WHITEPAPER

storageMAP

MODERN UNSTRUCTURED DATA ARCHIVING

BENEFITS AND BEST PRACTICES

As data creation rates have soared, the timeframe for active use of data has shrunk. The end-state result is data flooding onto storage systems at a rate that has required near relentless storage system expansion in an increasingly futile attempt to keep up. And while data archiving is not a new subject it is taking on an increasing level of importance within the topic of unstructured data management.

Before we delve into the details and benefits of archiving, let's cover a few items that are suggested as archiving solutions which, in fact, are not.

ARCHIVING VERSUS TIERING

What about tiering? Isn't that the same as archiving? In a word, no. Archiving and tiering of data are two different things. Archiving involves the relocation and removal of data from primary storage systems. Tiering, on the other hand, is not archiving but rather falls into the category of Hierarchical Storage Management (HSM). Tiering/HSM can appear similar to archiving because data is relocated to alternative storage systems and is removed from the primary storage. However, in place of the original data, an artifact (i.e., a stub or link) is left behind that should provide access to the content that was previously relocated to another tier of storage.

The problems with tiering/HSM (and why the HSM market did not fare well) are numerous. Here are a few of the problems:

1) The tiering/HSM solution owns the data. Without the tiering solution it will be difficult or impossible to recall data.

2) The tiering/HSM solution is partially in-band meaning that cold data can only be accessed by the tiering/HSM solution. If the tiering/HSM solution is down (or even removed from the environment) you cannot recall data.

3) At the time of asset refresh or retirement, the tiering/HSM solution must be used to migrate data. Your options for migration are now limited to the tiering solution which may not handle the complexities often found in full migration scenarios nor operate at the speed and scale required during a mass migration.

4) The tiering/HSM solution stubs are loosely coupled with the relocated data. Over time stubs that reference non-existent data or relocated data with a missing stub become a problem.

The bottom line is that you'll be better off in the long run by implementing a true archiving strategy instead of the short-term fix provided by tiering.

ARCHIVING VERSUS GATEWAYS

Since a NAS gateway device maintains all file related metadata in its global file system, it can be claimed that the device also serves as an archive front-end. Since the gateway's global file system relocates files that are aging, those files are considered archived. But just because the NAS gateway considers the files archived you shouldn't view it the same way. Remember, the metadata (location, file name, etc.) is stored in the NAS gateway so any access to that data in the event of recall is arbitrated by the gateway.

You want your archive data stored in an open archive where you could even recall file(s) using a simple S3 browser. In other words, you don't want your data, or your metadata, owned by another solution. What if that solution gets decommissioned or the vendor goes out of business? Long term retention is in play here so the data in the archive needs to be accessible by any reasonable means over the long term as well.

THE "ACT" OF ARCHIVING VS THE ARCHIVING STORAGE PLATFORM

Another potential point of confusion to clarify is the "act" of archiving data versus the archiving storage platform itself. The "act", or action taken, to archive data is distinct from the storage platform upon which the archive data will be stored. While this may seem to be common sense, it can be a point of confusion given that many vendors talk about their storage platform as an "archiving solution" when, in reality, it is the archive destination.

The "act" of archiving data involves more than the storage platform to which the archive data is relocated. The candidate data needs to be identified, and this identification needs to be done in a variety of ways. For example, do you look for files based on the length of time since a file was last accessed? Or based on the length of time since a file was last modified? Or maybe you want to archive data for a given user ID – either one that was associated with a former employee or even a service ID that owns the data associated with an application that has been retired.

Identifying these proverbial needles in the haystack when there are billions of files distributed among multiple storage systems is no small challenge without the right solution at your fingertips.

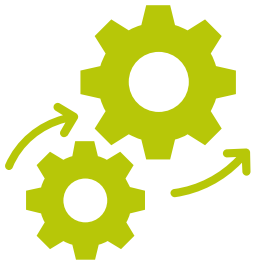
BENEFITS OF TRUE ARCHIVING

Now that we have covered options that are not true archiving, let's cover the benefits of modern archiving where content is copied and removed from the original storage system in an open and transparent format. There are three big benefits to be realized:

1) COST MANAGEMENT/REDUCTION

a. Relocate data to lower cost storage for long term retention
b. Free up valuable performance-oriented storage
c. Extend asset use and avoid expansions
d. Enrich the archive data with metadata tags

2) OPERATIONAL EFFICIENCY



a. Avoid storage system performance degradation (as systems fill up, they tend to slow down)
b. Have smaller snapshots
c. Realize faster backups
d. Have less data to replicate

3) SUSTAINABILITY IMPROVEMENTS

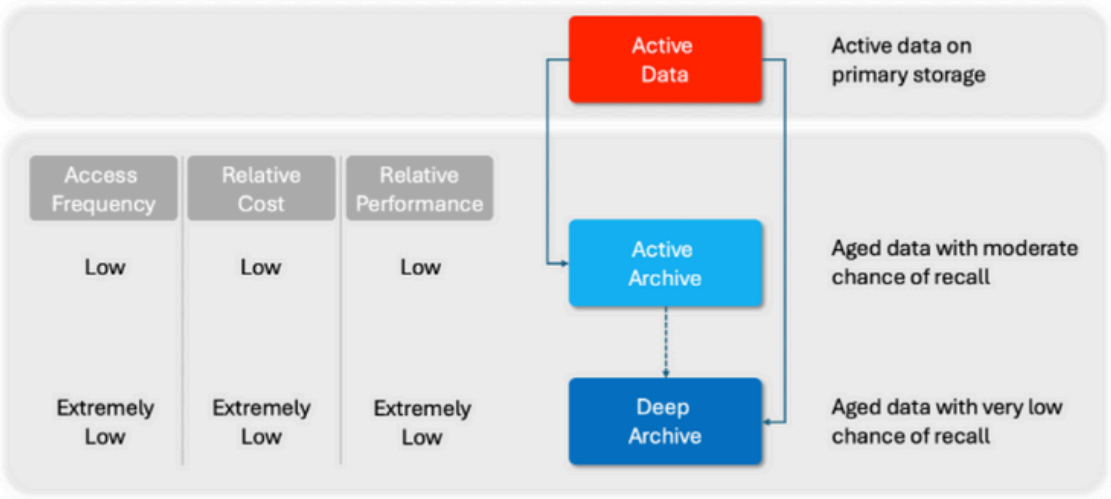
a. Lower CO2 emissions associated with the data being stored
b. Negate HVAC impact by avoiding storage system expansions
c. Potentially consolidate storage systems for smaller footprint

An additional benefit you get by truly archiving data is reducing the "blast radius" involved with a breach. By archiving data, the footprint exposed to exfiltration or corruption by a bad actor is reduced.

ARCHIVING STRATEGY

When creating your archiving strategy, you'll want to consider whether to leverage an active archive, a deep archive, or a combination of the two.

An active archive is an archive that is used for data that has a modest or reasonable chance of needing recall. A deep archive is an archive that is used for data that has very little chance of needing recall but must be retained for either regulatory compliance reasons or for internal governance reasons. The deep archive can also become the next location for data in the active archive that has passed a threshold defined by corporate policy dictating its movement into the deep archive.



Clever use of the active and deep archives can generate higher cost savings, but access requirements need to be considered. Essentially, there is a balancing act between the frequency of access, the cost of storage, and the performance of the recalls.

INSIGHTS - USE YOUR METADATA

A critical step in archiving is to get insights into the aging profile of your files. Policies can be created that dictate files that meet certain criteria get relocated to the archive platform. For example, perhaps files that have not been accessed nor modified within the last 3 years are relocated to the archive platform. Many organizations are surprised to learn that upwards of 60% of their stored data falls into this category. And with petabytes of data and billions of files being stored this can really add up.

Conceptually, this may seem trivial but getting insight at scale when there are multiple storage systems and cloud services deployed (probably from different vendors) storing billions or even tens of billions of files is impossible without the right analytical capabilities.

AVOID LOCK-IN

Finally, the files copied to the archive platform should be done so in an open format. In other words, the files should be accessible without the original archiving software. By requiring the original archiving software to be used for recalls, the data is essentially locked-in to the archiving solution.

Many archiving solutions claim to write data in an open format, but they do so by writing the files in their original format to proprietary containers or other mechanisms that render the files unreachable to any software except their own. You want to use a solution that will store your archive data in an open and accessible format.

CONCLUSION

In this paper we've covered why tiering is not the same as archiving , why NAS gateways do not fulfill the full need of archiving, , why the "act" of archiving is not the same as the archiving platform, and looked at the benefits of true archiving (cost management/reduction, operational efficiency, and improved sustainability).

Of course, to realize these benefits you need an unstructured data management solution that can handle the scale and complexity of your environment. StorageMAP delivers the benefits of true archiving by providing the analytics necessary to find the archive candidates lurking among billions of files based on a variety of criteria.

Beyond that, StorageMAP's Unstructured Data Mobility Engine (uDME) has the power to handle the relocation of data to the desired archive storage platform without locking you in.

The net result is a far more manageable environment with the ability to continue to identify and relocate data as it continues to age and cross the archiving policy thresholds.