MODERNIZING UNSTRUCTURED DATA PROTECTION AND MANAGEMENT

by George Crump

Storage Switzerland, LLC

CHAPTER 1: WHY IT NEEDS A NEW DATA PROTECTION STRATEGY FOR UNSTRUCTURED DATA

Unstructured data has always been a sore spot for the data protection process. The growth in the number of files that make up unstructured data sets and the capacity that they consume now threatens to break the data protection model completely. Considering that every indicator suggests the growth in the unstructured data will not only continue but also accelerate IT needs a new strategy so it can stay ahead of the problem.

Unlike most data protection conversations, the problem with unstructured data protection does not revolve around restore speeds. The time it takes to restore an individual file is roughly the same between products and storage types. Even restoration speed of a single file from the cloud is generally not a cause for concern anymore.

The problem with unstructured data protection is everything else; making frequent backups, retaining and organizing the protected copies and finding the exact copy needed for restoration. Solving this problem correctly sets an organization up for success not only with data protection, but with all the other uses of a backup set: retention, restoration, and archive.

WHY TRADITIONAL DATA PROTECTION SOLUTIONS FALL SHORT

Traditional data protection solutions typically back up unstructured data by scanning or "walking" the file system directory structure, indexing that information, looking for files that have changed since the last backup. If the file has changed, it copies it to backup storage.

The advantage of the file walk approach is that the backup system has specific knowledge of each individual file and versions of that file, which it is protecting. However, the problem is that each of these files and versions are files contained within the backup job. Retention and compliance policies can only be granular to the job. If the organization wants to remove an individual instance of a file, it has to remove the job, and any other files that may be in that job. Conversely, if the organization wants to ensure that it retains certain files for a period longer than a default policy for the job, then it cannot meet this requirement either.

The only potential workaround is to have special jobs for each file type or retention type, which means multiple jobs walking the file system but this approach is not viable at scale. The ideal way to handle this problem is to classify data based on tags which can be automatically or manually set. Then jobs can be set to only backup file of a certain classification and have specific retention policies within that job

WHY MODERN BACKUP SOLUTIONS STILL FALL SHORT

Modern backup solutions have addressed the unstructured data backup problem by doing some form of an image based block level incremental (BLI) backup. A BLI backup is much faster because it does not interface with the file system, rather it operates below it and is only looking for blocks that have changed since the last backup, and then copies those blocks to the backup device. Even though it is image based, most modern backup solutions can provide file level restores by transparently mounting the volume on the backup device and interfacing with it.

Image or block based backups present several problems. First, it can only maintain a finite number of incremental backups prior to either performing another full backup or running a consolidation job. Both of these efforts take time. Additionally these solutions provide even worse granularity for setting a specific retention of file data. Essentially, it cannot do it. Organizations need to implement another solution to meet their compliance and retention demands.

THE FREQUENCY PROBLEM

Another challenge in both the file system walk method and the block image method is the frequency with which the solution can protect data given the time required and limitation on the number of incrementals. New threats like ransomware have the potential to strike at any moment, and unstructured data is the prime target. Once a night backups of unstructured data is no longer acceptable given the risk.

THE SECONDARY STORAGE PROBLEM

Another challenge facing unstructured data protection is the secondary storage requirement. The secondary storage system has to maintain at least one copy of the primary storage and in almost all cases, it stores at least two copies. In reality, most organizations find that their backup storage is 5 to 10X the size of primary storage. It can be even worse if organizations are making additional copies for other purposes such as retention or archive.

While secondary storage systems have capabilities like compression and deduplication to alleviate some of this capacity requirement, there is no question that it is still a major issue. Cost of these secondary storage systems is of course a real concern but a bigger concern is the data center floor space that they consume. The cost of storage may be continually declining, however, the cost of a new data center is continually rising.

THE LACK OF AN EXIT STRATEGY

The final problem is that both of these unstructured data protection methods do not provide any means for escape. The problem will just continue to get worse as unstructured data grows and unless the data protection solution can lay the foundation for archiving old data off primary storage, IT will be like the hamster on the wheel, never getting ahead of the problem.

WHAT IT NEEDS

IT needs a new way to handle the protection of unstructured data. First, unstructured data protection needs to return to its more granular roots. Image backups were a band-aid to solve a performance problem but sacrificed any means of compliance and retention. As both compliance and retention become more critical, lack of those capabilities is no longer acceptable.

Of course, the granular understanding of the files IT is protecting cannot result in weeklong backup jobs either. The solution is an agent like solution that can monitor the file system and make copies of changing files at specific and narrow intervals. The solution should make these copies to secondary storage or to the cloud, but it should also self encrypt those files so that they are not exposed to an accidental or purposeful breach of the cloud account.

This type of solution also lays the foundation for archive. Any archive process must first start with creating a known good copy of data on a secondary storage device. Once in place IT can remove old data, either manually or programmatically, with the comfort of knowing it is stored safely on less expensive storage.

Both legacy and so-called modern solutions for unstructured data protection have run into a perfect storm. Not only are the number of files and capacity requirements growing, the demands to ensure data retention or removal based on regulations are becoming more prevalent. Unstructured data protection no longer can remotely access the file system; it must be on the file system and be able to interact with it and make copies of changed data more frequently. Unstructured data protection storage also needs to be more native so that individual policies can be set and data repurposed.

CHAPTER 2: UNSTRUCTURED DATA'S COMPLIANCE AND RETENTION GAP

In most data centers, unstructured data now consumes more storage capacity than all of the organization's structured data combined. Yet organizations still often treat unstructured data like a second-class citizen when it comes to data protection. Because of its size and the sheer number of files, organizations tend to protect their unstructured data store with legacy backup solutions and outdated best practices. As a result, most unstructured data protection strategies fail to meet the current requirements for compliance, retention and multi-cloud support.

> "...it may be more pragmatic to improve backup so it can fulfill two of archive's most important responsibilities; retention and compliance.

UNSTRUCTURED DATA PROTECTION NEEDS BUILT-IN COMPLIANCE

Other than email data no other data set is the target (or source) of regulatory and legal requirements like unstructured data. It is necessary to identify, segregate and in some cases set aside unstructured data to meet an ongoing statute or a new legal hold requirement based on a discovery request.

In theory, the organization should have a separate archiving process to support these complaints and demands, but the reality is they don't. Establishing a separate archive has proven itself to be expensive and complicated to adhere to over a long course of time. Organizations have been slow to adopt archive because of its requirements for a separate silo from backup, as well as the specialized storage and software needed to achieve the archive system's goals.

In practice, most organizations count on the backup solution to be the archive. Trying to extend backup to be the organization's archive, especially with legacy software, creates even more challenges. The design of most backup systems does not meet compliance standards. They have no way to classify data by category or to make a dynamic backup of a particular data set to meet a legal hold. They also have challenges at scale. A backup solution may need to store metadata information about millions if not billions of files, and each version of those files. This

combination leads to a backup database of massive proportions, which is susceptible to corruption and presents a backup challenge of its own.

Closely related to compliance is retention. However, the retention use case is broader. Organizations may decide to retain information for reasons other than compliance. There is, of course, the "keep it just in case" use case that leads to rarely deleting files from a file server or NAS. But, there is also the legitimate need to retain information for possible future data mining needs. There is also a need to ensure the verification of retained data to ensure it does not degrade over time. The problem is the organization has no idea exactly what information needs to be retained and for how long. Many organizations retain all data both on primary storage and on backups. The challenge is particularly problematic for protected data sets since again, it stores multiple copies of the files and multiple copies of each version of the files. As a result, protection storage is often 5 to 10X the size of primary storage and consumes large amounts of data center floor space, as well as organizational budgets.

FILLING THE COMPLIANCE AND RETENTION GAP

There is a need to fill unstructured data's compliance and retention gap. Governments and ruling bodies are passing specific laws and regulations around data governance. Counting on an organization to adopt and implement a separate archive strategy is too optimistic. Archiving solutions have been available for decades and their adoption rate, especially compared to data protection, is too small to measure.

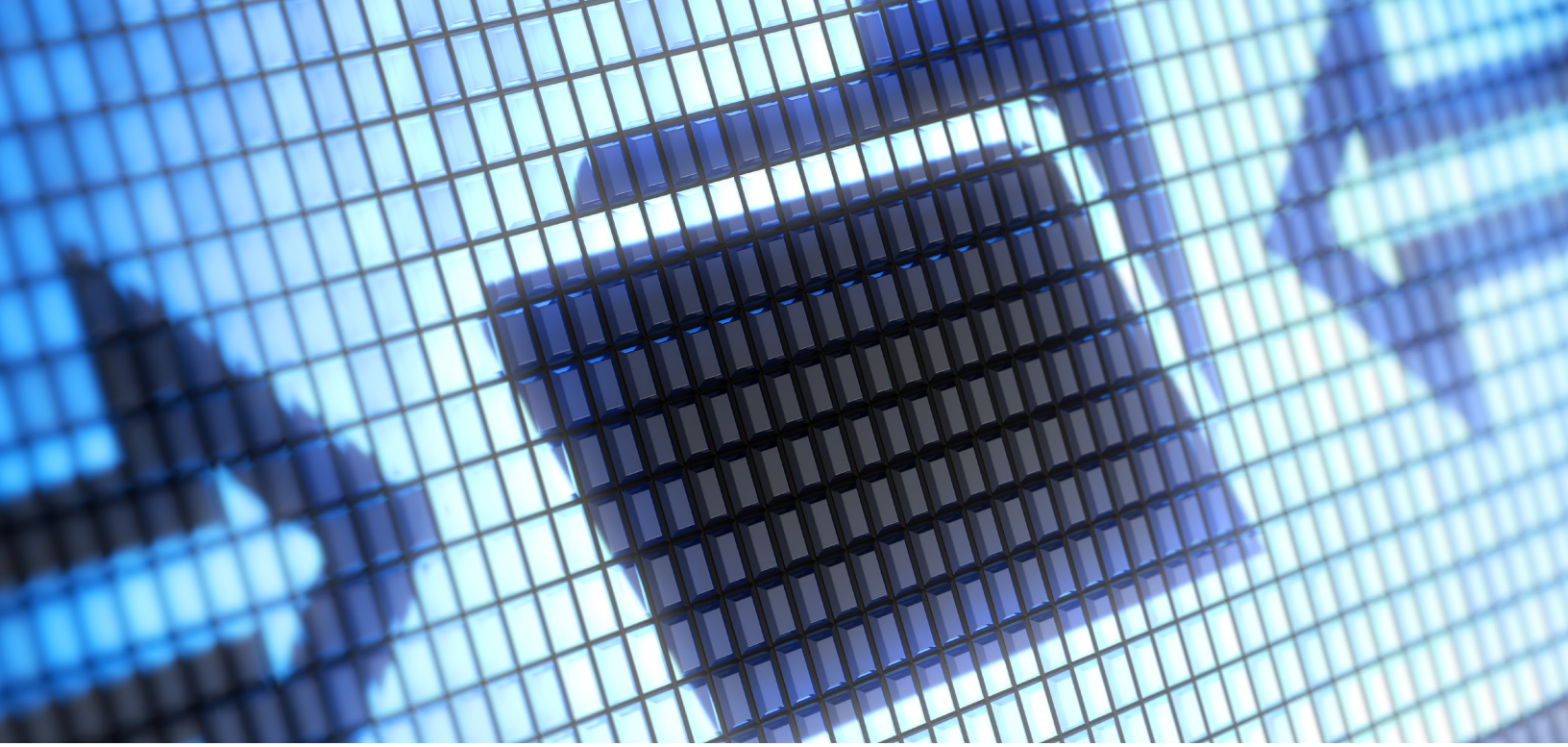
DATA-DRIVEN BACKUP

The mantra has always been "backup is not archive," but given an organization's willingness to invest in data protection versus archive solutions, it may be more pragmatic to improve backup so it can fulfill two of archive's most important responsibilities; retention and compliance. Most backup solutions are "job-driven" in that they backup a given mount point without regard to the type of data within that mount point.

Since most file servers and NAS systems have a wide variety of data types in them, each with their own compliance and retention needs, it makes sense to change backup from "job-driven" to "data-driven". While a data-driven backup could just protect a file server or NAS as a whole, it can also be designed to backup data by type. A data-driven backup will require a classification capability so it can organize data by type and/or location. The backup will automatically create the classes based on file type or directory location, or organizational requirements can manually tag items.

With the tagging in place, the backup process can protect the file server or NAS and each protection pass is organized and performed according to these tags. The tags can have specific retention and compliance requirements associated with them, allowing the organization to meet both internal and external standards quickly and easily.

It's undeniable that unstructured data is growing in almost every data center. Since this data is now useful for many purposes, it needs to be stored for a long-term period. It is ironic that the way of protecting and storing data has not changed. Unstructured data protection may require a fresh approach; one built around the protection of actual data as opposed to the servers on which that data resides.



CHAPTER 3: UNSTRUCTURED DATA PROTECTION SHOULD INTEGRATE ARCHIVE

Backup and archive have always been on opposing ends of the data management spectrum. Conventional wisdom suggests that the two should never meet and cries of "backup is not archive" fill the air. The reality is, the "backup is not archive" mantra is based on the limitations of old technology and doesn't take into consideration, the capabilities of modern protection software and hardware.

WHAT HAPPENS IN ARCHIVE

When describing archive, it typically is defined as a process that identifies old data, classifies that data and then moves it to some low-cost storage device with terrible recall performance. The reality is that nothing can be further from the truth. First, archive seldom actually "moves" anything. The first step in an archive process is to COPY data not move it. The archive software will wait for a data set to reach certain criteria, typically not accessed for a predefined period, and then make a copy of that data to the secondary storage device. At some user defined point, the data is deleted from the original target, which frees up primary storage and makes the copy available only on secondary storage. It is expected that the archive software will provide some way to quickly find these files, by either name, or group or tag.

Except for waiting, backup software does the same thing. It copies data once per night to secondary storage. As we discussed in the last chapter, protection software designed for unstructured data will often make copies of data, at the point of creation or as it changes, to secondary storage. Legacy backup software can also find files based on name or backup job.

Legacy data protection solutions fall short when trying to also be an archive solution. The problem is archiving is an afterthought. The way the legacy solution backs up data, either via a backup job or via an image, is at odds with the way archive needs to work with discrete files. As a result, organizations that want to manage data are forced to implement a separate archive system that requires an additional and separate scan of unstructured data and often a separate storage architecture.

For an unstructured data protection solution to handle unstructured data archiving as well, it must resolve something which legacy solutions are particularly bad at, managing large backup indexes. Part of the solution to this is simply to use a indexing architecture designed for long-term scale, which most backup solutions do not. Another part is for the software to collapse intelligently, the number of copies/versions it has of data. Over time, having every single version of a file becomes unnecessary, and in most cases, the organization needs only the final copy. Deleting unnecessary prior file versions from secondary storage could possibly reduce the size of the solution's database.

The last remaining need is for the backup solution to add a data grooming feature. Here it could have an advantage over traditional archive. It could make sure that no grooming takes place unless there is X number of copies on protection storage and Y number of copies are available off-site (or in the cloud). Since most archives have no integration into the backup solution, they are unaware of the protection status of a file.

For a data center considering this integrated strategy, data grooming is not a "must have now" feature. As long as the protection vendor has created a foundation to add it later, then it will take at least months or probably a year before the organization needs to start grooming data from production storage.

IT professionals have resisted archiving for decades, choosing instead to keep buying more and more primary storage, even though most of the data on that storage is inactive. Those same IT professionals HAVE bought data protection solutions from day one and continue to buy them. Perhaps, instead of forcing IT to buy a separate solution, it's time to integrate archive into the protection process. Backup is, after all, the first step in any archive process. An unstructured data protection solution with the right core capabilities designed into the architecture from the beginning could serve as an excellent foundation for implementing an archive strategy that works.

CHAPTER 4: THE REQUIREMENTS FOR MODERN UNSTRUCTURED DATA PROTECTION

Unstructured data presents two challenges that organizations need to deal with; the sheer volume of data and the quantity of files in the data sets. Storing this data is a problem in and of itself, but protecting it is an entirely new problem, which most legacy data protection solutions are ill equipped to handle. A new wave of data protection solutions is on the way to the data center, IT planners need to make sure they understand modern unstructured data protection requirements to see if these new solutions are up to the challenge.

Requirement 1 – Fine Grained Backups

The first requirement of a modern data protection solution is to provide fine grained backups. Most legacy backup solution have tried to work around the file quantity issue, discussed in chapter 1, by doing image based backups. While it's true that image based backups, especially when combined with changed block tracking technology, are a fast efficient way to backup millions of files, image based backups lack the fidelity needed to manage that data.

To recover an individual file from an image based backup requires that image be mounted, examined and an individual file or files extracted from it. If the administrator knows exactly which file, they are looking for and which backup job contains the version of the file they want, then recovery is relatively straightforward. The reality is though that most unstructured data recoveries look nothing like this. Most of the time a request to recover unstructured data is more like restoring all data related to project X or restore the third version of this file but without knowing which backup job contains that version.

The modern unstructured data protection solution needs to backup and store data so that a recovery request can search across all the files and all the protection instances.

> "...instead of forcing IT to buy a separate solution, it's time to integrate archive into the protection process. Backup is, after all, the first step in any archive process.

Requirement 2 – Frequent and Rapid Backups

Unstructured data changes frequently throughout the day and especially in the modern data center, terabytes of new information can be added to the unstructured dataset within hours. Much of this data can't be recreated as it is the recording of conditions at a specific date and time. Unstructured data is also particularly vulnerable to user error and cyber attacks like ransomware.

Because of this vulnerability, protection of this new and updated data needs to occur more frequently than the typical once per night backup. But, that backup frequency can't break the first requirement of fine grained backup detail. The problem is that typically the only other alternative to image backup is a slow walk of the file system that identifies data requiring protection. In an era where millions of files are commonplace, a file system walk approach is impractical.

The modern unstructured data protection solution needs to deploy via a driver or agent that resides on the protected file-server or interfaces with the NAS API. After the initial backup is complete, the solution needs to create and manage a journal like system in order to quickly identify and protect modified files within seconds, throughout the day.

Requirement 3 – Cloud Support

Secondary or protection storage is typically 5X the size of production storage. Given the current capacities and growth rate of unstructured data, the floor space requirements of the protection storage infrastructure may require its own data center. A third requirement for modern unstructured data protection is to provide the option to leverage cloud storage as the secondary data store. The approach should be hybrid so that some of the data can be stored on-premises, for rapid recoveries of the most recently modified data, while older data is stored in the cloud for cost effective, long-term storage.

Requirement 4 – An Archiving Future

While data protection is the immediate battle for the unstructured data, data management is the war. A fourth requirement is that unstructured data protection solutions lay the groundwork for an archiving future where data can be migrated from primary storage to less expensive storage. Integration of archiving with data protection makes sense, since policies can be architected to make sure that data is not removed from production storage until, not only has it not been accessed for a specified period of time, but also that the data has been protected (copied) a specific number of times.

The first three requirements are not only necessary for unstructured data protection, they are also the necessary foundation for the fourth requirement, archiving. Without it, integrating archive doesn't make sense and is the reason that for years we've been told that backup and archive are two separate processes.

It comes as no surprise to IT professionals that unstructured data is dramatically different in capacity, quantity and how it is used, than in years past. Remarkably, the attitude towards protecting and managing unstructured data has not changed. As unstructured data continues its meteoric growth path, it is time to rethink how to protect and manage it. An unstructured data protection solution that meets these requirements will not only position the organization to protect this data but also to manage it.

CHAPTER 5: MEETING THE FIVE UNSTRUCTURED DATA BACKUP REQUIREMENTS

In the last chapter we laid out the five requirements for unstructured data protection; fine-grained backups, frequent and rapid backups; cloud support, data classification, and an archiving future. Aparavi is one of the first data protection companies specifically focused on the problem of backing up unstructured data stores and they address each of the five requirements.

STRIKING THE BALANCE – FINE GRAINED BACKUP VS. RAPID BACKUPS

The trend in modern data protection, at least for solutions not focused on unstructured data, is to backup unstructured data using an image-based backup. Although, image-based backup allows these solutions to meet the second requirement of rapid backup, it leaves them unable to meet the first, fine grained backup. Aparavi uses a more traditional file-walk method to create the initial baseline of files but also creates a catalog from this walk, and then uses this catalog to check for new or modified files quickly. Unlike legacy backup solutions that walk the file system every time, Aparavi only does it once. The result is that Aparavi gets the file level detail of the file system walk method without sacrificing backup speed.

FREQUENT BACKUP IS MORE THAN JUST CHANGED BLOCK

Image based systems backup just the changed blocks, which allows them to backup rapidly and since the backup completes quickly, those backups can occur frequently. However, all these backups need to traverse the network. Aparavi provides sub-file level backup, which enables it to provide rapid and frequent backups. But Aparavi also uses an intelligent mix of targets protecting data on the file-server first, then to an on-premises appliance and then ultimately to the cloud.

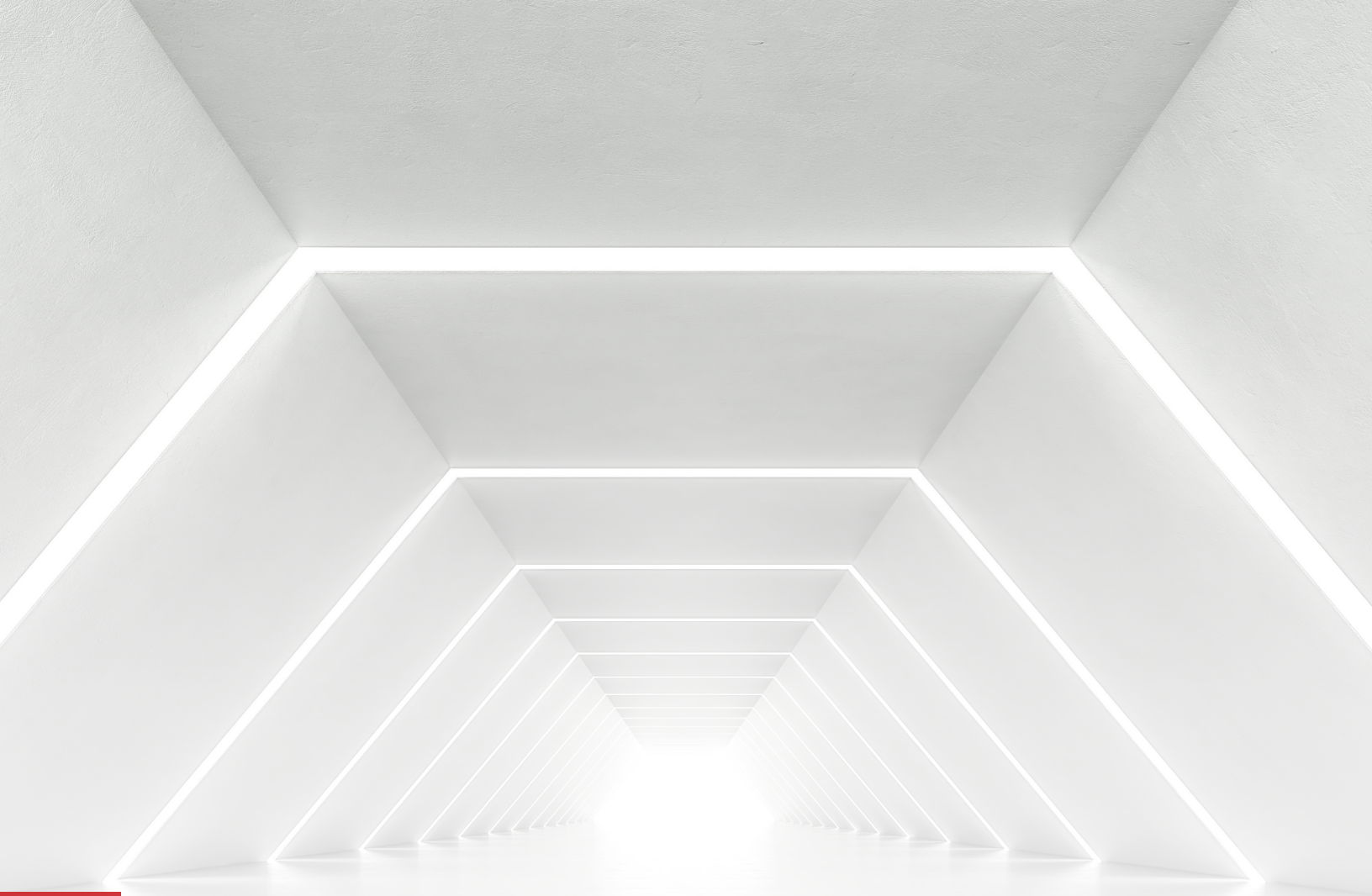
CLOUD STORAGE

To alleviate the capacity requirement of unstructured data copies made by the protection process the solution also needs to support cloud storage but in an efficient manner. Aparavi uses cloud storage for two purposes, first as a disaster recovery copy for any on-premises data. Second, Aparavi also uses cloud storage as a tier so that organizations no longer have to continue to purchase on-premises secondary storage. Most legacy solutions only use cloud storage to create a disaster recovery copy, they do not use it as a tier to relieve on-premises storage requirements. Aparavi's sub-file object storage and active pruning of retired data uses cloud capacity in a highly efficient manner.

DATA CLASSIFICATION

Understanding and organizing data within unstructured data sets is critical. If data can't be found, it might as well not be stored. Aparavi allows customers to organize data by type, size, as well as create, modification and access dates. Additionally, customers can create their own custom tags to organize data by the device that created it (cameras and IoT) or specific projects and use cases.

> As unstructured data continues its meteoric growth path, it is time to rethink how to protect and manage it.



AN ARCHIVING FUTURE

Archiving can describe many different processes. Historically, it is the process of making a special copy of data prior to removing the original data from production storage. The first step in creating an archive is creating that special copy, which if the backup is fine grained, is something the backup process could deliver and does normally. The next step is to classify this data so policies can be set for retention and eventual data movement. A third step is to report and provide analytics of the protected data so that IT can make decisions on what to do with it. The final step is to execute the remove (because the copy already exists) process based on those decisions, thus freeing up production storage capacity.

Aparavi has delivered on the first three steps; fine grained backup, data classification and reporting/analytics and shortly will deliver the last component, the actual removal of files from production storage. Nothing actually has to be moved again, since the backup process already sent it to cloud storage. The timing is ideal, since most organizations will want to run the data protection component and build up backup history prior to any data removal occurring.

For the data driven organization, unstructured data is as critical as data in production databases and today typically represents 80% or more of an organization's total data, but modern protection of this critical asset is lacking. Given its size and criticality, organizations need to take deliberate, well considered steps to protect and manage unstructured data. They need to compare their legacy solutions to the requirements listed in chapter 4 and then see if more modern solutions like Aparavi are a better fit.

ABOUT US

Storage Switzerland, LLC

Storage Switzerland is an analyst firm focused on the storage, virtualization and cloud marketplaces. Our goal is to educate IT Professionals on the various technologies and techniques available to help their applications scale further, perform better and be better protected. The results of this research can be found in the articles, videos, webinars, product analysis and case studies on our website storageswiss.com

APARAVI™

Aparavi Active Archive helps organizations master out of control unstructured data growth with protection, retention, and archive. Delivering both on-premises and multi-cloud mobility, Aparavi delivers true storage independence, and together with an open-data format removes vendor lock-in forever. Aparavi slows secondary storage growth by 75% with guaranteed availability regardless of how long data is retained, and pays for itself in reducing sedondary storage spend. A pay-as-you-go model based on usage eliminates up-front expenditures for a better return-on-investment.

George Crump is President and Founder of Storage Switzerland. With over 25 years of experience designing storage solutions for data centers across the US, he has seen the birth of such technologies as RAID, NAS and SAN. Prior to founding Storage Switzerland he was CTO at one the nation's largest storage integrators where he was in charge of technology testing, integration and product selection.