

A case for cloud-integrated archival solution, powered by a highly available, multi-cloud integration gateway -oneCloudStorage -The production implementation of the granted US Patent US9128626

Athinio Technical Paper

Peter Chacko.

In recent years, growth in the digital assets and the resulting storage requirements has grown dramatically. Business of all sizes grapple with un-controlled growth of file level(un-structured/semi-structured content) business information including customer files, scanned documents, e-mails, data in various cloud services, machine generated files such as log files in disparate sources, design/IPR, audio and video files which must be stored, and preserved and may need to be discovered from multiple locations of any business.

To stay competitive in business, companies have to stay compliant with regulations, such as HIPPA that requires years of retention of patient information. Legislation such as Sarbans-Oxely Act[2002] raised the bar of corporate governance and accountability as they force businesses to retain documents for a long period of time, in an immutable and tamper-resistant manner. Even individual consumers have a strong motivation to store their personal data such as photos, personal financial records for long term storage. Large enterprises have dedicated IT in place and can implement in-house storage investments to store various data for long term that can survive decades. This is simply impractical for SMBs/SOHOs of the world or power consumers of digital revolution. ROBOs (Remote offices, Branch offices) of Enterprises also need a cost effective ways of archiving data for a long term, without worrying about the data protection, or employing the IT personal to keep the lights on. Outsourcing archival workloads to a third-party data center such as public cloud storage or hosted private cloud storage is now a viable alternative. Cloud storage is the best solution to implement reliable, cost-effective archival solution as the provider amortizes the cost of added redundancy, security, availability over several clients. But unfortunately, out-sourcing data means losing control on their data in some way in terms of vendor dependence, service outages, data security, and visibility, thereby losing the peace of mind for the data owner.

Enter oneCloudStorage, the game changing archival storage platform, getting the best of cloudstorage in and getting the negatives of cloud storage out for a variety of use cases, with flexible tools, interfaces to integrate customer data, implemented on top of third-party cloud storage services with no cloud-adoption issues as vendor-lockin, cloud outages or data security.

The following sections will explain the purpose of Archiving, the special needs of archival workloads and how Athinio OneCloudStorage uniquely positioned in the market as the archival storage delivery platform.

Purpose of Archiving :

Staying compliant with regulations/IT retention laws is the primary objective. Management simplification and new initiatives such as Future projects that may need a historical data/past transaction or big data mining on all stored data at one place motivate many business to archive all their digital assets in one place. Cost reduction on storage resources motivates SMBs, as it frees up the primary storage and administrative burden on the IT staff. Archiving to a safer, long term cloud storage service means assured availability of data in a low cost tier. For individuals, it's mostly driven by their emotional attachments to "yesterdays", like the photos of their kids, honey moon snaps and for keeping their personal compliance to government regulations.

Special characteristics issues of Archival storage system.

Archival storage systems are optimized for write-once, read-maybe usage pattern as it emphasizes on the long term preservation of their data. Data life time of conventional, primary storage assets are measured over months and years, while long term archival systems are measured for decades. There are many security threats that can shake the secrecy, integrity and availability of the archival storage, as data is kept for long periods of time, where conventional assumptions on data protection do not hold water in the context of long term archival context.

- ***Encryption challenges for long term storage:***

Encryption keys are the single point of a failure, when it comes to archival storage as people who know the key can be no more, hardware/software when files need to be retrieved after decades !!. Did I say about the possibility of the only person who has access to key being no more ? When key becomes unavailable, data is lost. Re-encryption at a future date is a good option, but not a simple process. Key-rotations and access revocation also are necessary. Cryptography is only computationally bound, and is not certain about the future of cryptosystems, in the context of technologies such as quantum/cloud computing. Long term archival storage system, that implement secrecy through encryption should adapt to those changes to computing and takes its own time.

- ***Data discovery/active archives:***

Sufficient mechanisms should be incorporated to an archival storage system, to discover the files easily and in a timely manner, with rich metadata tagging, intelligent structuring of various data objects. If an e-Discovery involves finding the right tape after collecting the storage allocations information from multiple stakeholders/Lines of business and then snapshot of an

archive, need a storage system to mount it, and then find the right snapshots for it to work, and then to search for the file in the file system while file names are forgotten, its essentially not an archive storage system. Active archival storage systems, enable a user to find files as if, its available on-line, “right now” and “right there”. Certain industry regulations may also demand a chain of custody audit trail of data.

- ***Integrity gurantees :***

Conventional storage components are not designed for storing data for decades and can silently corrupt the data due to the hardware errors(Disk electronics/Bit rots, storage system errors, user errors, software errors, data damage by a malicious attacker, ransomware/cyber threats etc). Integrity of the data should be checked on regular basis, and self-healing software must be in place. Integrity checks or disk-scrubbing procedures use strong cryptographic hashes such as SHA-1 to detect errors which has practically no collisions.

- ***Migration and recovery***

As data is stored for decades, data must be relocated on a regular basis, to a different system for protecting it from the media failures, vendor mishandling, and also to move to another place for security reasons.

- ***Cloud storage un-availability***

The cloud provider that stores the data can increase the price, or reduce the service quality or can go out of business as we are talking about decades .Looking at the history, more than 70% of fortune companies of the past 25 years are gone or changed hands. Sudden demise of a famous cloud storage startup Nirvanix is an example. Large companies can also suddenly stop the cloud storage services such as EMC Atmos. As such, depending on a single vendor for long term storage simply doesn't work and vendor dependence is no longer an option. It should be implemented in such a way that data can be easily accessed, while the vendor providing the service has issues with their business. When a file needs to be accessed, cloud service can be un-available. A cloud integrated storage system should have mechanisms to contain this un-availability either short-term or long term.

- ***Scalability of the storage system and integration with Big Data infrastructure such as Hadoop:***

An archival storage system should not be limited by any components that can run out of memory or disk-capacity. Many software systems are designed for a single server in mind, having internal memory based data structures as the building block such as a host based data base. Some systems use internal/embedded databases, which are inherently not designed for archival class, scalable metadata serving. A smart archival storage system should have a seamless integration with big data infrastructure such as Spark/Hadoop, as archiving and big data analysis are now the two sides of the same coin.

- *Difference between backup from archival, and what it means to an archival storage platform.*

Backup is implemented for near-term data protection meant for mission critical, application recovery or disaster recovery, while archiving is for long term data storage meant for compliance purposes, litigation responses, data forensics which are usually meant for helping companies stay in business. Backup doesn't save disk capacity as it is not a primary copy, while archiving to cloud storage moves the data to another lower cost tier. Backup is another copy of data which may not be accessed for regular business operations, when there is no data loss, while archive is the primary copy that may be accessed for business operations.

- *Lack of cloudOnRamp tools to cloud storage.*

Many a times third-party cloud storage services use proprietary interfaces and customers find it difficult to push data to cloud storage, without developing a special app for that. Some market leaders such as Amazon now offers gateway solution, but only help customers onramp to S3 only. Customers data located at disparate sources, such as desktop PCs, branch office file servers, hosted servers in the cloud, or in various cloud storage services. Aggregating, consolidating all these data to steer into archive is a nightmare.

- *Geo-Preference and Geo-specifics of archival storage location*

Some of the archival work-loads are subject strict compliance/regulations/data sovereignty mandates and force customer to store in the same country. HIPPA requires US patient information is stored in the US, and many Indian government rules also stipulate the same, which make cloud storage based archiving difficult as none of the cloud services offers country-based storage placement. The new regulation which comes into force on May, 2018 require specific requirements for EU citizens data as it relates to data privacy, data ownership and data availability.

Benefits of OneCloudStorage, archival platform.

oneCloudstorage is a purpose-built, Multi-cloud storage integration gateway, propelling cloud storage as the archival destination solving all the the above mentioned challenges for data archiving and is designed from the ground-up for propelling third-party cloud storage services such as Amazon, Google Cloud platform etc as the reliable destination of long term data archiving, without the overheads of cloud outages, vendor-locking or cloud security issues. Athinio oneCloud is an overlay cloud storage, which is built on top of the positives of various third-party services, and presents itself as a "virtual, storage cloud storage. It use Exabyte-scale data structures for scalable metadata serving,

Did it say, onecloudstorage technology granted a US patent (Invention name is “Distributed Virtual Cloud) in 2015 ?

Multi-cloud storage virtualization technology is the crux of the backend core. This technology virtualize storage at the granularity of cloud storage level, and multiple cloud storage services are aggregated, virtualized, like a RAID controller virtualize storage across independent disks, to provide storage availability in the context of a cloud outage. oneCloudstorage has the ability to transparently migrate the data, for data relocation or vendor hand-off and has in-built cloud data verification mechanisms providing integrity protection. Data security is provided by state of the art, information dispersal technologies based on reed-solomon erasure coding, with or without using encryption technologies, while key management overheads is avoided completely. With information dispersal, no data in a single cloud storage location reveal any meaningful information, offering information-theoretical security. All data can be seen on-line, and can be searched, and recovered without any administrative intervention. Athinio has built its on cloud storage module, which can enable a partner to launch a cloud storage service which athinio can steer data into, if a specific customer want the archived data stored in a country where there is no cloud storage service available or if the customer want the archive be in their own data center. As in the figure, athinio has built many surrounding customer data integration modules, to help customers easily push their data in flexible ways to oneCloudstorage system. Figure shows on top, 4 interfaces which are,

- **Archival app:** This is a down-loadable agent which can be installed in a windows PC in 2 minutes. Customers can then right-click their mouse, and select “send to archive” and file are pushed to onecloudstorage, which are available for search and retrieval later.
- **Archival gateway/archival file system:** Hosted servers in the cloud can download and mount the file system to steer the data to oneCloudstorage seamlessly, for off-site backup or replication purpose. Archival gateway can be down-loaded into a standard PC/server converting itself as a gateway to oneCloudStorage for a seamless access to archival platform, over a standard NFS/CIFS interface.
- **Cloud storage services plug-ins:** These plug-ins pull inactive cloud data such as google docs and dropbox to oneCloudstorage based on various policies, metrics for archiving, on a cloudstorage-to-cloudstorage model.
- **CDMI compliant APIs allow third-party ISVs,** or partners develop their own storage apps, onto oneCloudStorage platform for differentiating their cloud-storage powered products with oneCloudstorage value proposition. In this way onecloudstorage intend to play in the market as a highly extensible, cloud storage delivery platform.

In the middle , oneCloudstorage platform is shown with two gateways(one as the primary and the other as the secondary) that replicated metadata in real time, having a backup copy to a metadata cloud, and send customer data directly to various cloud storage services. Customers can download a recovery agent, using which they can access their data directly, without any access to onecloud storage.This feature eliminates any single point of failure for the platform.

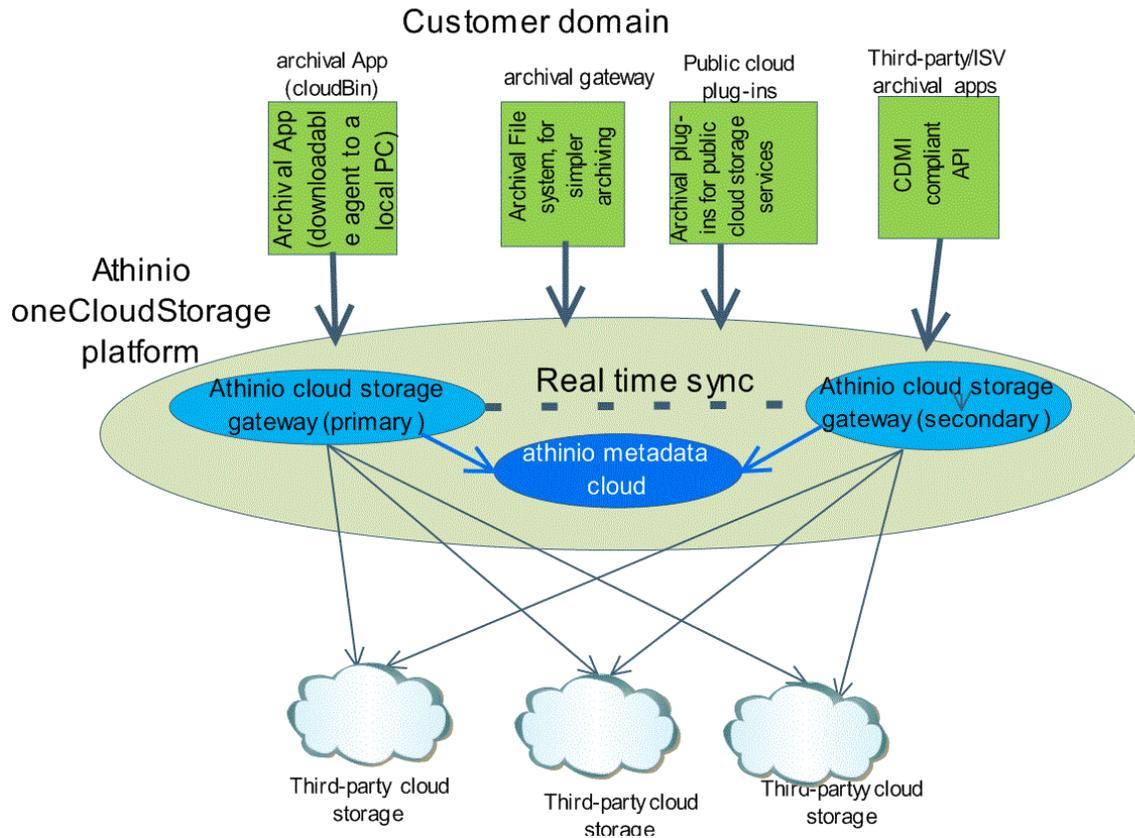


Fig 1

Please visit www.onecloudstorage.com for more information and to get a real sense of how it works !