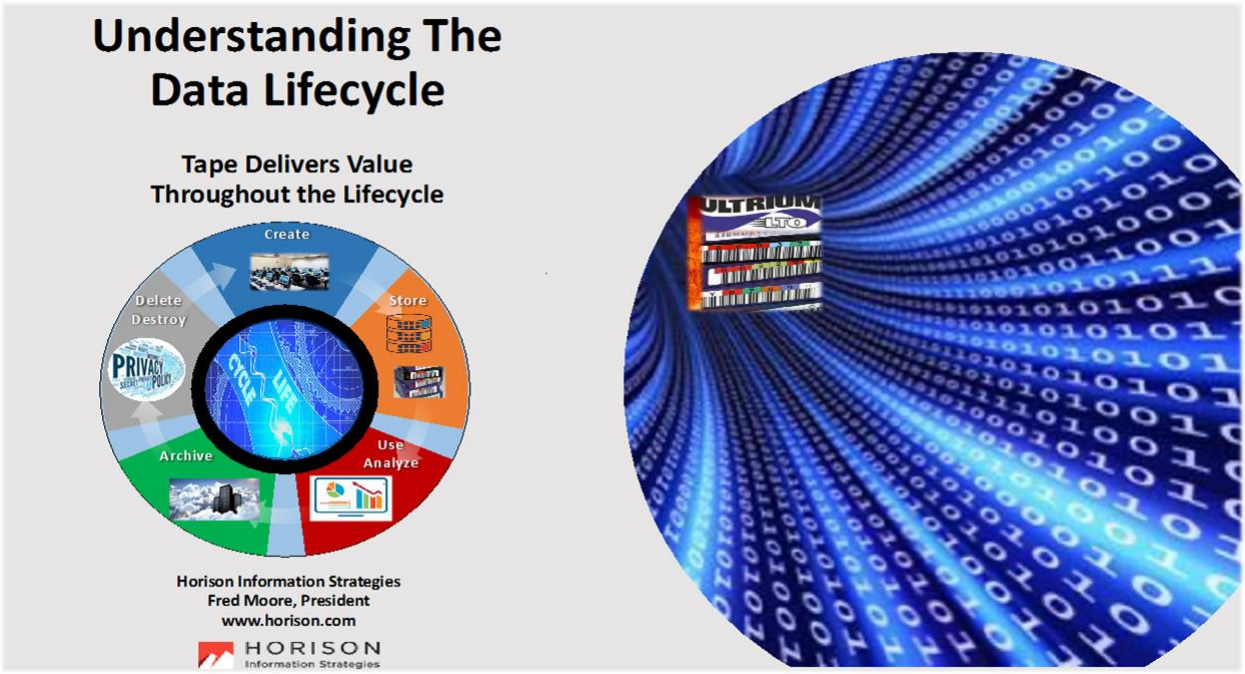


Introduction

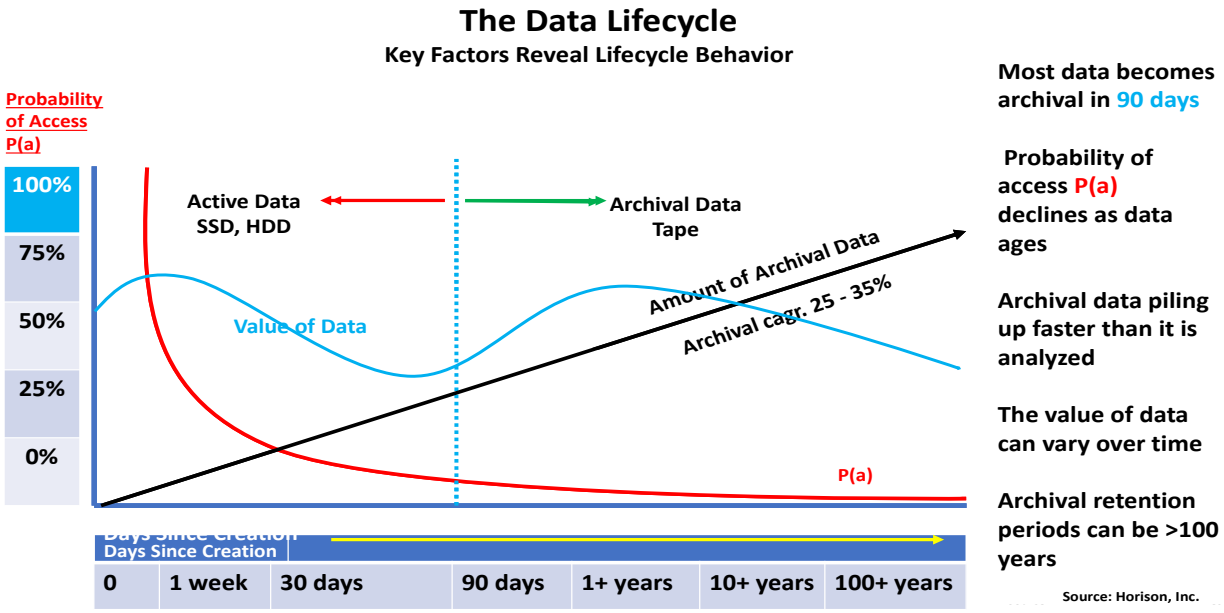
Data creation is on an unprecedented growth trajectory as data has become the most valuable asset for many companies. Today's largest companies are primarily based on digital services like Alphabet (Google), Amazon, Apple, Facebook and Microsoft, none of which were on the top ten list in 2010. Nearly 90 percent of the world's digital data was generated over the past ten years, however, much of this data may not be touched for years, flooding the archives. Given this trajectory, the traditional storage paradigm will need to transform itself – quickly. Balancing a storage architecture to effectively manage data through increasingly longer lifecycles for a wider variety of workloads while minimizing costs presents *the* next great storage challenge. As a result, the role tape plays in the data lifecycle as the most cost-effective storage solution for the ongoing archival data pile up couldn't be more important.

The Digital Transformation Takes Off

The amount of stored data is projected to reach 8.9 ZB worldwide by 2024. In addition, over 60% of all data is archival and it could reach 80% (~7.12 ZB) or more by 2024, making archival data by far the largest storage class. *For most data types, the vast majority of their lifecycle will be spent in archival status.* There are an estimated 2.5 EBs of data beginning their digital lifecycle journey each day and that pace is accelerating with the growth of the sensor based IoT and the wide reach of the internet. IDC predicts that there will be over 41 billion IoT devices connected to the internet by 2025. There are presently an estimated 4.13 billion internet users worldwide which means that more than half of the global population of 7.8 billion is connected to the world wide web. There are also an estimated 2.87 billion smartphone users worldwide and nearly 100 million people started using smartphones in the past year – all of these sources creating new digital data!

When Does Data Reach Archival Status?

The data lifecycle typically begins with data capture, acquisition or ingest and data will spend most of its active useful life on SSDs or HDDs. With an understanding of I/O activity profiles and access patterns, the storage architecture to best address each stage of the data lifecycle can be implemented. For the majority of data types, the probability of access *P(a)* begins to fall during the first month and typically falls below 1% after 90 to 120 days of inactivity reaching archival status. Some data becomes archival upon creation and can wait years for any reference or analysis adding to the archival pile-up. The biggest archival challenges are managing the sheer amount of raw data and discovering its unknown value. Today the most cost-effective archive solutions are high-capacity robotic tape libraries deployed in local, cloud and remote locations. The largest tape libraries can store over 1 EB ($1x10^{18}$) of non-compressed data.



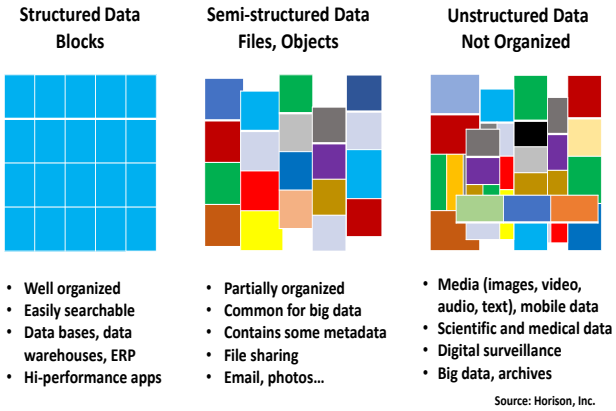
Three consistently observable profiles have evolved that describe typical data behavior over its lifecycle. See chart above:

1) **the probability of reuse *P(a)*** of most data declines as the data ages
2) **the value of data** to a business can change over time based on a variety of circumstances
3) **the amount of data** steadily increases as it ages since more data is being kept for longer periods of time than ever before, pushing data into archive status and leaving it there indefinitely

Since data has become the most valuable asset for most businesses, managing and protecting data throughout its lifecycle has become the most critical storage management task. With or without intelligent management tools, most organizations should be able to identify their most critical applications and have a clear understanding of the value they provide to the success of the business. From a security perspective, the storage archive copy is often the only copy of the data. If a second copy of the archive was created, the total amount of archive data stored could potentially double.

What Types of Data Are Being Created?

Data is created and organized using structured, semi-structured and unstructured formats and is stored in blocks, files or as objects. Big data is a combination of these formats that can be mined and used in AI and ML projects, predictive modeling and advanced analytics. The increasing amounts of semi- structured and unstructured data represents much of the enormous archival storage challenge.



Structured data is the easiest to organize and search because it is neatly contained in rows and columns and its elements can be mapped into fixed pre-defined fields in a database. Structured data is stored in a block format and is best suited for OLTP, random access and performance critical applications residing on SSDs or HDDs.

Semi-structured data is a form of unstructured data having *some* classifying or metadata characteristics but doesn't conform to a rigid database structure. The actual content is unstructured, but it contains some descriptive data such as name, address, the time and date sent of an email, or a digital photograph which is date, time and location stamped, but the image itself is completely unstructured. Semi-structured data can be stored in file or object formats. File systems often use the file name as metadata and are still the predominant format, but files simply can't scale to the level of objects. Object storage has become the de-facto standard format for storing data in the cloud and several new software products enable tape to cost-effectively store billions of files a single name space. Storing objects on tape has become a significant advancement for archival storage.

U*nstructured data* represents the largest percentage of data generated today and it can't be contained in a relational row-column database. The lack of structure makes it difficult to search, manage and analyze. There is a wide variety of unstructured data types including images, text files, sensor data (IoT), scientific data, audio, video, pictures, and social media data. Archival data is normally stored as unstructured or semi-structured data.

Stages of the Data Lifecycle

Data Creation and Capture

In this stage, data comes into an organization, usually through data entry, acquisition from an external source, or signal reception transmitted from IoT sensors. Data classification processes defining the value of data, service levels, and identifying the owners and stakeholders is performed here. This stage is where governance requirements are assigned as data may be public, private, sensitive, restricted or top security.

Data Storage

Data storage refers to where the data will optimally reside in the tiered storage hierarchy (SSD, HDD, tape or cloud) based on user defined attributes including price, performance, capacity, activity levels, data value and retention periods. Intelligent software has made the sizeable TCO advantages of tiered storage a reality and creates an optimized storage infrastructure by storing the right data in the right place. In this stage, data protection, security and availability requirements including backup methods, RAID, erasure coding, snapshot, CDP, and de-duplication are assigned to data sets based on business criticality and impact. Tape and HDDs combine to play a significant role here as they are both used for backup storage with tape often used as the offsite or geographically remote backup target.

Data Use and Analysis

This lifecycle stage represents the active life of data and supports the organization's key applications, business objectives and overall operations and as long as data remains active it will stay in this stage. Much of this data will be used in compute, performance and data-intensive workloads such as OLTP, large databases and Big data analytics using ML and AI to uncover trends, patterns, and correlations in large amounts of raw data to improve making data-informed decisions.

Data Archiving

The data archiving stage manages data either before it becomes active or after its active life. In this stage, data is usually moved from active production environments to archival storage or directly from the IoT into a data lake where it is securely stored (archived) for future reference. Archive data is no longer present at the source often making archived data the last copy of data and requires strong protection, preservation, integrity, long media life and redundancy methods. A growing percentage of digital data will be retained forever such as global news, entertainment, historical events, sports championships and scientific discoveries. LTO tape is the most cost-effective archive solution available given it has the lowest $/TB price and TCO of any storage option. With a media life of 30 years or more, LTO is best suited to address archive requirements among today's storage choices. The active archive enhances the archive by integrating HDDs as a cache buffer with tape libraries to provide faster access to archival data.

Data Deletion, Eradication, Degaussing and Destruction

Data deletion or destruction is the final stage of the data lifecycle and can take several forms. When data is deleted, it is removed from a file folder or catalog preventing future application access but physically remains on the media and is recoverable with special software. With data deletion, the released space can be re-used to store new data. For WORM (Write Once, Read Many) HDDs and tape media, the space cannot be reused or re-written when data is deleted.

Eradication (data wiping, erasing or overwriting) is performed by special software that ensures data is rendered inaccessible without destroying the physical drive and media markers. Eradication writes a pattern of ones and zeros over the existing data fields rendering it meaningless, but the drive or media capacity can be reused. Bad SSD and HDD sectors usually cannot be overwritten but may contain

recoverable information. The most effective eradication solutions ensure that all user data, including data in overprovisioned spaces that are inaccessible to the user and host(s), is physically erased and verified.

Degaussing is the process of totally destroying data by demagnetizing or eliminating the unwanted magnetic field (data) stored on disk or tape media. Degaussing uses a powerful electromagnetic field to render your media and drives inaccessible causing permanent damage to the special servo control data that is written on the media at the factory by the manufacturer. Note that degaussing doesn't work on SSDs as they don't store data magnetically, so applying a strong magnetic field won't eliminate anything.

Destruction or shredding completely destroys data on SSD, HDDs and tape and is highly secure, fast and efficient. Shredding reduces electronic devices beyond repair into tiny fragments no larger than 2 millimeters and the remains are eligible for recycling and electronic waste management services.

Summary

Managing data throughout its lifetime is now required discipline to optimize its value, ensure its availability and minimize its potential for cybercrime attacks or natural disasters. The LTO Consortium has pushed tape capacity, reliability and media life to record levels. Advanced laboratory demonstrations signal steady advancements with relatively few limitations in tape recording technology for the decade ahead pushing native cartridge capacities beyond the 300 – 400 TB range further increasing tape's archival value. As a result, the role of tape has become a fundamental and critical component of data lifecycle management and will become even more so as the amount of data created and preserved continues its relentless growth trajectory.

Sponsored by the LTO Consortium

Linear Tape Open (LTO) Ultrium is a high capacity tape storage solution developed and continually enhanced by Hewlett Packard Enterprise, IBM and Quantum and promoted by the LTO Program. Linear Tape-Open LTO, the LTO logo, Ultrium and the Ultrium logo are registered trademarks of Hewlett Packard Enterprise, IBM and Quantum in the US and other countries.

About the author

Horison Information Strategies is a data storage industry analyst and consulting firm specializing in executive briefings, industry seminars, market strategy development, whitepapers and research reports encompassing current and future storage technologies. Horison identifies disruptive and emerging data storage trends and growth opportunities for end-users, storage industry providers, and startup ventures.