28 Days Later

Managing the challenges posed by long-term retention backup.

Introduction

Within a data backup and recovery environment there are typically two types of data retention, namely:

- Short-term or "operational" retention, and
- Long-term or "compliance" retention.

While day-to-day backup and recovery operations are typically focused on the short-term retention, the long-term retention backup data can easily represent the largest portion of data held by a business, particularly in situations where that data is held for years.

For many businesses, short-term retention data is held for a period of either four weeks (28 days), or a calendar month. Common exceptions can include:

- Reduced retention for development or test systems (e.g., 14 days)
- Increased retention to offer additional recovery options (e.g., keeping daily incremental backups for 28 days, and weekly full backups for 13 weeks)
- Increased retention to reduce the need for long-term retention of certain data sets (e.g., keeping short-term retention backups for 90 days and relying on in-application data retention for most long-term retention requirements).

However, it would not be unreasonable to assume that a significant number of businesses with a requirement for both short-term and long-term backup retentions will use the 28-31 day retention period for their short-term backups.

Initially, there might be no functional difference between a short-term and long-term retention backup at the time they are taken. This offers operational advantages by allowing long-term retention backups to be recovered using the same methods used for recovery from short-term retention backups during the defined operational recovery window. However the management of and future recovery from long-term retention backup can create challenges months or years later depending on the format and method used for data capture.

Using a backup and recovery system for both short- and long-term retention backups introduces seven key challenges that a business must be mindful of in initial planning, operational processes and lifecycle refresh activities. These are:

- Cost.
- Future recoverability.
- Testing.
- Platforms and media management.
- Organisational change.
- People and processes.
- Legal and auditing.

This whitepaper will review these challenges and mitigation techniques that might be considered.

Challenges posed by Long-Term Retention Backups

Cost

Backup and recovery systems are never free. Even in situations where free software is used, there will be other costs associated with the backup and recovery service. Following is a list of just some of the costs that can be observed in a backup and recovery solution:

- **Software licensing**, which may be permanent or subscription-based, and either feature, front-end TB (FETB) or socket-based. In some instances, software licensing may be absorbed into platform costs for integrated appliances.
- **Storage**. Regardless of whether backups are held on-premises, off-site or in a public cloud environment, there will be a cost associated with the storage of those backups. Where removable media is used (e.g., tape), additional storage costs may be incurred for transport and storage of the media.
- **Support/maintenance**. Particularly in situations where on-premises hardware and perpetual software licenses are used, there may be periodic charges incurred for access to

vendor support databases, software patches and updates, and hardware repairs.

- **Retrieval, or data egress**. Particularly in the case of public cloud environments, executing a recovery may result in additional charges for retrieving the data.
- **Operational**. This will include all other costs associated with running a backup and recovery environment, including but not limited to staffing and their associated costs, power, cooling, etc.

While some businesses will struggle to accurately calculate the cost associated with a backup and recovery service, it is correct to say that known or unknown, there will always be a price, per GB protected, of a backup and recovery service.

The advent of data deduplication platforms can seemingly skew this cost. Deduplication can be considered to be a form of compression – but rather than applying only to a single file or archive, it can be applied to a broader set of data. For some vendors this might apply to all the backups written in a single session, all the backups written for a single host within a certain time-frame, all data of a specific type (e.g., database vs filesystem), or all data written to a single storage appliance (physical or virtual).

Deduplication introduces the differentiation between *logical* and *actual* data stored. Logical data refers to the *source* data size, whereas *actual* refers to the amount of target storage occupied. For example, a 1 TB virtual machine written with 10:1 deduplication is 1 TB logical and 100 GB actual. As deduplication can be applied to more than a single backup, this calculation becomes more complex over time. For instance, a mix of short- and long-term retention backups written to a Dell Technologies Data Domain 9300 as of August 2021 shows logical backups of 16,811 TB stored on 599 TB – a deduplication ratio of approximately 28:1.

Should the cost then be calculated against logical or actual data stored? The answer to this lies within another question: does the business recover actual data or deduplicated data?

Of course, deduplicated data if recovered in that format would represent unusable gibberish for the business. The business may utilise a deduplication platform to reduce overall storage costs, but will recover data in its original format (a 'rehydrated' recovery). Thus, cost comparisons for a backup environment should always be modelled against the logical data stored rather than the actual data stored.

To understand the cost impact posed by long-term retention backups, we must first understand the difference between the size of logical short-term and long-term backups. Consider an environment comprising of 500 TB of primary systems data that is protected by a backup environment offering:

- Daily incremental backups, retained for 4 weeks (28 days),
- Weekly full backups, retained for 4 weeks (28 days), and
- Monthly full backups, retained for 7 years (84 months).

Additionally, we will assume there is an average 3% daily change rate in data, and an average annual growth rate of 15% in the data. The daily change rate represents the effective size of the incremental backups, whereas the annual growth rate, amortized to a 1.17% monthly growth rate, affects the size of the monthly full backups. The table below outlines the data volume growth over 7 years, shown per month (the long-term retention period).

M1	M2	M3	M4	M5	M6
500.00	505.86	511.78	517.78	523.84	529.98
M7	**M8**	**M9**	**M10**	**M11**	**M12**
536.19	542.47	548.83	555.26	561.76	568.34
M13	**M14**	**M15**	**M16**	**M17**	**M18**
575.00	581.74	588.55	595.45	602.42	609.48
M19	**M20**	**M21**	**M22**	**M23**	**M24**
616.62	623.84	631.15	638.54	646.03	653.59
M25	**M26**	**M27**	**M28**	**M29**	**M30**
661.25	699.00	676.83	684.76	692.78	700.90
M31	**M32**	**M33**	**M34**	**M35**	**M36**
709.11	717.42	725.82	734.33	742.33	751.63
M37	**M38**	**M39**	**M40**	**M41**	**M42**
760.44	769.35	778.36	787.48	796.70	806.04
M43	**M44**	**M45**	**M46**	**M47**	**M48**
815.48	825.03	834.70	844.80	854.37	864.38
M49	**M50**	**M51**	**M52**	**M53**	**M54**
874.50	884.75	895.11	905.60	916.21	926.94

M55	M56	M57	M58	M59	M60
937.80	948.79	959.90	971.15	982.52	994.03
M61	**M62**	**M63**	**M64**	**M65**	**M66**
1,005.68	1,017.46	1,029.38	1,042.44	1,053.64	1,065.98
M67	**M68**	**M69**	**M70**	**M71**	**M72**
1,078.47	1,091.10	1,103.89	1,116.82	1,129.90	1,143.14
M73	**M74**	**M75**	**M76**	**M77**	**M78**
1,156.53	1,170.08	1,183.79	1,197.65	1,211.68	1,225.88
M79	**M80**	**M81**	**M82**	**M83**	**M84**
1,240.24	1,254.77	1,269.47	1,284.34	1,299.39	1,314.61

At the end of 7 years, the logical backups stored will be:

- Short term retention: 4 x weekly full (1,314.61 x 4) + 24 x daily incremental (1,314.61 x 3% x 24), totalling 6,835.97 TB.
- Long term retention: 70,850.69 TB.

The total logical backup storage will be 77,686.66 TB, with the long-term retention backups comprising 91% of the data volume stored.

If a cost can be assigned to logical backup storage per GB, either per-month or for the life-time of the backup, it is obvious that the vast majority of the cost of the backup environment can reside within long-term retention data, particularly if the same storage medium and platform is used for both short- and long-term data.

With this in mind, the following cost considerations should apply to a backup system that includes both retention types:

- Wherever possible, long-term retention should be applied using a granularity that is sufficiently precise to avoid unnecessary costs. Many legacy retention policies come from tape-based environments where the management effort of separating data into multiple pools for different retention is (a) a high overhead and (b) not always reliable. This approach should not be carried through to a modern backup and recovery service – particularly if tape is no longer used.
- While long-term retention policies should keep data for the length of time dictated by regulatory compliance, policies should ensure the automatic deletion of backup data when it ages beyond this time to help define an upper-bound on cost.
- If medium-term retention can be introduced to limit the volume of long-term retention data without impacting compliance requirements, it should be. (E.g., instead of keeping monthly backups for 84 months, it may be legally permissible to keep monthly backups for 12 months, and annual backups for 7 years.)
- Wherever possible, the system should support moving long-term retention backups to a storage platform or medium that presents a lower cost-per-GB than the operational recovery tier once the data has passed beyond the short-term retention period.
- Where data can be either archived (removing it from backup cycles entirely) or stored in protected, immutable storage (e.g., primary systems retention-lock), this should be considered as an alternative to long-term retention within a backup system.

Future Recoverability

As noted previously, for many environments there is no practical difference between a short-term and long-term retention backup during the short-term recovery window. This is best explained through a standard schedule. Consider a backup retention/scheduling policy whereby:

- Incremental backups are taken in the evening, Saturday through to Thursday (inclusive) and retained for 4 weeks
- Full backups are taken on Friday evenings and retained for 4 weeks
- On the last Friday evening of each month, a long-term retention backup replaces the weekly full backup and is retained for 7 years.

In this scenario, for the first 4 weeks after the long-term retention backup is taken, it will operationally function in place of the standard weekly full. The only difference would be that after 4 weeks, instead of being eligible for deletion, it continues to be kept – for a total of 84 months.

It would be a highly unusual scenario if, 28 days after a backup was taken, there were no longer systems able to read the backup data, or platforms able to meaningfully use the data. However, can

this be said of data that is recovered seven years after it was first backed up?

The recoverability challenges fall into three primary considerations:

- Media integrity
- Backup compatibility
- Logical compatibility

Media integrity is often considered to be a removable tape problem, and while it is true that we might most notice this issue with tapes, it is not necessarily limited to that storage medium. The challenge herein is whether backup media written on day X can be can be reliably read at some future point X+*n*, where *n* is the number of days that have passed. Where *n* is small (e.g., 28 days), the risk is relatively low, so long as the media has been adequately handled. As *n* increases though, there is an increased risk of integrity failure. While tape vendors often advertise 30+ year lifespans for media, this is in optimal conditions. Tape media that has been improperly handled or poorly stored will deliver a less reliable outcome. Other offline media can similarly present problems – a system making use of removable hard-drives for instance may encounter stiction on a hard-drive that was removed from use several years ago and not powered up since. Equally, optical media (CD-ROMs, DVD-ROMs, Blu-ray discs) may also degrade over time, reducing or destroying the integrity of the data stored on them.

The challenge of media integrity can be broken into two categories:

- Offline media integrity, and
- Online media integrity.

Offline media integrity (tapes, optical discs, etc.) requires a process whereby aging media is periodically recalled for testing. Additionally, at least two copies should be generated for all media that will be stored offline to provide a level of protection against media failure. As media ages, recovery tests should be performed against the backups stored on that media. In the event of a failure, the alternate copy should be recalled and used to generate new copies.

Online media integrity should be maintained by using storage platforms that perform appropriate ongoing integrity checking – at minimum, this would be RAID or erasure-coding with periodic validation. Again, at least two copies should be maintained so that in the event of a failure, new copies can be generated.

Next, one must consider backup compatibility. This falls into three primary considerations:

- Assuming the same backup product is in use today as was used to generate the backup, can it still understand the format of the historical backup?
- If a different backup product is in use today from the product used to generate the backup, how can the historical copy be read?
- If the backup was written to removable media (or indeed other media that has simply been offline for an extended period of time), is there a device ready and able to read the data?

For the first consideration: it is unusual for backup products to drop support for reading older copies of their backups. This should be the lowest-risk concern of the three considerations, but does warrant an approach whereby upgrade procedures (and therefore change approvals) are contingent on:

- Verifying from release notes that there are no documented compatibility issues with reading older backups,
- If there are highlighted compatibility issues, planning the required steps to mitigate them. Two options would be to either (a) stand up a 'legacy' environment that can be used to perform the recovery, or (b) migrating the legacy backups to a compatible format, and
- Randomly testing recoverability of legacy backups following major version upgrades.

There are two key reasons where there is a risk of businesses encountering backups written in an alternate format. These are:

- The backup platform was previously replaced (e.g., as a result of a tender), or
- As a result of company mergers/acquisitions.

For the former, this speaks to a need for

businesses that transition from one backup product to another to consider what options they have for guaranteeing recoverability of long-term retention backups. As it is practically unheard of for vendor *A* to support the recovery of backups written by vendor *B*, some form of migration or format mitigation technique will be required for any long-term retention backups. (It is usually assumed that short-term retention backups can be 'aged out' – i.e., allowed to gracefully expire by virtue of their limited retention period.)

Techniques for this can include:

- Migrating the backups through recovery and creation of new backups. This is typically a time-consuming and costly exercise and is rarely undertaken. If not handled correctly, this can also introduce legal headaches, regarding providing proof of the original backup's date, and that no data was changed between recovery and subsequent backup.
- Engaging an 'escrow-like' service that holds the previous format backups for a nominal monthly fee and can be leveraged (usually at a cost) for any compliance recoveries required.
- Using an archival service and consultancy process that scans the backup content, recovers a granular subset of only those records that are required and stores them with single-instancing in a 'neutral' format, or even their original format. (For instance, if the same spreadsheet appears, without change, in 12 x monthly backups, only one instance of the spreadsheet might be kept, but linked to each block of time.)

In the case where businesses acquire, merge with or are acquired by other companies, much of the focus on IT-systems integration is on the "here and now" – how to merge or otherwise consolidate two operational environments. However, businesses should be mindful in these situations of the risks of simply switching all backups over to a single format when there are long-term retention copies. In essence, unless the business decides to run both backup systems independently with systems kept separate (which in itself could be a costly decision), the decision to consolidate onto a single backup platform should trigger the same considerations for platform replacement as cited above.

Finally, where backups are written to removable media and long-term retention is used, there creates an obligation on the business to factor in backup migration between media formats as successive generations are used. For instance, since most LTO formats provide generation *n-2* backwards compatibility for reading, switching from LTO-*n* to LTO-*n+1* does not create any immediate obligation. However, once a hardware replacement programme is initiated to switch to a tape format that is no longer compatible with previously written media, that programme should include appropriate mitigation techniques – such as cloning or replicating backups from the old tape format to the new. (LTO-8 broke the *n-2* convention, only offering backwards compatibility for reading LTO-7 media.)

Moving beyond media integrity and backup compatibility, the final consideration for future recoverability is logical compatibility – assuming there exists functionality to read a long-term retention backup from the media it is held on, are there systems (operating, application) that can subsequently *access* the data and provide it in a usable way to the requesting user?

Examples where this issue may occur include:

- Where the business has transitioned from one application type to another (e.g., long-term retention backups from 4 years ago were of Oracle databases, but 3 years ago the business transitioned to PostgreSQL),
- Where the business transitioned between hardware platforms (e.g., from a little-endian to a big-endian system),
- Where the business transitioned between different software-defined infrastructure (e.g., changing from Hyper-V to VMware, or VMware to the public cloud),
- Where an application provider or a software-defined infrastructure provider changed their data format and has, over time, dropped support for older formats.

Logical compatibility is potentially the most problematic of the future recoverability issues as the groups responsible for these major changes

might have little oversight of or involvement in backup and recovery operations within the business. It is important however the business remains at least aware of this risk – at a minimum, vetting the implications of the long-term compatibility effects of these sorts of changes should be the responsibility of the IT change board, and it must be understood that *resolving* the challenge should be part of the budgetary considerations for enacting substantial change. Mitigation considerations for these challenges can be a mix of previously discussed techniques – such as passing data across to a migration or archival service, or standing up a 'legacy recovery' area.

Additional challenges for logical recovery include expertise (which will be covered in *people and processes* later), licensing and software. Is there any point, for instance, in keeping backups from a database platform from 6 years ago if the business no longer has backups of the software installers, or the licenses required to activate the software to access the recovered data? This highlights the need to ensure that long-term retention backups are generated and kept for *all* details relevant to successful restoration. Not just the data itself, but the licenses, the software installation kits, the recovery procedures and so on. In fact, the licensing aspect may be particularly tricky to resolve:

- **Perpetual licenses** may have been tied to specific servers (e.g., via IP address, hostname or some relatively unique identifier such as the host ID, or an application-generated identifier). In short, recovery of the license may not be enough.
- There may be **questionable legal right** to use a software license (or install the software) if the business does not have an up-to-date maintenance contract.
- **Subscription licenses** typically cannot be used beyond the subscribed dates and it may not be possible to install such software at all without purchasing an entirely new subscription.

Finally, businesses that switch from one product to another while maintaining long-term retention backups may find themselves in a particularly challenging situation if the vendor for the original product goes out of business, or discontinues the product. In these situations, even if the business wants to purchase new maintenance to facilitate a recovery or resolve a recovery issue, it may be impossible to do so.

An alternate consideration for logical compatibility issues is to minimise the potential impact *at the time of backup*. It was previously mentioned that functionally there may be little difference between short-term and long-term retention backups when they are first taken. However, to avoid future logical incompatibility, there can be merit in generating more neutrally formatted data in long-term retention backups. Examples of this include:

- Using hypervisor image-based backups for short-term retention backups, but leveraging agent-based backups for long-term retention ones
- Generating database dumps or even database exports rather than conventionally integrated online database backups – either solely, or concurrently with the conventional backups
- Eschewing 'online' protection methods such as storage-level snapshots for long-term retention copies
- Using conventional filesystem mounts rather than NDMP backups for long-term retention backups os NAS storage

These techniques are not without their own challenges. In particular there are three key potential issues that must be considered if this approach is to be used:

- **Performance variability**. Many current backup techniques are designed to deliver a backup using the appropriate mix of performance and efficiency. This variability may not just affect the backup environment, but also the primary systems as well. For instance, image-based backups of virtual machines typically delivers high-speed backups, and also reduce the overall resource consumption on the hypervisor servers during the backup process. I.e., image-based backups of 1,000 virtual machines will generate a

significantly different performance impact than agent-based backups of those same 1,000 virtual machines.

- **Manageability**. Using two entirely different backup configurations (one for short-term retention, one for long-term retention) will by necessity increase the management overhead for the solution.
- **Storage efficiency**. Particularly when deduplication storage is used, generating an alternate format *X* backup for a workload normally protected using format *Y* may result in reduced storage efficiency and consume more space, negatively increasing the cost of the solution.

Much of the decision making process for using an alternate backup format for long-term retention will come down to the questions:

- What is the legal or financial impact of being unable to perform a recovery for this particular workload or dataset from a long-term retention backup? In essence, is there an external compliance requirement for recoverability of this data, or a preferential business-internal requirement that has no legal impact if not met?
- How frequently are these recoveries required? If recoveries from long-term retention backups are performed frequently for a particular workload, generating them in a neutral format may be useful. (Conversely, if they're generated in a neutral format, but the business never changes the accessing application, it may introduce a disproportionate overhead.)
- What are the known future plans for the business for this workload? While the future cannot always be predicted, businesses often have forward-looking plans for their IT environments, such as say, shifting from one database type to another, or closing down a datacentre and moving to the cloud. Such strategic directions may affect the forward-planning on a backup solution.

In essence, this becomes a risk-vs-cost decision, which is quite typical in data protection activities. The challenge is to balance to risk of being unable to successfully recover long-term retention data (and any associated costs – such as fines) with the operational costs of generating alternate-format backups that may or may not be needed for recovery later.

Testing

It was mentioned earlier that an alternative to long-term retention backups is to move data into a suitable archive platform, thereby removing that data from the backup solution entirely. An appropriately configured archive platform should not require ongoing operational backups (let along long-term retention backups), so long as it meets the following requirements:

- **Immutability**. Once data is written to the platform it cannot be erased or altered, other than automatic deletion once it ages beyond a set retention period. Note that immutability here refers to *regulatory approved* immutability and should have achieved third-party verification.
- **Redundancy and fault tolerance**. There is always more than one platform that holds the data (e.g., archives replicated between storage systems on two separate sites), and each platform offers independent and comprehensive fault tolerance against individual component failure.
- **Automatic fault detection and error-correction**. The archival storage system should be capable of detecting a data fault and recovering from it through appropriate reconstruction techniques (e.g., checksum based parity).

So long as these conditions can be met, data can be transferred entirely from primary storage systems into archive platforms and the archive platforms will not need backup and recovery services.

However, in all other cases, backups that have long-term retention should have a programme for periodic testing that is no less robust than the programme used for periodic testing of short-term retention backups.

In some cases, this testing can be combined with other testing – for instance, if removable media is

used, periodic testing of tapes can meet both requirements. But even if removable media is not used, there should still be periodic testing of long-term retention backups. This testing should be:

- **Scheduled**. Testing should not rely on someone remembering to periodically run a test, but should be scheduled as part of standard business operations.
- **Random**. Sufficient randomisation should be built into the testing process to avoid a situation where only a small subset of the long-term retention workloads are ever tested.
- **Documented**. Not only should testing be conducted according to accurate recovery procedures (and those procedures updated whenever required), but the testing results should be adequately documented so that company can respond to audit requests.

Except in the most cynically operated businesses, backups are not performed for the sake of appearances. We backup in order to recover when necessary, and that means the recoverability of the data should be regularly verified. It makes no difference whether the data was backed up yesterday or 6 years and 364 days ago – if the business is obligated to retain compliance copies of backup data for 7 years, the copy generated 6 years and 364 days ago is just as important as the copy generated last night.

Platforms and Media Management

Particularly when there are long-term retention copies held, a backup and recovery system can represent an operational investment for a business that spans many years, if not decades.

The longevity of use can present several platform and media challenges that need to be considered, and equally manifest problems when a backup and recovery solution is replaced. The considerations here are:

- Media lifecycle aging
- Platform and application currency
- Legacy systems support

Media lifecycle aging has previously been discussed. However, it is worth reiterating that where removable media formats are used (e.g., tape), there should be well-defined policies for:

- Tracking ages of media (individually, and by batches)
- Recalling and testing aged media
- Migrating aged media to prevent data loss from degraded media quality, and avoid a situation where media required at a later date cannot be read because there is no compatible device available

While modern removable tape formats such as LTO offer a relatively lengthy shelf-life if stored correctly, the media management challenge is more often than not defined by newer generation devices being unable to read from older tapes. From the IEEE article, "The Lost Picture Show: Hollywood Archivists Can't Outpace Obsolescence", Marty Perlmutter observes[1]:

> The problem with LTO is obsolescence. Since the beginning, the technology has been on a Moore's Law-like march that has resulted in a doubling in tape storage densities every 18 to 24 months. As each new generation of LTO comes to market, an older generation of LTO becomes obsolete. LTO manufacturers guarantee at most two generations of backward compatibility. What that means for film archivists with perhaps tens of thousands of LTO tapes on hand is that every few years they must invest millions of dollars in the latest format of tapes and drives and then migrate all the data on their older tapes – or risk losing access to the information altogether.

Many businesses that make use of removable media in their backup environments do not apply such rigor to the management of media lifecycle as the film industry appears to, instead relying on infrequent recovery requests and the 'luck' of being able to find compatible hardware to recover long-term retention backups from. Assuming long-term retention backups are kept due to legal or financial obligations, such approaches are clearly risky and should be avoided. Instead, it is preferable to ensure that where removable media is used, lifecycle replacement programmes factor in not just the cost of new tape autochangers, drives and media, but also the effort required to recall and migrate media from old formats to the new before the old format becomes obsolete.

The issue of platform and application currency refers to the tendency in some organisations to

deploy a backup and recovery service and perform little-to-no updates on it for its intended lifetime. While modern security practices are chipping away at this behaviour, there is still some way to go before businesses on the whole accept that backup and recovery solutions require the same rigor for patching and updating as other production systems.

Long-term retention within a backup environment creates a *stretching limit*. The backup environment must be able to support the latest applications and platforms deployed by the business, but must simultaneously also support recoverability from the long-term retention backups of platforms and applications that may be many years old. This is described as a *stretching limit* as the problem resembles the issue of stretching a rubber band – the stretching process will work for a while, but if not stopped, risks breaking the environment.

This problem is exacerbated when businesses are slow at primary systems replacement. For instance, despite Microsoft ending Windows Server 2003 support in July 2015, it was not uncommon to find businesses in 2019 and 2020 still completing programmes to replace their remaining Windows Server 2003 systems with more recent operating systems.

Simultaneous support for Windows Server 2003 and Windows Server 2019 represents a considerable stretch for backup software drivers and agents. As system libraries and functions change, software that interacts with it also changes, and simultaneously supporting functions across platforms or applications that may be decades or more apart is non-trivial. More often than not the 'fault' of this lack of support is placed at the feet of the backup vendors, with an expectation that they should resolve this 'risk' situation by providing functional agents across such a large spread of time – despite primary vendors ceasing support.

The *stretching limit* also applies to the third platform/media management consideration, that being legacy systems support. In here, we define a legacy system as any operating system, application or software package that is no longer supported by the vendor who created it. Legacy systems can occur throughout the entire infrastructure stack and might include:

- Operating systems (e.g., Solaris 2.5, Windows 2000).
- Databases (e.g., SQL Server 2000, Oracle v8).
- Hypervisors (e.g., VMware ESX Server 3.5).

The true risk introduced in these situations is when legacy systems act as an 'anchor', holding back upgrades to the rest of the environment. E.g., a solution might be found that enables a product released in 2021 to protect a physical Windows 2000 server. However, if that *prevents* any further upgrades of the backup software or backup storage platform, the risk might be considerable. I.e., the business considers it a risk if an application platform 15 or 20 years old can't be backed up – but what about the risk that supporting such a platform might *prevent* infrastructure upgrades?

In addition to the stretching limit, legacy systems can also pose significant headaches to businesses that want to replace their data protection solutions. Imagine the scenario, for instance, of wanting to take advantage of modern data protection offerings for the overall infrastructure only to be held back because the preferred platform can't protect an operating system that hasn't been supported for 15 years!

Businesses will often cite a high calculated cost as the reason for not migrating off legacy platforms. For instance, it might be that some platform of service built up around an Oracle 7 database would be deemed too costly to upgrade, and the business accepts the risk this creates. But these cost calculations are rarely accurately determined, because of the tendency to ignore the costs imposed on the *rest* of the environment. If a business is to accept the risk of keeping a legacy system operational, it must accept all the risks and caveats therein, including:

- Using a less-integrated backup strategy (e.g., abandoning agent-based backups in favour of operating-system or application exports to shared storage)
- Adopting lower service level agreements relating to backup performance, recovery time objectives and recovery point objectives

- Less-granular recovery operations – particularly in situations where a legacy system has been virtualised, accepting that file-level recovery from image-level backups may no longer be possible
- Deploying isolated, unsupported data protection platforms.

Most of these options meet friction and objections within the business, but these *are* the genuine risks associated with continuing to use legacy systems within an environment, and it is not the responsibility of data protection vendors to assume that risk on behalf of businesses. (It is not uncommon to have businesses object to the notion of deploying unsupported versions of data protection platforms just to protect legacy platforms – legacy platforms that themselves are not supported by their primary vendor.)

Over time, it may be that leaving unsupported software running within the environment may become a quaint anachronism, with change forced by the continuing growth of cyber attacks. Indeed, running unsupported software is increasingly gaining the attention of regulators and government security practices. The Australian Essential Eight Maturity Model for Cyber Security[2], for instance lists three maturity models, with the first level unattainable unless:

> Operating systems that are no longer supported by vendors are replaced.

There is no easy solution to the stretching limit problem caused by legacy systems – either in relation to recoverability of long-term retention backups, or simultaneous ongoing backup support for a broad spectrum of product versions. However, backup systems will typically interact with more applications, operating systems and platforms within an IT infrastructure than any other function bar networking itself. As such, it is critical that all businesses include consideration for the impact of upgrades and platform changes (or a lack thereof) on the data protection posture of the company particularly when long-term retention backups are used.

Organisational Change

Organisational change of course can impact both short-term and long-term retention backups. While the impact on short-term retention backups will be obvious and more likely than not directed at the immediate operational changes, the affects on long-term retention backups may be more difficult to predict – but failing to consider them will guarantee future challenges. Examples of organisational change that can impact long-term retention backups include:

- Moving responsibility for backups between teams (e.g., from a storage team to a broader infrastructure team)
- Outsourcing IT operations
- *In*sourcing previously outsourced IT operations
- Pivoting the business to or from the public cloud
- Mergers and acquisitions.

There are no golden rules for handling these particular types of situations, and considerations around long-term retention backups will typically be a relatively minor facet of the overall scenario. However, if long-term retention backups *aren't* considered in these types of changes, the business may find itself in a risky legal situation years or decades later if it is discovered that recovery from long-term backups has been compromised and cannot be repaired.

People and Processes

When backup data is retained for years or even decades, the people and process side of management and recovery should not be forgotten.

Part of the challenge here has already been discussed in *future recoverability*, and relates to situations where the business changes application or infrastructure software used. It may be necessary to recover an Oracle database from 6 years prior for legal reasons – but if the business moved all databases from Oracle to Microsoft SQL Server 3 years ago, this can present three distinct problems:

- Licensing
- Installation software

- Expertise

Licensing and installation software challenges were discussed in *future recoverability*, but expertise warrants consideration here. Even if the business has the licenses and installation software required to install and activate the appropriate version of Oracle, as a result of the database transition it may no longer have staff who can execute Oracle recoveries. This problem might be mitigated by ensuring documented processes (e.g., recovery procedures) also have long-term retention applied to their backups, but even documentation that outlines the recovery may be insufficient if the operator does not know how to use the underlying software.

The logical implication of this is that if the business decides to keep long-term retention backups for a particular data set, it needs to retain *more* than just that data set. In fact, long-term retention backups should be approached with the same rigor as planning for *business continuity* activities. At minimum, the business must ensure that it retains (for the same length of time as the data set) backups of:

- Software licensing
- Software installers, including all patching in use up until that point
- System configuration details (either formal CMDB records or appropriate system excerpts)
- Requisite passwords
- Installation processes
- Recovery processes for the data set
- Disaster recovery processes for the backup application.

Data (and backups of that data) do not exist in isolation, and failing to capture at minimum the above alongside long-term retention backups may render the retained data unusable. Additionally, where businesses change platforms, they should understand the risk or cost associated with sourcing expertise in the previous products if later required. (This might be compared to the scramble many businesses went through in the late 1990s to find COBOL experts to help mitigate the Y2K issue.)

Legal and Auditing

Three primary legal issues should be considered in relation to long-term retention. These are:

- What data has to be retained?
- Data handling when recovering from long-term retention backups.
- Data deletion when the required retention time has expired.

IT workers are not, on the whole, legally trained. They would not, for instance, be expected to review taxation legislation to determine the impact of changes to progressive tax brackets when it comes to employee salaries. Yet despite this being a reasonable assumption for most businesses, successful interpretation of legal retention requirements, and correct classification of data often does seem to be expected of IT teams. This leaves IT teams frequently having to 'guess' long-term retention requirements for business data, which may result in either (a) storing too much data for too long a period of time, or (b) failing to keep backups the company is legally obligated to.

In short, businesses that know they are subject to legal data retention requirements should be prepared to invest in the appropriately skilled resources to correctly define the retention policies that will be applied to long-term backups. To be certain, this may require involvement with IT teams to consolidate the settings, but IT cannot drive this process.

As more legislation is developed around data privacy and records retention, this places new challenges onto the IT teams responsible for long-term backup retention. Consider the European General Data Protection Regulation (GDPR). Among other protections, this legislation allows affected citizens to request their data be deleted under certain circumstances. While deletion of online data (i.e., data within primary storage and application environments) should be relatively straight-forward, it may not be possible to automatically delete individual records from backup data – particularly long-term retention backup data. In such situations, it may be necessary for the company to granularly delete relevant records from retrieved data following the

completion of a recovery from long-term backup. In "Backups and the right to be forgotten in the GDPR: An uneasy relationship", the authors note[3]:

> A major issue arising from the obligation for erasure requests under the GDPR RtbF [right to be forgotten] concerns the case where personal data have already been backed up or archived. The issue is increasingly occupying the IT industry since any noncompliance may cause high sanctions.

Automatic deletion of restored data may not be possible, since it inevitably requires interaction between a backup product from one vendor and an application from another vendor, and processes such as data masking or in-stream data analysis of backup and recovery sessions is rare due to its inherent performance impact and relative unreliability. As such, companies that either currently or in future find themselves under the purview of privacy and data security legislation should be mindful of this when establishing procedures relating to recovery of long-term retention data. In the case of legislation like GDPR, this may be as simple as having a centralised register of deletion requests providing sufficient detail relating to the requesting entity and systems/data sets affected. This register could then be consulted by recovery and application operators to ensure that retrieved data is appropriately vetted and deletes performed before making the data set available.

Of course, this issue does not just apply to GDPR. There is increased attention to data privacy legislation and the right to be forgotten developing in many jurisdictions. Additionally, data that has been backed up into long-term retention may subsequently be determined to hold sensitive data, such as information that falls into PCI DSS (Payment Card Industry Data Sensitive Standards) or other personally sensitive/identifying information. Where it is demonstrably impractical to cleanse this data from backup systems, legal exemptions may exist allowing the backups to be left as-is, so long as they are sufficiently secured and appropriate post-recovery scrubbing can be guaranteed. However, lacking requisite legal training, IT teams alone cannot safely make this policy determination for the business.

Finally, when the business elects to hold long-term retention backup data, the retention needs to be honoured both in the holding of the data for the required time *and* the deletion of the data once that time has expired. Failure to delete this data at the allotted time may result in additional operational costs to the business relating to the storage of the backup, but this is a relatively minor problem compared to the potential legal impact. In many jurisdictions, if a business still holds data – even if that data is older than the required retention time – the data can be subpoenaed in a legal situation. That is, a business might retain financial backups for 7 years under taxation legislation, but forget to enact records disposal on those backups. If the business still has the backups after 10 years and a legal discovery is executed, the business may have to still produce them as part of the discovery process, and find the records used against them in legal proceedings.

In Summary

In the movie "28 Days Later", Cillian Murphy awakes 28 days after an apocalyptic virus is unleashed on the population of England and has to survive the ensuing horror.

In data protection environments, more often than not somewhere around the 28 day period, backups shift from being about short-term/operational recovery to long-term retention (if they are not deleted). This shift imposes entirely different operational requirements on the business that can have implications stretching into years or decades.

If appropriate processes and operational considerations are not applied to these long-term retention backups, the business may later find itself dealing with a horror situation that can result in excess costs, compliance issues or legal headaches. In short, choosing to use data protection systems to retain long-term copies of backups creates obligations on the business and its IT teams that cannot be cavalierly dismissed or avoided.

About the Author

Preston de Guise has been working with data protection and backup/recovery systems for more than two decades and is the author of three books on the subject, and co-author of a fourth. A complete bibliography can be found at https://nsrd.info/books.html.

Additionally, Preston has been running the Data Protection Blog (https://nsrd.info/blog) since 2009, providing insights into the operation of various Dell EMC data protection products and general data protection theory.

This whitepaper reflects on topics outlined in Chapter 25, "Long Term Retention" of Data Protection: Ensuring Data Availability (2nd Edition, May 2020, Auerbach Publications, ISBN 9780367256777). If you found this whitepaper useful, please consider purchasing the book, available from the publisher at: https://www.routledge.com/Data-Protection-Ensuring-Data-Availability/Guise/p/book/9780367256777

1. IEEE Spectrum, "The Lost Picture Show: Hollywood Archivists Can't Outpace Obsolesence", Marty Perlmutter, 28 April 2017. https://spectrum.ieee.org/the-lost-picture-show-hollywood-archivists-cant-outpace-obsolescence

2. "Essential Eight Maturity Model", first published June 2017, July 2021 update quoted. Australian Signals Directorate. https://www.cyber.gov.au/sites/default/files/2021-08/PROTECT%20-%20Essential%20Eight%20Maturity%20Model%20%28July%202021%29.pdf

3. "Backups and the right to be forgotten in the GDPR: An uneasy relationship", Eugenia Politou, Alexandra Michota, Efthimios Alepis, Matthias Pocs, Constantinos Patsakis, September 2018, https://www.researchgate.net/publication/327639998_Backups_and_the_right_to_be_forgotten_in_the_GDPR_An_uneasy_relationship