

Best Practices in Designing Cloud Storage based Archival solution

**Sreenidhi Iyengar & Jim Rice
EMC Corporation**

Cloud storage facilitates the use case of digital archiving for long periods of time by transparently providing scalable storage resources. With ever increasing amount of data to be preserved for legal and compliance reasons, cloud storage when designed correctly, can provide a low cost solution in a geographically distributed environment.

This presentation highlights the key considerations while developing an archive product using cloud storage based on REST interface. It goes on to highlight the design choices while developing a file based archiving solution to cloud storage using EMC Atmos as an example. The aspects covered in the slides are – security, performance, using vendor neutral APIs, developing portable application irrespective of the backend cloud supported, taking advantage of geographically spread cloud storage nodes, faster searches and an efficient disaster recovery mechanism

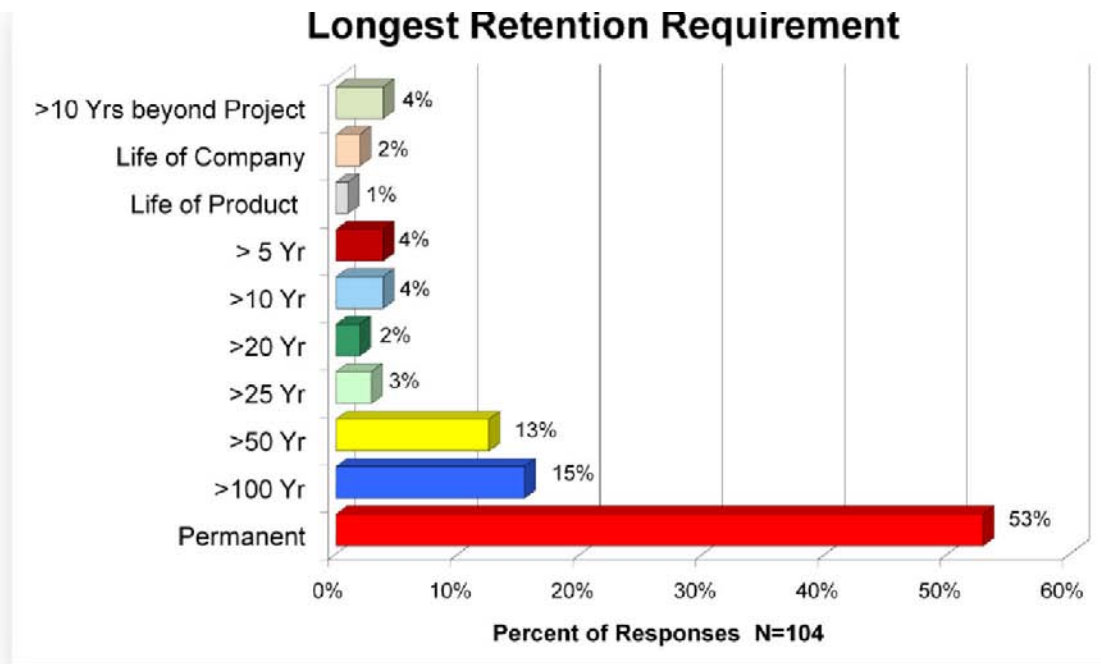
Learning Objectives

- ❑ Considerations while developing an archival application with cloud storage
- ❑ Security and performance aspects while designing an application for cloud storage
- ❑ Leverage cloud vendor provided capabilities in your application

Agenda

- ❑ Archiving
- ❑ Cloud Storage
- ❑ EMC Atmos
- ❑ Development considerations
 - ❑ Performance
 - ❑ Security
 - ❑ Metadata feature
 - ❑ Flexibility
- ❑ References
- ❑ Acknowledgements
- ❑ Q & A

- SNIA - A collection of data objects, perhaps with associated metadata, in a storage system whose primary purpose is the long-term preservation and retention of that data.



Source: 100 Year Archive Requirements Survey

- SNIA - Delivery over a network of appropriately configured virtual storage and related data services, based on a request for a given service level

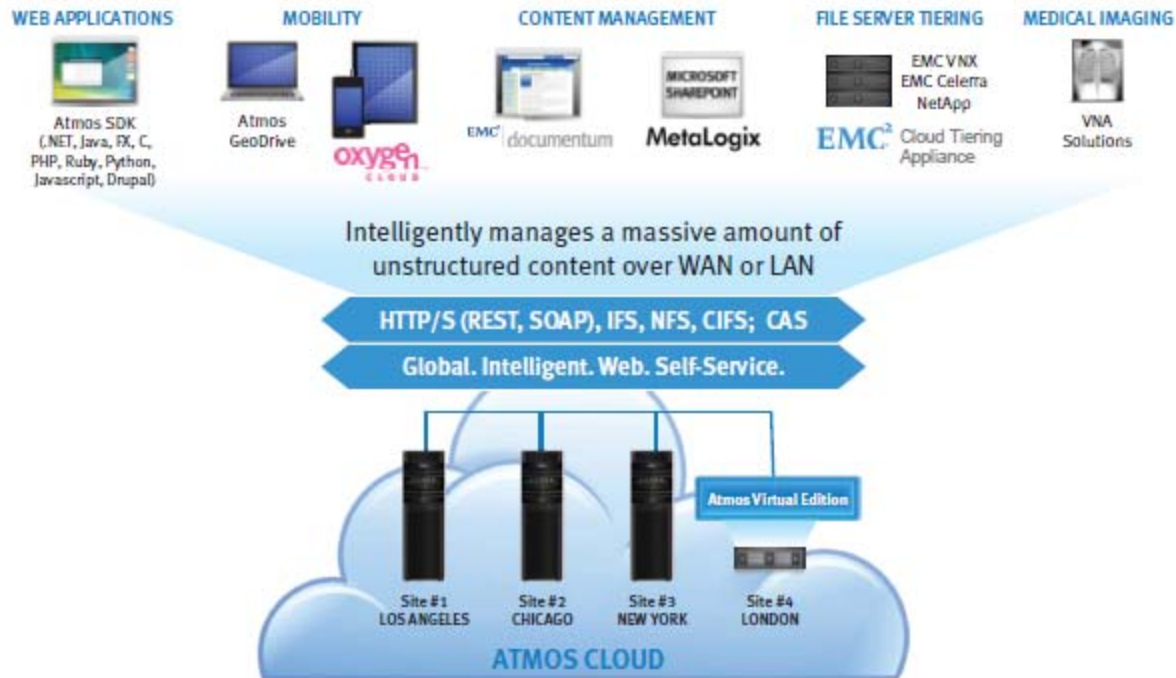


Image Source: www.indianweb2.com

Pay for what you use
No Storage devices
Maintenance offloaded to service providers

- EMC Atmos
- Amazon WebServices
- Windows Azure

- EMC Atmos is globally accessible cloud storage platform to manage content in modern content-rich applications, high-scale content infrastructures and cloud service provider environment.



- ❑ Stands for REpresentational State Transfer
- ❑ Standard programming interface for Apps integrating with Cloud
- ❑ Methods
 - ❑ POST
 - ❑ GET
 - ❑ PUT
 - ❑ DELETE

Key considerations in cloud storage based archival solution

- ❑ Performance
- ❑ Security
- ❑ Retention
- ❑ Searching
- ❑ Flexibility

- ➔ Issues to be considered by application
 - ❑ Archiving application could be far away from storage; high latency
 - ❑ Data access in reasonable amount of time
 - Amazon S3 - Peak requests: 200,000+ per sec (source: www.datacenterknowledge.com)
 - ❑ Low cost ≠ low performance
- ➔ Design consideration for performance optimization in data transfer
 - ❑ Asynchronous Transfer of Packets
 - ❑ Load balancing – Read & write from multiple nodes in round robin fashion
 - ❑ Optimum buffer size for chunk transfer
 - ❑ Send header and data together with small files

→ Data access considerations

- ❑ Data resides somewhere else
- ❑ Hacks during data transfer
- ❑ Avoiding blocked/unavailable port numbers

→ Provide data protection options, such as,

- ❑ Facility to use HTTPS for secure data transfer
 - ❑ Based on security level of the network, option to choose HTTP or HTTPS
- ❑ Expose cloud storage data at-rest encryption choice to users through the application based on rules
- ❑ Securing REST messages with signatures (hash of the request and secret ID)
- ❑ Port number modification

- ➔ Persisting access restriction to the cloud
 - ❑ Access from anywhere at anytime
 - ❑ Access only to legitimate users
- ➔ Data protection
 - ❑ Authenticating by signatures (combination of User ID and request header)
 - ❑ User level authorization using ACL
 - ❑ Pre-authenticated URL to an object

- ➔ Aspects to address archival needs for
 - ❑ Fast search with high accuracy
 - ❑ Disaster recovery
- ➔ Mechanism to identify objects written
 - ❑ Expose “Name = value(s)” pair tagging to end user
 - ❑ Tag acts as index for the object
 - ❑ Fast search by querying tag/token
 - ❑ Disaster recovery by tagging unique-ID with every object written
 - ❑ Versioning of objects (files)

→ Data availability considerations

- Retain data >50 years or permanently (Source: 100 Year Archive Requirements Survey)
- No DUDL
- Inconsistent policies across applications

→ Aid in data life cycle

- Specific value in “Name = value” pair activates exact policy
- Policies could provide various
 - Retention
 - Replication
 - Striping across servers
 - Deletion

- ➔ Multi-Vendor issues to be considered by application
 - ❑ Design to support multiple vendors
 - ❑ Moving data to different vendor – No feature compromise
- ➔ Use vendor neutral APIs
 - ❑ Separate business logic from REST/SOAP communication
 - ❑ A layer above commands
 - ❑ Features exposed to user independent of vendor
 - ❑ Language independence
 - ❑ Vendor provided SDK

- ➔ Optimizations for accessing cloud storage objects
 - ❑ Sub-directory specific Disaster recovery
 - ❑ Fast reading
- ➔ If available, use flexible view of files to optimize access
 - ❑ Writing files in namespace (file system hierarchy)
 - ❑ Query under specific path
 - ❑ Fast access by reading with object ID

References

1. SNIA Definitions - <http://www.snia.org/education/dictionary>
2. EMC Atmos Programmers Guide - <https://community.emc.com/docs/DOC-3481>
3. 100 year archiving requirements survey
http://www.snia.org/sites/default/files2/100YrATF_Archive-Requirements-Survey_20070619.pdf
4. EMC Atmos <http://www.emc.com/storage/atmos/atmos.htm#!>
5. Cloud Archiving
http://www.snia.org/sites/default/education/tutorials/2010/spring/cloud/PaulField_Cloud_Archiving.pdf

Acknowledgement

We are thankful to

Vinodraj Daniel (vinodraj.daniel@emc.com)
- Engineering Manager, EMC Corporation

for their ideas and suggestions

Questions?

You can email queries to

jim.rice@emc.com or
sreenidhi.iyengar@emc.com