Texas Conference on Digital Libraries

Archiving in the Cloud: Tackling Security, Scale and Savings

Alex Corley
acorley@amazon.com
Senior Solutions Architect
What is archival?: Backup vs. archive

Backup:
Data retained to support near-term business continuity

Archive:
Data retained for the long term – for compliance or potential future reference
Why cloud?

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

- Metered usage: Pay as you go

- Avoid opex and risks of physical media handling

- Control your geographic locality for performance and compliance
Perspectives on scale: storage

Year Over Year View of Total Number of Objects Stored in Amazon Simple Storage Service (S3)

Peak requests: 1.5M+/sec

---|---|---|---|---|---|---|---
2.9 Billion | 14 Billion | 40 Billion | 102 Billion | 262 Billion | 762 Billion | >1.7 Trillion | >3 Trillion
We’ve announced price reductions 47 times since our inception in 2006. Recent price drops included:

- **34%**
  - Amazon ElasticCache reduces prices for cache nodes by an average of 34%
  - March 26, 2014

- **51%**
  - Amazon S3 reduces prices for Standard and Reduced Redundancy Storage, by an average of 51%
  - March 26, 2014

- **20%**
  - Amazon Route 53 lowers prices for both Standard Queries and Latency Based Routing Queries by 20%
  - July 31, 2014
AWS Global Infrastructure

11 Regions
28 Availability Zones
52 Edge locations

You choose where your apps and data go!

AWS Worldwide Public Sector
Architected for Enterprise Security Requirements

Certifications and accreditations for workloads that matter

AWS CloudTrail - AWS API call logging for governance & compliance

Log and review user activity
Stores data in S3, or archive to Glacier
Used by educational institutions worldwide
TCO – On-premises cost considerations
Scale

• **Object Store!**

• **Interact directly with a service instead of an intermediate layer**
  – Storing data in Glacier, but have a product that interfaces with S3
  – Storing data in block storage and backing by Object Store without any added caching or deduplication

• **Leverage a Storage Gateway**
  – AWS Storage Gateway, NetApp SteelStore, CTERA, panzura
Integrating AWS with your existing on-premises infrastructure

- Active Directory Shibboleth / SAML
- Network Configuration
- Encryption
- Backup Appliances

---

- Users & Access Rules (IAM)
- Your Private Network (VPC)
- Encryption (S3, RDS, HSM)
- Backups (Storage Gateway)

---

Your On-Premises Apps

---

Your Cloud Apps

---

AWS Direct Connect

---

Legacy Data Centers

---

AWS Worldwide Public Sector
Upload and Storage Best Practices

- Checksum 1
  - File 1
- Checksum 2
  - File 2
- Checksum 3
  - ...

Aggregate Index
- File 1
- File 2
- File 3
Maintaining your own index

• Reduce cost and improve download speed with Ranged Gets

• Many options: DynamoDB, RDS:Oracle, RDS:MySQL, local DB, BerkeleyDB, S3
Download API Best Practices

Reduce costs with Ranged Retrievals

Example 1.2 GB archive
- Retrieved using a standard 4 hour job = 300 MB/hour peak retrieval charge
- Retrieved over 24 hours using 6 consecutive jobs = 50 MB/hour