



WHITE PAPER

HOW TO STORE AND PROTECT YOUR COMPANY'S DIGITAL ASSETS FOR DECADES



CONTENTS

- About Quantum3
- Introduction3
- Requirements of a 100-Year Archive4
- Considerations for the High-Performance Stage of the Data Lifecycle4
- Best Practices for Protecting Data for Decades.....5
- File versus Object Storage Formats.....7
- Storage Technologies for Archive Storage7
- The Need for Data Classification.....8
- On-Premise Versus Public Cloud Considerations9
- Next Steps11

ABOUT QUANTUM

Quantum has been building and servicing the largest digital archives in the world for over two decades. We've built industry leading technologies for the purposes of capturing, processing, analyzing, and preserving large unstructured data sets. Learn more at www.quantum.com.

INTRODUCTION

The most valuable asset of many companies is their digital data—their files. Extracting value and insight from digital data is driving business today and will drive the markets and economies of tomorrow. This unstructured data—stored as files and objects—is growing exponentially, and it is estimated that by 2025 this 'unstructured' data will represent 80% of all of the data on the planet.

Businesses have an imperative to start treating this data like the valuable asset that it is.

- These assets need to be maintained for decades – and the cloud is not the answer.
- The data cannot be lost. It is too valuable of an asset, so it must be protected and stored in multiple places to protect against disaster.

Because of the growth rate, these data sets can quickly grow into petabytes and exabytes, adding another layer of complexity when it comes to preserving and protecting data at this scale, for decades. But storing and protecting the data is only the beginning. You need to ensure the people and applications using this data can access it—to gain insight, to produce products, to speed time to market. So the data must be managed, easy to search, and easy to analyze.

How do you store the original copy of *Star Wars* or *Steamboat Mickey*? How long do you keep the film footage of Babe Ruth hitting a home run? How long do you store patient records, MRI images and CAT Scans? Surveillance footage used in a trial? Genome sequencing images used to design and produce a new drug?

Many of our customers refer to this as the “100-year archive” problem, and over the next few years most large enterprises, agencies, and research organizations will need to plan for a 100-year archive.

In this paper, we will explore:

- **The Requirements of a 100-Year Archive**
- **The Anatomy of a 100-Year Archive**
 - **Considerations for the High-Performance Stage of the Data Lifecycle**
 - **Best Practices for Protecting Data for Decades**
 - **File versus Object Formats**
 - **Storage Technologies for Archive Storage**
 - **The Need for Data Classification**
 - **On-Premise versus Cloud Considerations**
- **What's Next**

REQUIREMENTS OF A 100-YEAR ARCHIVE

Let's start with the basic requirements and attributes that a 100-year archive needs to have:

Attribute	Explanation
Ability to scale to exabytes as data grows.	Devices like video cameras, genome sequencers, satellites, drones, and connected cars are creating unstructured file and object data at unprecedented rates. It is not uncommon for unstructured data sets to grow into many hundreds of petabytes and even exabytes in the matter of a few years. Any 100-year archive plan needs to be able to handle exabyte scale.
Storage services that span from the fastest "hot" storage to "cold" storage.	Any 100-year archive starts with data that needs to be ingested quickly, and worked on. In this stage, speed matters—and technologies like NVMe flash drives are being used to accelerate data pipelines. On the other end of the spectrum, large cloud providers are increasingly deploying tape technology for long term cold storage. And of course, hard drives will continue to play a role. We'll talk about DNA storage later. Any 100-year archive will need to use a combination of storage technologies.
Data must be protected, it cannot be lost.	Once this data has been stored, it must be protected. We will explore considerations and best practices for protecting data at this scale.
Data must be easily searchable and accessible, sometimes years or decades later.	Someone once said "the data in the archive is not valuable until it is." Data that was created and stored years and even decades ago will often need to be retrieved, and when this happens is not always predictable.
Ability to migrate data between storage technologies and generations in a way that is non-disruptive to users.	Since the data must be kept for longer than many storage technologies, and in many cases longer than some applications, any 100-year archive must provide means for data to be migrated without disruption to any users accessing the archiving.

With these basic requirements in mind, let's look at the anatomy of a 100-year archive, including some of the chief considerations and best practices.

CONSIDERATIONS FOR THE HIGH-PERFORMANCE STAGE OF THE DATA LIFECYCLE

When data is first created, it needs to be uploaded and ingested quickly, and because the files are large they require a very high-speed file system. People and applications then work on the data for some time, and when data is actively being worked on it is typically in file format.

Today these 'fast' pools of storage are built on NVMe flash storage with options to access that data using file, block, or object interfaces. Key requirements are very fast streaming performance, as well as I/O intensive workloads that drive performance requirements. The use of GPU clusters adds a capability for massive parallel processing, and traditional storage systems can't feed the GPUs fast enough.

Traditional scale-out NAS designs and network protocols such as TCP/IP, are starting to break down in the face of both the volume of data being created, and the rates at which the data needs to be ingested and shared. Increasingly, software-defined file system clusters running on NVMe flash are the preferred method, with special clients that integrate with application and GPU clusters.

BEST PRACTICES FOR PROTECTING DATA FOR DECADES

Backing up a database or a virtual environment is a simple proposition. The servers are backed up over the network using a backup application, or snapshots are used for continuous data protection. This data is retained for 7 years for compliance purposes, then expired. Protecting very large data sets that need to be kept for decades is a very different proposition.

Challenge	Considerations
Data is “too big to backup”	Trying to backup this data using a batch backup method puts too much strain on the networks, and in many cases there is no backup window large enough.
Replicating the data is prohibitively expensive	Many users try to protect this data by replicating between scale-out NAS clusters. With both product and service costs, this becomes very expensive.
The cost of keeping data on flash or disk for decades adds up quickly	Although flash and disk are great for high performance and nearline storage, storing data on either medium for decades is very costly and unreliable.
Protecting against localized disasters	To protect against any localized disasters, it is highly recommended that the data is copied to two or even three locations.
Protecting against ransomware and cyberattacks	The threats of ransomware, malware, and other forms of cyberattacks are on the rise. The data in any 100-year archive needs to be protected against the threat of ransomware.

Erasure Encoding as the Preferred Protection Method

Erasure encoding, or erasure coding, has emerged as the best method for protecting very large data sets at scale. With erasure encoding, objects are split into chunks and spread across multiple nodes, and in some cases even multiple sites. Erasure coding algorithms can be adjusted for protection and safety, as well as for storage efficiency.

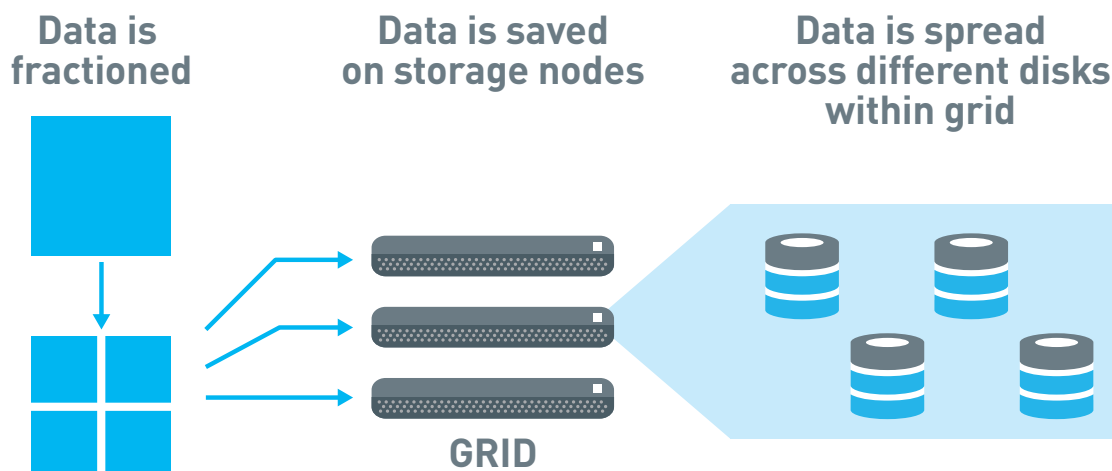


Figure 1 – Illustrating Erasure Encoding

Data that is stored and then erasure encoded is protected—it doesn't need to be backed up.

Geo-spread Erasure Coding is Most Efficient TCO for Three Copies

The best practice for storing data in a 100-year archive would be to keep 3 copies of the data, in 3 locations, which protects against for hardware and software issues, and localized disasters. Geo-spread erasure coding refers to software that can spread erasure codes across multiple geographically distributed sites. The example below illustrates the efficiency of geo-spread erasure coding:

Table 1 – Comparing Usable Capacity as a % of Raw Capacity

Protection Method	Single Copy of Data	Two Copies of Data	Three Copies of Data
Block storage with RAID and replication	84% Using RAID	42% Using RAID and replication	28% Using RAID and replication
Geo-spread erasure-coded object storage*	80% Using erasure coding	40% Using erasure coding	61% Using Three-Geo erasure coding

*Based on Quantum ActiveScale™ software with BitSpread™

Ransomware Protection and the Need for an Offline Storage Copy

Ransomware attacks have become more common and more sophisticated. Malware can reside on a server and lay dormant and replicating or backing up corrupted data doesn't help—the data is still corrupted.

For this reason, many companies are adopting a tried and true method for protecting against ransomware—using tape. Tape is unique in that data stored on tape is offline, or air gapped from the network. Data that is stored on a tape is much more secure.

Tapes stored in libraries are "offline"
and "air gapped" from the network

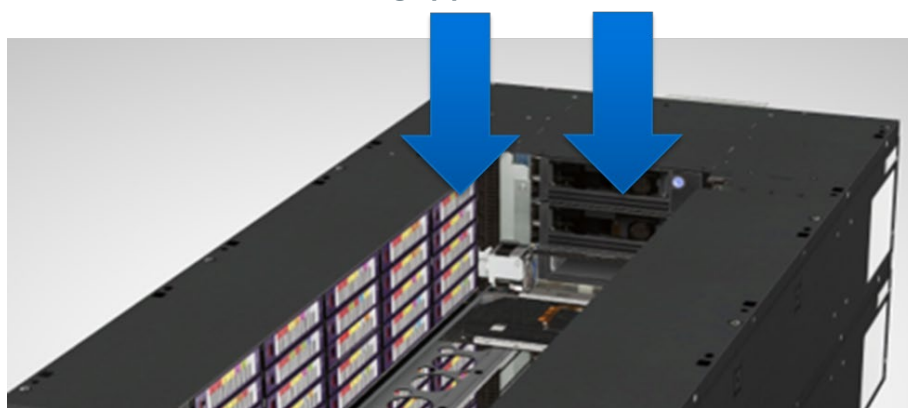


Figure 2 – Illustrating the "Offline" or "Air-Gapped" Attribute of a Digital Tape Archive

FILE VERSUS OBJECT STORAGE FORMATS

One of the current considerations when it comes to 100-year archives is which format to store data, file or object. Increasingly, the lines are blurring and many vendors offer both very fast storage that supports a file or object interface, as well as “very safe” storage with either a file or object interface.



File System

`C:\folder\music.mp3`

- Data stored as **files**
- **Hierarchical** organization
- Addressed via file interface



Object Storage

`GET /object/Kbg18n7qepo`
`PUT /object/Kbg18n7qepo`

- Data stored as **objects**
- **Flat** namespace
- Addressed via “S3” interface

Figure 3 – Illustrating the Difference Between File and Object Storage

However, storing data in object format for the long term has some distinct advantages:

- **Enables massive scale:** Object stores can easily scale to billions of objects and exabyte scale.
- **Easy to search and index at scale:** A flat namespace, and the fact that metadata is included as part of the object, make object stores and the object format easier to search at very large scale.
- **More “cloud friendly”** to enable users to run cloud services on the data stored in their archives.

STORAGE TECHNOLOGIES FOR ARCHIVE STORAGE

The table below explores the most common alternatives available today for archive storage.

Technology	Pros	Cons
Flash	<ul style="list-style-type: none">• Low power• Some “cold” options exist	<ul style="list-style-type: none">• Expensive
Hard Drives	<ul style="list-style-type: none">• Lower cost than flash• Erasure coding for protection	<ul style="list-style-type: none">• Tech roadmap reaching limits• Requires power
Tape	<ul style="list-style-type: none">• Lowest cost• Requires almost no power• 30+ year data life• More reliable than disk• “Greenest” option• Air-gapped storage copy	<ul style="list-style-type: none">• Difficult to manage• Difficult to access / retrieve
Optical	<ul style="list-style-type: none">• Low cost	<ul style="list-style-type: none">• Not widely deployed or proven
Cloud (same tech options as above)	<ul style="list-style-type: none">• Easy to ‘write and forget’	<ul style="list-style-type: none">• Data access fees• Data security
Synthetic DNA	<ul style="list-style-type: none">• Low cost, low power, very dense	<ul style="list-style-type: none">• Not yet viable technically or commercially

As noted in the table above, tape is a key technology to consider and use for 100-year archive. Tape has many distinct advantages over other storage technologies, until that time that Synthetic DNA storage becomes commercially and technically viable.

Tape's advantages include:

- Lowest cost
- Requires almost no power
- 30+ year data life
- More reliable than disk
- "Greenest" option
- Air-gapped storage copy to protect against ransomware

In addition, tape technology, i.e. magnetic media, has a viable technology roadmap to continue to improve areal densities—unlike hard drive technology.

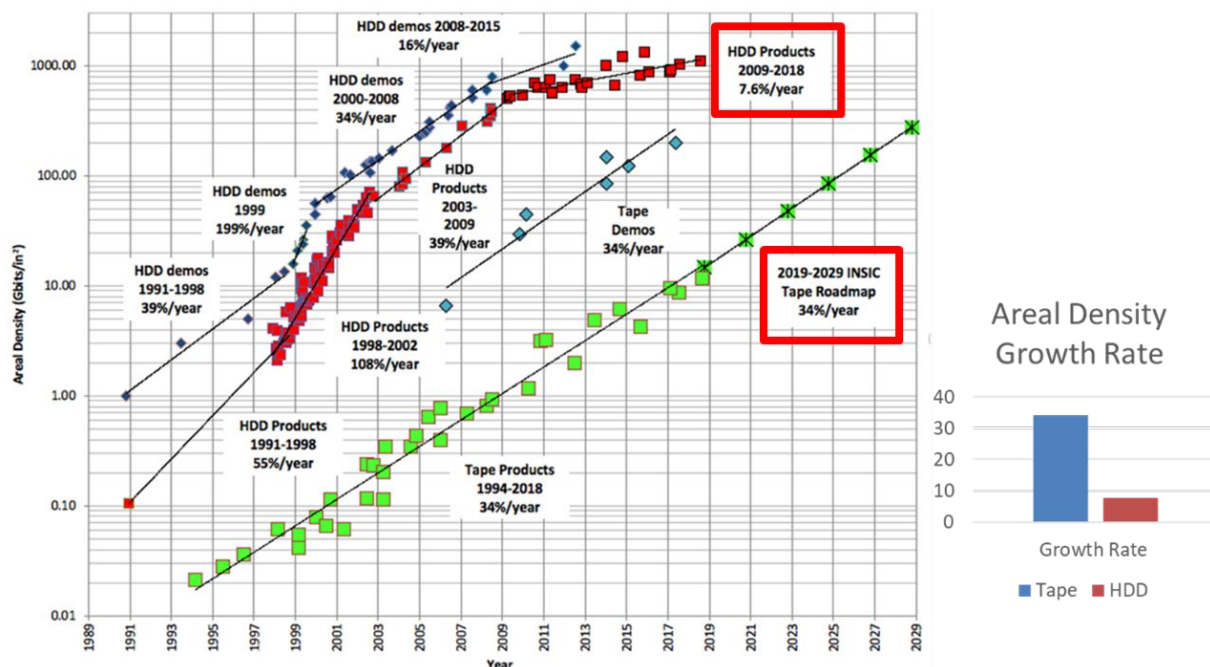


Figure 4 – Comparing Areal Density Roadmaps Between Digital Tape and Hard Drives

THE NEED FOR DATA CLASSIFICATION

So far we've talked about different storage technologies, and some of the chief considerations of how to build a 100-year archive. But the biggest challenges are how to make the archive accessible and searchable, and the first step to achieve this is **zero-touch real-time data classification on ingest**.

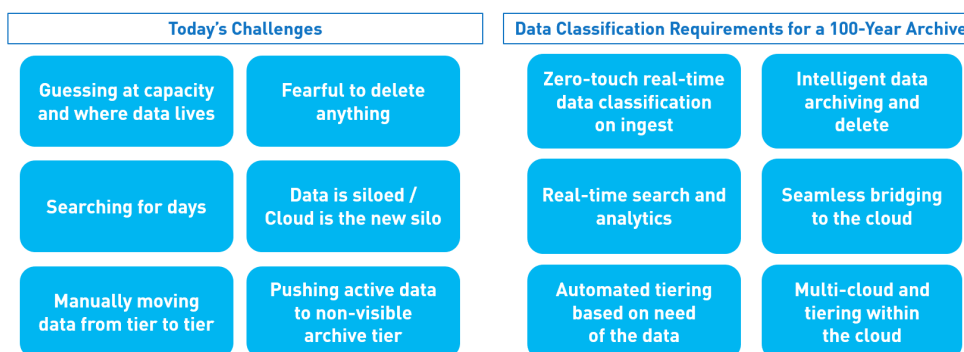


Figure 5 – Common Challenges and Data Classification Requirements

As illustrated below, a properly designed data classification engine needs to include powerful, real-time analytics, transparent access to the archived content, and a scalable elastic search engine.

This will enable users to build an intelligent archive so data can be placed where and when you need it, between on-premise and cloud infrastructures, and ultimately build an archive where data can be quickly turned into insights and results.

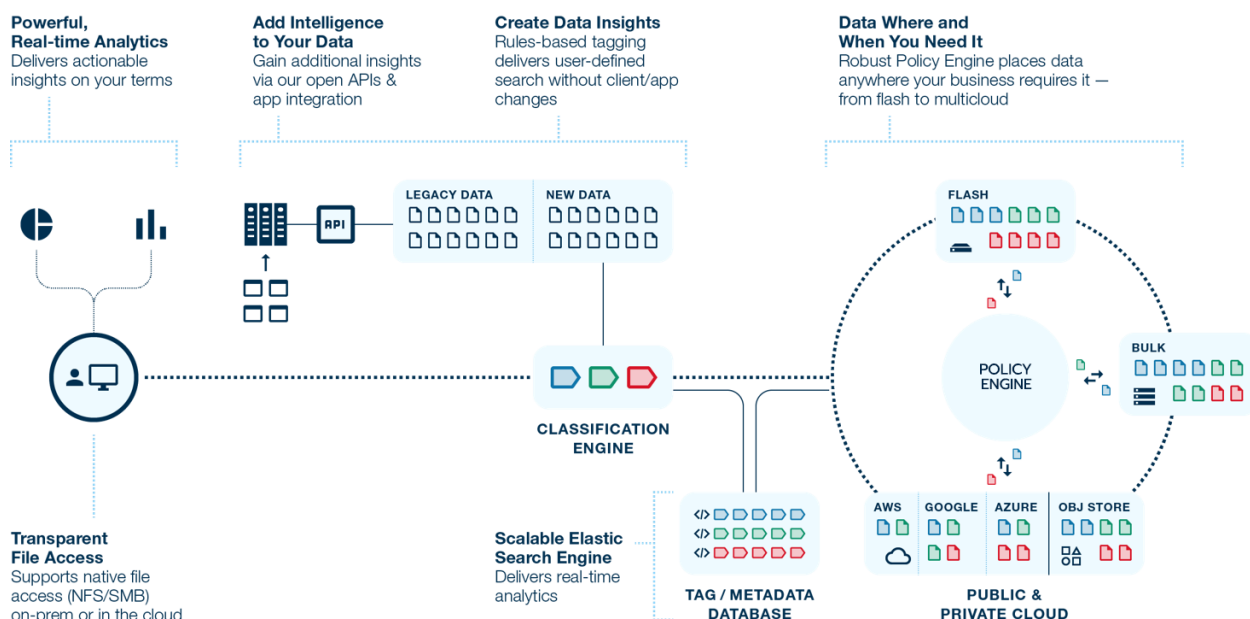



Figure 6 – Illustrating a Real-Time Data Classification Engine

ON-PREMISE VERSUS PUBLIC CLOUD CONSIDERATIONS

Many organizations are considering and using the public cloud for backup, archive, and cold storage. And the public cloud providers have done many things very well, including set a new bar for ease of use, and ultimately have abstracted users from the underlying technologies. We see the future as both hybrid- and multi-cloud, with some of the key considerations outlined below:

Consideration	Explanation
Fees to access and use the data	Most cloud providers provide a very low-cost storage service to store large amounts of data, but retrieving this data or accessing it becomes very expensive very quickly. And since archived data is “not valuable until it is” these access and retrieval fees are very hard to predict and hard to budget for.
Risk of data silo	Companies are becoming increasingly aware that the cloud providers are another form of data silo. It is difficult and expensive to switch providers, and data managers give up a level of control when storing their data in the cloud.
Desire to use services from multiple cloud providers	New services are being added to the public cloud providers’ marketplaces all the time. This competitive landscape will continue to develop rapidly in the coming years, and organizations will want to leverage the best services from various cloud providers. This reinforces the consideration to not be siloed or locked in to a particular cloud provider.
Data security and control	Even public cloud providers are not immune to ransomware attacks, and may or may not be following the best practices for data protection.



A Seamless Bridge Between On-Premise and Multiple Clouds

The public cloud providers use the same fundamental storage technologies that have been outlined in this paper, but have developed an excellent user experience that abstracts the users from the underlying technologies. Further, they have developed large marketplaces filled with excellent services that managers of a 100-year archive will want to leverage.

This is why we believe the best architecture is for large enterprises, content producers, and anyone managing a 100-year archive to build their own. The economics are better, and it gives these organizations the freedom and flexibility to move their data between cloud providers.

With the right data classification engine, and ability to move data seamlessly between on-premise and various cloud providers, users will get access to the best services at the best economics while maintaining control, security and data protection for their critical digital assets.

NEXT STEPS

In summary, we see the “100-year archive” problem as the chief problem that many CIOs, data scientists, content creators, and security integrators will need to tackle in the next 5-10 years.

Here at Quantum, we have been helping our customers solve this problem for close to two decades. Your favorite movies and TV shows are archived digitally on Quantum. Your favorite sports moments from the last 100 years are likely archived digitally on Quantum. Government agencies across the world have built their digital archives on Quantum; for national defense, to help study the planet and the effects of climate change, and to explore space.

We have built a portfolio of technology designed to solve this problem, from the fastest NVMe storage systems in the world, to ActiveScale object storage software, to the lowest-cost, most secure ‘cold’ storage in Quantum tape. We can bring these technologies together with real-time data classification and tagging, and the ability to place data across all of these tiers to build a hybrid- and multi-cloud 100-year archive.

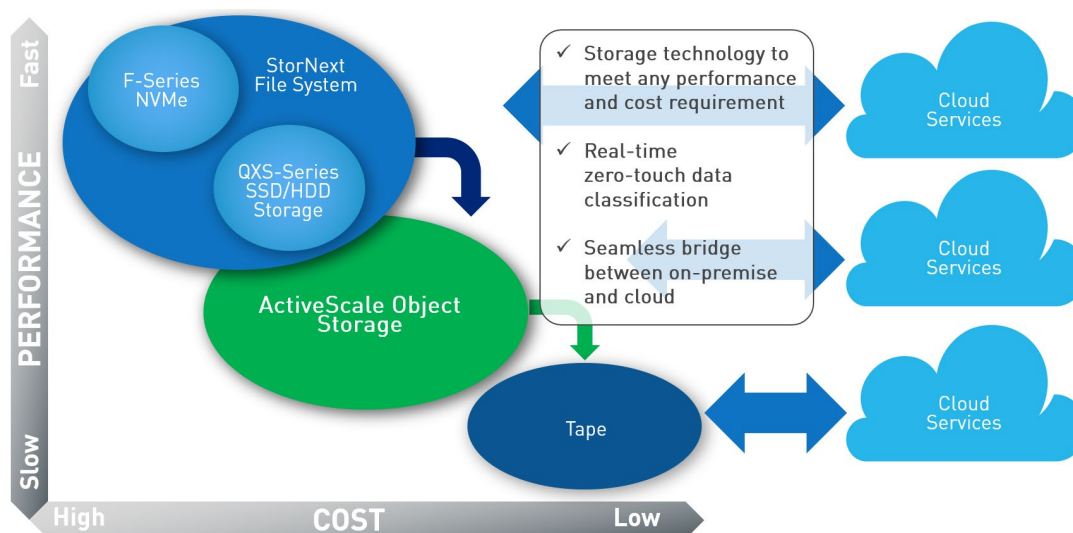


Figure 7 – Quantum Technology to Solve the “Forever” Content Archive

To learn more, visit www.quantum.com/objectstorage.

Quantum®

ABOUT QUANTUM

Quantum technology and services help customers capture, create and share digital content – and preserve and protect it for decades at the lowest cost. Quantum's platforms provide the fastest performance for high-resolution video, images, and industrial IoT, with solutions built for every stage of the data lifecycle, from high-performance ingest to real-time collaboration and analysis and low-cost archiving. Every day the world's leading entertainment companies, sports franchises, research scientists, government agencies, enterprises, and cloud providers are making the world happier, safer, and smarter on Quantum. See how at www.quantum.com.

www.quantum.com • 800-677-6268