Backup and Archiving to AWS S3

Introduction

AWS S3 (including S3 Glacier) can provide very cheap object storage for **some on-premise backup and archiving use cases**. But, be careful. There are pitfalls:

- S3 is an object store, not a file system. You will need to make sure that the tools you use to accomplish backup/archive are S3-savvy. Since S3 is not a file system, some features you might expect are missing or can be costly to replicate.
- S3 storage can be very cheap indeed, but you need to be careful that the tools you are using don't end up costing you a lot for S3 API operations used for checking object hashes and collecting other metadata from objects.
  - E.g., it costs money just to get the MD5 hash or creation date for an S3 object. It's not much money, but it can add up when dealing with hundreds of thousands or millions of objects.
- Remember AWS S3 isn't the only cloud storage available. Azure, Google, and Wasabi are other options.

> If you aren't sure that AWS S3 is right for you, this Cornell tool can show other alternatives: https://finder.research.cornell.edu/storage

Framing Your Use Case

Here are some questions that may be valuable to answer when thinking about backup and archiving:

- What do you want to restore from your backup or archive? Specific files? All files as of a specific date?
- How fast do you need it? I.e., what is your Recovery Time Objective (RTO)?
- What is your Recovery Point Objective (RPO)? I.e. how far out-of-date can objects be when restored?
  - E.g., A server RPO may be 24 hours, meaning that its OK to have restored files be as much as 24 hours old, but no older.
- Should different versions of backed-up/archive objects be kept? Or, do you always want the latest version?
- How often do you envision having to restore data?
  - Some service pricing is fairly expensive when you actually need to restore data, especially in short timeframes.
- What are the basic metrics of the data in scope for your use case?
  - Total cumulative size of target data?
  - Total number of target files/objects?
  - Total number of target files/objects < 128KB?
    - Some services handle smaller objects differently than larger objects
  - Estimated number of target files/objects deleted before 90 days
    - Some services require a minimum object lifetime and will charge you for storing the object for the entirety of that period, even if the object is deleted before reaching that age.

General Approaches

There are a ***lot*** of pathways to get on-premise files to S3 or other AWS services. Picking the right one will depend on your use case, budget, ability or desire to tinker and monitor costs, and palatability of deploying additional on-premises resources.

CLI Tools

- AWS CLI contains a basic sync command.
- rclone is a CLI tool in the same vein as rsync, but it is savvy about cloud object stores like S3.
- s3cmd is a another third-party option

True Backup Software Using S3 for Storage

Many backup software solutions can use S3 for backend storage. An example of this for smaller-scale deployments is MSP360 Managed Backup (formerly CloudBerry Backup).

AWS Tools and Services

AWS has a lot of tools and service options to make it easier to move/sync data from on-premise sources to AWS. These services can be a great solution if they do exactly what you need and you have the budget for them. Tools in this category include:

- DataSync
- Storage Gateway

Anti-Patterns

Data Already in AWS

Don't roll your own backup/archive solution if your data is already in AWS. Use built-in AWS services and features:

- AWS Backup
- EBS Snapshots
- RDS Snapshots
- S3 Replication
- ...

Resources

Internal

- SFTP and Cloud Object Store Clients - OS X and Windows
- S3 Sync Analysis – Analyzes S3 API-costs of various rclone settings

External

- Backup versus Archive
  - https://www.computerweekly.com/news/1369092/Data-backup-vs-archiving-Whats-the-difference
  - https://www.backblaze.com/blog/data-backup-vs-archive/
  - https://www.ironmountain.com/blogs/2020/5-major-differences-between-backup-vs-archive
- AWS Pricing Calculator
- rclone
- s3cmd
- S3P - Massively Parallel S3 Copying
- https://wasabi.com/ – Provides S3-compatible storage
  - Cornell has a contract with Wasabi and we expect to roll out Cornell unit access to Wasabi under that contract in the future (no specific timeline set as of 18 Oct 2021).