Introduction to Cloud Storage

Kevin Luu

Virtualization and Cloud Specialist

Kevin.luu@northwestern.edu

Research Computing & Data Services

Northwestern

INFORMATION TECHNOLOGY

After this session, you will be able to...

- Learn the advantages of getting a Northwestern cloud account and how to request one.

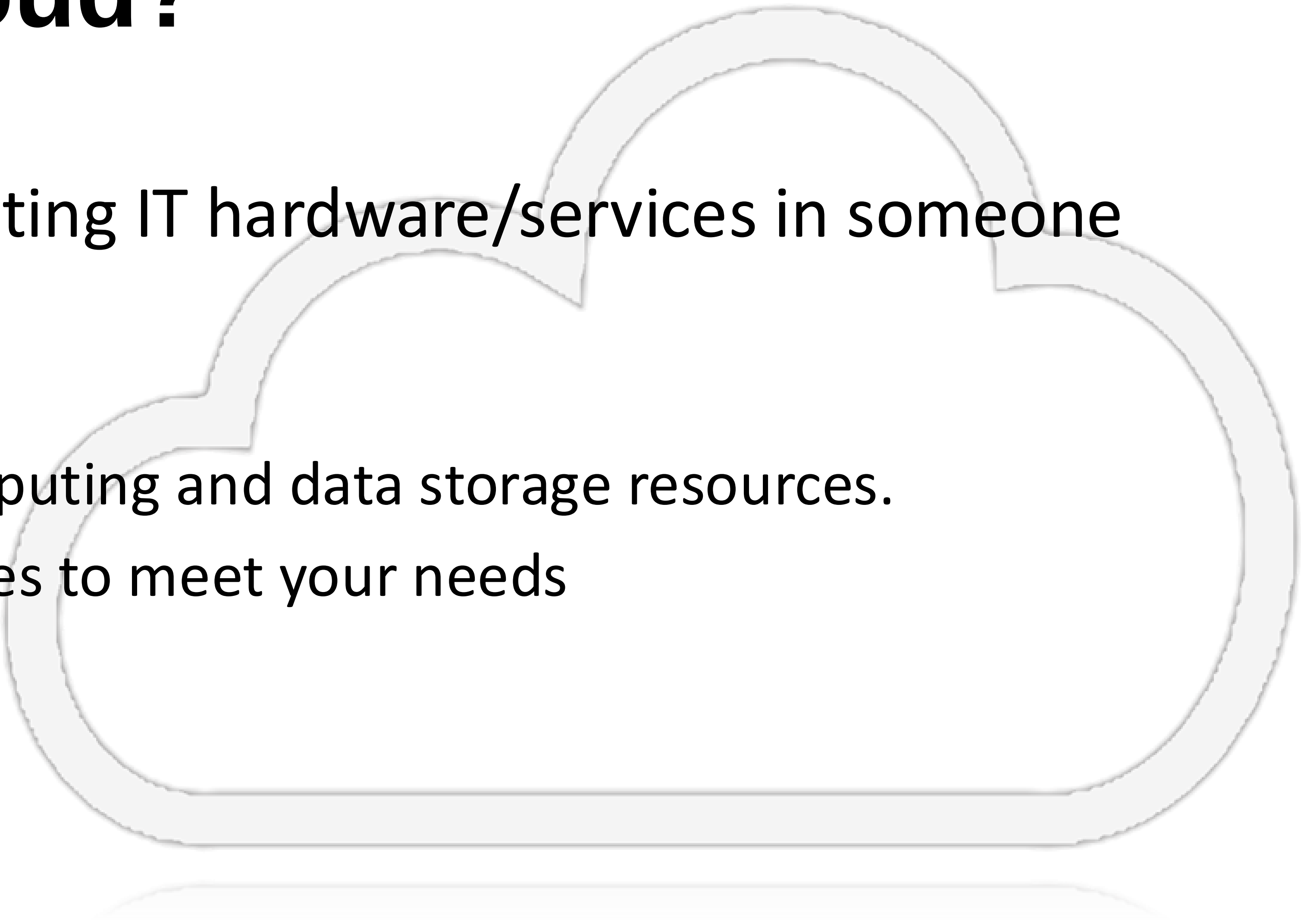
- Understand what **cloud object storage** is and why you would use it

- Learn how to
  - Create a bucket for storing your data.
  - How to choose the correct storage class for your data.
  - Configure your bucket's lifecycle policy.
  - Set up data replication

What is the cloud?

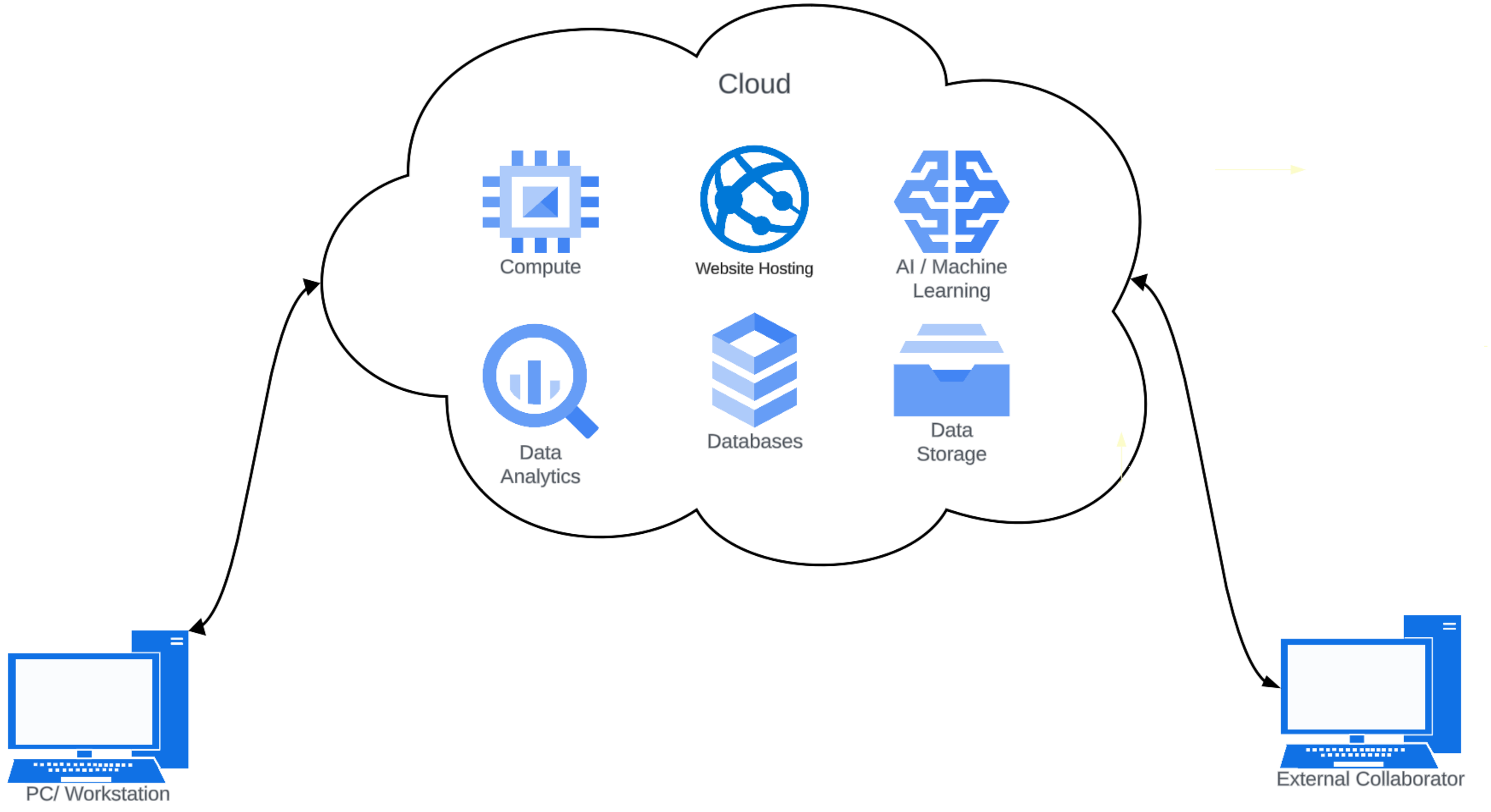
Using the cloud is like renting IT hardware/services in someone else's data center.

- **On-demand** access to computing and data storage resources.
- **Flexible & scalable** resources to meet your needs
- **Pay for what you use**



How do you access the cloud?

Access a wide range of IT services such as computing and storage over the Internet from any device.



Does this sound familiar?

- Is the cost of your current data storage too high?
- Do you feel like your storage solution is no longer up to standard.
- Do you have large amount of outdated data that needs to be archived?

Cloud Storage could be the solution.

Why Cloud Storage

- Cloud Storage, when properly configured, can result in lower costs.
- Data is encrypted and has ease of access for approved users.
- You can streamline data archiving and enhanced data durability.

When to consider using Cloud?

Easy Access Anywhere: Cloud storage enables researchers and collaborators to access files from any device with an internet connection.

Scalable: Cloud flexibility allows you to scale computing and storage resources up or down as needed, ideal for varying demand.

Guided Support: We offer support to help you navigate and understand cloud options, advising you on whether to use cloud services or rely on existing university resources.

Cloud Providers

aws

Largest of the three platform

Extensive global infrastructure

Azure

Intergration of Microsoft's ecosystems like Office 365

Access to OpenAI Models

Google Cloud

Considered to be a leader in data analytics tools and services

Integrates with Google's ecosystem (Google Drive, Collab)

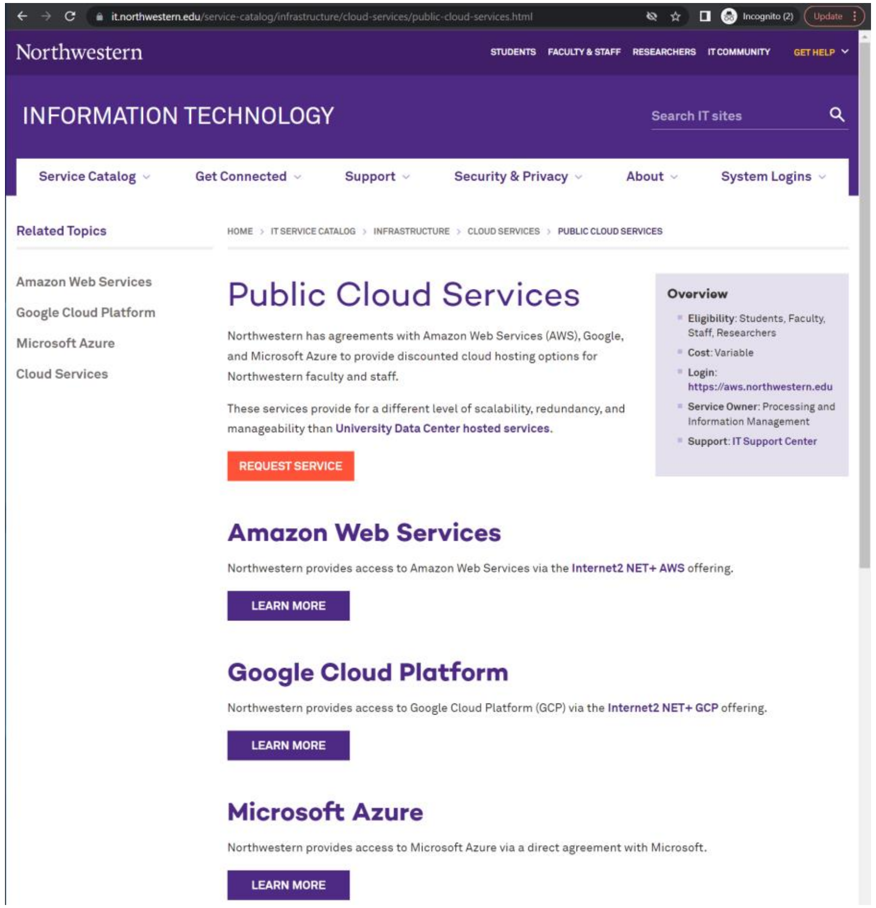
Each cloud provider generally offers similar services, such as virtual machines and storage. At Northwestern, AWS is the most popular, followed by Azure and Google Cloud

Where to Start at Northwestern

Public Cloud Services

- Northwestern has agreements with AWS, GCP, Azure to provide benefits for cloud hosting options
- Available to NU faculty, researchers and staff
- You will need a Blanket PO

Request Public Cloud Account



Public Cloud Services

Northwestern has agreements with Amazon Web Services (AWS), Google, and Microsoft Azure to provide discounted cloud hosting options for Northwestern faculty and staff.

These services provide for a different level of scalability, redundancy, and manageability than University Data Center hosted services.

REQUEST SERVICE

Overview

- Eligibility: Students, Faculty, Staff, Researchers
- Cost: Variable
- Login: https://aws.northwestern.edu
- Service Owner: Processing and Information Management
- Support: IT Support Center

Amazon Web Services

Northwestern provides access to Amazon Web Services via the Internet2 NET+ AWS offering.

LEARN MORE

Google Cloud Platform

Northwestern provides access to Google Cloud Platform (GCP) via the Internet2 NET+ GCP offering.

LEARN MORE

Microsoft Azure

Northwestern provides access to Microsoft Azure via a direct agreement with Microsoft.

LEARN MORE

Northwestern Cloud Benefits

Amazon Web Services (AWS)	Google Cloud Platform (GCP)	Azure
• Direct billing to chart strings in NU Financials • A waiver of data egress fees for most applications • Single sign-on with Northwestern credentials • Discount of 5%-10% for services • HIPAA-compliant environments		

Each Northwestern cloud account is provisioned with account monitoring and a financial tracking tool Kion, which can help track spending and budgets.

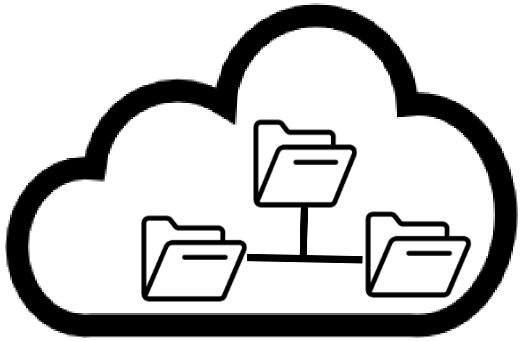
STRIDES

- Project that are funded by a NIH Grant, are eligible for additional benefits through STRIDES Link

Amazon Web Services (AWS)	Google Cloud Platform (GCP)
• 9% discount on all AWS services except S3 and Marketplace • 14% discount on AWS S3 storage services (except Glacier Deep Archive) • Free access to enterprise-level AWS support (Documentation)	• 25% off compute and standard storage • 10% off coldline and archive storage • 19% off managed services such as BigTable and BigQuery

Types of Cloud Storage

File

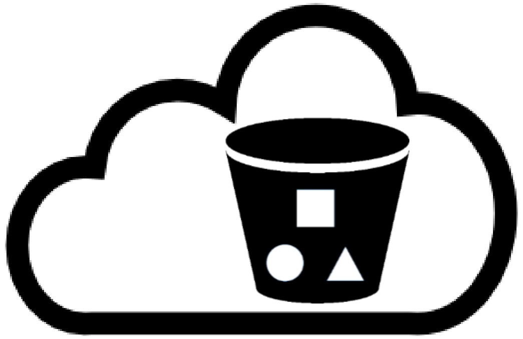


- File Server where multiple device / machines can touch the data at the same time
- Ex. Elastic File Storage

Block

- High Speed storage
- Designed to be attach to a server.
  Ex. Elastic Block Storage

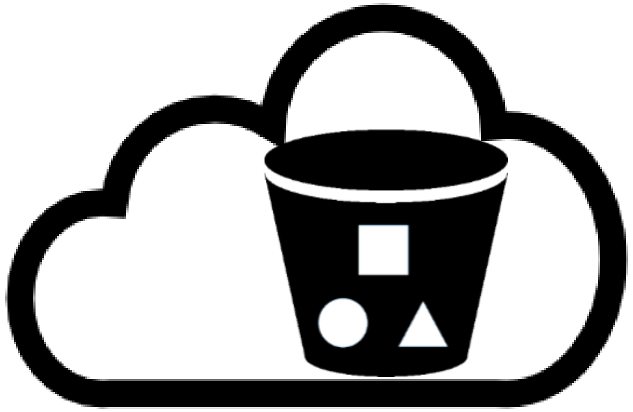
Object



- Large amounts of unstructured data
- Stores your data in buckets
- multiple tier of storage class
- Ex. Simple Storage Service

Object Storage

- Inexpensive
  - Pay only the storage you use
- Scalable
  - Theoretically unlimited
- Accessible
  - Accessible via web portal, Globus, and CLI.
  - Enables efficient data sharing.
- Long Term Storage
  - Cost-effective for long-term data storage

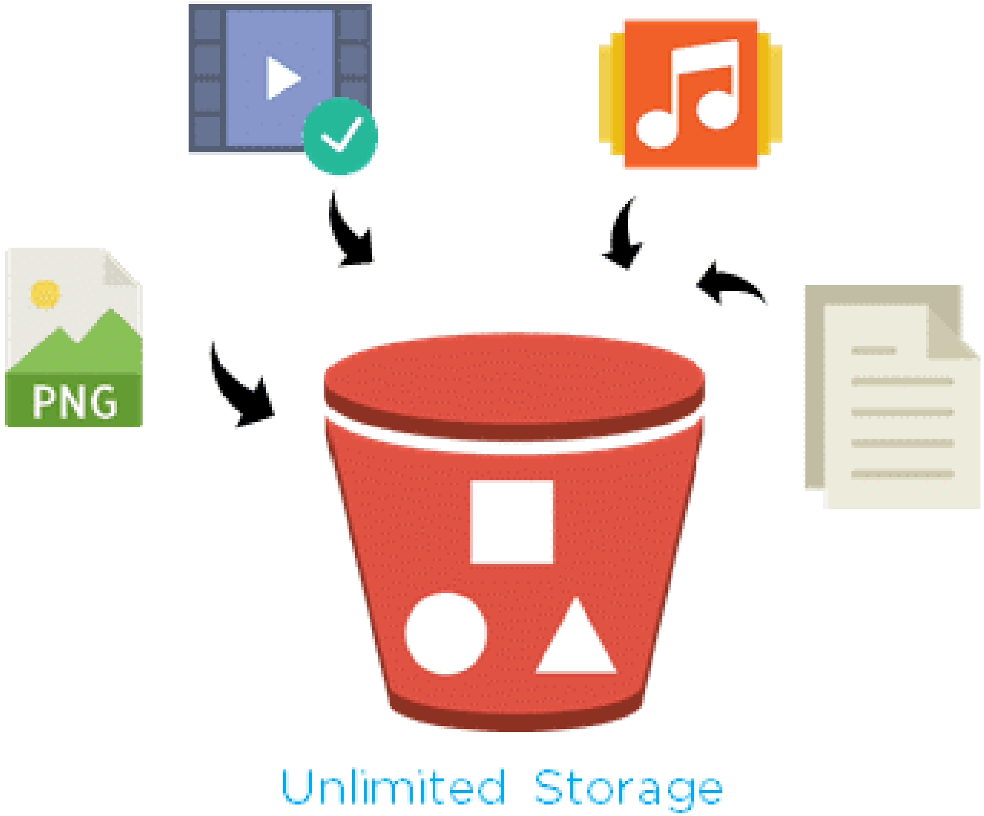


In object storage, the data you upload is referred to as an 'object.' Each object includes the data itself, associated metadata, and a unique identifier

AWS - Simple Storage Service (S3)

What makes a bucket?

- **Bucket -** A bucket is a container that stores objects, similar to how folders store files.
- **Object** - Any type of data + Tag
- **Tag** – A tag is an optional identifier used to organize data within a bucket.



Any File that has an extension that can be opened from a file explorer can be stored in an Object Storage

Understanding Tagging in Object Storage

- What is Tagging?
  - Tagging involves attaching labels to objects to provide additional metadata. These labels can describe the content, status, or any other characteristic of the data
- Why is Tagging Useful?
  - Enhances the searchability of data within cloud storage, allowing users to find specific items quickly based on tag criteria.

Think of any time where you can't find a file but you are sure it exists and can't find it. With Tags you can search by tag results.

General Storage Tier/Class

Storage Tier	Usage	Cost/TB/month	Min days
Hot	Frequently accessed data (multiple times a day)	~$20	none
Cool	Infrequently accessed data (monthly), needs to be fast when accessed	~$10	~30
Cold	Accessed 1-2 times a year, slower to access	~$4	~90
Archive	Not expected to access, data retained for compliance, slowest retrieval	~$1	180-365

Storage Pricing at a glance

Storage Class	Per Gb/ Month	Per TB/ Month	Per TB/ year	Per 5TB/ year
S3 Standard (Hot)	$0.023	$23	$276	$1380
S3 Glacier deep archive	$0.00099	$1	$12	$60

Data Retrieval Fee and time

- Changing the data's storage tiers from cold to hot incurs varying retrieval fees, depending on the tier and retrieval speed.
- There are 3 Types of Retrieval

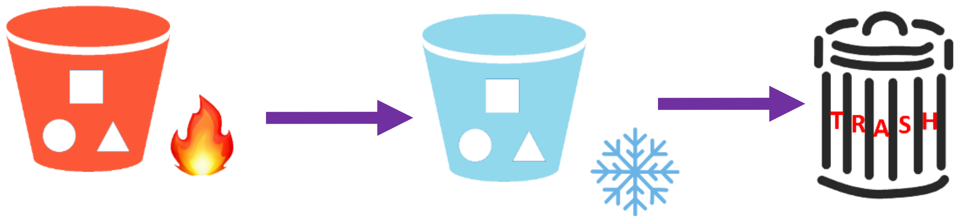
Retrieval from Deep Glacier Archive	Cost (per GB)	Time
Expedited (Expensive And Fastest)	Not available for DGA	N/A
Standard (Balanced Price and Speed)	$0.01	Within 12 Hours
Bulk (Free/Cheap and Very Slow)	$0.0025	Within 48 Hours

Data Retrieval is the process of making a file available to copy or download

Data Lifecycle & Intelligent Tiering

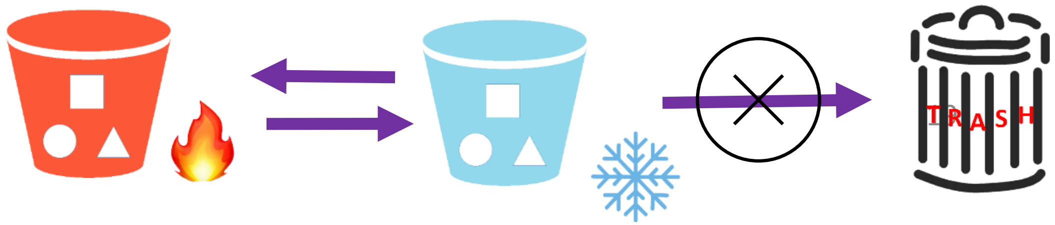
Data Lifecycle Rule Policy

- A rule that you configure that automates the transition and expiration of objects in an S3 bucket based on lapsed time.
- Retrieval fees and times are still applicable



Intelligent Tiering Storage Class

- A Storage class that will automatically move files between different storage class based on your access patterns
- No Retrieval Fees and retrieval times are immediate.



Data Versioning

- Creates multiple versions of the same object.
- Prevents accidental deletions by retaining previous versions.
- Easy retrieval and restoration of previous versions of objects.
- Providing a complete history of changes.
- Versioning can be turned on or off
- Versioning costs may quickly add up

Keeping your Data Secure

Encryption In Transit

- All communication via the web console or the command line interface, take place over encrypted HTTPS connections.

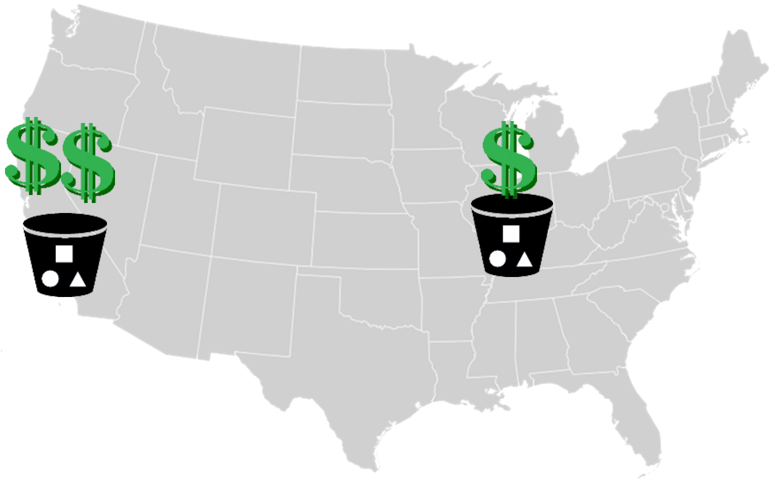
Encryption At Rest

- All S3 buckets have default encryption
- All S3 buckets have a policy attached disallowing non-SSL(Secure Sockets Layer) connections

Location Considerations

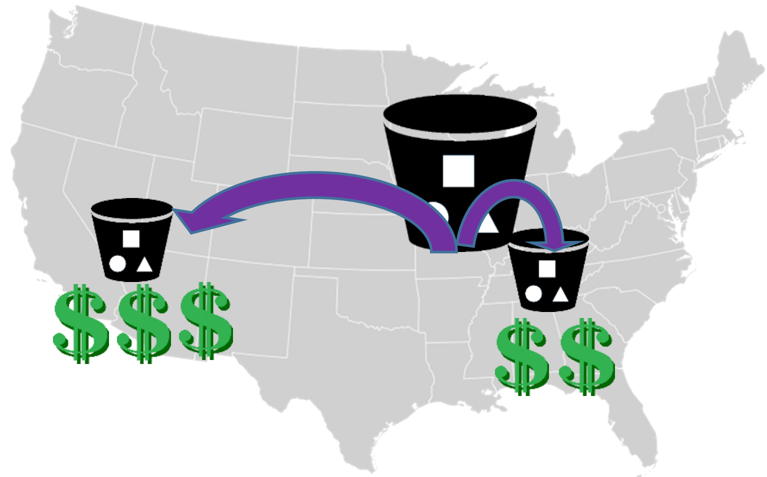
Regional Pricing

- Different regions may have different storage costs



Redundancy

- Increasing data availability & resilience will increases cost



Data Access Frequency

Operations on Data

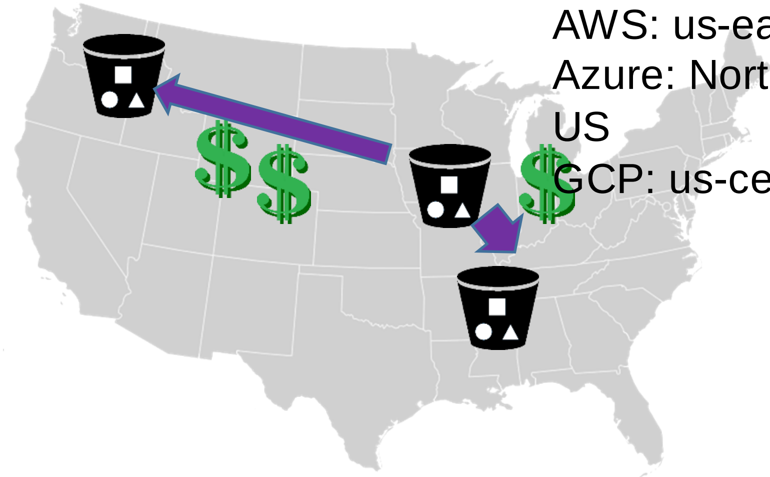
- Operations on data have costs.
- Minimize operations on cooler storage tiers to reduce costs
- **Example:** listing the contents of a bucket with a 250K files costs $1.35: $1.25 for listing and $0.1 for reading."

Operation type	Price per 1K operations
Write	$0.005
List	$0.005
Read	$0.0004

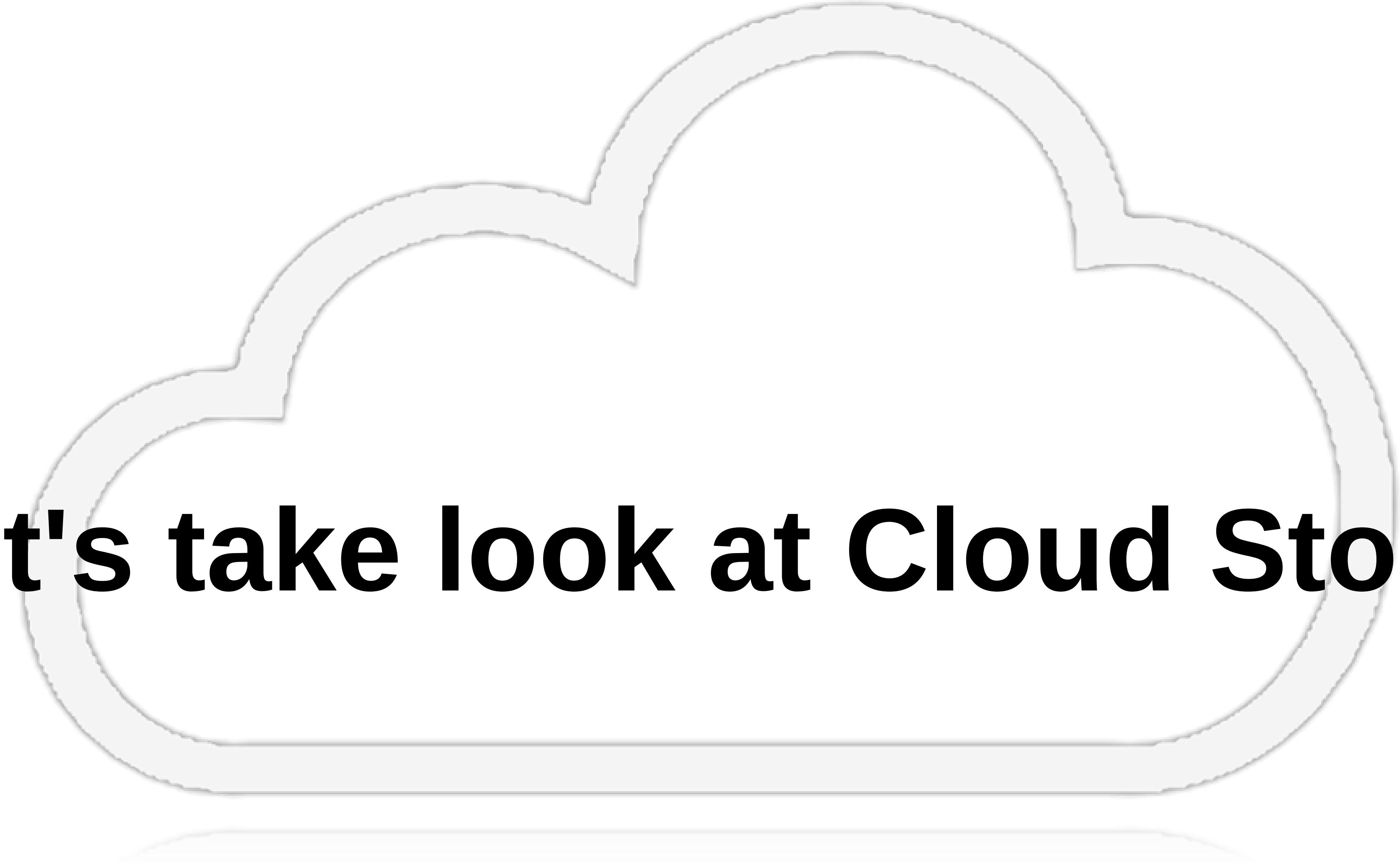
Network Usage

Minimize data distance

- Different regions
- Distant geographic locations



AWS: us-east-2 (Ohio)
Azure: North Central US
GCP: us-central1



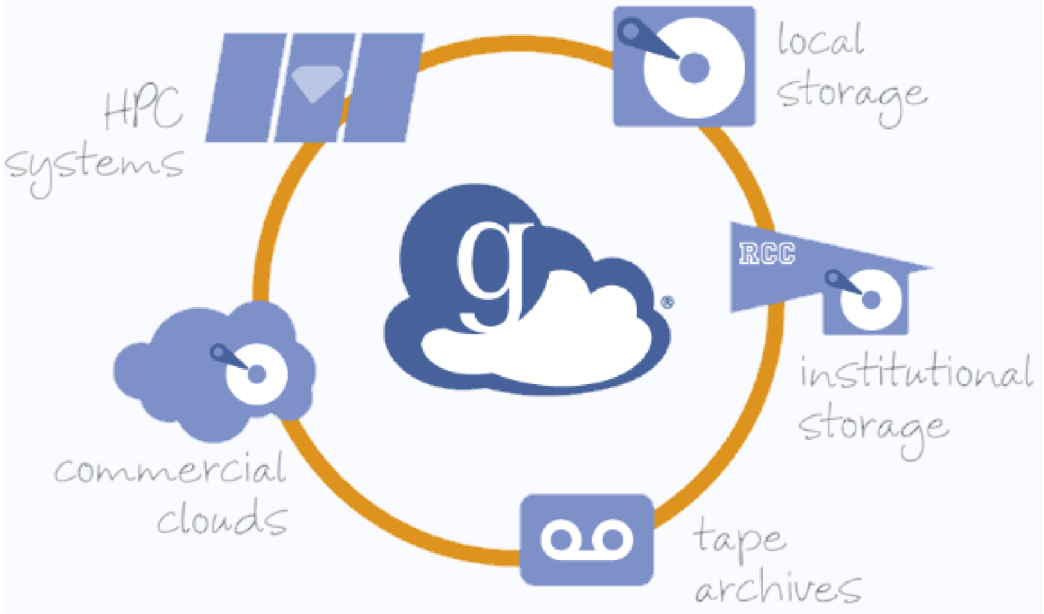
Let's take look at Cloud Storage

AWS SSO login

https://nu-sso.awsapps.com/start#/

How to Migrate to Cloud Storage

- Northwestern IT's supported tool for transferring and sharing data stored in S3
- All storage and compute solutions at Northwestern can communicate with Cloud Resources with Globus

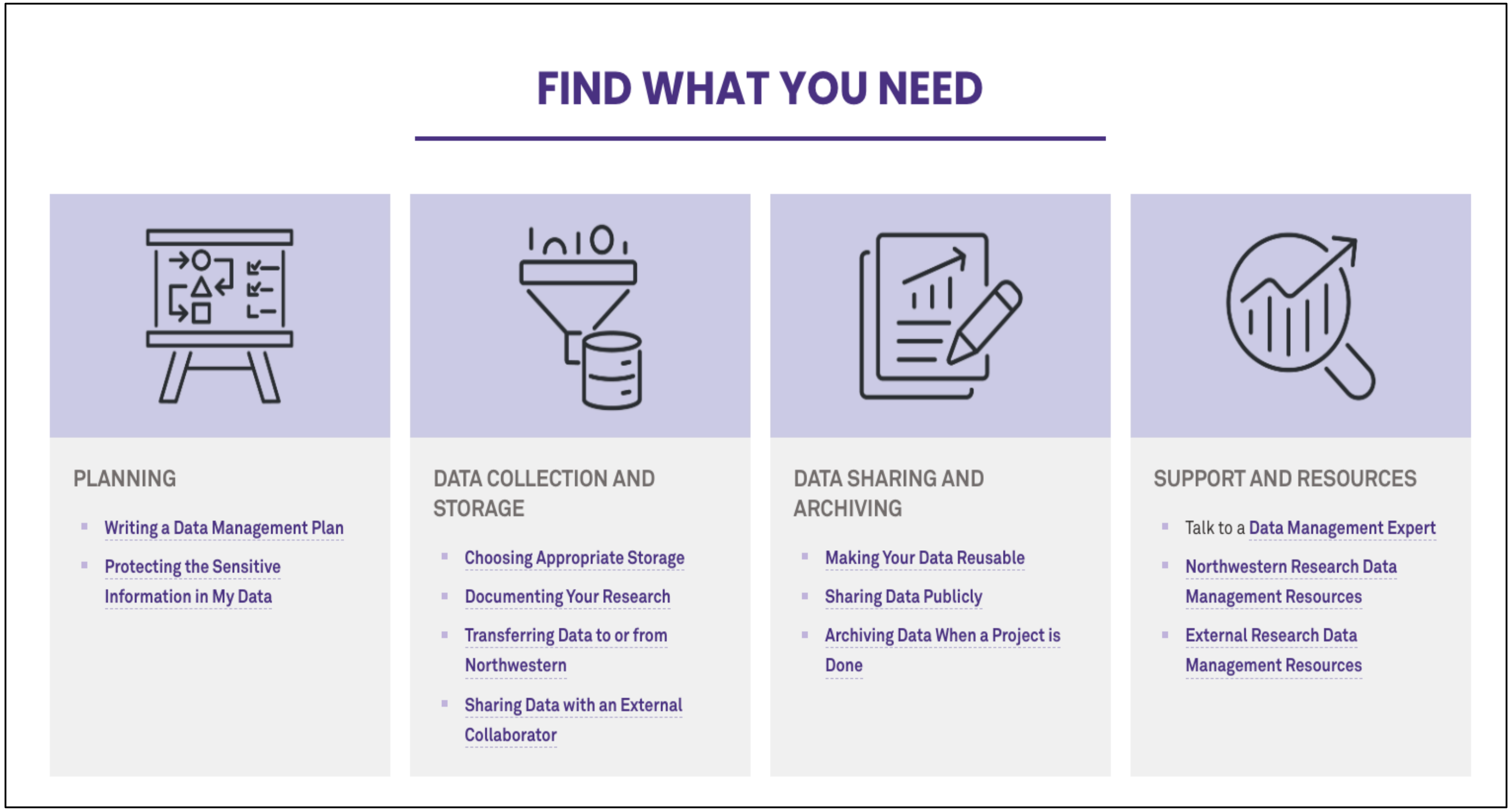


Want to Learn or Sandbox?

- **Northwestern AWS Sandbox Access**
  - Northwestern IT operates a “sandbox” AWS account that provides faculty and staff an environment to learn about and experiment with AWS products and services.
  - AWS sandbox account access can be requested by contacting the AWS Cloud Operations Group

- **NIH Cloud Lab (NIH Grants Funded projects only)**
  - ***Cloud Lab***
  - Cloud Lab removes barriers to cloud adoption by providing no-cost, customized, and scientifically relevant training, making it easier for researchers to learn about and explore the cloud with confidence.

Reach out!



FIND WHAT YOU NEED

PLANNING

- Writing a Data Management Plan
- Protecting the Sensitive Information in My Data

DATA COLLECTION AND STORAGE

- Choosing Appropriate Storage
- Documenting Your Research
- Transferring Data to or from Northwestern
- Sharing Data with an External Collaborator

DATA SHARING AND ARCHIVING

- Making Your Data Reusable
- Sharing Data Publicly
- Archiving Data When a Project is Done

SUPPORT AND RESOURCES

- Talk to a Data Management Expert
- Northwestern Research Data Management Resources
- External Research Data Management Resources

Research Data Management Website

Office Hours: Every Monday
3 p.m. – 4 p.m.
Mudd Library,
Genomics Lab

Emails: researchdata@northwestern.edu

Consultation Calls

Conclusion

In this workshop we explored

- Understand what **object storage**
- Learn the advantages of getting a Northwestern cloud account and how to request one
- Created and configure a bucket.

Useful Links:

- Northwestern research data management webpage
  https://www.it.northwestern.edu/departments/it-services-support/research/
- Northwestern IT - Research Computing Services
  https://www.it.northwestern.edu/research/index.html
- Northwestern IT - Cloud Operations & Public Cloud Accounts
  https://www.it.northwestern.edu/service-catalog/infrastructure/cloud-services/public-cloud-services.html
- Northwestern Cloud Community of Practice
  https://www.cloud.northwestern.edu/resources/cloud-community-of-practice/

Cloud Object Storage Documentation

- AWS S3 documentation
  https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html
- Microsoft Azure Blob documentation
  https://docs.microsoft.com/en-us/azure/storage/blobs/
- GCP Cloud Storage documentation
  https://cloud.google.com/storage/docs

Northwestern Cloud account benefits

- Microsoft Azure
  - https://services.northwestern.edu/TDClient/30/Portal/KB/ArticleDet?ID=1856
- Google Cloud Platform
  - https://services.northwestern.edu/TDClient/30/Portal/KB/ArticleDet?ID=1855
- Amazon Web Services
  - https://services.northwestern.edu/TDClient/30/Portal/KB/ArticleDet?ID=1854