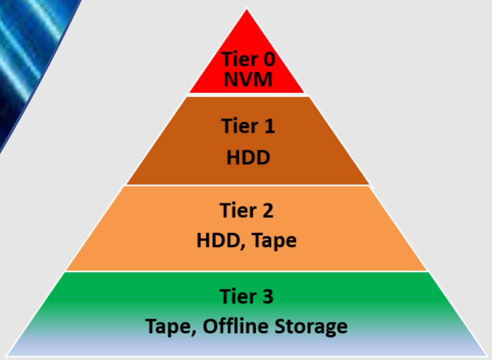
The Anatomy of the 100-Year Archive

Preparing for Infinite Data Retention



Horison Information Strategies
Fred Moore, President
www.horison.com

HORISON
Information Strategies

Introduction

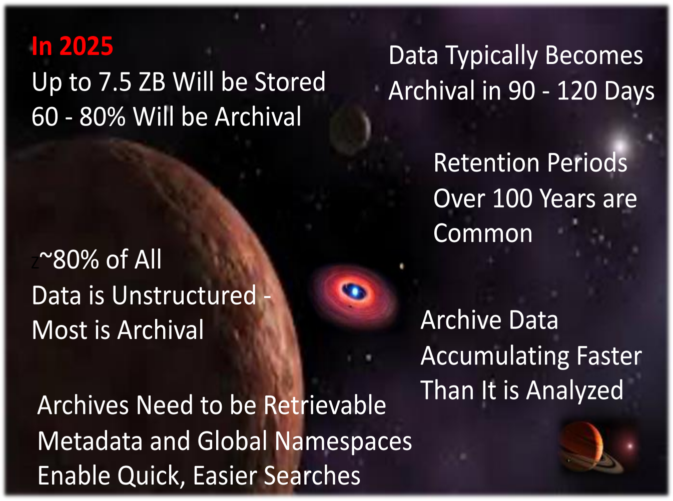
Relentless digital data growth is inevitable as data has become critical to all aspects of human life over the course of the past 30 years and it promises to play a much greater role over the next 30 years. Data retention requirements vary widely based on the type of data, but archival data is rapidly piling up everywhere. Much of this data will be stored forever hoping that its potential value will be unlocked. Not only is data creation on an unprecedented trajectory, it has become the most valuable asset in many companies. For example, in 2006, oil and energy companies dominated the list of the top five most valuable firms in the world, but today the list is dominated by firms based on digital content like Alphabet (Google), Apple, Amazon, Facebook and Microsoft. Equifax, a U.S. based consumer credit reporting agency that collects information on over 800 million individual consumers and more than 88 million businesses worldwide, suffered a data breach in 2017 affecting 143 million consumers. They're now facing a class-action lawsuit of up to $70 billion, representing the perceived value of the data at risk. Movies, sports events, legal, medical and pharmaceutical data, top security government and national security data are a few examples of data that may be retained forever. Many businesses plan to retain data in some digital format for 100 years or more making long-term retention necessary. As a result, today's archives demand a more intelligent solution that leverages the advanced capabilities of intelligent data management software and high-availability, scale-out hardware. For many organizations, facing petabytes and potentially exabytes of archive data for the first time can force the redesign of their entire storage strategy and infrastructure making it a key strategy for enterprises and a required discipline for hyperscale data centers. As a result, the dawn of the 100-year archive is rapidly approaching.

How Much Data is Archival?

Industry estimates vary but the amount of data to be stored in 2025 is projected to be ~7.5 ZB according to IDC's 2018 Data Age report. Approximately 1.1 ZB of total storage capacity was shipped in 2019 across Non-Volatile Memory devices (SSDs), HDDs, and magnetic tape media with HDDs making up the majority (over 80%) of the shipped capacity.

Today at least 60% of all digital data can be classified as archival, and it could reach 80% or more by 2025, making it by far the largest and fastest growing storage class while presenting *the* next great storage challenge. Most archival data have never been monetized as the value of data remains unknown, but companies are just now realizing that digital archives have great potential value. Companies looking to be relevant between now and 2025 will need to understand the role archive data can play in their organization's long-term success and how data archiving strategies will evolve during that period. Given this trajectory, the archival storage paradigm will need to reinvent itself.

The Digital Universe



In 2025
Up to 7.5 ZB Will be Stored
60 - 80% Will be Archival

Data Typically Becomes Archival in 90 - 120 Days

Retention Periods Over 100 Years are Common

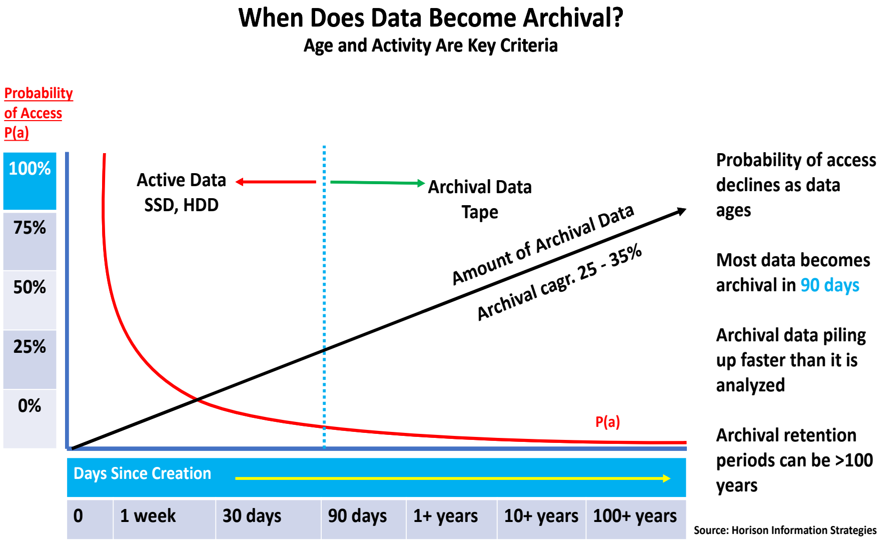
~80% of All Data is Unstructured - Most is Archival

Archive Data Accumulating Faster Than It is Analyzed

Archives Need to be Retrievable
Metadata and Global Namespaces Enable Quick, Easier Searches

When Does Data Reach Archival Status?

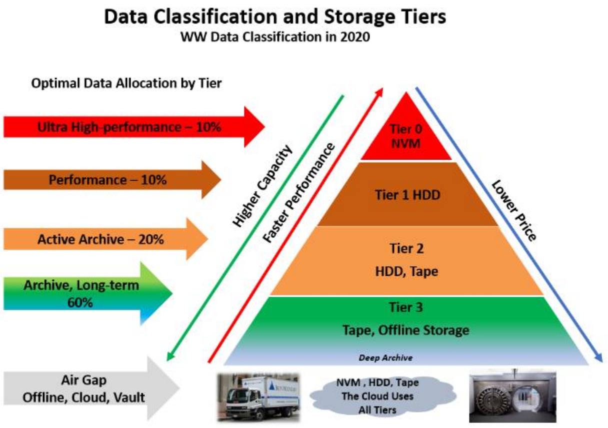
Archival data will continue to be the largest and fastest growing data classification segment. For most data types, the probability *P(a)* (probability of access) begins to fall after one month and typically falls below 1%, most often between 90 - 120 days. Some data becomes archival upon creation and can wait years for reference or further analysis adding to the archival pile-up. Today the most cost-effective solution for archival data are high-capacity tape robotic libraries used in local, cloud and remote locations. The 100-year archive will add significant capabilities to this foundation. See adjacent data lifecycle chart.


When Does Data Become Archival?
Age and Activity Are Key Criteria

Source: Horison Information Strategies

Data Classification Guidelines

The data classification process is critical to effectively manage data and becomes more critical as data usage patterns are in constant flux and the storage pool constantly grows. Though you may define as many storage tiers as you want, four de-facto standard tiers of classifying data are commonly used: Ultra-high-performance data (OLTP - Online Transactional Processing), Performance Data (Mission critical), Active Archive (lower activity data) and Long-term archive which is the largest and fastest growing storage segment. Data classification enables the alignment and mapping of data characteristics with the optimal storage technology tier. Moving as much data as possible to the lowest cost storage tier is the core key ingredient for modern archiving strategies and returns the greatest economic value.

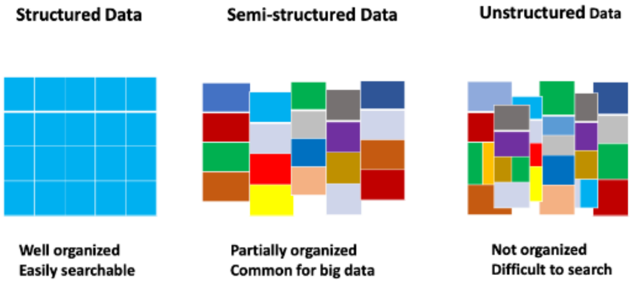

Data Classification and Storage Tiers
WW Data Classification in 2020

Looking ahead, the ideal archive solution will classify data and create metadata upon ingest. The top four challenges to unlocking the value of archival data are:

1) making archival data accessible at ingest (by assigning classification, index, catalogs, metadata)
2) managing the long-term archival storage infrastructure
3) ensuring that only the potentially needed archive data is actually stored
4) ensuring the security and availability of archival data.

Data Classifications – Structured, Semi-structured and Unstructured Data

The rapid increase of semi- structured and unstructured data today is important to understand as these classifications represent much of the archival data challenge. The software that manages and analyzes these classifications is a critical component of the 100-year archive.



Source: Horison, Inc.

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. Structured data is most often stored in a database and isn't archival.

Semi-structured data has some classifying characteristics but doesn't conform to a rigid database structure. The actual content is unstructured, but it also contains some structured data such as name and address of the email sender and recipient, and time sent, or a digital photograph which is date and time stamped, geo tagged, and may have a device ID, but the image itself is completely unstructured.

Unstructured data is the largest percentage of all digital data; however, it can't be contained in a row-column database. The lack of structure makes unstructured data difficult to search, manage and analyze.

Data Formats - Understanding Blocks, Files and Objects

Files, blocks, and objects are the storage formats that hold, organize, and present data in different ways, each with their own capabilities and limitations. Block storage chunks data into arbitrarily organized, evenly sized volumes; File storage organizes and represents data as a hierarchy of files in folders; and object storage manages data and links it to associated metadata. For archival data, object storage is quickly becoming *the preferred format*.

Object storage evolved out of the need to optimize capacity scaling capabilities of large volumes of unstructured and archival data and to store and retrieve any amount of data from anywhere using the Internet. The S3 cloud service provides object storage access through a web interface and has grown to become the de-facto standard format for on-premise archival and cloud storage services. This capability allows for a seamless integration between cloud storage and other block and file data storage solutions.

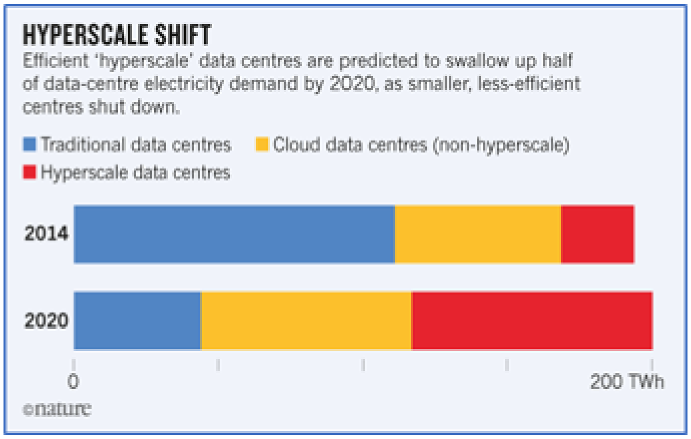
Object storage uses metadata that can be generated automatically for each object or defined by the user, enabling faster search and a wide variety of analytics. The best object storage solutions can scale to hundreds of petabytes in a single namespace without any performance degradation. Object storage is highly compatible with today's emerging technologies such as AI and the IoT which are further fueling the archival data pile up. Objects are also immutable and can't be modified without re-writing the entire object making object storage ideal for the long-term retention of unstructured archival data. Metadata is *the* key to unlocking the archives and includes details like age, privacies/securities and access contingencies.

	Block Storage	File Storage	Object storage
Type of Data	Structured	Semi-structured	Unstructured
Usage Profile	Performance Intensive	General Purpose	Archival, Immutable
Use Case	Databases, Mission Critical Apps, OLTP	Hierarchical System, Folders	Unlimited Capacity Scaling
Capacity	Lower Capacity	Medium to High Capacity	Very High Capacity
Metadata	None	Limited Metadata	Considerable Metadata
Storage	DAS, SAN (SSD, HDD)	NAS, Tape	HDD, Tape, Cloud
Performance	μ, milliseconds	ms, secs, mins…lengthy search times	ms, secs, mins …better search times

What We've Learned from Cloud and Hyperscale Data Centers (HSDCs)

HSDCs face insurmountable growth of disk farms which are devouring budgets, overcrowding data centers and creating enormous energy and carbon footprint problems, forcing data migration to more cost-effective tape solutions. Cloud archive storage services are relatively inexpensive, but cloud data retrieval/transfer (bandwidth) costs soar as the amount of data transferred increases. Amazon Glacier, Amazon Glacier Deep Archive, and Microsoft Azure are examples of hyperscale (green) cloud storage services for large-scale archiving. *If you want to have ongoing cloud savings, you need to manage the lifecycle of your cloud data. That means managing your cloud archives with an effective archive strategy.*

There are projected to be as many as 570 HSDCs worldwide in 2020 representing the fastest growing type of data center. HSDCs appear to be the epicenter for modern archiving strategies. Data centers and information technology consume about 2% of the world's electricity currently and is expected to [soar up to 8%]() by 2030. For the hyperscale world adding disk is tactical – adding tape is strategic. Tape cartridges spend most of their life in a library slot or on a shelf and don't consume energy when not mounted in a tape drive making tape the ideal archival storage choice. Advanced, easily scalable air-gapped tape architectures that can support erasure coding, geo-spreading with exascale capacities while providing the lowest TCO, highest reliability, and improved cybersecurity protection will play an increasing role to address and contain the enormous HSDC storage challenges that lie ahead. HSDCs and large enterprises are primed to take advantage of the rich functionality of a 100-year archive strategy.



HYPERSCALE SHIFT

Efficient 'hyperscale' data centres are predicted to swallow up half of data-centre electricity demand by 2020, as smaller, less-efficient centres shut down.

The 100-Year Archive - Building the Archive of the Future

The optimal archival architecture of the future will have the capability to store, protect, preserve, retrieve and easily scale into exabyte-level capacities spread across multiple locations. For most enterprises and service providers, a single location may not be sufficient to deliver a high availability data protection strategy since the entire data center might go offline or its access be blocked by a natural or man-made disaster, cyber-attack, EMP or other catastrophic event.

The Anatomy of the 100-Year Archive

Source: Horison, Inc.

Key components of the 100-year archive will include intelligent active archive software with smart data movers, data classification and metadata capabilities, a highly scalable tape library technology, RAIL (Redundant Arrays of Independent Libraries) architectures, Elasticsearch, erasure coding and geo-spreading data across zones in different locations for higher fault-tolerance, redundancy and availability.

Active Archive Software

The key to the 100-year archive is intelligent, cloud-scale, object storage software that geo-spreads unstructured and object data to manage and move mixed file and object workloads. Advanced data availability and integrity capabilities are used for a world-class archive and cloud infrastructure. The 100-year archive can optimally be deployed in a multi-site configuration to geographically disperse data for extreme (high) availability.

Geo - spreading

Geo-spread erasure coding software spreads erasure codes across multiple geographically distributed availability zones for highly efficient site redundancy. Advanced availability and integrity are essential for a world-class archive infrastructure and can effectively be deployed in a three-site configuration to geographically disperse data for extreme availability. Even with a complete data center outage, the three-site configuration delivers continuous data availability for uninterrupted operations.

Erasure coding

Erasure coding is a method of data protection in which data is broken into fragments (shards or small chunks), expanded and encoded with redundant data pieces and stored across a set of different locations or storage media. Erasure coding only requires subsets of the original data to recover data essentially making the legacy backup process unnecessary. Erasure coding coupled with geo-spreading is best suited for data grids, object storage and large archival storage making it ideal for the 100-year archive.

Data classification catalogs, tags, metadata creation

Data classification, identification and tagging procedures label data based on its relevance to the enterprise and to make it easily searchable and trackable. It also eliminates multiple duplications of data, which can reduce storage and backup costs while speeding up the search process. The ideal data classification software creates metadata on ingest creating a detailed inventory of an organizations data assets to quickly store and retrieve with elastic search for analytical or business purposes.

System analytics and insights capabilities

Analytics proactively monitor and model storage and user trends over time to better manage large scale storage systems. Analytics can run in place without creating concerns about data movement or ETL (extract, transform and load) overheads. This increases efficiency because there is no overhead for data movement and no need to maintain redundant data copies saving storage expense.

Elasticsearch

Elasticsearch is a distributed, open source search and analytics engine for all types of data, including textual, numerical, geospatial, structured, and unstructured archival data. Since its release in 2010, Elasticsearch has quickly become the most widely used search engine.

Optimal archive technology

Modern tape offers the lowest TCO, highest reliability, 30+ year media life, exabyte+ capacity scaling, minimal re-mastering, and the lowest acquisition cost available. Modern tape libraries can offer exabyte capacity levels with optimized robotic movements. Today’s tape technology is nothing like its predecessor.

RAIL (Redundant Arrays of Independent Libraries)

RAIL provides significant improvements in tape data transfer time and much higher tape availability. Like RAIT (Redundant Arrays of Independent Tape), with RAIL data is striped across tape cartridges, but each cartridge is in a different robotic library. The 100-year archive includes geo-spreading tape libraries in separate geographic locations for even higher degrees of fault-tolerance. Ideally a 3-site tape library configuration can deliver the optimal mix of performance and fault-tolerance protection.

Data protection

Tape is inherently “air-gapped” meaning it can’t be hacked, providing cybercrime protection through encryption and WORM for end-to-end data protection. Advanced data classification improves data protection capabilities by facilitating proper security responses based on the type of data being retrieved, transmitted, or copied. Geo-spreading delivers further availability with multi-site redundancy.

Non-disruptive data migration between storage technologies for archive access

Since most archive data must be kept longer than the life of many storage technologies and longer than some applications, a 100-year archive must provide capability for data to be migrated to a new technology without disruption, to any users accessing the archives.

Conclusion

Many businesses plan to retain data in digital form for 100 years or more mandating the emergence of a more intelligent and highly secure, long-term storage infrastructure. The 100 year-archive defines a comprehensive strategy to store and protect enormous amounts of archival data while enabling its value to be realized. Classification procedures that label data based on its relevance to the enterprise are quickly becoming critical as the size of digital archives is now reaching the order of petascale ($1x10^{15}$), exascale ($1x10^{18}$) and will approach zettascale ($1x10^{21}$) capacities in the foreseeable future. The hardware, intelligent software and management components to implement a cost-effective archive are now in place – sooner or later, chances are high that you will be forced to implement a solid and sustainable archival plan. The extreme growth of archival data and need for better access requirements present data centers with major obstacles to overcome and could mandate implementation of a 100-year archive strategy in the next few years to meet the challenge.

About the Sponsor

Quantum.

Quantum focuses on creating innovative technology and solutions to help our customers get the most value from their data. With 40 years of storage know-how, Quantum's technology, solutions, and services help customers capture, create, and share digital content – and preserve and protect it for decades. Whether it's unlocking the potential of digital content, powering breakthrough innovations, creating entertainment that enriches lives, or keeping nations secure, Quantum works with customers and partners to make the world a happier, safer, and smarter place.

About the Author

HORISON
Information Strategies

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.

WP00261A-v01 Sept 2020