Exploring Cloud Resources for Research in the NIH Cloud Lab

Cloud Lab Team

STRIDES, NIH Center for Information Technology

August 2024

Exploring Cloud Resources for Research in the NIH Cloud Lab

Agenda

STRIDES Overview

NIH Cloud Lab

Bringing NIH Cloud Lab to External Users

Discussion

The Opportunities & Challenges of the Cloud

Cloud is an expanding frontier for biomedical research: the potential benefits are considerable, but so are the challenges.

THE OPPORTUNITIES

Simplifies testing and iteration

Always available on demand

Pay as you go for only what you use

Democratizes access to scientific data

Easily scales up and down

Provides a rich set of tools & services

THE CHALLENGES

Acquisition is complex and time-consuming

Security protections are unclear

Costs are unpredictable

Training deficit is substantial

NIH offers two programs that help funding recipients **address those challenges.**

NIH STRIDES Initiative

Purchase long-term commercial cloud services and support.

NIH Cloud Lab

Temporarily try the cloud at no cost to test your use case.

The NIH STRIDES Initiative

STRIDES: Science & Technology Research Infrastructure for Discovery, Experimentation, & Sustainability

Overview

Serving **both NIH staff researchers and NIH awardees**, the STRIDES Initiative accelerates biomedical research in the cloud by:

- Simplifying access
- Reducing costs
- Lowering technological barriers
- Standardizing administrative & financial processes

Core Motivations

1. **Democratization of computational research & data science**
   Leveling the playing field for those traditionally underrepresented in biomedical research

2. **Cost savings & efficiencies for the research community**
   More usage begets more savings and greater discounts for all

3. **Strong partnerships with cloud providers**
   Resulting in collaborative R&D engagements and more direct focus and support on research

Partnerships with: aws   Google Cloud   Microsoft Azure

STRIDES Overview

Value to Participants

STRIDES participants benefit from a variety of exclusive features, from competitive pricing to training expertise.

Competitive pricing & financial benefits

Professional service consultations

Flexible business model

Expanded communication reach

Expert support from cloud providers

Reach-through to additional partners

Training expertise and scaling capacity

Impact to Date*

339+ PETABYTES OF DATA

675M+ COMPUTE HOURS

2,420+ RESEARCH PROGRAMS

$113M+ COST SAVINGS

194+ EXTRAMURAL PARTICIPANTS

*as of June 30, 2024

NIH Cloud Lab: Experiment in the Cloud

NIH Cloud Lab is a no-cost, 90-day program for NIH intra- and extramural researchers to try commercial cloud services in an NIH-approved environment. Cloud Lab provides training and guardrails to protect against financial and security risks.
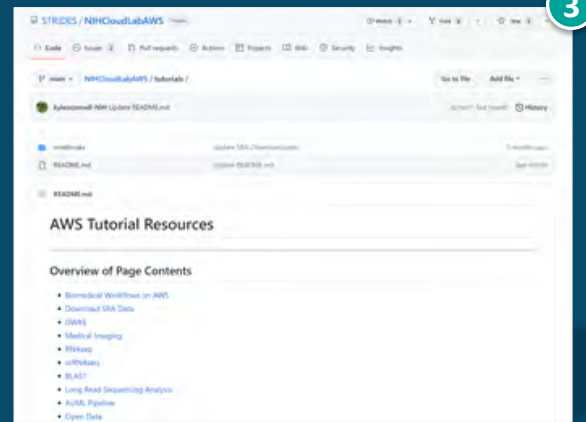
How It Works

1. **Fill out** interest form
2. **Get** account and $500 of credits
3. **Access** tailored cloud trainings
4. **Practice and learn** for 90 days

NIH Cloud Lab Sign Up Page
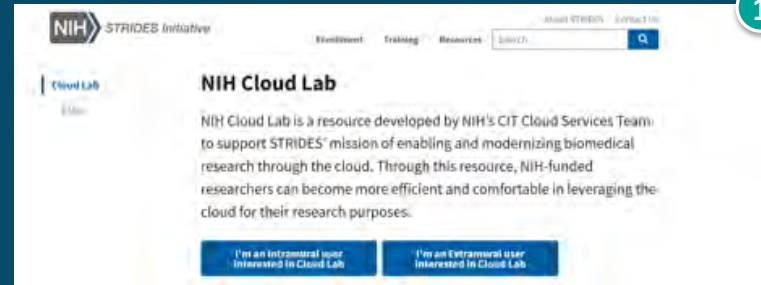
NIH Cloud Lab AWS Tutorial Repository

Example of NIH Cloud Lab Use Case

NIH Use Cases

Evaluate Utility & Cost

Provides an easy route to evaluate the cloud's utility/cost for a project without major time or financial commitments

Develop New Tools

Allows experienced teams to prototype new architectures and evaluate software and hardware combinations

Share Ideas

Connects NIH'ers from across ICs to share ideas on how to conduct biomedical research in the cloud

Learn New Skills

Simplifies access to tools and cloud environments that participants can use for training purposes

Interactive Tutorials

Cloud Lab offers a **suite of interactive tutorials** designed to help participants perform viable research in the cloud.

Scientific

- Accessing Open Datasets
- AI/ML
- ATACseq (NIGMS)
- Biomarker Discovery (NIGMS)
- BLAST
- FHIR
- Genomic Pathway Analysis (NIGMS)
- Genomic Variant Calling
- GWAS
- Long Read Genomics
- Medical Image Analysis (NIGMS)
- Medical Image Segmentation
- Metagenomics (NIGMS)
- Proteomics
- RNAseq
- scRNAseq
- Radiogenomics
- SRA Data Interaction
- VCF Query with SQL
- ...and more

General Cloud Ops

- Access Marketplace Offerings
- Access Public Datasets
- Command Line Tools / SDKs
- Disk Images
- Ingest and Store Data
- Introduction to Cloud
- Jupyter Notebooks
- Serverless Functionality
- Virtual Machines

Computing & Code

- Conda Environments
- Container Registries
- Git Repos
- HPC Clusters
- Kubeflow
- Serverless Functionality

NIH Cloud Lab AWS Tutorial Repository
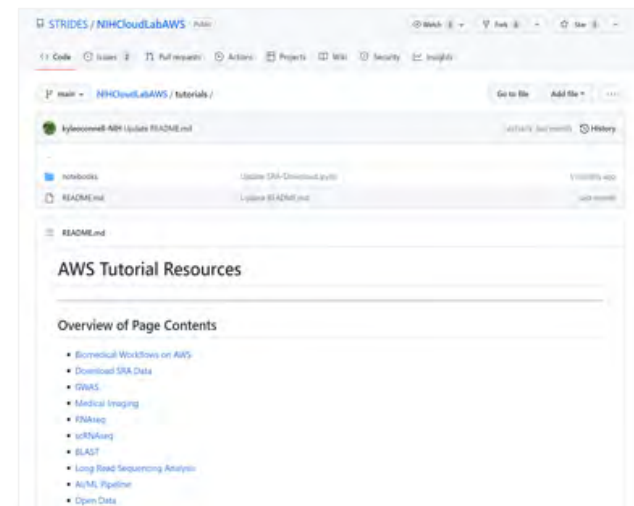
NIH Cloud Lab | External Users

As Cloud Lab grows, we envision including more crowdsourced training. To start, we integrated twelve interactive cloud-based learning modules developed by NIGMS grantees into our training offerings.
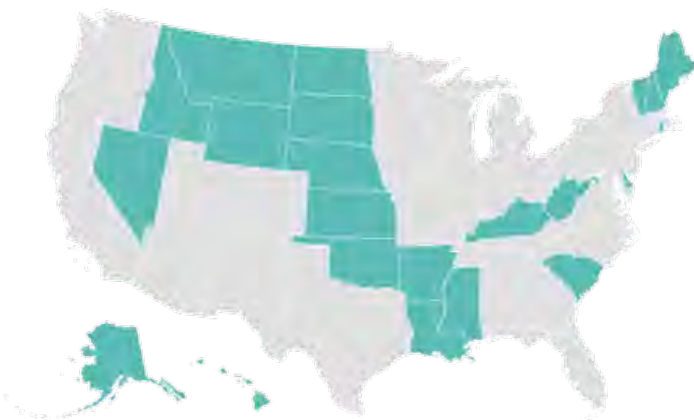
Data Science For Biology (San Francisco State University)

Integrating Multi-Omics (University of North Dakota)

Metagenomics Analysis (University of South Dakota)

Fundamentals of Bioinformatics in a Terminal Environment (Dartmouth)

Consensus Pathway Analysis (University of Nevada, Reno)

RNA-Seq Analysis (Maine's Public Universities — University of Maine System)

DNA Methylation Sequencing Data Analysis (John A. Burns School of Medicine, University of Hawaii)

Biomarker Discovery (University of Rhode Island)

Transcriptome Assembly, Refinement, and Analysis (MDI Biological Laboratory)

ATAC-seq and Single Cell ATAC-seq Analysis (University of Nebraska)

Biomedical Imaging Analysis Using AI/ML (University of Arkansas)

Proteomics Analysis (UAMS — University of Arkansas for Medical Sciences)

RE-CYCLING KNOWLEDGE

A total of twelve modules were developed in 2022 through an NIGMS Notice of Special Interest (NOSI) supplement and are now permanently available to Cloud Lab participants via NIH GitHub.

This exponentially increases the value of NIH's investment.

Cloud Lab for Workshops & Courses

NIGMS Sandbox modules can be tailored for student workshops, code-a-thons, and courses.

Previous Workshops and Courses

University of Hawaii

A weeklong bioinformatics conference utilizing the NIGMS Sandbox modules as the core of the curriculum.

Oklahoma University

A month-long summer course utilizing the NIGMS modules on the fundamentals of bioinformatics, consensus pathway analysis in the cloud, and metagenomics analysis of biofilm-microbiome.

University of South Carolina

Two 2-week seminars, the first focusing on Python programing for bioinformatics and the second on R programming for bioinformatics.
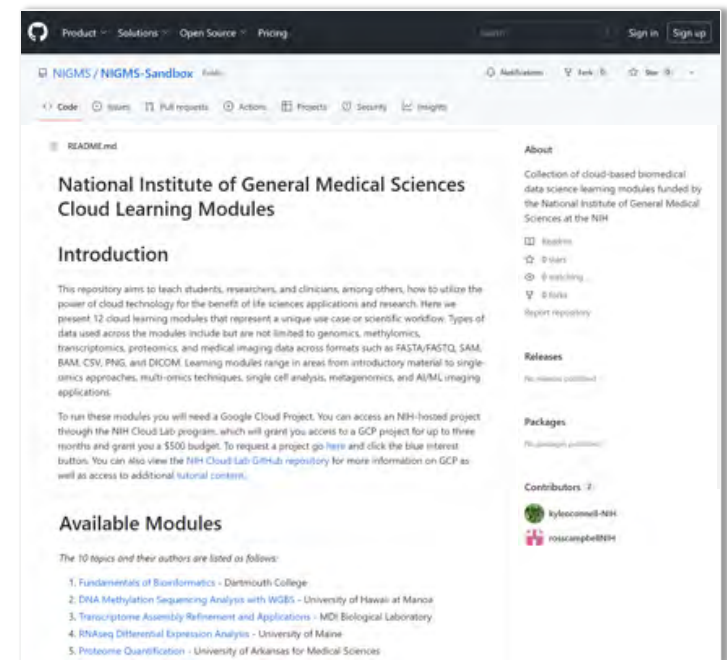
NIGMS Sandbox GitHub Repository

The NIH Cloud Lab User Expansion

NIH Cloud Lab began as a sandbox for internal NIH users to test out cloud services.

Internal NIH Cloud Lab sandbox

Desire to expand Cloud Lab access to additional external users & institutions

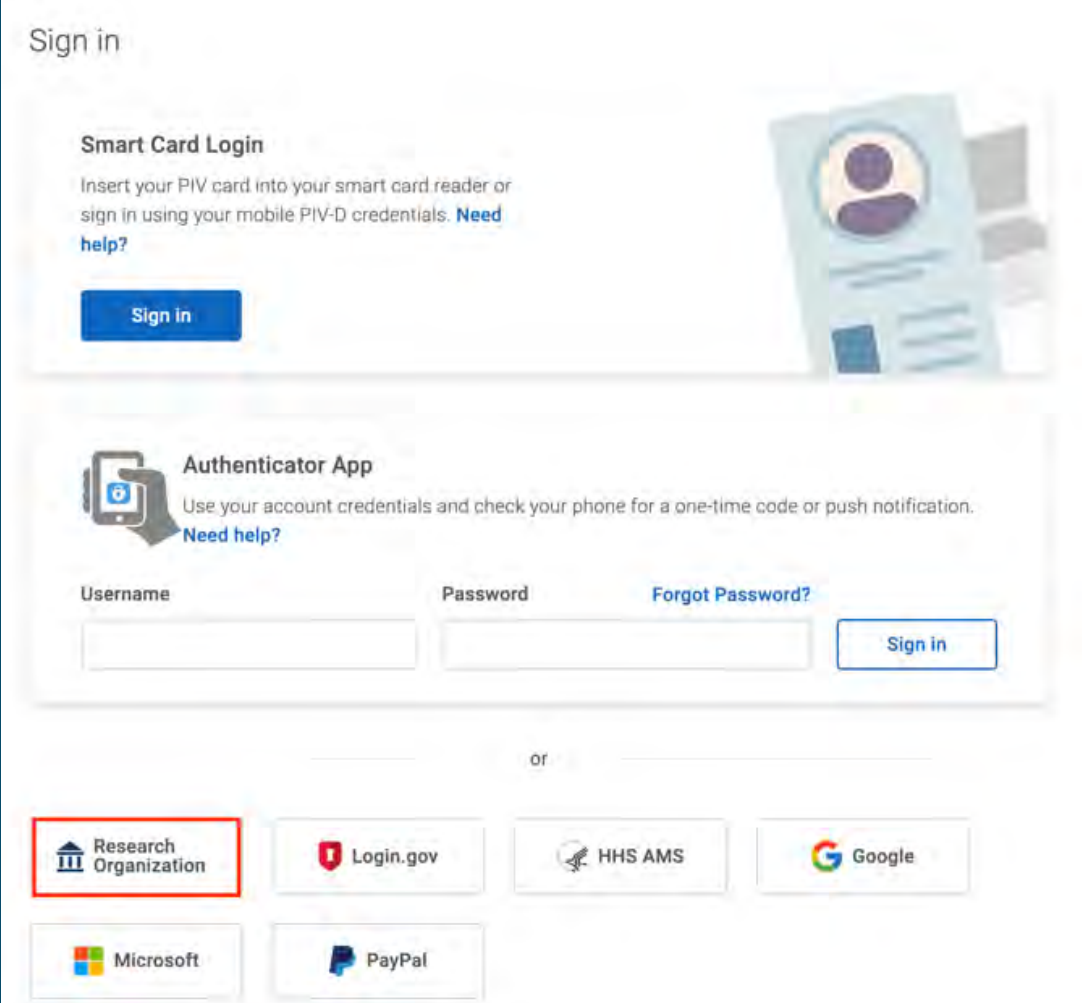**Cloud account access issues for non-NIH users**
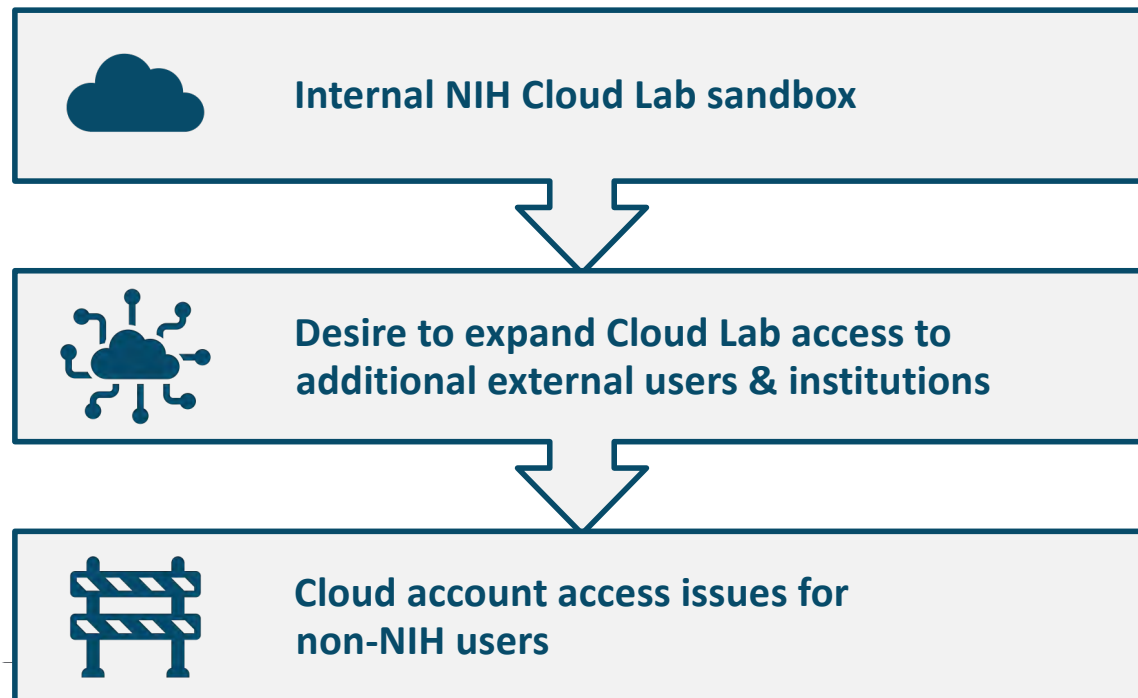
THE **OBSTACLES**

All users need a NIH login with multi-factor authentication to access Cloud Lab resources

Global users need access to U.S.-based resources

Different Solutions are required for each cloud service provider

Federated Login Solutions

- **InCommon** – federation provider for U.S. based university and research institutions.
- **eduGAIN** – multinational federation alliance of which InCommon is a member of.
- **Login.gov** – secure federation provider used to access government resources.

Bringing NIH Cloud Lab to External Users

Cloud Lab Federated Login – Account Registration
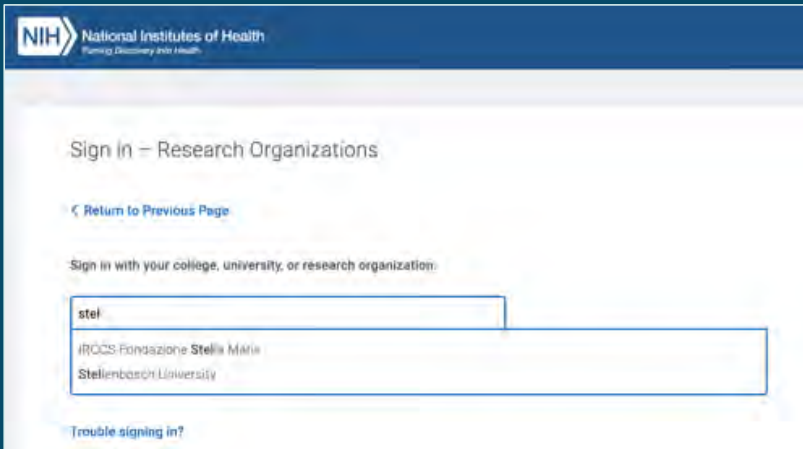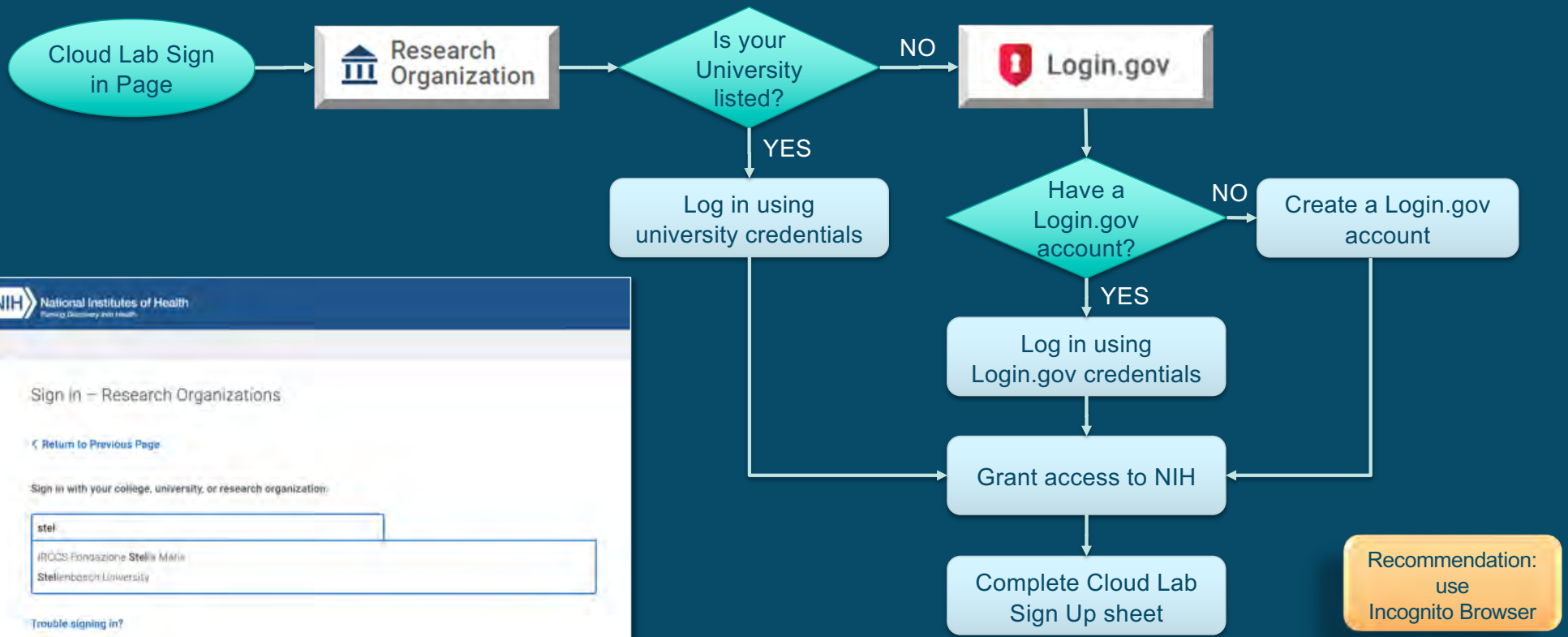
Extramural Federation Development

Google Cloud

Google Cloud Platform (GCP)

- Developed shadow identity mapping process leveraging InCommon and Login.gov federation.

- This process involves mapping a user's institutional email address to a specially provisioned NIH email (restricted access) that has a Google Identity attached and allows access to Cloud Lab's Google console.

- To avoid conflicts with any other NIH applications that researchers use, we created a unique NIH login pathway and use unique URLs for Users to access their Google projects.

aws

Amazon Web Services (AWS)

- Developed a VDS mapping process to replace SailPoint for extramural users. Manually maps research institution email address, the extramural AWS account and the Cloud Lab user role so extramural users can access the AWS console

- Currently working with IAM to develop an API tool to automate the extramural federation mapping process.

Microsoft Azure

Microsoft Azure Cloud Platform (Azure)

- Multilateral Federation (InCommon and Login.gov Federation) is not natively supported by Azure, so we use Azure external Guest accounts to provide access.

- Federation is provided by Microsoft/Azure via Microsoft Active Directory (AD) or Azure Entra for Microsoft based organizations.

- For organizations that are not Microsoft based users create a Microsoft Online accounts and are federated via Microsoft .

Using Shadow Identities to Provide Google Access

CHALLENGE:
You need a Google Identity to access Google Cloud and most university email addresses don't have one.

SOLUTION:
Shadow Identities

What is a Shadow Identity?

- A mapping between a secure internal NIH email addresses to an institutional email address.

- Allows piggyback/transitory access to Google Cloud

- Is secure so research user can only utilize the shadow identity to access his provisioned Google Cloud account.

- Has a defined life cycle so identities only exist for limited duration minimizing any security issues

- Uses a unique login pathway to recognize that institutional email address has a special mapping.

Implementing Azure Federation

CHALLENGE:

Azure is not compatible with multi-lateral federation (InCommon, eduGAIN or Login.gov)

SOLUTION:

Microsoft Federation (Azure Guest Accounts)

How does Microsoft Federation Work?

- Microsoft Federation leverages Microsoft Active Directory (AD) or Azure's AD Domain Services (DS) (now called Entra Domain Services)

- It allows federation by organizations that utilize the on-prem or cloud version of AD or requires users to establish a Microsoft Online account using their institutional email address.

- An invitation is sent out to join NIH's AD system as a Guest User

- The User accepts the invitation either through their organization's AD service or their Microsoft Online account and accesses their NIH Cloud Lab Azure subscription.

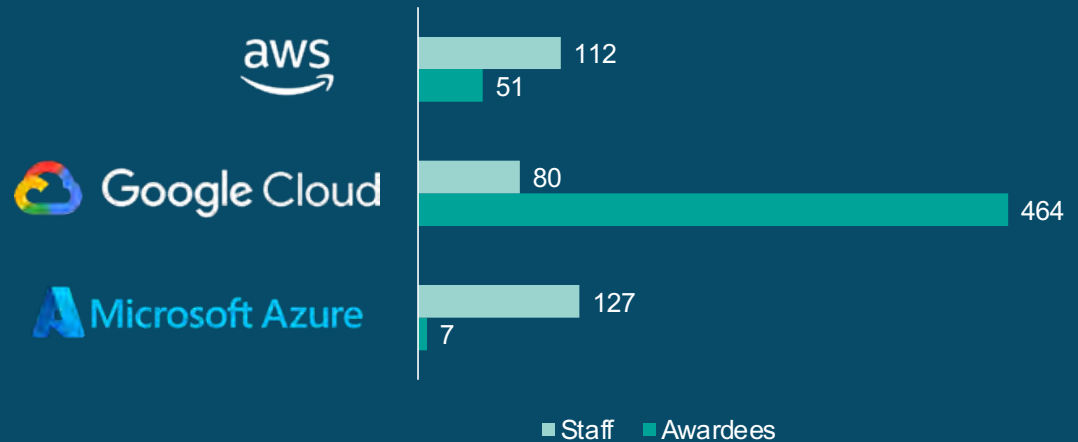- Guest account life cycle policy is set at an organizational level and expires after a predefined period
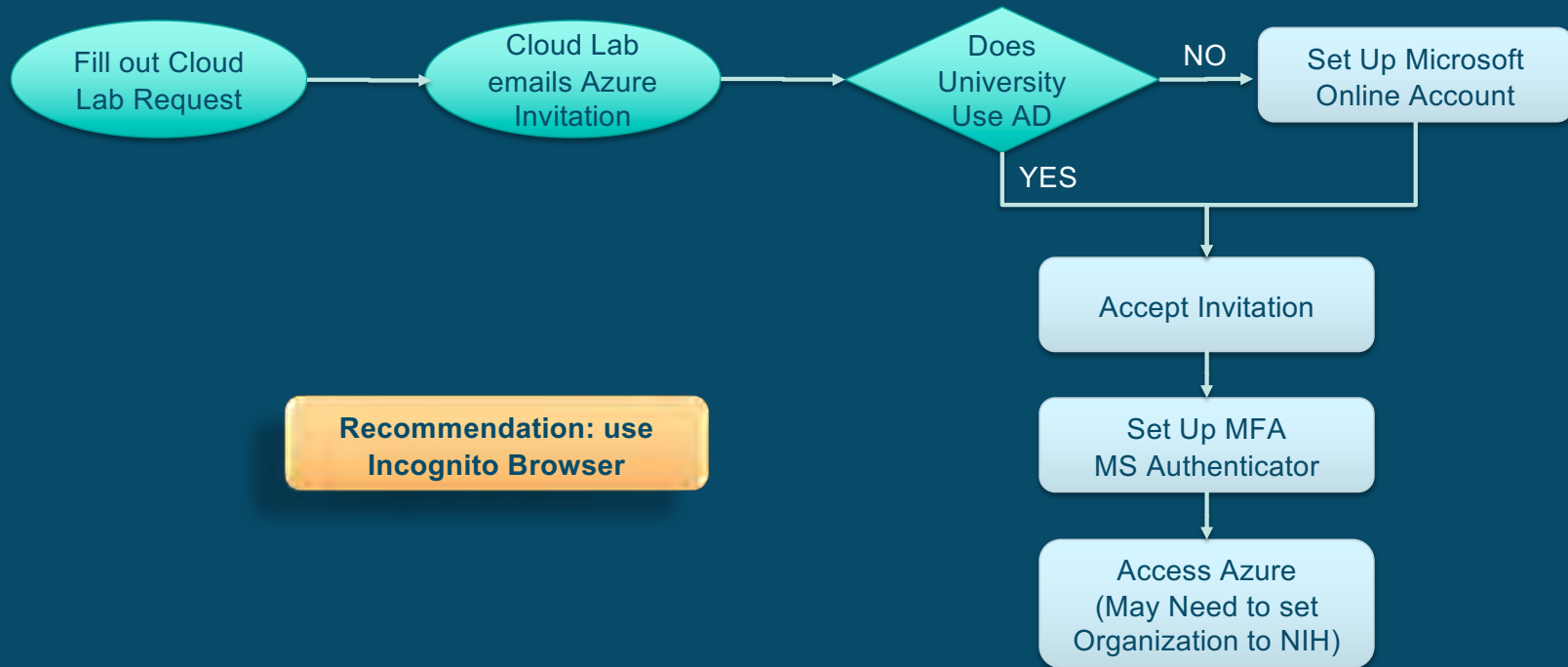
Discussion

Want to learn more? Visit our website at cloud.NIH.gov/resources/cloudlab or contact us at CloudLab@nih.gov