Long-term Data Archiving with Amazon Glacier

Henry Zhang, Senior Product Manager, Amazon Glacier

May 2016
Audio archives – SoundCloud

- World’s leading social sound platform
- Audio files transcoded and stored in multiple formats
- Stores PBs of data
- Transcoded files served from Amazon S3
- Originals moved to Amazon Glacier for long-term retention
- Media distribution backbone (Ve.nue platform)
- OTT broadcast service
- PBs of media assets
- Assets to be archived and retained for decades
Patient data – Philips Healthcare

- HealthSuite digital platform powered by AWS
- 15 petabytes of patient data
- Archived for decades (beyond the lifetime of patients)
- Uses AWS HIPAA eligible services in the BAA
Public sector – King County

- Most populous county in Washington state
- Replace tape solution for backup from 17 agencies
- Meet compliance requirement
- Saved $1MM in first year, no more tape refresh or management churn
Data archiving needs are growing everywhere

- Media assets, 4K, 8K
- Health care / life sciences
- Financial services
- Regulated industries
- Oil and gas / geospatial
- Digital preservation
- Long-term backups
- Logs

Archive:
Data retained for the long term, for compliance or potential future reference
Traditional archiving approaches

- Storage arrays / disk arrays
- Tape silos / tape libraries
- Tape drives (LTO-X / DLT / etc.)
- Virtual tape libraries (VTLs)
- Tape out / vaulting
- Specialized software and personnel
How can AWS help with your archival?

- No capital investment
- No commitment
- No risky capacity planning

Metered usage: Pay as you go

Avoid risks of physical media handling

Control your geographic locality for performance and compliance
Amazon Glacier is a low-cost storage service for archival data with long-term retention requirements.

$0.007/GB per month  3-5 hour data retrieval

Financial records  Medical PACs images  High Res Media Assets
How can Amazon Glacier help with your archival?

Extremely low-cost archive storage service, starting at $0.007 GB/mo

Allows you to retrieve data within 3-5 hours

99.999999999% of durability (7 orders of magnitude higher than 2 copies of tape)

No data migration, **no hardware/infrastructure investments**

Infinite scale and pay for what you use

Access to on-demand compute resource on AWS
Amazon Glacier – key concepts

- **Account** – Access AWS services, view billing/usage, manage security
- **Vaults** – Container for archives, up to 1000 vaults per account
- **Archives** – Files and records, write-once, 40TB max, unlimited archives
- **Inventory** – Cold index of archive properties refreshed every 24 hours
Amazon Glacier – 3 ways to Access

• Direct Glacier API/SDK

• S3 lifecycle integration

• Third party tools and gateways
Amazon Glacier concepts: Uploading data

1. Create vault (films)

2. Configure access policies
   - ArchiveApp user policy
   - Effect: Allow
   - Resource: arn:aws:glacier:<accountId>:vaults/Films
   - Action: glacier:UploadArchive

3. Upload archives
   - UploadArchive(data) -> Archive ID
Amazon Glacier concepts: Retrieving data

1. Initiate Job
   Archiveld: AE99F…
   Vault: Films -> Job ID

2. 3-5 hours for job completion

3. Job completion notification

4. Download output
Amazon Glacier – Third-party tools and gateways

• Consumer grade: less than $50
  • Example: Cloudberry, FastGlacier, Arq (Haystack Software)

• Small / medium business: $500 - $1,000
  • Example: Synology, Veeam, QNap

• Enterprise grade gateway (price varies)
  • Example: NetApp AltaVault
Example: Backup software integration

- CommVault – Native Integration with Amazon S3 & Amazon Glacier
- Deduplication & encryption
- Single console management
Object Storage Options

- **S3 Standard**
  - Active data
  - Milliseconds
  - $0.03/GB/mo

- **S3 Standard - Infrequent Access**
  - Infrequently accessed data
  - Milliseconds
  - $0.0125/GB/mo

- **Amazon Glacier**
  - Archive data
  - 3-5 hours
  - $0.007/GB/mo
Data lifecycle management

- Transition Standard to Standard-IA
- Transition Standard-IA to Amazon Glacier
- Expiration lifecycle policy
- Versioning support

Data access frequency over time
Archiving older videos over time

**Lifecycle Rules**

Step 1: Choose Rule Target
Step 2: Configure Rule
Step 3: Review and Name

**Action on Current Version**

- **Transition to the Standard-Infrequent Access Storage Class**
  - 30 Days after the object's creation date
  - Standard-Infrequent Access has a 30-day minimum retention period and a 128KB minimum object size. Lifecycle policy will not transition objects that are less than 128KB. Refer [here](#) to learn more about Standard-Infrequent Access.

- **Archive to the Glacier Storage Class**
  - 365 Days after the object's creation date
  - This rule could reduce your storage costs. Refer [here](#) to learn more about Glacier pricing. Note that objects archived to the Glacier Storage Class are not immediately accessible.

- **Expire**
  - Days after the object's creation date
  - Versioning is enabled on this bucket and Expiring Current Version will generate new versions. If you wish to permanently delete all versions of your objects, you can combine the Expire current version action here with the Permanently Delete previous versions action below.

**EXAMPLE:**

<table>
<thead>
<tr>
<th>October 1</th>
<th>October 31</th>
<th>September 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022</td>
<td>2022</td>
<td>2022</td>
</tr>
</tbody>
</table>

[Feedback] [English] © 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use
Save money on storage

- 1 PB of storage and growing
  - 1 PB for S3 Standard = $31,457 per month for storage cost
  - Videos older than 30 days are watched less than once a month on average

- 1 PB for S3 Standard-IA = $13,107 per month for storage cost, **58% saving over S3 Standard**

- 1 PB archived in Amazon Glacier = $7,340 per month for storage cost, **44% saving over S3 Standard-IA**

* Assumes the highest public pricing tier
Amazon Glacier Vault Lock allows you to easily set compliance controls on individual vaults and enforce them via a lockable policy.

- Time-based retention
- MFA authentication
- Controls govern all records in a Vault
- Immutable policy
- Two-step locking
Vault Lock for compliance storage

- Non-overwrite, non-erasable records
- Time-based retention with “ArchiveAgeInDays” control
- Policy lockdown (strong governance)
- Legal hold with vault-level tags
- Configure optional designated third-party access and grant temporary access
Amazon Glacier received a third-party assessment from Cohasset Associates on how Amazon Glacier with Vault Lock can be used to meet the requirements of SEC Rule 17a-4(f) and CFTC 1.31(b)-(c).
Example control: 1 year record retention

- Deny delete archive operation
- From anybody (root, administrators, users, business partners)
- When ArchiveAgeInDays is \( \leq 365 \) days

Archive age computed from the time an archive lands in a vault
Example control: 1 year record retention

![Vault Lock policy for BusinessCritical](image-url)
Vault Lock: Two-step locking

- **InitiateVaultLock**
  - Effectuates a retention policy for testing (in-progress state)
  - Returns a unique lock ID (expires after 24 hours)

- **AbortVaultLock**
  - Deletes an in-progress policy
  - Ability to modify a policy before locking it down

- **CompleteVaultLock**
  - Locks down the vault with the appropriate lock ID
  - Vault Lock cannot be aborted afterwards
AWS storage options

- Amazon EFS: File
- Amazon EBS: Block
- Amazon EC2 Instance Store
- Amazon S3: Object
- Amazon Glacier

Data Transfer

- AWS Direct Connect
- AWS Snowball
- ISV Connectors
- Amazon Kinesis Firehose
- Amazon S3 Transfer Acceleration
- AWS Storage Gateway
Transfer Petabyte scale datasets with Snowball

- Ruggedized case
  “8.5G Impact”
- E-ink shipping label
- 80 TB
  10G network
- Rain & dust resistant
- Tamper-resistant case & electronics
- All data encrypted end-to-end
### What does it cost?

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage Charge per Job</td>
<td>$250.00</td>
</tr>
<tr>
<td>Extra Day Charge (First 10 days* are free)</td>
<td>$15.00</td>
</tr>
<tr>
<td>Data Transfer In</td>
<td>$0.00/GB</td>
</tr>
<tr>
<td>Data Transfer Out</td>
<td>$0.02/GB</td>
</tr>
<tr>
<td>Shipping**</td>
<td>Varies</td>
</tr>
<tr>
<td>Amazon S3 Charges</td>
<td>Standard storage and request fees apply</td>
</tr>
</tbody>
</table>

* Starts one day after the appliance is delivered to you. The first day the appliance is received at your site and the last day the appliance is shipped out are also free and not included in the 10-day free usage time.

** Shipping charges are based on your shipment destination and the shipping option (e.g., overnight, 2-day) you choose.

Transfer 1 PB with 13 devices in parallel in 1 week!