Archive and Preservation in the Cloud - Business Case, Challenges and Best Practices

Chad Thibodeau, Cleversafe, Inc.
Sebastian Zangaro, HP
SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
Abstract

**Cloud Archive Challenges and Best Practices**

- This session will appeal to Storage Vendors, Datacenter Managers, Developers, and those seeking a basic understanding of how best to implement a Cloud Storage Digital Archive and Cloud Storage Digital Preservation service. In addition, we will discuss how these approaches result in a “greener” implementation versus traditional in-house implementations.

- This session will examine current challenges within the Public Cloud Storage Industry, delve into some specific services profiles, and address some best practices for utilizing cloud storage for archive and preservation needs.
Agenda

- What is the problem?
  - Challenges of Private vs. Hybrid vs. Public Cloud Storage

- Backup and Archive and Preservation Defined

- SNIA Cloud Archive and Preservation SIG

- Solution – Services Profiles
Paradoxes of Archive & Preservation

- Data continues to grow
  - Terabytes ➔ Petabytes ➔ Exabytes

- Data will be lost!
  - Dropbox and iCloud breaches

- Migration does not scale

- Access & use models keep changing

- Cost overwhelms everything that complexity does not

- Ever-increasing and changing regulations
Additional Challenges

- Lack of uniform semantics and standard interfaces
- Interoperability between public cloud providers
- Managing data format changes over time
- Authenticity verification
- Compliance and Governance
  - HIPPA
  - Sarbanes Oxley
  - J-SOX
  - SAS 70
- Risk Management & Litigation
Backup (BU): A collection of data stored on (usually removable) non-volatile storage media for purposes of recovery in case the original copy of data is lost or becomes inaccessible.

Disaster Recovery (DR): The recovery of data, access to data and associated processing through a comprehensive process of setting up a redundant site (equipment and work space) with recovery of operational data to continue business operations after a loss of use of all or part of a data center.

Digital Archiving: A storage repository or service used to secure, retain, and protect digital information and data for periods of time less than that of long-term data retention.

Digital Long Term Preservation: [Long Term Retention] Ensuring continued access to, and usability of, digital information and records, especially over long periods of time.

Source: SNIA Dictionary
Cloud Digital Archive Service:

- A cloud-base service providing a specialized online storage repository for the purposes of compliance, litigation support, and/or retention for extended periods of time, not including “long-term.”
- Can be utilized as a component of a complete digital preservation service.
- Does not necessarily provide adequate services to accomplish digital preservation.
Cloud Digital Preservation Service

- A cloud service providing digital preservation of information and data.
- A digital preservation service includes a comprehensive management and curation function that controls:
  - Supporting Infrastructure
  - Information
  - Data
  - Storage Services
Digital Preservation Framework

Record Sources
Applications, Email, User-Apps

Identification & Classification Process

Digital Curation Content Manager
Collect, Index, Filter, Search
Retention Mgmt, Audit, Distribution
Preservation Services Operations

Create Preservation Objects

Compute and Storage Resources

Federated Storage Local & Remote

Cloud

Data & Storage Services

Recipient Policy

Mailbox Store

Policies, SLAs

Web-Based Access

DIP

SIP

AIP

OAIS Terms

Source: www.ltdprm.org
Forms of Deployment

» **Private Cloud**: [Services] Delivery of SaaS, PaaS, IaaS and/or DaaS to a restricted set of customers, usually within a single organization (and under its complete control).

» **Public Cloud**: [Services] Delivery of SaaS, PaaS, IaaS and/or DaaS to a relatively unrestricted set of customers.

» **Hybrid Cloud**: A composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability.
Data Migration Today

Cloud A

Data over WAN via vendor specific API’s

Cloud B

Users

File System
db Records
Disc Blocks
Tape Blocks
Compliance

- Corporate Legal, Records Management and Data Security policies require companies to keep data for long periods of time.
  - HIPAA: Personal Medical Records - Lifetime + 2 years
  - SOX: Audit correspondence - +4 years
  - SEC 17a-4: Trading account records – Account Life +6 years
Cost Drivers

- Operating costs are higher when using in-house (more capacity than data, redundancy, backups, administration costs)

- Cooling equipment consumes about 45% of power delivered to data center

- Storage consumes 13% of total data center power, with 15% for servers
Cloud Storage is not Going Away

Archiving in the Cloud 2009-2015

IDC. Worldwide Storage in the Cloud 2011-2015 Forecast: The expanding role of Public Cloud Storage Services
Benefits...

Cloud-based storage is 74% less expensive than in-house ("File Storage Costs Less in the Cloud Than In-House, Andrew Reichman, Forrester 2011")
Advance the use of public, private and hybrid clouds for archival services and long term retention
- CDMI
- Market Education
- Best Practices
- Services Profiles
- Standards Promotion
- Industry Liaison
- Interoperability Demonstrations/Certifications and Plugfests
- Implementation Reference Model

Participating companies:
Archive vs. Preservation

Digital Archive

- Specially designed system / repository to store digital data
  - Systems management
  - Physical security
  - Data security
  - Data backups
  - Disaster recovery
  - ISO 9001 certification
  - Manifest verification
  - Virus check
  - Format verification
  - Fixity check

Digital Preservation

- Process to ensure long-term data availability
  - Refresh
  - Migration
  - Replication
  - Emulation
  - Metadata Attachment
  - Sustainability
  - Timeless
What is already standardized?

**Benefits of Industry standards:**

- Allows storage vendors and developers to easily integrate with any cloud infrastructure.
- Allows Data Object Migration between heterogeneous systems:
  - End User site to Public Cloud
  - Public Cloud A to Public Cloud B
  - From Public Cloud back to the End User
- Standards already exist such as Self-contained Information Retention Format (SIRF) and CDMI (The Cloud Data Management Interface)

**SNIA’s Cloud Data Management Standard (CDMI)**

- Standardized Data Path (Access) to the Cloud
- Standardized metadata to express the Archive requirement for the Data put in the cloud
- Immutability
How does this work in CDMI?

- Standarizes the access to data in the cloud
- Uses RESTful principles
- Can be implemented on top of the provider’s own interface.

Cloud Client needs to discover what archiving capabilities are provided by the cloud

- CDMI does this though Capabilities – a type of resource that acts like a service catalog for the functions that the cloud offers customers
- If the cloud offers the capability, the customer marks the data objects and containers with metadata (Data System Metadata) that specifies the requirements
- Lastly the Cloud provider has a way of expressing what is actually being provided also through metadata
An Analogy

- **Standard physical archival box**
  - Archivists gather together a group of related items and place them in a physical box container.
  - The box is labeled with information about its content e.g., name and reference number, date, contents description, destroy date.

- **SIRF is the digital equivalent**
  - Logical container for a set of (digital) preservation objects and a catalog.
  - The SIRF catalog contains metadata related to the entire contents of the container as well as to the individual objects.
  - SIRF standardizes the information in the catalog.
SIRF and CDMI

- Cloud Data Management Interface (CDMI) specifies a standard API for clouds.
- CDMI API can be used to access the various preservation objects and the catalog object in a SIRF-compliant container.

Example:
- Assume we have a cloud container named "PatientContainer" that is SIRF-compliant.
  - each encounter is a preservation object
  - each image is a preservation object
  - the container has a catalog object

We can read the various preservation objects and the catalog object via CDMI REST API as follows:

- GET <root URI>/PatientContainer/encounterJan2001
- GET <root URI>/PatientContainer/chestImage
- GET <root URI>/PatientContainer/catalog
Cloud Digital Archive

Storage Services
- Snapshot – type
- Replication – type/class
- DeDuplication – type/class
- Data Integrity

Data & Information Services
- Retention Period
- Permanent Deletion
- Confidentiality/Encryption
- Security – Access, Audit logs
- Physical Migration
- Indexing/Searching
- Litigation Hold
Cloud Digital Preservation

Storage Services
- Snapshot – type
- Replication – type/class
- DeDuplication – type/class
- Data Integrity
- Fixity computation

Data & Information Services
- Retention Period
- Permanent Deletion
- Confidentiality/Encryption
- Security – Access, Audit logs
- Physical & Logical Migration
- Indexing/Searching
- Litigation Hold
- Digital Auditing
- Preservation Objects
- Provenance
Evaluating Tools and Providers

✧ Backup & DR
   ✦ SLA and Performance is key
   ✦ Insist on Proof of Concept
     › Validate in your environment
     › Perform established backup routines in parallel
     › Perform different sizes of restores
     › RTO/RPO objectives

✧ Archiving & Long-Term Preservation
   ✦ Management is key
     › Preservation of file attributes (metadata), ownership
     › File access with multiple search techniques
     › Content management
     › Security and auditing compliance
Digital Archive and Preservation Services are becoming more prevalent and a basic requirement for businesses beyond traditional libraries and content repositories.

Cloud-based digital archives and preservation services can offer significant advantages regarding: ease-of-use, power/cooling, datacenter footprint, security, and high-availability.

Companies can take advantage of “green” cloud technologies for their digital archive and preservation requirements in place of relying solely on their own internal infrastructure – achieving >70% savings.
The SNIA Education Committee would like to thank the following individuals for their contributions to this Tutorial.

Authorship History
Chad Thibodeau
Sebastian Zangaro

Additional Contributors
Bob “Mister” Rogers
Chris Marsh
Michael Peterson
Mark Carlson
Ray Clarke
In Memory of Don Post

Please send any questions or comments regarding this SNIA Tutorial to tracktutorials@snia.org
We need a vision
Private vs. Public Cloud

Private Cloud

- Lower latency
- Power, cooling costs
- Administration costs
- Migration costs
  - Format
  - Storage platform
- Backup
- New technology adoptions
  (e.g. dedup)

Public Cloud

- Higher latency
- Service provider costs
- WAN costs
- Migration costs
  - From one provider to another.
Cloud Peering

Models for Cloud Ecology

Cloud Federation

Computing Cloud

Object Storage Cloud

Multiple Distribution Points

Distribution Cloud

Cloud Peering

Data Usage

Cloud Peering

Object Storage Cloud

Data Usage
Information Governance Reference Model

UNIFIED GOVERNANCE

BUSINESS
Profit

POLICY INTEGRATION

VALUE
Create, Use

DUTY
Hold, Discover
Retain Archive
Dispose

LEGAL
Risk

RIM
Risk

ASSET
Store, Secure

IT
Efficiency

PROCESS TRANSPARENCY
CDMI Reference Model

Figure 5 – Cloud Storage Reference Model