Executive Insight
Business Fit Assessment

FREEFORM DYNAMICS

and

SITUATION PUBLISHING

in association with

komprise

Manage your data, not just your storage

Don’t get burnt by keeping ‘cold’ data on expensive ‘hot’ storage systems

Freeform Dynamics Ltd, 2020

Introduction

Most organizations are keeping an ever-increasing amount of data, yet only a small amount of that data is live or 'hot'. Some of the rest is occasionally accessed, but the vast bulk is 'cold' – data that's rarely if ever accessed but is kept in case it is required in the future. It might be needed as evidence if a dispute arises, say, or for regulatory compliance, or because external market changes mean we need to analyze something we could formerly ignore.

Keeping cold data on the same storage as hot data means wasted capacity, cost and effort.

Keeping that cold data on the same storage as the hot data wastes capacity and cost – and time too, if it means you repeatedly backup cold data that doesn't need it. Yet for many organizations, sorting the hot from the cold is like trying to unmix a warm bath.

Others know differently, and have resorted to a range of approaches in order to move cold data to cheaper locations. In this paper we look at the challenges that cold data presents, at techniques and technologies that can help with the problem, and at the advantages organizations can gain from a smarter approach to data management.

The cold data challenge

It's been claimed that data is the new oil, and that it may be the most valuable thing that an organization owns. Partly as a result of these claims, partly because almost everything now generates a data-trail, and partly just because they can, organizations are accumulating more and more data. Worse, the data growth rate for many organizations is exponential, rather than linear.

The advent of cloud storage and SaaS applications has added a whole new layer of complexity. Now, not only have we got more data being generated in new places, but we may not have good visibility into it. Even if we do have visibility, that cloud data exists separately from our on-site data. It's almost as if it lives in a different world.

How much of all that data was used in the last few days, or even in the last year? Do you know which elements are cold and which are hot?

Whether your data is on-site, cloud-based or hybrid, it's little wonder then that many organizations have decided to ignore the problem and simply keep everything, leaving IT to sort out the technical aspects. However, even for the best IT teams there comes a time when adding yet another disk array or NAS 'filer' simply becomes unsustainable from a management perspective.

And then there's the cost aspects. How much of all that data was used in the last few days, or even in the last year? Even if you know that a lot of it is cold, and could usefully be moved somewhere cheaper, do you know which elements are cold and which are hot?

And if you do move cold data to cheaper storage, what happens in a few months' time when a user needs to access it – will they have to wait while it's restored or 'rehydrated', and will that process be automatic or manual? Data only generates business value if it can be turned into information, which means it needs to be both readily visible and easily accessed.

What are you really managing?

Part of the problem is the management approaches and tools that we use. The words we use to describe things also tell us about the often-subconscious thinking behind them. Take Storage Resource Management, Cloud Storage, and Software-Defined Storage, for example: do they imply a focus on the infrastructure, or on the stuff that our organizations actually care about – the data?

In other words, are you looking at the individual letters, or at the sentences and paragraphs that they make up?

Cutting the storage knot

There are three necessary steps or capabilities before you can look at reducing storage costs by solving the cold data challenge:

- Finding out what data you have,
- Gaining visibility into it,
- Taking action on what you learn.

As discussed above, the problem of data management has only gotten worse with the addition of cloud storage into the mix. It means that not only do you need to find solutions that will give you discovery, visibility and control, but they need to be capable of operating in a hybrid infrastructure. That mean that the software needs to 'understand' cloud costs, limitations, advantages and so on, alongside the same factors for on-site storage infrastructure.

Remember that public cloud is not the answer to every data challenge. Depending on the use-case, modern on-site storage or private cloud may be more practical and/or cost-effective.

It is important to remember though that 'the cloud' is not the answer to every data challenge. There's a saying that if all you have is a hammer, then everything looks like a nail, but as an IT or storage professional you have a complete toolkit. Business managers' fascination with public clouds might well be your catalyst for change and the availability of funding, but depending on the use-case, modern on-site storage or private cloud may be the more practical and/or cost-effective option.

Ask questions, act on the answers

The list of questions prompted by the first two steps above – *How much data have you got? Where is it? How fast is it growing? Who is using it and when did they last touch it?* – makes it clear that this is a challenge of analytics as much as discovery and measurement. Any effective system for managing cold data must therefore include deep data analytics, along with the ability to convert its insights into actions.

Examples of those actions include moving data into an archive, deleting it, or at a more granular level, moving it to less expensive data storage. The most common technique used for the latter is HSM (hierarchical storage management), which dates back to the 1970s and 80s but remains widely implemented. As the name implies, HSM treats the storage hardware as a hierarchy, with each tier providing a different balance of cost and performance.

The original hierarchies had two or three tiers: magnetic disk for working data, tape for bulk or long-term storage, and maybe a middle layer of optical media. Modern versions can have even more layers, for example they might add layers of Flash (SSD) or non-volatile memory (NVM) on top, and a cloud tier and/or local object storage at the bottom for long-term archival.

In every case the aim is the same: move files that have not been used for a certain length of time to cheaper, but slower storage. If a file on a lower tier is reused, it is restored to the higher tier. In most cases the user will not notice the delay, because only the most rarely used files will have migrated to the very slowest tier.

Tiering without tears

There are many ways to implement tiering, whether HSM-based or otherwise. As well as software solutions, many arrays and storage systems offer tiering in hardware as a native feature. The challenge they all face is how to make it as transparent as possible.

As well as software solutions, many arrays and storage systems offer tiering as a native feature. The challenge they all face is how to make it transparent to the end-user.

Of course, it’s possible for the storage administrator to manually move old data to an archival tier. In some cases, such as where the chunk of data to be migrated is relatively large – a completed movie project, say, or perhaps a sequenced genome – this may be the simplest option. Users will know, or can be advised, that the project is finished and now resides elsewhere.

However, this will not be practical for most business applications – for reasons of continuity, they need both transparency and automation. In the case of storage systems, this is often achieved by tiering at the block level, for example moving cold data from fast but expensive SAS drives to a cheaper tier of SATA drives. The system maintains its own block map, so when older data is requested it knows which tier to find it on.

However, this process is internal to the array or NAS server. Other tiers can be added, such as cloud or an object store, but then the file must be restored or 'rehydrated' on the primary tier before it can be used, which introduces a delay and eliminates storage savings.

Pointing the way

Another approach, popular with data management software solutions, is to tier transparently at the file level by leaving a pointer, link or 'stub' in the file's original location. This reports that the file has been moved and says where it can now be found. Accessing the file in its new location normally needs no intermediate restore or rehydrate.

Most of today's software solutions use **stubs**, just as early implementations of HSM did. The stub is a small file containing the redirection information. This is a relatively efficient and flexible approach, but one that is specific to the data management solution that created it. This means the stub can only be accessed through that software or via a proprietary software agent.

Symbolic links are standard to NFS and SMB, so no special interfaces or agents are needed.

The main alternative route is the **symbolic link** (or symlink). This is a file system object that simply points to another file system object – normally this will be the target file or directory. The biggest advantage of symlinks is that they are standard to both the NFS and SMB protocols, and are therefore transparent to application, so no special interfaces or agents are needed.

With the concepts and requirements of cold data management and file tiering now explained, let's look at how data management software fits within a business environment such as yours, and at the advantages file tiering can bring.

Cold data management: a real-world example

As our example we'll use Komprise Intelligent Data Management, from the eponymous sponsor of this paper. While nothing we say here should be interpreted as endorsing or recommending this solution, talking around a specific offering enables us to move beyond the theory, and illustrate how some of the key principles we have been discussing can be translated into operational reality.

The IT or storage administrator defines policies based on the data analysis, and the appliance uses those policies to move cold data from primary NAS to secondary storage.

The business needs to take control of spiraling data volumes and storage costs, but of course it must do so with minimal change and disruption. A solution such as Komprise Intelligent Data Management can meet these needs by combining data analysis tools with policy-driven data movement and replication capabilities, all deployed as a grid of virtual appliances.

The concept is simple: you point the Komprise Observer appliance at your

NAS servers, it analyses the file shares and the data they contain. The IT or storage administrator defines policies based on this analysis and the data profiles, and the Komprise Director appliance uses those policies to move cold data from primary NAS to secondary storage such as a capacity-optimized NAS system, cloud storage, or a local object store.

This all helps to answer a second business requirement, which is to bring in and take advantage of new storage technologies, such as object stores and cloud storage, again doing it as seamlessly and non-disruptively as possible.

A key feature that can help with this storage evolution is what Komprise calls its Transparent Move Technology™ (TMT), which uses industry-standard symbolic links to transparently redirect file access to the secondary store. These links are dynamic and resilient – they can be rebuilt if need be, and allow the data to be moved multiple times without changing the symlink on the primary store. This allows Komprise to be storage-agnostic and to sit outside the primary (hot) data path, with no need for software agents and the like. Moved files remain accessible on the secondary store even if Komprise is offline.

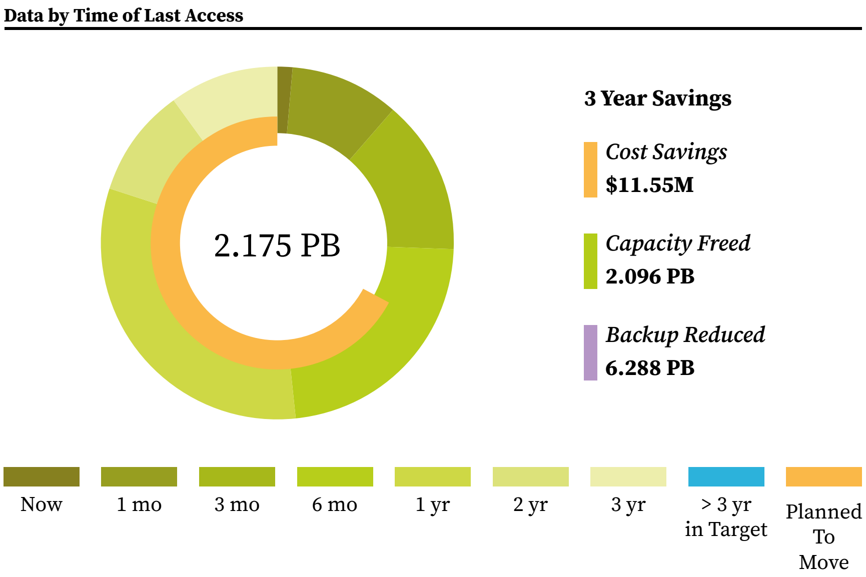
Why deep analysis matters

From a business perspective, deep and high-resolution data analysis is key. Not only can Komprise tell you what file types you have, the distribution of file sizes, and how much data is inactive, it also provides very granular analysis of file access and ageing. This information goes right down to the owner, group and directory level, which allows custom data migration and copy policies to be tailored to specific users, data sets and so on.

One of the biggest reasons for managing data is to cut the cost of storage, so you need to model how migration will change those storage costs.

It can also help with Big Data and AI/ML projects too. Finding the right data is typically the most time-consuming element, so the ability to quickly search and file data, then create virtual data lakes with it, can be extremely valuable.

Komprise analytics of data usage and savings.



Data growth analytics are vital, too. It is not enough to know how fast your file storage is growing, for example. For proper costing (including internal charge-back) and capacity management, Komprise's analytics can also tell you where it is growing, who 'owns' the growth, how much it costs, and so on. The option to set up alerts and run 'what-if' budget and savings assessments can also be very useful here.

And because a major reason for managing data is to cut the physical cost of storage, Komprise's analytics can model how migration will change your storage costs and data access times. This is especially – though not only – relevant when planning to include cloud storage, which can have differing fees for different tiers, for API access, for data egress versus ingress, and so on.

Replicate, move, migrate

As well as identifying cold data for migration, Komprise's data analytics can be used to identify and profile hot data. For instance, if you are replicating to the cloud for disaster recovery, you might choose to replicate only a certain subset of your active data. Alternatively, you might have data that is not currently on top-tier storage but ought to be.

And because Komprise TMT moves the file hierarchy and metadata along with the data, the data is then accessible both from the source NAS via symlinks and directly on the target storage, with all the same access and security controls. This enables it to be used for NAS migrations, among other things.

Conclusion

Properly managing cold or inactive data, and separating it from your hot or active data, can be advantageous for any organization. As well as saving on the cost of primary storage, it can save on administrative costs and significantly reduce the amount of data that you must backup regularly.

Good quality data management can cut costs, seamlessly introduce new storage technologies, and make it easier to convert data into business value.

It is also an ideal opportunity to take advantage of the new capabilities and costing models offered by the likes of object and cloud storage. That's especially true if you can connect them seamlessly to your existing data storage, as technologies such as dynamic symbolic links allow you to.

Lastly, data management and data analytics are essential in the wider business context. Just as storage's real importance is the data it holds, data only matters once you turn it into actionable information and insights. That means your data needs to be both visible and accessible from anywhere, which is something that the right data management tools can enable with ease.

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 research library, visit www.freeformdynamics.com.

About Situation Publishing

Situation Publishing brings you the latest news and views on the global tech scene from its offices in London, New York, San Francisco, Sydney — and through the Anseatic Alliance — Europe.

We deliver an unrivalled mix of fiercely independent, analytical, and original stories and events to technology professionals worldwide, giving our readers what they want.

For more information, visit www.situationpublishing.com.

About Komprise

At Komprise, our mission is to help businesses handle the incoming deluge of data with analytics-driven data management that helps customers know, move and control their data to save costs and extract more business value—without disrupting access.

For more information, visit www.komprise.com.

Terms of Use