quocirca

Are you one eDiscovery away from a budgetary meltdown?

The use of cloud nearline storage for data archival may look attractive at first glance. Watch out for hidden costs – and the impact it all could have when faced with an eDisclosure notice…

Key Points:

Public clouds are offering deep data storage at cents per gigabyte. While this may be a great financial draw at first glance, there are many hidden costs with many cloud offerings that make it very hard to exactly what the costs of recovering required data would actually be.

When combined with poorly implemented means of identifying files and records that are actually required, the costs of eDiscovery can be so expensive as to have a severe impact on a business' bottom line.

Using an online object storage platform using global namespaces can provide full data protection, high availability and easily managed eDiscovery in a way that fully supports the business.

Report Authors

Clive Longbottom
Tel: +44 118 948 3360
Email: Clive.Longbottom@Quocirca.com

Marcus Austin
Tel: +44 7973 511 045
Email: Marcus.Austin@Quocirca.com

Commissioned by:

HGST
a Western Digital brand

"The dog ate it" defence

Governments are getting hot on electronic data. They want to get their sticky mitts on your data – and where they have the requisite warrant, the need to ensure that you provide that data to them within a requisite timescale is vital.

Gone are the days of being able to say "we had a flood/fire in our storage centre – all the paper was destroyed". Similarly, you won't get away with prevarication around "We know it is on a tape – somewhere. We'll find it – eventually"; followed by "Here's three large trucks of stuff – we'll let you find it". A modern eDiscovery notice will be pretty focused, setting out exactly what needs to be disclosed – and when.

Different verticals and regions will have different laws and requirements around data lifecycle management. As an example, the UK data protection laws state that information on a person should not be stored for 'longer than is reasonable' (well, that's precise), whereas the US IRS mandates, amongst other things, that all information should be kept for 3 years since a tax return was filed, or two years from when the tax was paid, or 7 years where a claim is made for losses due to worthless securities.

An example of the confusion that reigns around data retention schedules can be seen in this UK government document, where data retention schedules are shown varying from 2 years to 'the life of the organisation'.

And it is not only governments that will be battering your door down to gain access to the data. The individual now has a lot greater legal access to all the data that an organisation stores about them. In the US, HIPAA forces any healthcare organisation or entity to disclose all required data to a patient requesting it within 36 hours.

Likewise, with areas such as the UK's Freedom of Information (FoI), where public organisations have to provide most types of information requested by a citizen within a set period of time – the need to easily and cost-effectively identify and recover information is pretty much a necessity these days.

This has led to many organisations going for a model of 'store everything, forever, just in case' – often resulting in old storage technologies being present that make data recovery difficult. Although Quocirca does not subscribe to this approach, advising organisations to create a well-planned and implemented information lifecycle strategy, there will still be a requirement for organisations to deal with large (and growing) storage volumes over a long period of time.

However, eDiscovery is seen as an insurance policy, and no-one likes paying too much for insurance. Archive storage on tape is dying away as a mass-market approach, as the speed of retrieval is seen as being too long, and the failure of tapes stored for long periods of inactivity is still seen as too high a risk.

Rush to the Cloud!

Therefore, Quocirca has seen a rush to cloud-based data archival systems such as AWS Glacier or Google Nearline. These systems use economy of scale and commodity storage systems to offer very low cost storage. These 'deep storage' systems are not offline – they are near line. Data can be pulled directly from them at a speed that is slower than on-line primary storage, but is a lot faster than having to go to tape – whether this be offline tape or a near line tape library. Headline pricing is cheap – so when looked at in the light of figure 1, it all sounds perfect, eh?

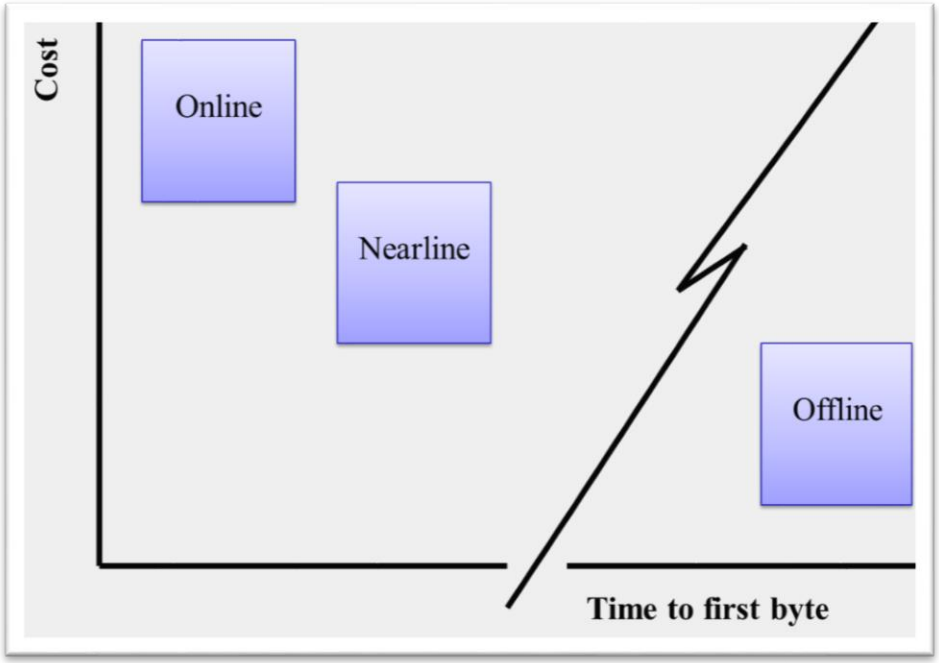


Figure 1

As long as you never need to touch that data again, then it may be a cheap way to waste money. However, for eDiscovery, it may well result in being the biggest, most expensive mistake that you have ever taken.

Why? Sure – storage costs are cheap – a cent or less per gigabyte (GB) per month. To put it another way, a petabyte (PB) of data will cost no more than $10,000 per month to be available for recovery – orders of magnitude cheaper than an organisation could generally manage themselves.

But the point is not around the storage costs – it is really around the recovery costs. That eDiscovery warrant lands on your desk, and from that 1PB of data, you find that you need to recover, say 100GB – hardly that much. Luckily, you have been using a suitable information management system that enables you to rapidly identify all the information from metadata records without needing to touch the files themselves. By doing so, you have created a list of files that now need to be recovered from the deep storage system. You have done this, no? Probably not – and you are not alone. Few organisations have a suitable system in place to deal with this sort of problem.

Anyway, let's assume that you have…

Keep on rolling…

Off you go, happy in the belief that you are meeting the requirements of the eDisclosure warrant. You carry out the retrieval – and find that although you have pulled back a load of files, there are associated files that should be held in records that have not been retrieved. Therefore, you have to carry out another search, and pull things back again. As it is an eDiscovery warrant, you can't just retrieve what was missing – the time stamp for creation and recovery will have to be the same on every file.

Even then, the authorities may not be happy. Some deep storage systems only present a copy of data – proving that this is the exact information that is held in the deep storage system may not be easy without allowing the authorities direct access to the system – in which case, every action they take will require further retrieval costs as well.

The meter just keeps on running, doesn't it?

Then, the next month, the bill arrives. That $10,000 isn't really mentioned. You see, you have been adding data (not unexpected), so the price has gone up. You changed some files – and the cloud storage provider has seen this as the deletion of a file and the creation of a new file, so you have been charged an 'early deletion' fee. If you move an item, the same 'early deletion' fee applies.

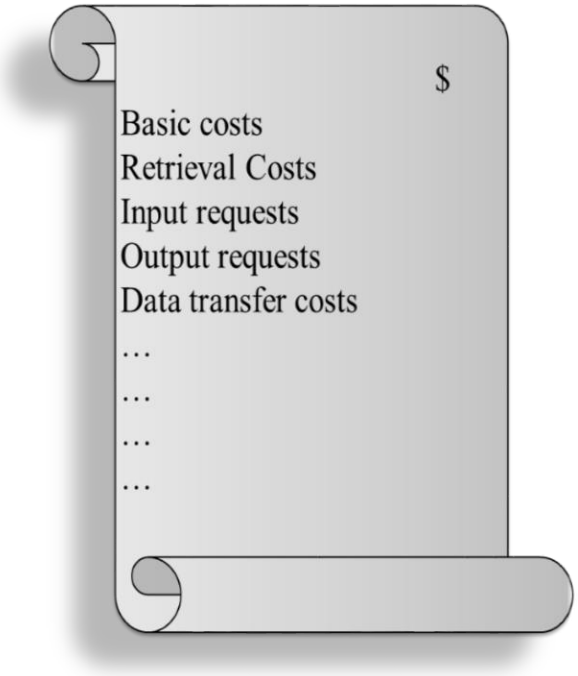
Then there is the cost of retrieving the data that you needed. On the cloud provider's page, the price didn't look that bad – maybe the same price as storage of the data – around a cent per GB. Ah – you missed the data egress charges that weren't spelled out. That could be an extra 12-20 cents per GB. Then there could be data exchange costs – anywhere from free within a region to 12 cents per GB for international data movements.

"Then there is the cost of retrieving the data that you needed. On the cloud provider's page, the price didn't look that bad – maybe the same price as storage of the data – around a cent per GB. Ah – you missed the data egress charges that weren't spelled out. That could be an extra 12-20 cents per GB. Then there could be data exchange costs – anywhere from free within a region to 12 cents per GB for international data movements."

That 100GB of data – a relatively small amount, has just cost you anything up to 33 cents per GB. Oh, and by the way, this is for just one retrieval of the 100GB, not the multiple attempts you might have had to make. But only $33 to retrieve 100GB – surely just a rounding error?

More additives than in candy

Ah – there could well be operation costs as well – how many 'puts' and 'gets' (Upload and Retrievals) did you carry out? And all of these individual costs had to be applied to the multiple recovery operations you had to carry out, as you iterated on the files you retrieved. Pretty soon that 33 cents per GB may start to look like a real bargain.



So, maybe not a financial breaker – but that was just for one retrieval. Now assume that 10% of your customers request all data on them per annum. What if you have to deal with a few legal eDiscovery warrants per year. How about that you have a data breach that requires full disclosure, and you need to search through pretty much all of that 1PB of data and recover a large chunk of it? A few shareholders request some data from a few years back. The auditors need to pull back a load of files to ensure that they can provide you with a clean bill of health. Your ISO accreditation comes up for renewal, and you need to show proof that you are storing information in a specific way – and you have to demonstrate your data retention policies, strategies and capabilities.

Suddenly, that $33 starts to get multiplied – and added on top of the $10,000 per month base storage costs. As an insurance policy, it no longer looks quite as good as it first appeared when all you were looking at was a cent per GB of storage.

Global namespace object store to the rescue!

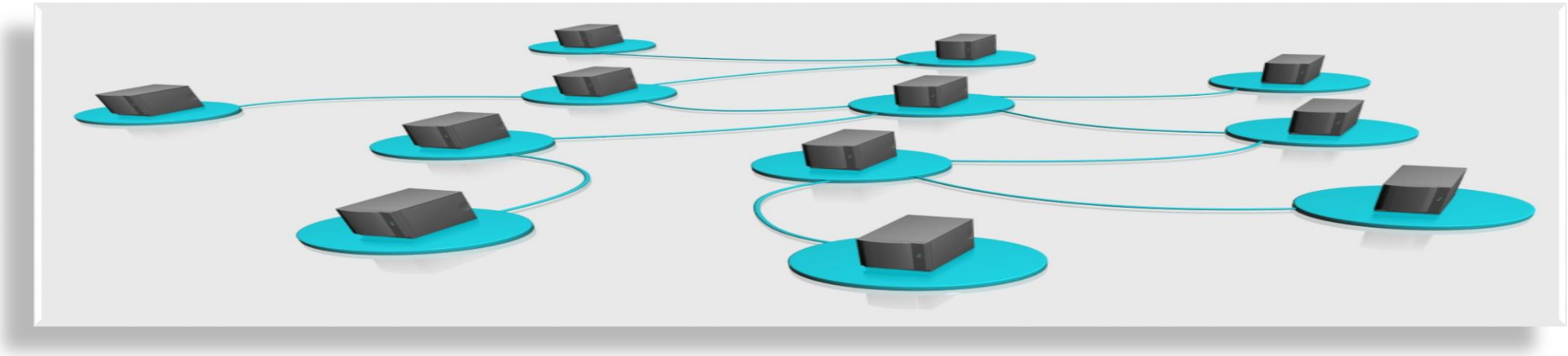
What could be done instead? Well, you could use an object-oriented primary store of enterprise-class disks, using a global namespace for information availability and redundancy. The upfront costs may be more; however, the variable costs would be quite low in being able to identify and retrieve exactly what is required as and when you need it – for whatever reason.

Use the metadata that object storage creates to maintain a full index of what is stored on primary storage. Use that library to easily identify the files and records that need to be recovered. Use the capabilities of object storage to provide a full audit trail of what has been done with and to the data.

Moving away from the variable, and often arcane, cost models of the deep storage service providers may be the best investment that an organisation can make. The costs of data recovery from many of the deep storage platforms can soon start to ramp up. If you do want to use the cloud, then choose a provider who offers a fixed cost model using object storage, where your global namespace can also use metadata from that system, enabling data to be identified and retrieved in a seamless manner across the two platforms.

The use of a dedicated object store – or at least one where the costs are laid out in full in a manner where an organisation can more easily calculate the costs any data recovery will entail – makes a much greater value case to an organisation.

Plus, you won't have to worry about an eDiscovery request turning into a budget busting event.



About Western Digital

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

HGST

HGST (@HGSTStorage), a Western Digital Corporation (NASDAQ: WDC) brand, helps the world harness the power of data. Our smarter storage solutions power the markets and companies that shape our lives—enabling possibilities for the cloud, enterprise and sophisticated infrastructures everywhere. For more information, please visit www.hgst.com #LongLiveData

About Quocirca

Quocirca is a research and analysis company with a primary focus on the European market. Quocirca produces free to market content aimed at IT decision makers and those that influence them in business of all sizes and public sector organisations. Much of the content Quocirca produces is based on its own primary research. For this primary research, Quocirca has native language telephone interviewing capabilities across Europe and is also able to cover North America and the Asia Pacific region. Research is conducted one-to-one with individuals in target job roles to ensure the right questions are being asked of the right people. Comparative results are reported by geography, industry, size of business, job role and other parameters as required. The research is sponsored by a broad spectrum of IT vendors, service providers and channel organisations. However, all Quocirca content is written from an independent standpoint and addresses the issues with regard to the use of IT within the context of an organisation, rather than specific products. Therefore, Quocirca's advice is free from vendor bias and is based purely on the insight gained through research, combined with the broad knowledge and analytical capabilities of Quocirca's analysts who focus on the "big picture". Quocirca is widely regarded as one of the most influential analyst companies in Europe. Through its close relationships with the media, Quocirca articles and reports reach millions of influencers and decision makers. Quocirca reports are made available through many media partners.

To see more about Quocirca's analysts, click here

To see a list of some of Quocirca's customers, click here

To contact Quocirca, please click here.