Northwestern
INFORMATION TECHNOLOGY

Archival Data Storage in Cloud

Tobin Magle

02-25-26

Goals

Learn why you want to archive your inactive data in the cloud.

Understand what drives archive costs and how to keep them low

How to prepare data for long-term storage

Transfer data into cloud archive

Know how to restore and retrieve archived data

Scenario
Your lab just published a paper

You found inactive data. Now what?

What is inactive data?

- Not used day-to-day
- Too important to delete
- Taking up space in active storage

Goal:

Move inactive data out of your working space to reduce clutter and keep it safe.

Why Archive?

Archiving inactive data isn't just about saving space it's about protecting your research and reducing costs.

Free up space and **k**eep your active data

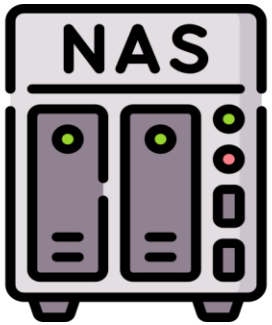
Cut costs - by moving inactive data to cheaper storage

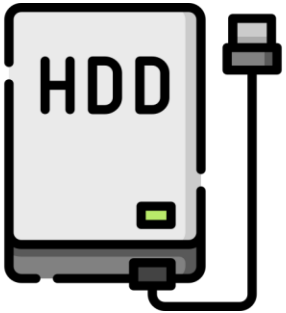
Protect data from loss on local devices

Archiving = Long-Term, Low-touch, Retrievable Storage

Common Archiving Options

- **External Hard Drives**
  - Simple and inexpensive
  - Easy to lose track of, can fail over time
- **NAS / Local Servers**
  - Centralized, familiar, fast
  - Require ongoing maintenance and hardware refreshes
- **SharePoint / Institutional Platforms**
  - Good for documents, small datasets, and team collaboration
  - Not suited for very large or long-term scientific datasets
- **Cloud Storage / Cloud Archive**
  - Durable, scalable, designed for long-term retention
  - Retrieval may take time, and costs depend on size / frequency of access

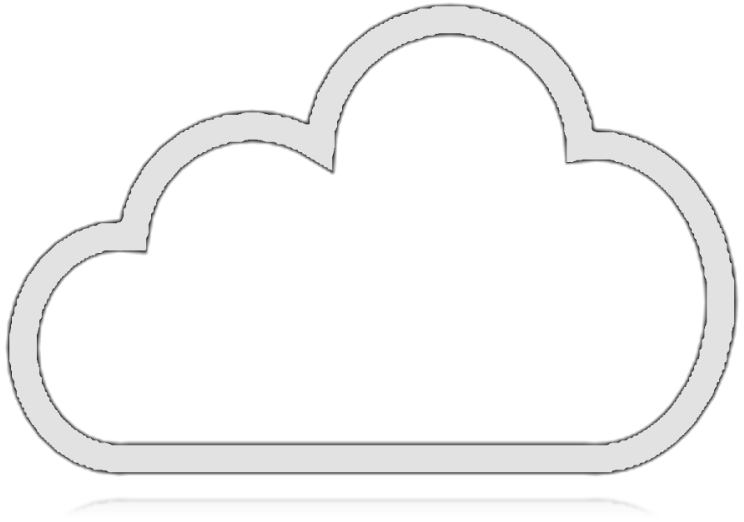




What is cloud?

Using the cloud is like renting IT hardware in someone else's data center.

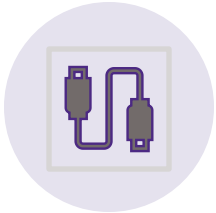
- On-demand computing and data storage resources
- Resources are flexible & scalable
- Pay for what you use



Cloud Lets you focus on your research while someone else manages the Infrastructure

Why Archive In The Cloud?

NO HARDWARE TO MANAGE:



MORE DURABLE AND RECOVERABLE

EASIER TO MANAGE OVER TIME

DESIGNED FOR INFREQUENT ACCESS

Amazon Web Services

Cloud Storage Tiers

Standard/Hot Storage

- For active, frequently accessed files
- Fast access, higher cost
- Best for daily research work

Deep Glacier Archive/ Cold Storage

- For long-term, rarely accessed data
- Very low-cost storage, slower retrieval
- Ideal for completed projects or compliance retention

Storage Tier	**Usage**	Cost/ TB/ month	Min days
Hot	Frequently accessed data	~$20	None
Archive	Not expected to access	~$1	~180 Days

Understanding Minimum Storage Duration

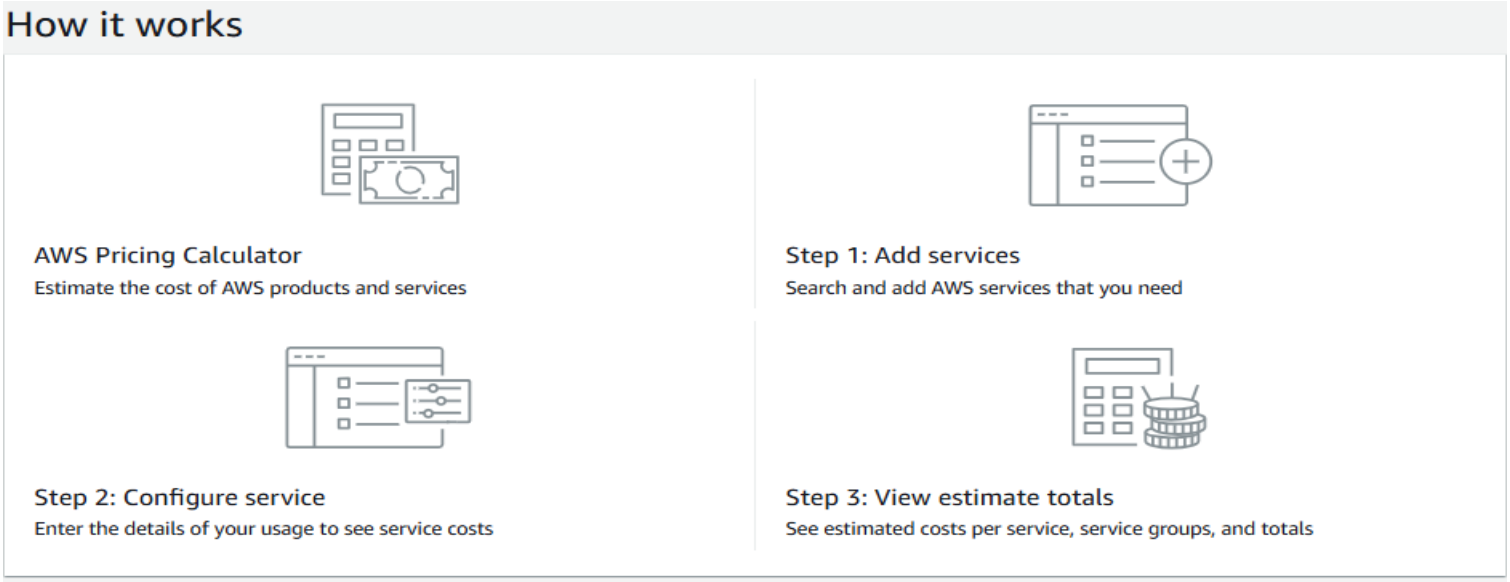
- Deep Archive requires data to remain stored for 180 days
- Moving or deleting data before 180 days will trigger the fee
- Fees are calculated as if the data stayed the full minimum duration

Steps Needed To Archive

- **Estimating Costs**: one-time and monthly
- **Preparing Data for Archival:** Document and package your data so it is usable later and more cost efficient
- **Archiving Data:** Move the data into Archive
- **Retrieving Data:** restore and access archived data later

AWS Cost Calculator

let's you estimate what it will cost to store data in S3 Glacier Deep Archive over time, based on how much data you plan to keep and how often you expect to retrieve it



Archive Costs Categories

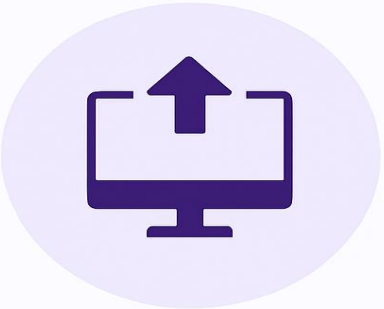
Three Types of Long-Term Data Archiving Costs

Archiving (One-Time)

Cost to transfer data into long-term storage

Storage (Ongoing)

Cost to keep data stored over time



Retrieval (As Needed)

Cost to access or restore archived data

What Drives Data Transfer Cost

- Uploading files into any S3 storage class costs $5.00 per 1,000,000 files.
- Transitioning files between storage classes also costs $5.00 per 1,000,000 files.

What Drives Ongoing Storage Cost

Inputs for Cost Estimation

Data Size

Determines storage cost; larger datasets cost more to keep long-term.

File Count

Affects request/metadata charges and transfer performance.

Storage (Ongoing)

Cost to keep data stored over time

Storing the data costs ~$12/TB/year
Storing metadata costs ~$2.28 per 1,000,000 files

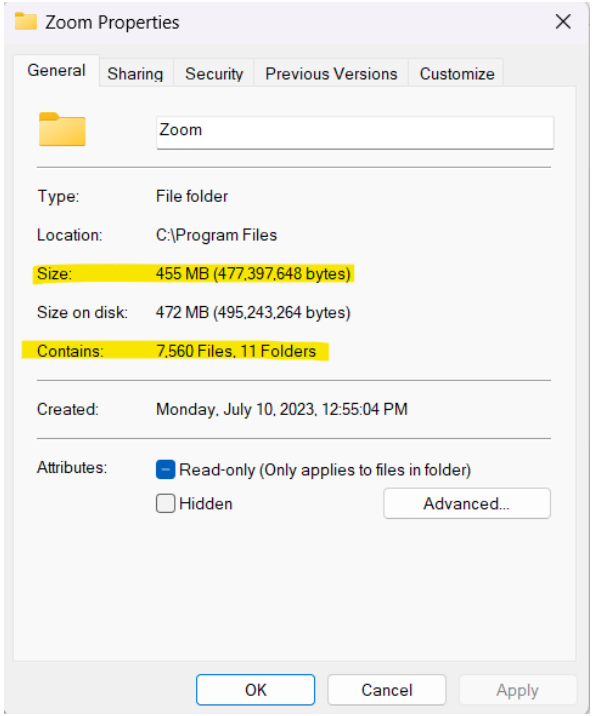
Windows | Size and File count

1. Open Windows **File Explorer** and navigate to the folder with your data.

2. **Right click** on the folder you want information about. Select **Properties**.

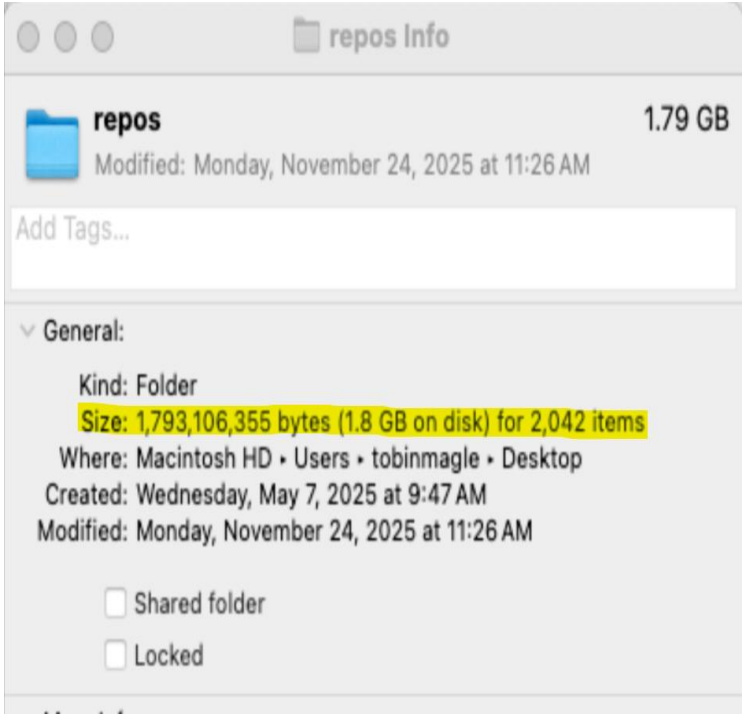
3. The line `Size:` will show you the amount of space your data uses.

4. The line `Contains:` will show you how many files there are.



macOS | Size and File count

1. Open a **Finder** window and navigate to the folder with your data.
2. **Control click** on the folder and select **Get Info**.
3. Check `Size:` for the total (bytes plus KB/MB/GB in parentheses), and use the number shown before “items” as the file count.



Estimate Average File Size

Total dataset size: **500 GB**
File count: **1,000,000**

Step 1: Convert GB to MB *(assume 1 GB = 1000 MB)*
500 GB × 1000 = **500,000 MB**
Step 2: Divide by file count
500,000 MB ÷ 1,000,000 = **0.5 MB per file**

Average file size ≈ 0.5 MB

Demo – Cost Calculator

What the Cost Calculator Shows

- Total data size stays the same, but
- file count can dramatically change your upload and storage cost
- Even if you aren’t archiving a large dataset,
- Preparing your data — bundling or compressing —
- helps reduce file count and saves money

Preparing Data for Archive

Archiving best practices

01

Only archive what you need: Not everything that you produce during the course of the research project needs to be archived.

02

Bundle small files together: To reduce the per file costs for archiving, bundle files together in a .tar file to reduce the total number of files.

03

Compress your data: To reduce the cost of storing and retrieving data, compress your files file to reduce the size of your data.

Tar VS Compress

Taring

Purpose: Combine many files/folders into a single package

What you get: One archive file

Size impact: Usually **does not** significantly reduce size

Compressing

Purpose: Shrink data to use less storage and transfer faster

What you get: A smaller file

Size impact: Reduces size (often a lot, depending on file types)

Time: Compressing takes a long time

Realities of Tarring & Compressing Data

Compression Takes Time

- Compression can take hours for large datasets
- Requires a machine that can remain awake and stable
- Best for when reducing size saves meaningful long-term storage cost

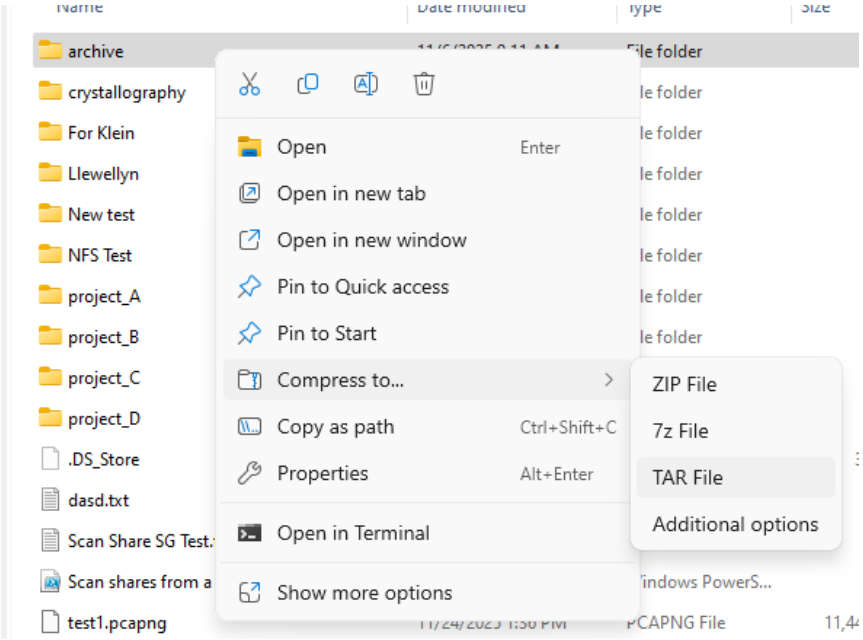
Tarring/Compressing Creates a Single File

- Since when you Tar or Compression makes a single file it's better to do them in logical chunks that would make transfer and verifying easier.

Windows | Taring

Windows

- Open a **File Explorer Window** and navigate to your files.
- Select the files/folders by **right-clicking** on the file(s) or folder(s) you want to archive.
- From the context menu, select **Compress to** and then choose **tar file.**
- Provide a name for your new `.tar` file and press `Enter`.



Mac | Taring

Mac

To create a `.tar` file on a Mac, in the **Terminal** program, navigate to the directory containing the folder(s) or files you want to compress.

Then run :

```
tar -cvf archive.tar archive_folder
```

The `-c` option creates the archive.

The `-f` option allows you to name the resulting tar file, in this case `archive.tar`.

The `-v` option will list the files it is adding to the terminal output.

`archive_folder`: Folders you want to compress. You can list multiple folder or file names separated by spaces.

You can use the `-z` option to compress and archive at the same time using `gzip`.

Compressing

Windows

- Locate the file or folder that you want to zip.
- Right-click the file or folder, select Send to, and then select Compressed (zipped) folder.
- OR
- Right-click the file or folder select Compressed to, and then select ZIP

Mac

- Locate the file or folder that you want to zip.
- Control-click it, then choose Compress from the shortcut menu.

Demo – Tar / Compress

Transferring Data to S3

Methods to Transfer to cloud

The AWS CLI (Command Line Interface) that allows you to interact with AWS services.

Globus is a secure file transfer and sharing platform designed to move large datasets reliably between computers, institutional storage systems, and cloud storage.

AWS CLI

```
aws s3 cp /path/to/folder/ s3://my-archive-bucket/your-folder/ --recursive --storage-class DEEP_ARCHIVE
```

Replace the following:

- /path/to/folder: Full path to your local folder
- my-archive-bucket: Your target S3 bucket's name
- your-folder/: (Optional) Destination folder or key prefix in the bucket

The flags included in the command do the following:

- recursive: transfers everything in the specified folder
- storage-class DEEP_ARCHIVE: flag ensures all files are immediately placed into the S3 Glacier Deep Archive class.

Demo-CLI upload

Verifying upload

- Once the upload is complete.
- Be sure to check that if the data is upload correctly and check if there was any error in the command

Scenario:
Months later, you need to pull archived data back into action.

What Retrieval Means

- Deep Archive data **cannot** be accessed instantly
- You must submit **a retrieval request**
- AWS restores a **temporary Standard-tier copy**
- You can download the data **after the restore completes**
- Retrieval can take **hours**, depending on speed
- You pay for r**etrieval speed + the time the data stays in Standard tier**

What Drives Retrieval Costs

- **Retrieval fee (based on speed)**: Based on how fast you want to be able to touch your data.
- **Standard-tier duration cost**: After a restore completes, you pay Standard storage rates for the amount of time the restored copy remains accessible.

Retrieval Speeds and Fee

- For Deep Glacier Archive there are 2 speeds
  - Standard
  - Bulk

Retrieval from Deep Glacier Archive	Cost (per GB)	Time	Cost per 1,000 Files
Standard (Balanced Price and Speed)	$0.01	Within 12 Hours	$0.10
Bulk (Cheap and Slow)	$0.0025	Within 48 Hours	$0.025

Retrieval Duration

- Restored Deep Archive data stays in the **Standard tier for a limited time**
- Data is automatically removed from the Standard tier after the window expires
- Extend the window if you need more time to download or process the data
- The longer the window, the more Standard-tier storage cost you pay

Duration	Cost for 1tb restored
7 Days	$5.60
14 Days	$11.20

Retrieval Demo

Key Takeaways

You now know when and why to archive inactive data

- freeing space, reducing cost, and improving data management.

You understand what contributes to archive costs

- strategies to keep them low, from file counts to retrieval choices.

You learned how to prepare data for long-term storage

- including organizing, bundling, and compressing to reduce cost and complexity.

You can transfer data into cloud archive storage

- using tools like AWS CLI or Globus.

You know how to retrieve archived data

- what to expect during the restore process, including timing and cost considerations.

Documentation on Archive

- RCDS Archiving Documentation

Reach out!

Visit the: Research Data Management Website
Email: researchdata@northwestern.edu
RCDS Consult Form
RCDS Cloud Consult form
Galter Data Lab Consult form
Information Security: Protect your research
Office Hours: Every Monday 3 p.m. – 4 p.m. Mudd Library Rooms 2202-2205 (2nd Floor across the bridge to Tech)