Introduction to Cloud Storage

Kevin Luu

Virtualization and Cloud Specialist

Kevin.luu@northwestern.edu

Brian Roland

Data Management Specialist

Brian.roland@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

- Perform the following tasks
  - Create and configure storage
  - Upload objects into the cloud using the Console, and CLI

What - is the 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 Overview
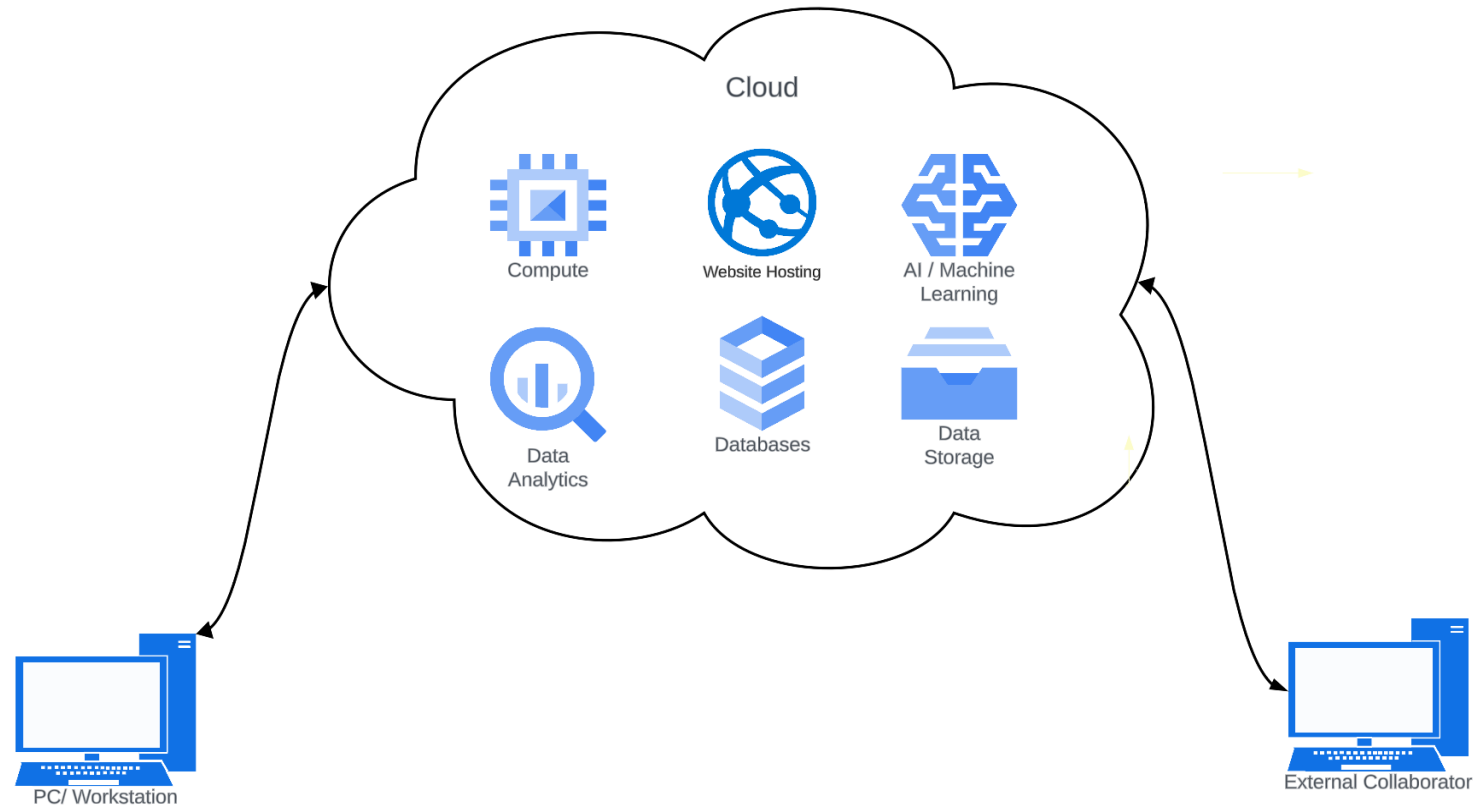
Does this sound familiar ?

- Is the cost of your current data storage currently too expensive ?

- Do you need more control of your files and folder such as who can read and write on them?

- Do you have too many old data that should be archived?

Cloud Storage could be the solution.

Why - Cloud Storage

- Cloud Storage if properly configured can result in lower costs.

- If you need granular permission on folders/ file you can create polices can on your storage to limit who has access.

- You can control spending by putting data into cheaper storage tiers designed for long term storage.

When - should you consider using Cloud?

Easy Access Anywhere: With cloud storage, researchers and collaborators can reach their files from any device with an internet connection.

Scalable: If you're uncertain about the required processing power or storage, the flexibility of the cloud makes it easy to scale up or down as needed.

Vendor Support: Unlike manually managing your current data management systems, here at Northwestern University we have a dedicated support across departmental IT and Vendors.

Cloud Providers

Amazon

Microsoft

Google

Largest of the 3 platform

Intergration of Microsoft's ecosystems like Office 365

Considered to be a leader in data analytics tools and services

Extensive global infrastructure

Access to OpenAI Models

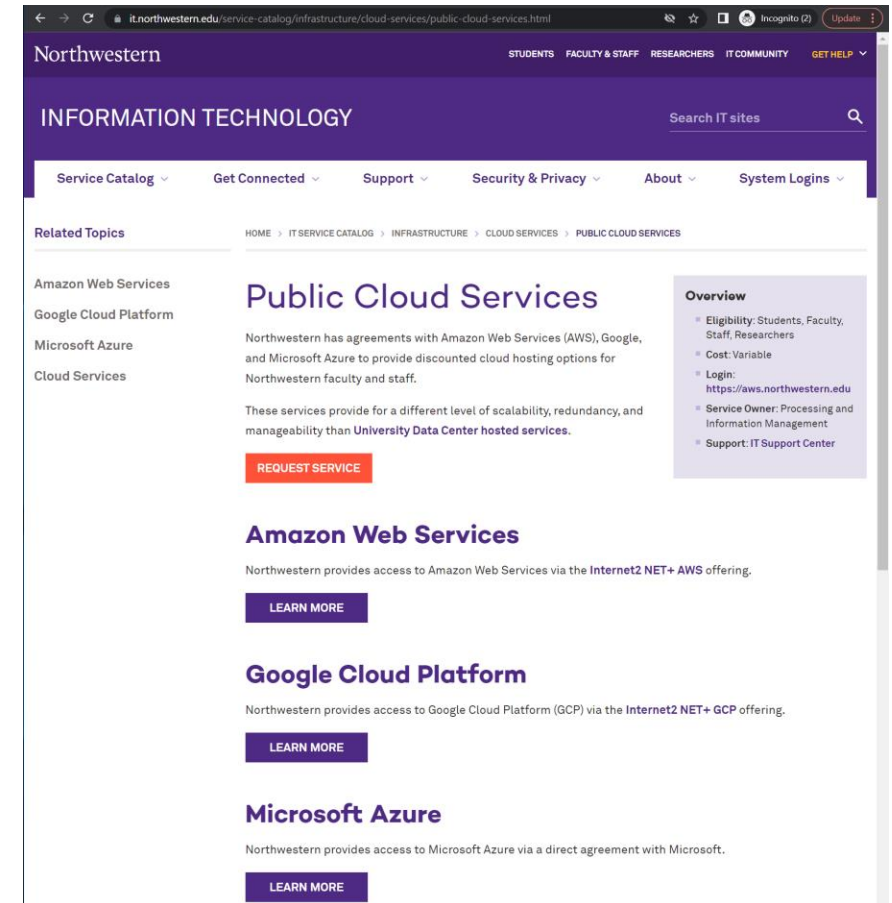Integrates with Google's ecosystem (Google Drive, Collab)

Each cloud providers in general do have the same services such as Virtual machines and Storage.
Here at Northwestern in terms of popularity AWS>Azure>Google

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

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		

NU cloud account, are setup with alerts ensure proper usage such notifying you when there is suddenly a large charges or improper configuration.

STRIDES

- If your project is funded by a NIH Grant, you may be able 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

Object Storage

- Inexpensive
  - o Pay only for what you use
- Scalable
  - o Theoretically unlimited
- Accessible
  - o Web portal, Globus, CLI
  - o Sharing data
- Long Term Storage
  - o Great for storing data long term for cheap.
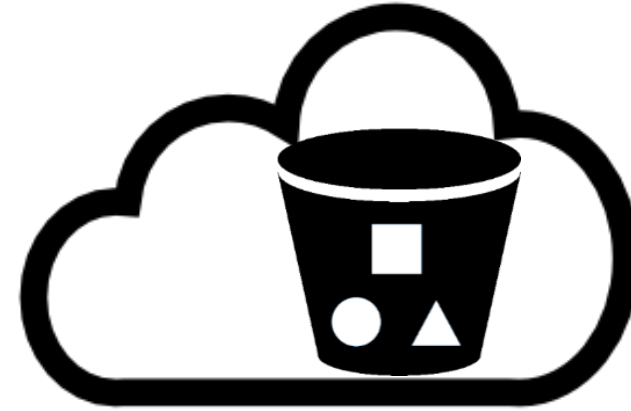
In object storage, The data you upload are referred to as a object and each object includes the data itself, metadata/s, and a unique identifier.
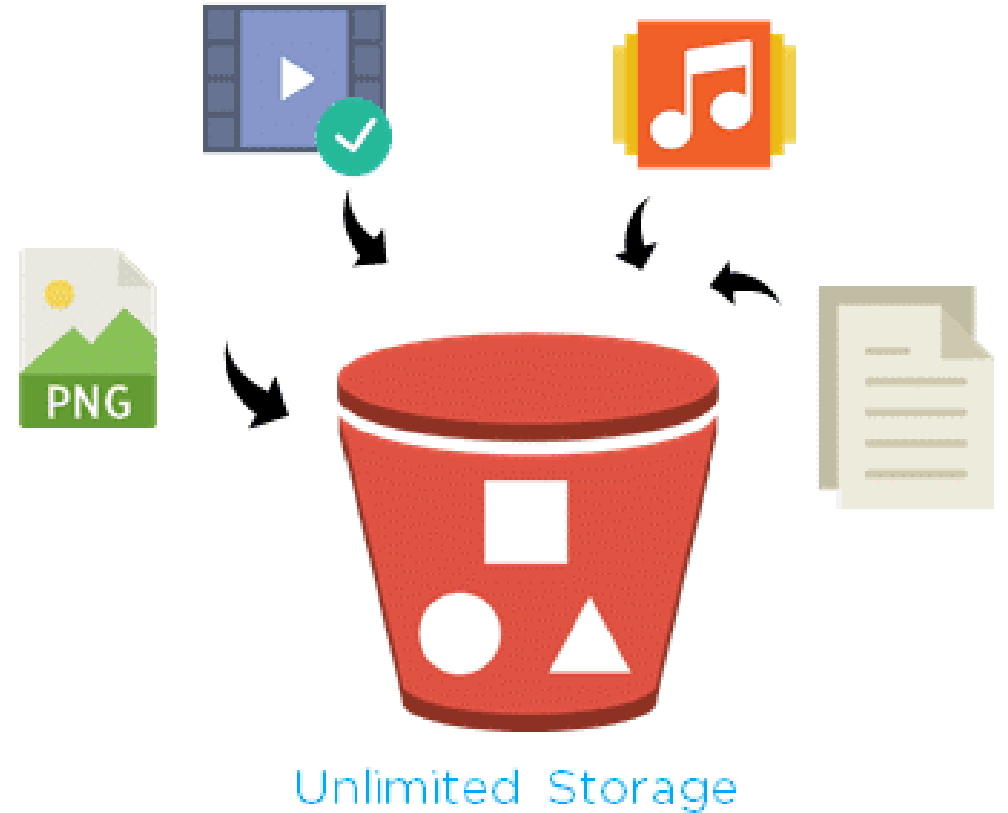
AWS - Simple Storage Service (S3)

What makes a bucket?

- **Bucket** – a container that stores objects

- **Object** - Any data type + Tag

- **Tag** – An optional identifier you can set for organization

Unlimited Storage

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

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 (weekly or 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

AWS – 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

- When changing storage tier from cold to hot tier you will incur retrieval fees the cost will vary depending on the storage tier and the retrieval time.

- There are 3 Types of Retrieval

Retrieval from Deep Glacier Archive	Cost (per GB)	Time
Expedited (Expensive And Fastest)	Can't Expedite on 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
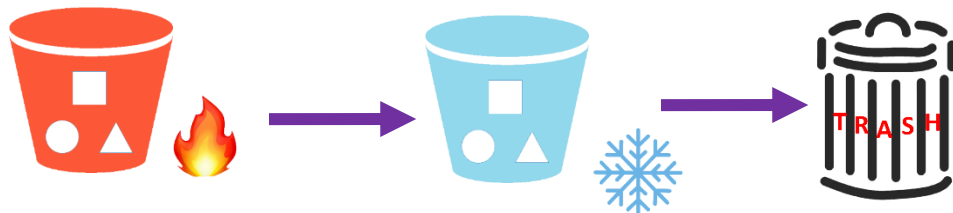
Intelligent Tiering Storage Class

- 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.

Object Storage: 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
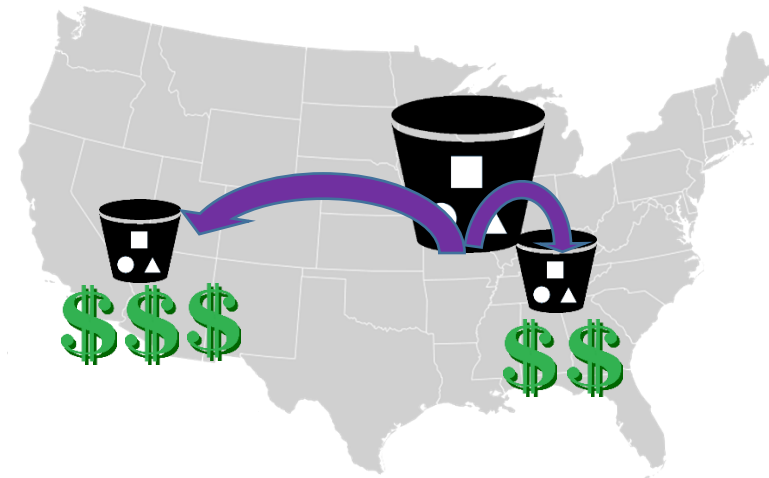
Object Storage: Other Considerations

Regional Pricing

• Different regions may have different storage costs
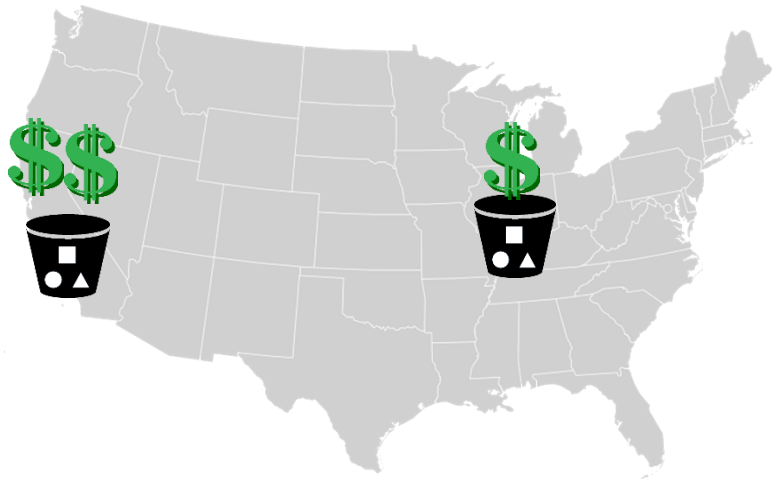
Redundancy

• Increasing data availability & resilience will increases cost

Object Storage: Other Considerations

Operations on Data

- Operations on data has cost
- Avoid operations on cooler storage tiers
- Impactful when tied to heavy Read or write applications
- Example: running an LS on your bucket with **1 million files** would cost
  - **$5.40 ($5 for List and $0.40 for Read)**

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
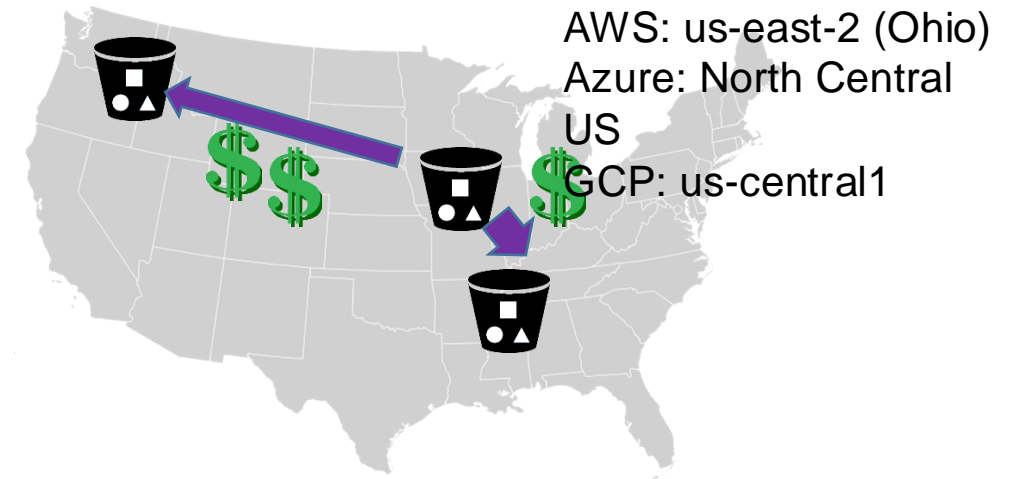
Object Storage Demo

Object Storage: AWS CLI - S3

AWS CLI (Command Line Interface)

- To interact with AWS on the command line
- Available in Linux, MacOS and Windows
  https://docs.aws.amazon.com/AmazonS3/latest/userguide/setup-aws-cli.html
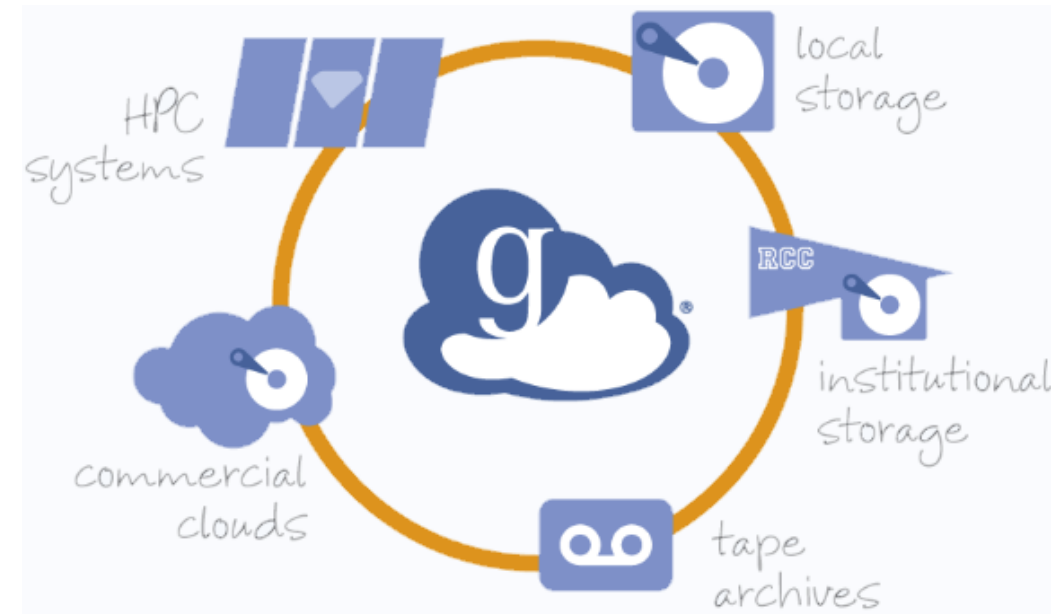
AWS CLI commands to interact with S3

- **aws s3api** : Set of commands mapping to low S3 application programming interface (API)
- **aws s3** : Set of high level commands to simplify S3 interactions

Globus: S3 Connector

- *Northwestern IT supported tool for transferring and sharing data stored in S3*

Have more questions?

Organize, Describe, Preserve, and Share

Research Data Management and Sharing
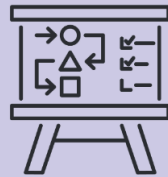
PLANNING
- Writing a Data Management Plan
- Protecting the Sensitive Information in My Data

DATA COLLECTION AND STORAGE
- Choosing Appropriate Storage
- 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

https://www.it.northwestern.edu/departments/it-services-support/research/data-storage/choosing-appropriate-data-storage.html

Reach out to RCDS

Research Computing Data Services

- Request a consultation to discuss your research needs.

https://www.it.northwestern.edu/research/consultation/index.html
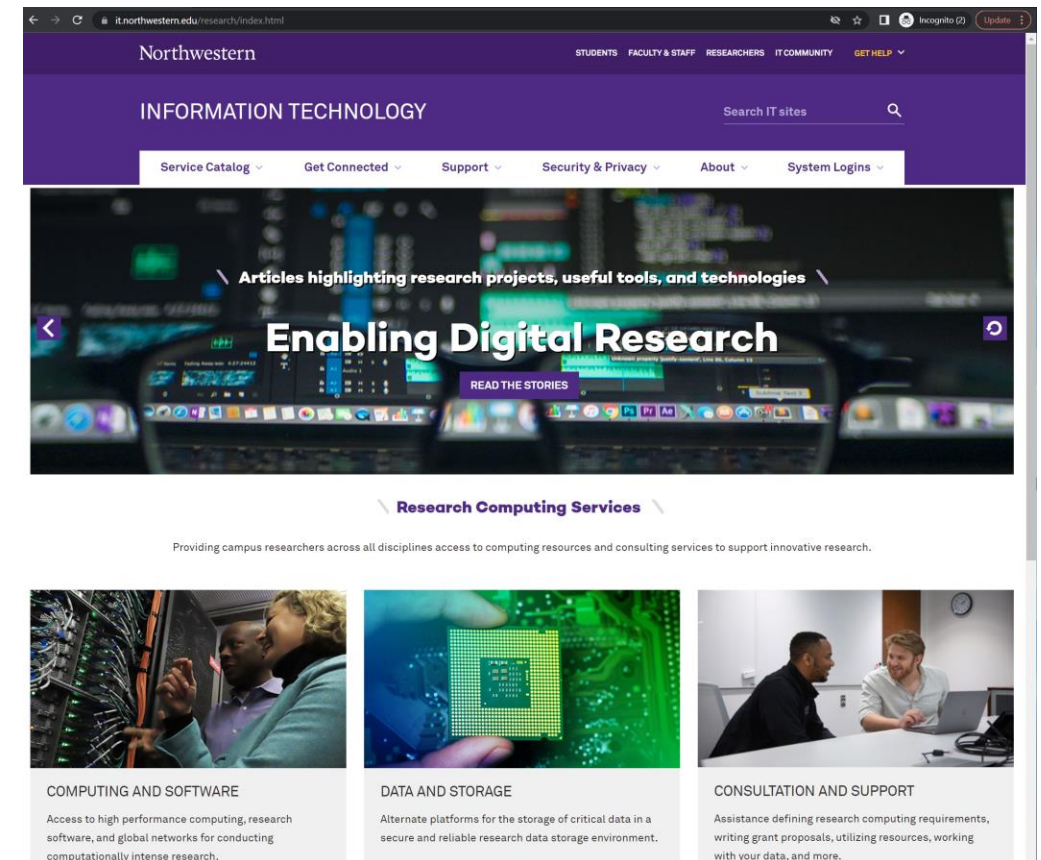
Conclusion

In this workshop we explored

- Understand what **object storage**

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

- Creating Bucket and basic configuration

- Transferring data via GUI, CLI and Globus.

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/

Useful Links

- 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

Useful Links: 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

Configuring AWS CLI SSO

Configuring SSO

aws configure sso

- o **SSO session name:** nu-sso
  SSO start URL: https://nu-sso.awsapps.com/start
  SSO region: us-east-2
  SSO registration scopes: Accept default "sso:account:access"

- You will configure your profile

  - o **Default Region:**us-east-2
    output format: Text
    Profile Name: Your Choice

Make your Profile Name memorable

When you run a command, you will need to append the follow to the end of your command.

--Profile <Profile Name>

AWS CLI SSO: Documenation