier storage hierarchy. Table 2 presents a comparison of performance and cost of the storage technologies listed in the four tiers.

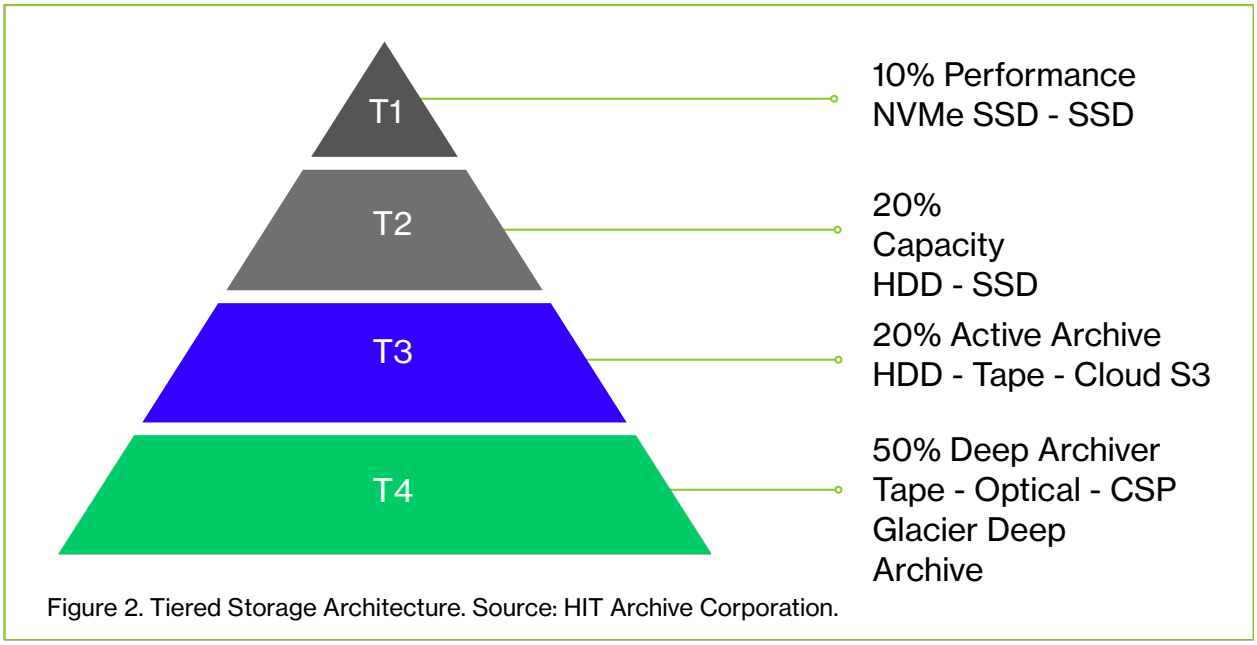


Figure 2. Tiered Storage Architecture. Source: HIT Archive Corporation.

The fastest, most expensive data storage is Tier 1. This is typically non-volatile memory in user machines or high-speed data processing machines.

Tier 2 is made up of solid-state drives (SSDs) and hard disk drives, typically in Networked Attached Storage (NAS) devices with RAID protection. High performance objects stores also fall in this category.

Tiers 3 and 4 are the focus of the rest of this paper. Data usually flows into these tiers when capacity is being reclaimed on devices in Tier 2, such as when a project or job is completed or when an accounting cycle is closed.

Table 2. Technology Performance and Cost Comparison

Technology	Performance			Challenges		Remarks
	Time to First Byte	I/O events/sec	Transfer Rate	Media Cost	Media Life	
NVMe SSD	10 µsec	500,000	3–20 GB/s	$100–300/TB	5–7 years	highest I/O and transfer rate
SSD	50–100 µsec	100,000–250,000	1–6 GB/s	$40–80/TB	5–7 years	cost efficienthigh I/O and transfer rate
HDD	3,000—12,000 µsec	75–100	80–160MB/s	$15–30/TB	4–6 years	Erasure code and RAID options to increase reliability and transfer rate
Amazon S3 Cloud Object Storage	less than 0.1 sec + network latency	3,500 -5,500	30 -300MB/s	$21/TB/month plus egress fees	not limited	High reliability. Backups at additional cost. Transfer rate is network dependent.
Tape (LTO)	1-2 minutes	up to 10	400 MB/s	$3/TB	20 years depending on use	Robotic + sequential accesshi transfer rate, E/C options
Optical	1-2 minutes	10 - 50	20 - 54MB/s	$6-10/TB	50 - 100 years	Robotic + direct accessPhysical WORM
Amazon S3 Glacier Deep Archive	less than 12 hours	n/a	n/a	$1/TB/month plus egress fees	not limited	Lowest cost under 300TBHigher cost for Petabyte storage12 hour SLA Time to First Byte

Note: NVMe SSD = Non Volotile Memory express. SSD = Solid State Drive. HDD = Hard Disk Drive. LTO = Linear Tape-Open. RAID = Redundant Array of Independent/Inexpensive Disks. WORM = Write Once Read Many.

Medium- and Long-Term Archive Solutions

Tiers 3 and 4 are for long-term data storage, and they are reserved for information which the user accepts longer retrieval times. Some Tier 4 storage systems may require up to 12 hours before the first data returns from a search or retrieval request.

When balancing cost against reliability and availability, most organizations get their best value from storage in two media: magnetic tape and optical disc storage. Many businesses use both to get all the benefits these formats have over each other.

Figure 3 shows how an organization might decide which storage is most appropriate for the data they archive, then use an archive management tool to direct it to different storage systems based on whether it is structured or unstructured, whether it needs to be overwrite-protected and permanent, and the likelihood of it being requested frequently.

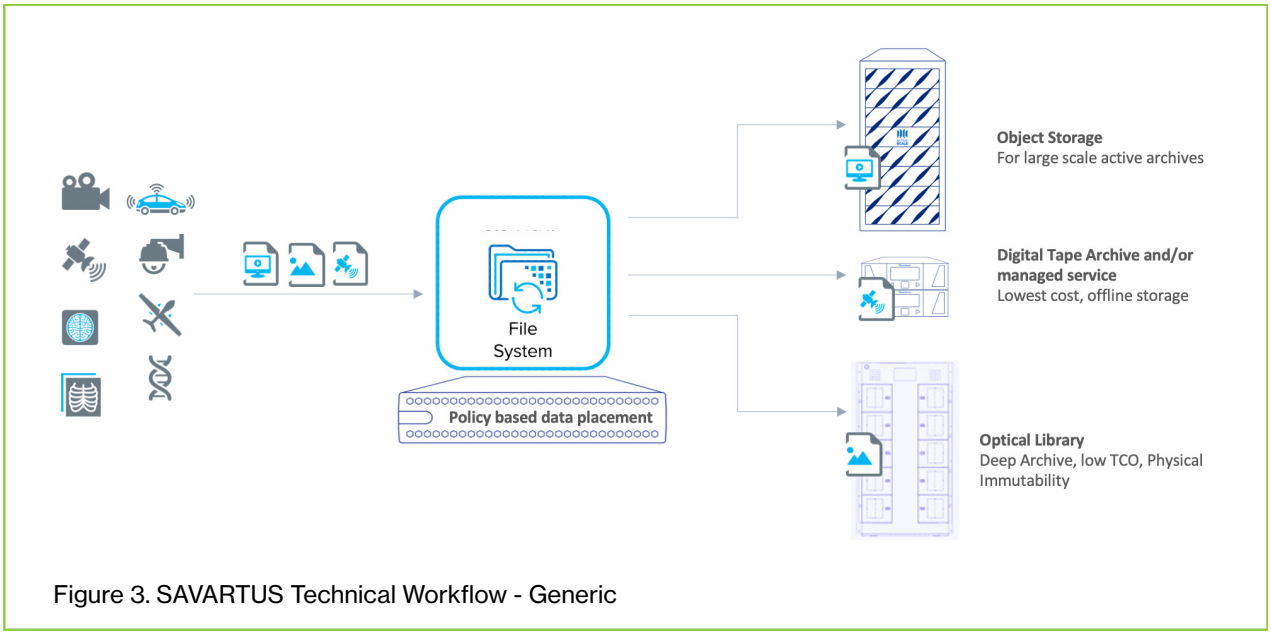


Figure 3. SAVARTUS Technical Workflow - Generic

Storage Media and Their Benefits

Optical Data Archives

Optical disc storage is secure, reliable, durable, and affordable.

The most advanced optical storage format for high capacity at low cost is developed by Folio Photonics (www.foliophotonics.com). Current Folio media capacities are at 500GB with a road map to 10TB/disc by 2030.

Folio Photonics' technical advances are the keys to greater capacity and further cost per Terabyte improvements. Improved data density and an increase in the number of recording layers per disc will support the increasing capacity per medium. Other technological advances in the optical pickups will support the increasing number of recording layers.

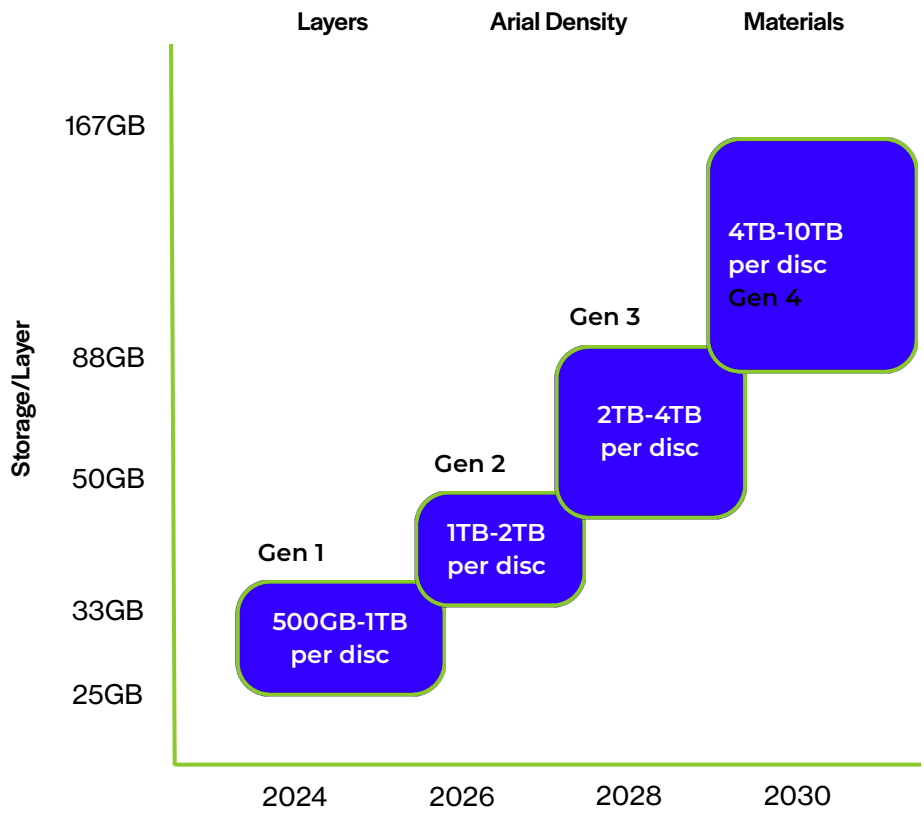
Figure 4 (see below) shows Folio Photonics projection of storage capacity in their next four generations of media and drives.

Storage Media and Their Benefits

SAVARTUS Enterprise Laser Storage

SAVARTUS Enterprise Laser Storage is utilizing the advances from Folio Photonics to offer the most reliable optical storage at the most attractive price.

Projects to 100TBs in a 10-disc cartridge



Layers
Multi-Layer architecture
- 8 layers per side today Scaling to
- 32+ by 2023

Areal Density
Crosstalk Cancellation
- Inter-Symbol Interference Cancellation
- Multilevel recording

Materials and Media
Multicolor multiplexing
Polarization multiplexing

Figure 4. Folio Photonics technology generations. Source: Folio Photonics.

Storage Media and Their Benefits

SAVARTUS Enterprise Laser Storage offers:

- **Energy, money, and planet savings:** Ultra-low power and cooling requirements significantly reduce the power needed to operate and house, cool, and retain optical data archives. Carbon emissions can be reduced as much as 60%.
- **Ultra-long lifespan and durability:** With a minimum lifespan of more than 50 years, optical media ensures a long archival lifetime and superior data integrity.
- **High use resilience:** The Folio Photonics optical format can be read frequently without degrading the disc.
- **Immutability:** Folio Photonics optical discs offer Write Once Read Many (WORM) characteristics. Once data is written, it cannot be altered by user mistakes, software errors, or cyberattacks.
- **Resilient to data disclosure:** Hardware-level encryption keeps your archive safe from unauthorized access or accidental disclosure.
- **Size:** SAVARTUS Enterprise Laser Storage systems are available as rackmount units or in full-rack packages.
- **Scalability:** Storage can be added as your needs grow. Future generation storage will be significantly less expensive for the same or better capacity and speed.
- **Backward Compatibility:** Folio Photonics optical discs support backwards read/write support for previous media generations. This greatly reduces the need for data migration.
- **Amazon S3 Interface:** Support for the Amazon S3 Glacier storage interface gives ELS optical archives much greater agility and flexibility to interoperate with cloud-based archives.
- **Immune to Environmental Factors:** Folio Photonics optical media are impervious to static and moisture.

SAVARTUS Enterprise Laser Storage is ideal where long-term data retention is required by regulation or where data integrity and long-term preservation are highly valued.

Storage Media and Their Benefits

LTO Magnetic Tape Storage

Linear Tape-Open (LTO) is a format for magnetic tape storage developed by the LTO Consortium – a group formed by Hewlett Packard Enterprise, IBM, and Quantum. They began working together in the 1990s to create an open standard for magnetic tape storage.
There are nine generations of LTO, with development and planning already underway on generations 10 through 14. Figure 5 shows the projected capacities.

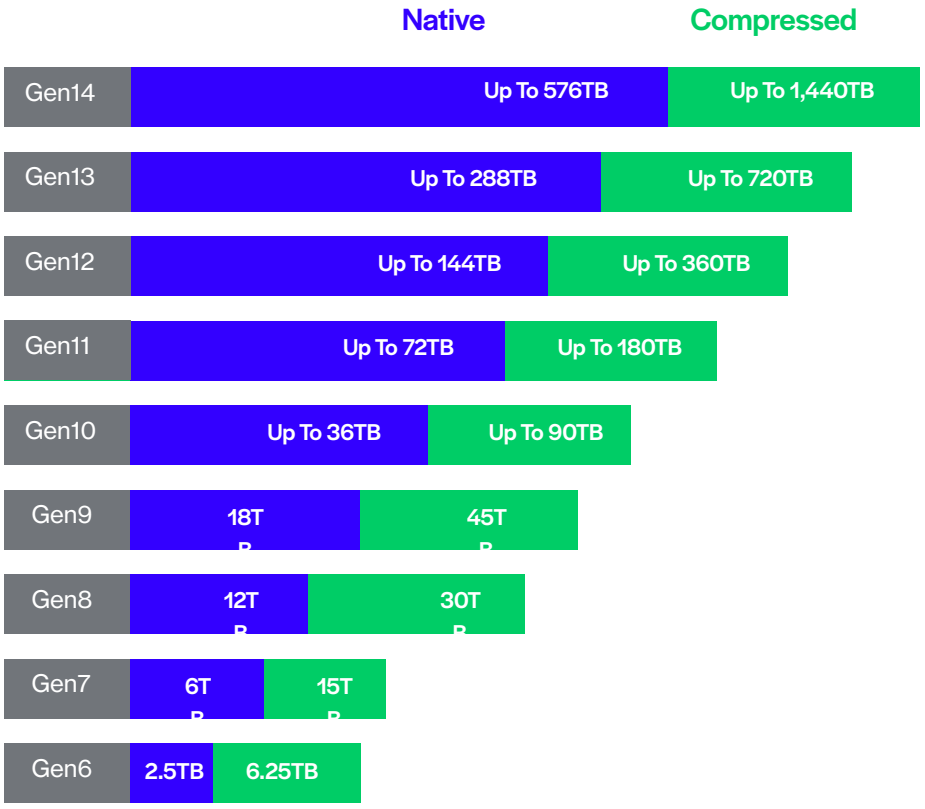


Figure 5. Projected LTO Capacities. Source: LTO Ultrium Roadmap by the LTO Consortium (Hewlett Packard Enterprise, IBM, Quantum)

There are nine generations of LTO, with development and planning already underway on generations 10 through 14. Figure 5 shows the projected capacities.

Storage Media and Their Benefits

Benefits of the LTO format include:

Highest capacity at low cost:
Magnetic tape storage using the LTO format has the highest capacity at the lowest cost of any reliable data format.

Exceptional data integrity and security:
The LTO format has the lowest Bit Error Rate on reading and writing. Tape-drive-level encryption and hash code techniques support data verification for best data protection and accuracy.

Mid-term lifespan for the least accessed data:
Magnetic tapes using the LTO format have a shelf life of 20 years. Tape durability in the 6th through 9th generations is 20,000 end-to-end passes.

Firmware-based data protection:
The Write-Once-Read-Many (WORM) functionality inherent to optical disc storage is accomplished with firmware controls within LTO tape drives. Attempts to overwrite tape already written are not honored.

Amazon S3 Object Storage on Tape:
Support for the Amazon S3 Glacier standard allows integration of on-premises archives with cloud storage for off-site backups, additional capacity, or to meet regulatory requirements for redundant storage.

Conclusion

Data generation is growing exponentially and the need or desire to archive unstructured data is growing, too. Reliable and affordable archive storage is crucial.

SAVARTUS Enterprise Laser Storage optical systems and LTO magnetic tape systems help businesses preserve data and ensure its integrity. We build sustainable, cost-effective, and efficient archives for our customers.

Our systems are designed and built for the greatest capacity and capability while offering the best lifetime value.

About SAVARTUS

SAVARTUS is a trusted partner with more than 40 years' experience helping organizations optimize operations, reduce storage costs, recover from cyber-attacks and system disasters, and ensure data integrity for long-term preservation and compliance.

We offer more than a data management solution. We are creating a revolution in data management and environmental responsibility to create a sustainable digital future. Our technology reduces humanity's carbon footprint, ensures data longevity, and increases security. These are not theoretical technologies, but systems used today in medical, military, government, and education sectors worldwide.

SAVARTUS isn't just about preserving data. It's also about preserving our planet for future generations.

Visit us at www.savartus.com.