



Inside Track Research Note

Sponsored by

**Western
Digital[®]**

Rethinking long term data management

The emerging role of
active archiving

December 2016

Most enterprises and IT professionals have not traditionally regarded archiving as either glamorous or high-profile.

Archiving's profile has risen significantly in recent years as a result of regulatory demands and a desire to unlock the value in historical data.

Data volumes are ballooning and the cost of storage is increasing as a result, even though the cost per gigabyte is falling.

Users want to be able to access everything, forever!

In a nutshell

Active archives aim to make data archiving more cost-effective, useful and practical. They do so by making it an integral and valuable part of the business. On the one hand, this can make archiving easier to implement and manage, and on the other it means the archive can provide much more benefit to the enterprise than merely a dusty long-term store.

No longer your grandfather's filing cabinet

In IT terms, and in life as a whole, archiving is not a new concept, but for most enterprises and IT professionals it has not been a glamorous high-profile project to get involved in. However, archiving's profile has risen significantly in recent years, thanks in large part to legal and regulatory demands for long-term data retention and protection – and of course for the ability to retrieve that data promptly.

A second key driver is the growing realisation that this long-term data may hold massive business value, which few organisations currently exploit. Add the fact that organisations are creating more data than ever before and creating it in new ways, and we can see that archiving offers considerably greater business potential than it used to, and needs a much higher profile within organisations.

Complicating this, however, is the common misconception that archives and backups are the same thing. Too many people assume that because they have backups, they have no need to archive, or less commonly vice versa. The truth is almost the diametric opposite: not only do you need an archive, but that archive may well require backing up or replication!

The reason is that backups and archives have quite different purposes, with one existing to support recovery from faults, errors or disasters, and the other providing a way to access and audit an organisation's long-term data assets, and leverage the deep value within them. Several modern storage systems can perform both services from the same data store, but they are still purposefully different services.

Triggers of change: something must give

Everyone finds it hard to find stuff. At the same time, data volumes are ballooning and the cost of storage is increasing as a result, even though the cost per gigabyte is falling. Worse, the more data you keep on your primary storage, the more you have to backup. Many organisations have reached the point where they simply do not have a long enough backup window – typically 'overnight', when system usage is low – to back everything up.

The result is that they must choose to leave data unprotected, or find some way to reduce the load, such as moving disused or infrequently-accessed 'cold' data to cheaper secondary storage. The former should be unacceptable in almost every case, both from a business perspective and in view of regulations and laws that require certain data to be protected. The latter tends to be resisted by users, who want to be able to access everything, forever, and always want to see someone else's data swept off before their own.

While the latest tape technologies let you search for data directly, it is a relatively slow process.

Cloud-based archiving is an option, but costs can spiral with scale – and archives are all about scale.

Active archiving software moves disused and underused data off expensive primary storage, onto media that is still online but less expensive to acquire and operate.

Because the archive is the sole copy of data, not a backup copy, it still needs protection.

At the same time, there is the realisation that all this information has potential value – both positive and negative. Talk of analytics and Big Data, and of the trouble directors can get into if they lose data or can't present it to regulators or courts on demand, or if they keep personal data that should have been erased, has therefore focused attention on how to ensure you can find what you need quickly and cost-efficiently.

Breaking from the past

The traditional way to do cold archiving was to use a robotic tape library. However, while modern tape is remarkably fast, cost-efficient and reliable, its access time is still measured in tens of seconds rather than milliseconds. So while the latest tape technologies do let you search for data on tape directly, it is a relatively slow process.

More recently we have seen specialist cloud archival services appear. These can present a good option for smaller or newer organisations which are seeking to avoid capital expenditure on IT systems, or who lack the relevant skills. However, cloud costs can spiral with scale – and archives are all about scale – plus they can be unpredictable because typically you pay to retrieve data and you never know in advance how much you will want to retrieve.

A better option for many purposes is to use active archiving software which moves disused and underused data off expensive primary storage, onto media that is still online but less expensive to acquire and operate. The availability of relatively cheap bulk disk storage, plus flash drives for caching, means that this is now a practical option.

In this approach, the archived data becomes the primary copy, and the archive can be accessed and searched directly because it remains online – cool rather than cold. An important element of all this is the introduction of software that takes over management of the data, taking it out of the direct control of whatever application created it. It does this by introducing a layer of abstraction; the application now talks to the abstraction layer, which determines whether to store specific data on primary, secondary or even cloud storage.

One way to create an active archive is to build your own. Low-cost commodity storage allied to a modern distributed object store can provide deep storage at a relatively low capital cost. Your archival software can then move that cold data according to whatever profile you give it. However, the DIY approach is not easy, in particular it can be very difficult to maintain, secure, and operate at scale.

Alternatively, you can buy ready-made active archive appliances, where the integration has already been done and where scalability has already been planned for. This in effect becomes a private cloud, and can even back up to public cloud services or tape.

Note that because the archive is the sole copy of data, not a backup copy, it still needs protection. However, it does not require the same level of protection as the 'hot' data on your primary storage. More importantly, it can dramatically reduce the amount of production data that your main backup processes must cover, which can also reduce the time required for that primary backup.

For a system to be cost-effective, it needs linear scalability up to multiple Petabytes, and an order of magnitude lower total cost of ownership.

Modern object storage systems – which can scale up and scale out - are a good foundation for an active archive.

Although archiving has a very different purpose from backup and recovery, it is feasible to service both from the same store of data.

Tape can often be replaced by low-power, dense and high-capacity disk.

Active archiving: some specifics

For a system of this type to be cost-effective, it needs linear scalability up to multiple Petabytes, an order of magnitude lower total cost of ownership than primary storage, and the ability to support multiple generations of storage technology as new ones become available. On the other hand, users can typically stand an order of magnitude worse latency (the waiting time for data) from a cool archive than from primary storage.

Under the covers it means encryption (for data security), resilience and scale – both scale up for performance and scale out for availability. Modern object storage systems are a good fit here, and are also self-healing and hot-pluggable, which allows failed devices to be replaced or storage added to the pool without service interruption. In addition, the layer of abstraction decouples the archive from the hardware, easing the task of future hardware migration.

However, while an active archive can minimise or even eliminate users' worries about their data becoming unavailable, there is the challenge of how to decide what data to keep, on which storage tier it should live, and how to make sure it is either erased at the correct time or stays around forever. The keys here are categorisation, lifecycle management and metadata, which is data about data – a kind of index, in other words.

On top of that, metadata can and does evolve. So the ideal active archive will be capable not only of indexing and managing data at ingestion, but also of re-indexing it post-ingestion, for example to generate new metadata to support new use cases or search requirements.

The last thing is of course that the active archive needs the ability to ingest existing data. There are many ways to do this but an increasingly popular option is to implement S3, Amazon's Simple Storage Service, which has become something of a standard for managing object storage.

Putting active archives to use

Traditional backup and recovery, DR

Although archiving has a very different purpose from backup and recovery, it is feasible to service both from the same store of data. Once data has been archived, metadata maintains the information required to construct recovery images. This can allow tape to be eliminated, with 'backups' consolidated onto the active archive.

Traditional archiving

Acting as a converged data retention service, an active archive appliance or cluster also replaces cold archives. Again, in all but the very largest applications (hundreds of Petabytes) tape can be replaced by low-power, dense and high-capacity disk. You could even repurpose the tape library to backup the active archive itself.

An active archive's relative simplicity and low latency means that cloud providers can deliver the performance and service levels that users now demand.

The advent of high capacity, low cost, scalable and reliable disk systems, allied to object storage that provides private cloud-like capabilities, has dramatically tipped the balance away from tape.

Archiving is fast becoming essential for the modern business, but only an active archive allows the creation of new value from old data.

Media and entertainment

The long-term archiving of film clips and the like has been one of the areas where automated tape libraries retained a presence. However, the ease of integration, management and expansion, plus its greater accessibility and speed, makes an active archive a seductive alternative.

Cloud & managed service providers

When you have cloud-based services that are data-heavy rather than transaction-heavy, active archive's relative simplicity and low latency means that providers can deliver the performance and service levels that users now demand.

Oil & gas, genomics research, etc.

Many modern fields of research spend large amounts of money and time to generate massive quantities of data. Not only must this be stored and processed as it is collected, but the researchers know that there is always the potential to learn more from it later, either as analytical techniques improve or simply as they think of new questions to ask.

That makes it essential to retain the data for the long term. In the past, the sheer volume involved has mandated the use of tape, but the advent of high capacity, low cost, scalable and reliable disk systems, allied to object storage that provides private cloud-like capabilities, has dramatically tipped the balance away from tape.

The bottom line

We have watched the evolution of integrated data preservation and protection solutions – that is, sophisticated software that provides backup, recovery and archiving from a single data store, with all the potential cost savings that implies – for several years now. More recently, other factors have converged to dramatically accelerate their maturity and uptake. Most importantly we have seen the growing maturity of object storage systems, the development of inexpensive multi-terabyte hard disks, and the use of tiers of flash drives to accelerate access to data.

Then there is the pressure on organisations to retain as much data as possible – which many interpret to mean 'all of it' - both for regulatory reasons and just in case it might have future value for Big Data and other analytical techniques, despite the cost of doing so.

On the other hand, there is the EU requirement to locate and remove certain data on request or after a time limit – such data may absolutely not be kept forever. Again, that means you need the ability to find every instance of that data and make sure all are deleted

The upshot is that archiving is not just advisable for the modern business, it is fast becoming essential. And yesterday's cold archive is no longer adequate – it is no longer a matter of keeping data 'just in case', instead that archived data can and will be needed and used. The faster access provided by techniques such as active archiving is therefore essential to allow the creation of new value from old data.

About Freeform Dynamics

Freeform Dynamics is an IT industry analyst firm. Through our research and insights, we aim to help busy IT and business professionals get up to speed on the latest technology developments, and make better-informed investment decisions.

For more information, and access to our library of free research, please visit www.freeformdynamics.com.

About Western Digital

Western Digital Corporation (NASDAQ: WDC) is an industry-leading provider of storage technologies and solutions that enable people to create, leverage, experience and preserve data. The company addresses ever-changing market needs by providing a full portfolio of compelling, high-quality storage solutions with customer-focused innovation, high efficiency, flexibility and speed. Our products are marketed under the HGST, SanDisk and WD brands to OEMs, distributors, resellers, cloud infrastructure providers and consumers.

Terms of Use

This document is Copyright 2016 Freeform Dynamics Ltd. It may be freely duplicated and distributed in its entirety on an individual one to one basis, either electronically or in hard copy form. It may not, however, be disassembled or modified in any way as part of the duplication process. Hosting of the entire report for download and/or mass distribution by any means is prohibited unless express permission is obtained from Freeform Dynamics Ltd or Western Digital. The contents contained herein are provided for your general information and use only. Freeform Dynamics Ltd provides no warranty or guarantee as to the suitability of this document for any particular purpose.