Cloud Archive – How can it benefit businesses?

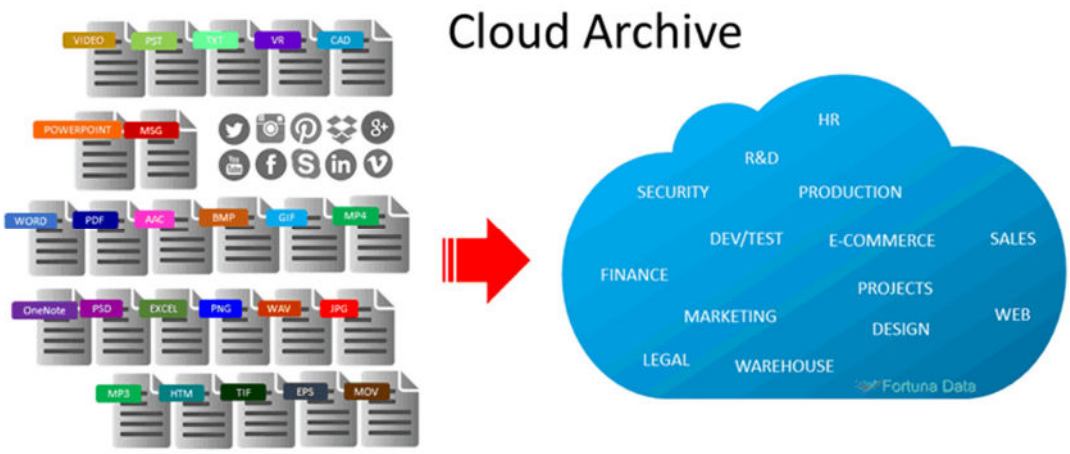
Overview

Data Archiving to the Cloud

A cloud archive is basically where you store infrequently accessed data that isn't essential to the day to day running of the business but could be required if a legal claim arose, engineers wanted to review plans or staff records need to be removed. Once the data resides in cloud archive the following should be followed:

1. Who should have access to this information?
2. Is it easily accessible?
3. Protected from tampering and overwrite?
4. Can the data be indexed and searched?
5. Full audit with date/time stamps, accessed by whom?
6. Data retention policy to ensure how long it is retained before deletion?

Most of the information that resides in a cloud archive will come from legacy storage systems such as EMC Centera, File Server, SAN/NAS storage and tape backups etc. The data that is held in the cloud archive is normally unstructured data and this typically accounts for 70% of all the data kept by the business.



It is estimated that 80% of files on a storage system haven't been accessed for 30 days!

How businesses identify unstructured data is of vital importance to find whether it has an intrinsic value or the next lawsuit waiting to happen.

https://en.wikipedia.org/wiki/Unstructured_data

Unstructured data consists of a variety of document types including Word, Excel, PowerPoint, Audio, Video, Text, Messaging, Pictures, .PDF, Images etc. Unstructured data is the Gorilla in our storage that is difficult to handle due to the severe diversity of files types created.

Data is a digital format that should be easy to find. Nine times out of ten it isn't since:

- New storage systems have been put in place and old systems removed.
- Data hasn't been backed up as expected.
- Data has been deleted accidentally or maliciously.
- Data has been copied to a different location that no one knows about
- Employing IT staff to find data costs money.

It is important that all archival data resides in a single location rather than spread across disparate storage platforms as this could lead to delays in resolving a dispute, project, or contract and this is where a cloud archive enables a business to classify its data to avoid future potential issues.

Lots of data but no idea what it is?

Sound familiar, you're not alone, housekeeping is a problem business of all sizes encounter. Storage is cheap and capacities are increasing so why not move everything across to the new storage platform and store it forever? Data is constantly being created, businesses acquire competitors, integrate them along with all their messaging, files and information, staff come and go, applications are phased in and out, so the cycle continues.

There are businesses who have Petabytes of data spinning 24x7, but no one can easily trawl this information to identify if it:

1. Needs to be kept or deleted?
2. Is legally required for compliance, legislation, or governance?
3. Cause the business embarrassment and fines?
4. Has value to the business but can't be identified?

Data is like elderly relatives, they tend to keep items that they no longer need or use. When they sadly pass away, it is up to you to decide and sort out the valuables, personal belongings or rubbish. In businesses there are no relatives to perform these tasks, it's easier to keep everything forever, regardless of cost and no one is going to decide whether to discard or keep information. Imagine if you could free up 50% of your storage space by carrying out housekeeping tasks. How many files reside on servers for people that left the company 5 years ago and shouldn't be kept?

How do you store information in the cloud archive?

Moving all your unstructured data to the cloud isn't a case of setting up a cloud link and pushing it to the cloud although you could do this you would have no idea what you are sending and the cost of putting unknown data in the cloud isn't good practise that could potentially cause the business concern further down the line. Start small by identifying the types of information that you know and understand. By starting small you could start moving documents relating to the legal department using the following criteria:

1. File extension
2. User
3. Last modified date
4. Last accessed date
5. File containing specific text.

Is your cloud archive part of your business domain or is a 3rd party maintaining your archive? This is important as you may at some time in the future want to move your archive somewhere else and there could be cost implications in doing this.

Once data rules have been created for moving data to the cloud archive you can then start to bring other departments into the system. As an example, you could migrate data from a NetApp Filer and leave a "pointer or stub" as to where the new data location resides. So, a user accessing drive J: would still see his file in this location even though the file now resides in the cloud archive. An alternative to do this would be to create a new drive letter K: and this would be where they search for aged data. Whilst they could recall the data and they modify a file the process of archive migration will start again.

Who manages your data?

Some cloud archiving providers use their own cloud to store data, whilst this might be fine for some organisations in our opinion, we would recommend that whenever possible your archive files are maintained under your cloud tenancy albeit public or private. There are many things that can go wrong if the data is not under your control.

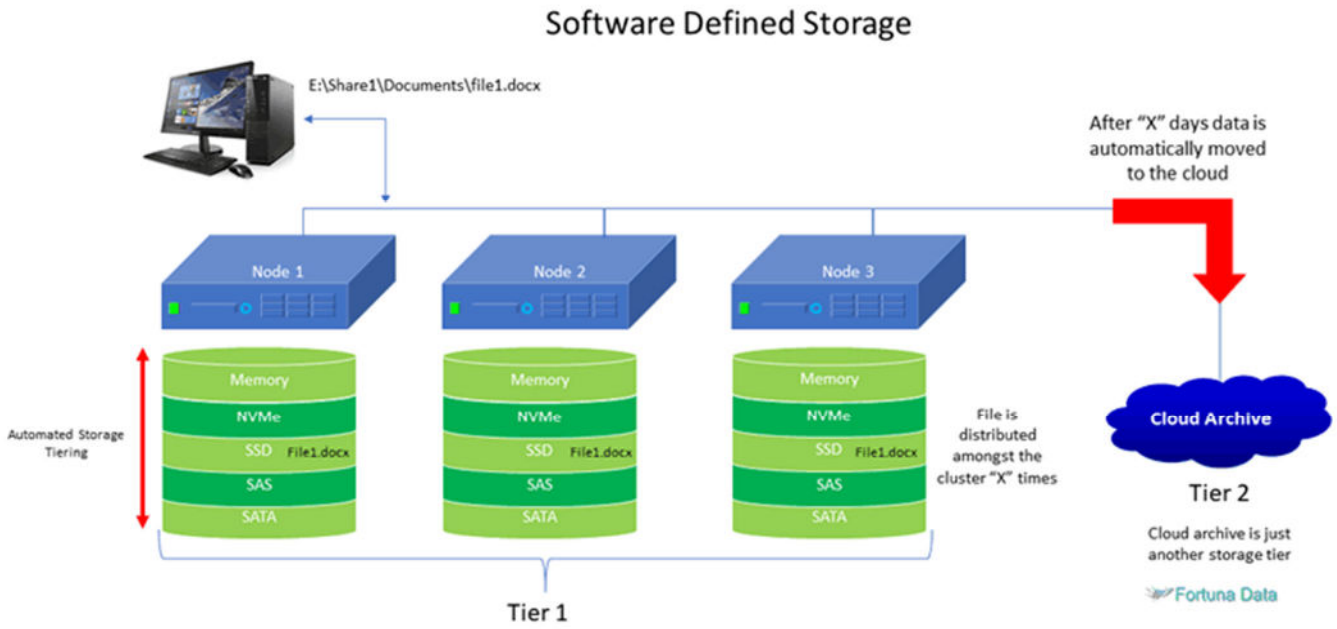
1. Cloud archive provider is part of a legal investigation or has financial issues.
2. Cloud security is lapse, and your cloud archive exposed.
3. 3rd party tampers or deletes your archive.
4. You enter a dispute with your cloud service provider.
5. 3rd parties have access to information that could be politically sensitive, compromise national security or secret projects and information made public.
6. Multi-tenancy whereby your data is stored on the same storage as others.

Fortuna Data

Cloud archive should be treated a storage tier.

A cloud archive shouldn't be a place where information is kept until it is required. A cloud archive should be an active location where information is continually added e.g. completion of a project, service delivery or design for a new ship.

The idea of storing archival data on tier 1 storage as it is the most expensive IT storage the business has. Whilst tier 1 storage provides the capacity and performance a business needs, storing archival data that consumes storage space, requires back up and fills the storage unnecessarily increasing the need to purchase more storage down the line. Filling it with information that is current and active way makes sense.



Backing up archive data residing on tier 1 storage also increases costs as you need additional expensive capacity-based licensing, increased backup device capacity whilst backups take longer to complete. In the event of a disaster, it is quicker to restore a server or image that isn't loaded with legacy archive data.

Cloud based email that isn't backed up after 30 days should be stored in the cloud archive. Store SharePoint data in the cloud archive if you exceed the 25TB storage limit.

Remember this is an active cloud archive and data is available 24x7x365!

Why archive content?

If you are intent on using cloud archiving, then the business has identified important information that must be kept and secured for many years. A typical cost for cloud archiving is based on the number of terabytes stored. Conversely as this is a long-term cloud archive you no longer need to maintain, upgrade, or replace on-premise storage or backup systems.

Archiving in native format

Ensure your archiving provider stores the files in native format as some solutions use proprietary formats to save files, compress, dedupe, multi-tenant and encrypt files. This could cause major issues in the future i.e., 10 years when accessing an archived file application or support is no longer available.

Accessing the archive

The archival location should be viewable and accessible as if where a mapped desktop drive. Access to the archive should be available to dedicated personnel with rights and permissions to view and copy content to external storage devices for access by legal entities or the authorities. Conversely the system administrator shouldn't be involved in recovering archive data unless something has gone wrong with the retrieval process. Some archival software will only allow data retrieval by the software administrator, in our opinion this is not ideal as you might not want staff performing searches on files relating to personnel or business.

Content Indexing vs File Archiving

There are some fundamental differences in the way archiving software works. Some software purely archives the files and has no knowledge of the document contents, author, keyword searches etc. Ideally your archival software should be able to perform content indexing with the ability to create a transcript from a video or audio recording that is searchable and indexed, full indexing of files with extensive document keyword search and reporting capabilities. The ability to view and search scanned documents and view them in another language i.e., Japanese to English.

What is EDRM?

The Electronic Discovery Reference Model (EDRM) is an eDiscovery model created by the EDRM industry association to establish standards and guidelines in the emerging eDiscovery market. The model outlines standards for the recovery and discovery of digital data in response to a discovery request.

Summary Description of the EDRM Stages

Information governance is the management of electronic data to mitigate risk and expenses should eDiscovery become an issue - from initial creation of ESI through its final disposition. Information governance is a proactive process intended to reduce overall cost and risk.

Identification - Locating potential sources of ESI and determining the potential data set's scope, breadth, and depth.

Preservation - Preservation of data, often referred to as litigation (or legal) hold, ensures that ESI is protected against inappropriate alteration or destruction.

Collection - Gathering all potentially relevant ESI for further use in the eDiscovery process (processing, review, etc.).

Processing - Reducing the volume of ESI and converting it, if necessary, to forms more suitable (and cost effective) for review and analysis.

Review - Evaluating collected ESI for relevance and privilege.

Analysis - Evaluating ESI for content and context, including key patterns, topics, people, and discussions.

Production - Delivering ESI to others in appropriate forms and using appropriate delivery mechanisms.

Presentation - Displaying ESI before audiences (at depositions, hearings, trials, etc.), especially in native and near native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

eDiscovery Challenges

The processes and technologies involved in eDiscovery are often complex, time consuming and costly because of the sheer volume of ESI that must be collected, reviewed, and produced in a legal case. Additionally, unlike hardcopy evidence, electronic documents are more dynamic and contain metadata such as time-date stamps, author and recipient information, traces of its passage, and file properties.

- Preserving the original ESI content and associated metadata is the priority of the eDiscovery.
- Process to eliminate claims of spoliation (destruction of evidence) or evidence tampering.

Data Preparation

Once data criteria are identified by the parties on both sides of a legal matter, all potentially relevant content (both electronic and hard-copy materials) is searched for across all potential data repositories and placed under a litigation hold – a process to protect the content (and metadata) from modification or deletion. After a preliminary culling process to remove obvious duplicates or other non-relevant system files, the remaining data is collected and analysed to further cull the potentially relevant data set. The remaining content is hosted in a secure environment and made accessible to reviewers who evaluate and code every file for privilege, confidentiality, or relevance to the lawsuit. All documents deemed responsive to the suit are then turned over to the opposing counsel.

Discoverable Data

For good cause, the court may order discovery of any content that is not privileged (or confidential) and is relevant to the subject matter involved in the suit – no matter where it resides. In layman's terms, if ESI is *potentially relevant to the case*, you may be ordered to produce it. You should always keep in mind that anything and everything is potentially discoverable.

Archiving and eDiscovery

The enforcement of discovery obligations has emphasised the need for businesses to better control all their electronic data, including email, in a more systematic way.

The result has been the development by organisations of information governance programs and the adoption of archiving technologies. Over the past decade, arguments have been made that ongoing archiving reduces the costs and risks inherent in the discovery process by significantly reducing the time, effort and uncertainties usually required in the identification, preservation, and collection stages of the EDRM process.

Archive Migration and eDiscovery Concerns

Many organisations have reached the point where they need to migrate their current legacy archiving system to another vendor or solution, due to changing IT strategies, lack of capabilities, rising costs, or End of Life. Unfortunately, most of the migration technologies available have not been designed to maintain the fidelity of the archived email message nor to provide a defensible chain of custody in the event of a discovery order.

Before planning a migration, IT teams and records managers should speak with their corporate legal department or outside law firm to discuss and address the following:

1. **Is the organisation currently involved in any lawsuits?**
    1. This question is critical as a sloppy migration process can adversely affect current litigation.
2. **Is any content in the legacy archive currently under a litigation hold?**
    1. Content secured under a litigation hold is considered by the courts to be legal evidence, and because of that, cannot be altered or destroyed (or lost).
3. **Is the organisation anticipating future lawsuits?**
    1. Just like the above situation where you are already in litigation, anticipation carries the same legal responsibilities – there is no legal difference between anticipated and current litigation.
4. **Is maintaining archived content in its original state, including metadata, a requirement due to current or anticipated litigation?**
    1. The migration process performed by the majority of archive migration vendors will alter the data, i.e., conversion to another format, loss of metadata, loss of attachments, etc.

5. **Can archived legacy data be migrated to a new platform while maintaining legally defensible chain-of-custody?**
    1. Chain of custody - the chronological documentation or paper trail showing the seizure, custody, control, transfer, analysis, and disposition of physical or electronic evidence in an unaltered state, can determine if evidence can be used during trial. Parties wishing to utilize ESI to back up their case may have to prove the data has not been altered in any way, an almost impossible feat if you are not able to demonstrate and document complete chain of custody.
6. **Did the current legacy archive utilise short-cuts/stubs as a storage management capability?**
    1. Many archiving systems automatically move email messages and/or attachments from the user's mailbox to the email archive and replace it with a pointer to the archive (known as a "stub" or "shortcut"). This process is designed to be transparent; end-users cannot tell which mailbox items include stubs and which ones are "whole" messages. During a migration project, incorrect migration of stubs causes them to stop working, generating error messages to the end-user, and dramatically impacting productivity. Message stubs can also include additional metadata generated from their movement from folder to folder within Exchange. This additional metadata could be relevant in litigation, so message stub metadata should not be arbitrarily deleted but instead re-combined with the original archived message.
7. **How will corrupt messages be identified and handled?**
    1. All archiving software creates some level of data corruption. Is the migration software capable of pinpointing the root cause and recovering "soft" corruptions? How will this information be captured and reported?

Whenever enterprise data is moved, migrated, or disposed of, care should be taken to ensure legal requirements are not being overlooked.

Ensure you work with a migration services provider that fully understands your legal responsibilities.

Litigation & Compliance

Any data you store whether current or past could have a positive or negative effect on the business depending on its content, a cloud archive should only contain business data and should provide all the following:

- eDiscovery
- Full audit and reporting
- WORM (write once read many) compliant.
- Ability to delete files otherwise you could be in breach of GDPR compliance.
- Capable of storing numerous file types
- Support PST ingestion with email attachments
- Store files in native format including metadata.
- Comply with SEC, GDPR, MiFID II rules.
- Full search analysis, including index, retention, and deletion.
- Export data with full compliance
- Secure access using identity management.

GDPR compliance

This legislation becomes law 25th May 2018 and was brought in by the EU to protect customer data and privacy.

The GDPR defines several roles that are responsible for ensuring compliance: data controller, data processor and the data protection officer (DPO). The data controller defines how personal data is processed and the purposes for which it is processed. The controller is also responsible for making sure that outside contractors comply.

Article on GDPR compliance http://www.cloudbackupservice.co.uk/gdpr-general-data-protection-regulation/

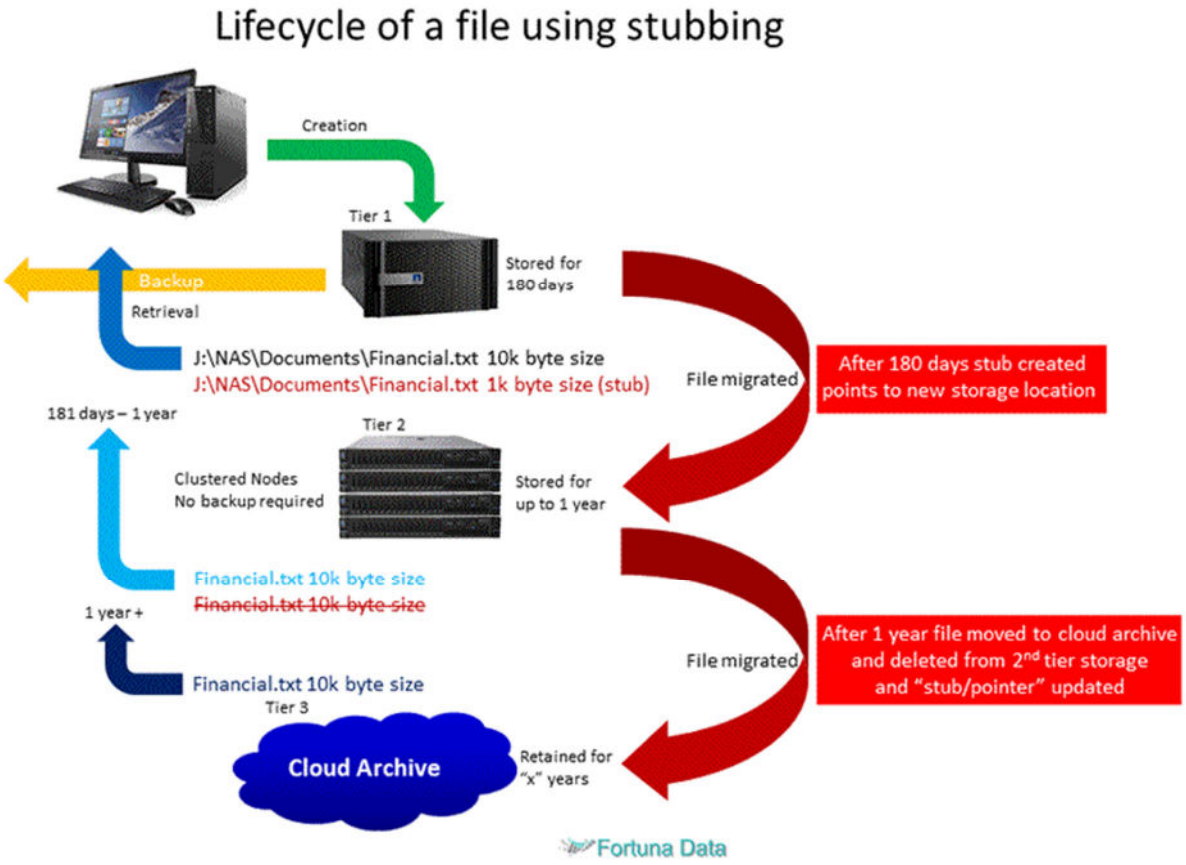
Organisations can be fined up to 4% of annual global turnover for breaching GDPR or €20 Million. This is the maximum fine that can be imposed for the most serious infringements e.g., not having sufficient customer consent to process data or violating the core of Privacy by Design concepts. There is a tiered approach to fines e.g., a company can be fined 2% for not having their records in order (article 28), not notifying the supervising authority and data subject about a breach, or not conducting impact assessment. It is important to note that these rules apply to both controllers and processors — meaning 'clouds' will not be exempt from GDPR enforcement.

The Cloud

The cloud whether public or private provides a simple, secure, and effective way of having access to your data and information 24x7x365 from anywhere with an internet connection. Simple access to unlimited storage, processing power and applications makes this a compelling reason to move. Businesses are increasingly moving to the cloud to reduce their own IT in-house systems, ease administration; the issue is they do not want to be tied to using expensive cloud storage.

The perfect balance for business would be to store as i.e., 6 months data locally and thereafter move it to the cloud but at the same time be always available. Or keep everything in the cloud and replicate 6 months of data locally. The reason is in the unlikely event you cannot access the cloud the business can continue to operate and run.

File stubbing



Some archive software packages leave a "pointer/stub" indicating where the file now resides. A stub is typically 1k in size against a file that might be 10MB. Although stubbing works well things can go wrong due to the following:

1. Backup software has difficulty backing up and restoring stubs.
2. If you replace storage and forget to update the archiving software that files are now at location "x".
3. Client's machine is replaced.

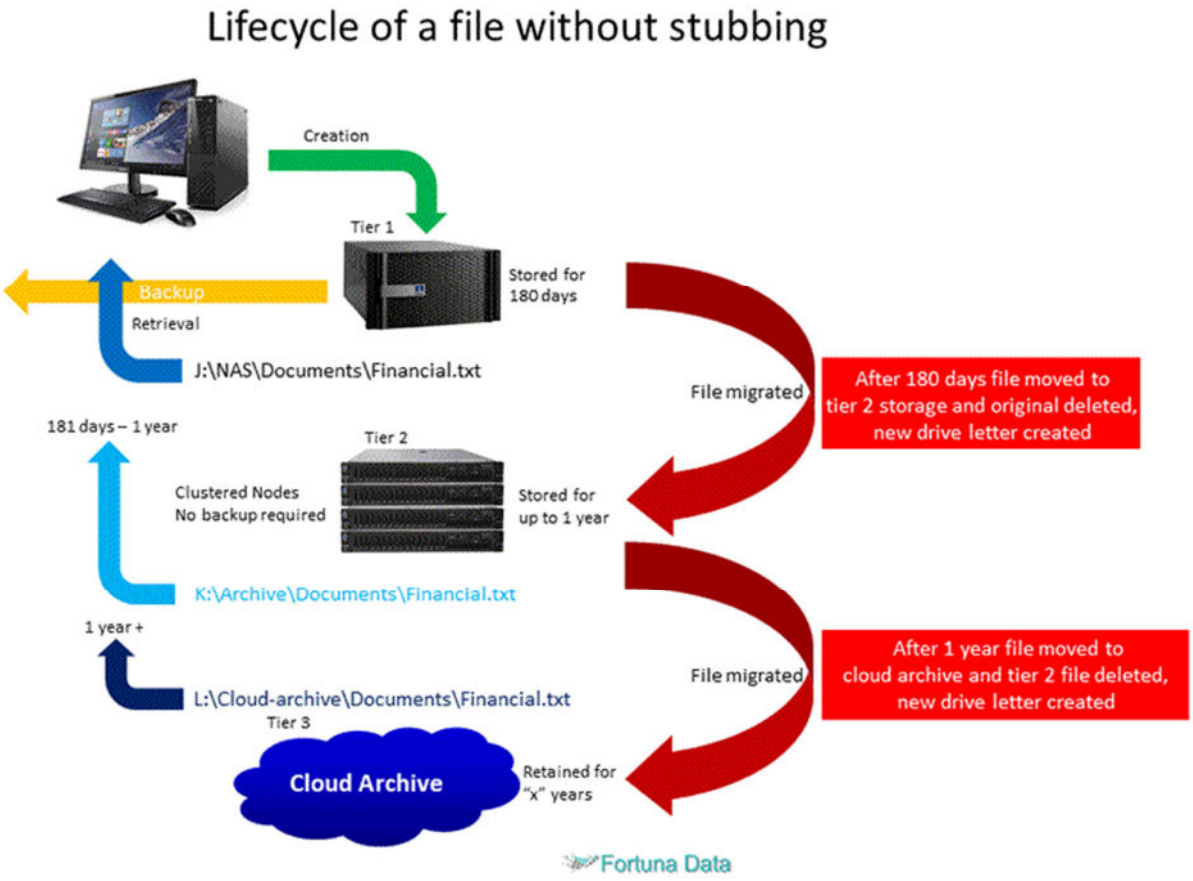
Your file server, storage or desktop could end up with lots of orphaned stubs with pointers going nowhere.

Stubbing works, but over time you will be forced to spend a lot of time and effort correcting file pointers and we don't recommend it.

What's the alternative to stubbing?

Many operating systems and software defined applications manage all the files and migrate them through storage tiers automatically. Therefore, a user who access drive letter J: is unaware that this pool of storage is made up of SSD/SAS/SATA disks and probably will not notice much of a performance difference when requesting the file as many of the aforementioned solutions have a high speed read/write cache.

Ideally if you are not planning any of the above, it would be better to physically copy all the files to the new storage location and re-map the user's desktop.



CAPEX vs OPEX using Cloud

Moving to the cloud is an OPEX cost, whereas CAPEX is an upfront or leased cost over a fixed period. With a CAPEX purchase the amount of storage you require can be an underestimate whereby the business growth and applications for storage and has been underestimated, therefore a request to purchase more is required. Conversely by purchasing more storage space than you need to have involved more expenditure than required. With the cloud as an OPEX you only purchase what and when you need, pay a recurring monthly charge based on the amount of storage you have use, you also could move data from fast tiered storage i.e., SSD to slower storage SATA to reduce cloud costs. It is this flexibility with none of the service, maintenance and running charges that makes cloud so attractive to businesses.

Backup vs Archiving

These are two different and distinct processes and should not be confused.

1. **Backup** – A backup allows for the instant restore of information residing on desktops, notebooks, servers, or storage. A backup is normally performed nightly although increasingly businesses are adopting point-in-time backups that have the capability to recover data lost within the past hour or minutes. A backup is time and business critical as its current and relevant.
2. **Archive** – An archive is where a single copy of the data resides that has intrinsic value to the business and can be proven to be the master file that hasn't been changed or tampered with since its creation as this is vital when presenting archive information to courts, regulators, or customers. Archival retrieval times are not as important as data normally resides on slower storage technologies.

Financial Savings and Cost controls

The realisation in many businesses is that is becoming increasingly difficult to control IT costs due to sheer volumes of data created, actioned and stored. The pressure to do more for less is only too apparent and by investing in new storage technologies that provide compression and deduplication alleviate the problem temporarily. Backup windows are shrinking, WAN performance is poor and expensive in the UK. Let's be honest, without IT there would hardly be a business as we rely on it so much, conversely IT shouldn't be the poor relative and should integrate seamlessly with the business demands.

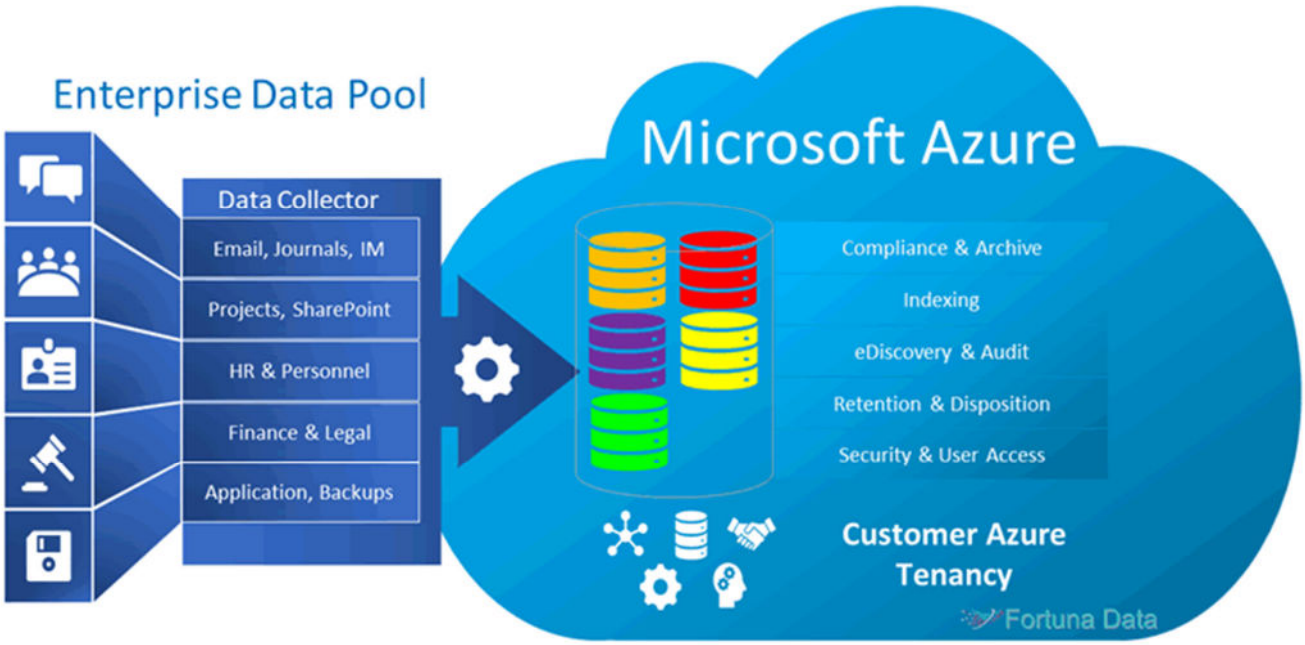
Whilst the cloud provides a centralised storage location for data there are concerns about security, unpredictable costs, and accessibility. In a perfect world this would work, unfortunately this isn't the case and so we must be creative in reducing IT costs.

- Data that doesn't change should be archived.
- Current and active business data should be put in the cloud for easy access.
- Backup expenditure needs to be re-thought.
- Reduce storage complexity and management.
- Involve IT in business decisions.

Earlier in this document we mentioned that it isn't ideal to use the cloud to dump everything, whilst this might be your only option. Moving 10TB's of data to the cloud is an OPEX cost, here is an example of dumping data to the cloud.

- Cost of putting 10TB of data in the cloud £400 per month recurring
- 2 months to setup data analysis rules
- We analyse the data and find that we can reduce our archive by 50%, thereby reducing our OPEX cost by £200 per month.
- Over 12 months you would save £2,000 and the data might be a value to the business.

Moving to a cloud archive makes sense if your existing IT infrastructure and systems need replacing or upgrading, unpredictable storage growth, your applications do not need sub millisecond access, tired of increasing software and maintenance charges or require 24x7 global access. You will also save on datacentre real-estate, no more floor space for systems that require constant cooling and power, significantly reduced backup software and hardware licensing, support and management.



The Solution

We provide a complete cloud archive service to provide cloud-based archiving of email, enabling organisations to meet demanding compliance requirements and address eDiscovery requests effectively and easily. Ideal for organisations using cloud-based email services like Office 365. The business data would transparently move to the cloud, and it would be indexed, catalogued and searchable the same as if your IT systems were still in-house organised, accessible, and secure.

We have a SaaS (software as a service) solution that provides all the advantages above in a simple, affordable, and easy to use application. We also provide Office 365 backup for customers as a separate solution and Exchange to Office 365 migration solutions.

Provide product demonstrations and site visits to discuss your requirements in more detail.

If you would like to learn more, please call us on 01256 331614 or email solutions@data-storage.uk for more information.