Public Cloud for Bioinformatics

Bioinformatics Workshop
April 25th, 2022

Data Sharing Needs for Biomedical Research

- Downloading large scale data to a local computer for analysis may be more difficult than bringing computational capability to where the data is located.
  - 1000 Genomes (>3,000, 80TB, > 2 days)
  - TopMed (>50,000, 2PB, weeks or months)
- Datasets are outgrowing local infrastructure, inhibiting researchers' ability to maintain them.
  - CCDG ~140,000 Whole Genome Sequence (WGS)
- Compute requirements to process these large datasets are exceeding local capacity, inhibiting analysis.
- Collaborating on research projects across organizations can be challenging due to differences in local IT environments.
  - Lack of standards and provenance
  - Batch effects, incompatible or inaccurate software
- Siloed and redundant infrastructure required for each lab, department, institution, etc.
  - Data Centers, Systems Administrators, Software Installation, User Support

Public Cloud Options

Direct Cloud Usage (IaaS)

(more features, more control, harder to use/manage)

Major Players:

1. Amazon Web Services (AWS) *(market leader)*
2. Microsoft Azure
3. Google Cloud *(GCP: aka "Google Cloud Platform")*

Indirect Cloud Usage (PaaS or SaaS)

(less control, application focused, easier to use/manage)

Examples:

- Terra (scientific computing; A joint Broad/Google Venture; see later discussion)
- DNAnexus (bioinformatics cloud platform; based on AWS)
- Seven Bridges Genomics
- Google Colab (scientific computing)
- Heroku / Tableau (websites & visualizations)
- RIS (they are working on "expanding" the cluster with the major cloud providers)

Choosing a Cloud Option

It's very hard to be "cloud independent". Inter-cloud operations are generally not encouraged by the cloud providers.

Much like choosing a programming language for your problem, choosing a cloud provider can impact your strategy, collaborations, and mindset.

Each provider has its own distinct features, tooling, and community.

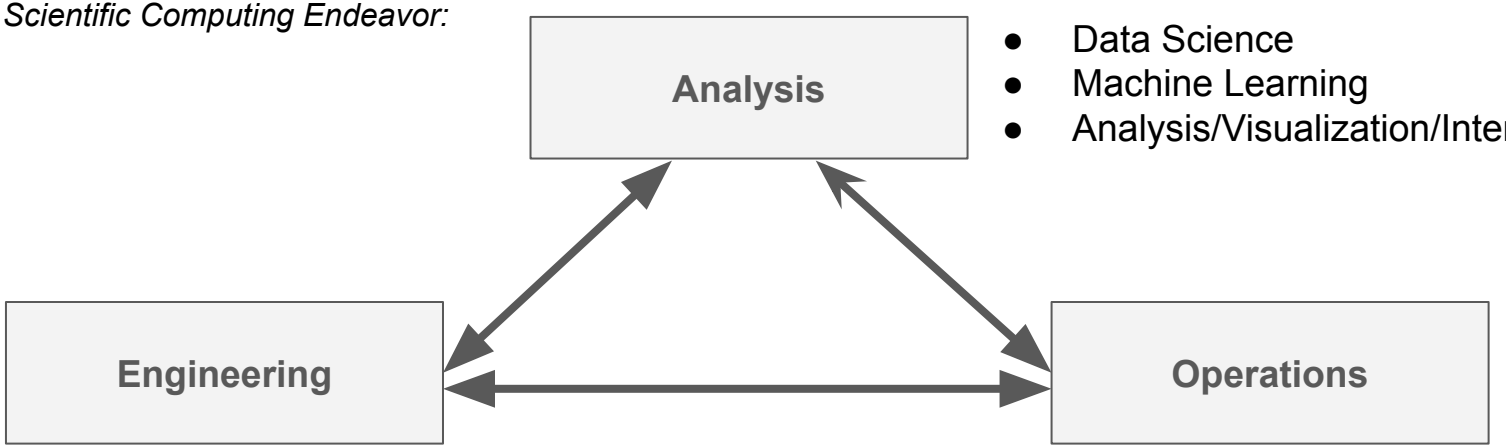
- Amazon
  - 80% market-share; very popular; more business-centric
- Microsoft Azure
  - enterprise-centric; WUIT preferred for protected data
- Google Cloud
  - expertise in Big Data
  - used by the Broad & Data Biosphere (bioinformatics tooling)
  - 3rd-place market-share; aggressively trying to increase it
  - WUIT-managed, dedicated fiber network to GCP for comparatively fast data transfers

Public Cloud & in-house compute

- Using the cloud is not the same as using the compute0/1 cluster
  - Not HPC-oriented (no LSF, à la `bsub`; no NFS-based file system)
  - Essence of scientific cloud computing:
    i. Create a VM
    ii. Download Input Data onto VM from Cloud Storage
    iii. Download software onto VM
    iv. Run computation(s)
    v. Upload Output Data on VM to Cloud Storage
    vi. Delete VM
- “research” on the cloud vs. local in-house compute
  - Exploratory compute on the cloud can get expensive very quick.
  - Try to do small-scale “research” on in-house compute, and then move to the cloud for large-scale “production” tasks
- infrastructural considerations
  - see the next slide

Infrastructural Considerations

A Scientific Computing Endeavor:



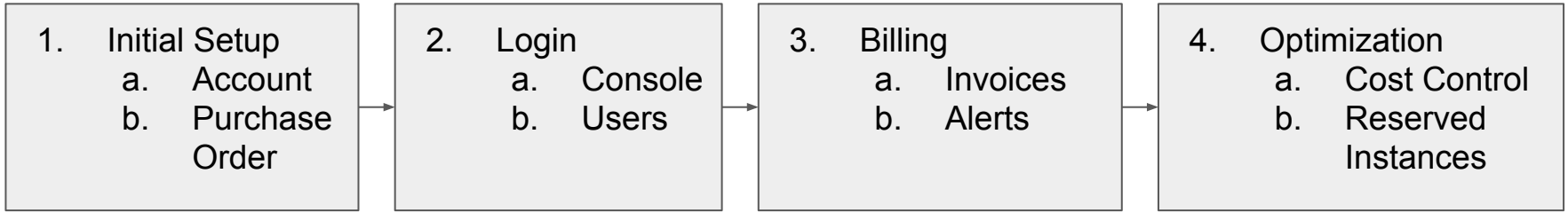
- Data Science
- Machine Learning
- Analysis/Visualization/Interpretation

- Pipelines
- Distributed Computing
- Data Management
- Scalability Concerns

- Monitoring
- **Security**
- Administration & Networking
- Policies

  - Currently, on the cloud you are responsible for the full infrastructure
  - Address either in-house or by purchasing 3rd party options -- in terms of personnel, tooling, security, and policy needs

WUIT Public Cloud: Overview



- Request Cloud account from WUIT
- Request purchase order from department

- Test login, user permissions
- Access console features
- Request additional users, permissions from WUIT (as needed)

- Routing of bills
  - Email
  - Workday
- Billing Resources
- Billing Alerts
  - Daily
  - Monthly

- Minimize Cost
- Spot/Preemptible Compute
- Workflow Refactor
- Storage Footprint

WUIT Public Cloud: Service Providers

- Google Cloud - Information Technology
  - Request new WashU Google Cloud Project - Information Technology

- Amazon Web Services - Information Technology
  - Request New WashU AWS Account - Information Technology

- Microsoft Azure - Information Technology
  - Request New WashU Azure Subscription - Information Technology

WUIT Public Cloud: Initial Setup

- Required Information
  - School/Department/Unit Name
  - Project/Lab Name
  - Lead Technical Contact: Name, Email, Phone
  - Business/Finance Contact: Name, Email, Phone
  - Admin Access List: List of @wustl.edu addresses (additional to Lead Tech Contact)
  - Blanket Purchase Order with suppliers
- Vendor/Supplier Specific
  - AWS & Azure - HIPAA or FERPA protected data
  - AWS - Business-Level Support (10% additional fee)

WUIT Public Cloud: Blanket Purchase Order (BO)

- Create through WorkDay/Marketplace
- The suppliers are:
  - Microsoft Azure - Insight Public Sector
  - Amazon Web Services - DLT SOLUTIONS LLC
  - Google Cloud - Burwood Group, Inc.
- Estimate your annual usage costs
  - Use a high estimate of what you expect to use
  - You decide the dollar amount, e.g. $500.00
- Monthly invoices paid against the BO
  - Available dollars are correspondingly reduced
  - Change Order to a BO can be done to add money if necessary.
- BO’s expire at the end of the fiscal year
  - Funds remaining on the BO go back into your department’s budget.
  - There is no roll-over to the next fiscal year.
  - A new BO is issued for the new fiscal year.
- Set the expiration date as 6/30 of the fiscal year
- Description of the product as “[Cloud Provider] for [someone].”

WUIT Public Cloud: Billing, POs, encumbering ...

- Billing contacts within each department open a Blanket Purchase Order (BO)
  - Billing contacts by Department
    - McDonnell Genome Institute
      - MGI_purchasing <mgi_purchasing@gowustl.onmicrosoft.com>
    - Medicine - may vary by PI
      - Ex. LaDonna Utley (ladonna.utley@wustl.edu)
- You may choose one BO per supplier or cloud/research project specific BOs
  - Check with Billing Contact for preference
- You will be asked periodically to encumber the BO for a certain amount based on how much you project to spend over the next 6-12 months
  - At this point you decide how much will be sourced to which fund numbers

WUIT Public Cloud: Reseller Relationships

- For billing and legal purposes WUIT has established business agreements with resellers
  - DLT (AWS)
  - Burwood Group (GCP)
  - Insight Public Sector (Azure)
- Interaction with these resellers can be quite limited
  - Does add some complexity to billing because what you see in the cloud billing console is not necessarily what you will be charged by DLT, Inisght, or Burwood (may be less or more)

WUIT Public Cloud: Cloud Computing Services Comparison

	WashU Private Cloud	Azure	AWS	Google Cloud
Signed University contract	✔	✔	✔	✔
University billing	✔	✔	✔	✔
WUSTL Key SSO login to admin console	N/A	✔	✔	✔
Approved for confidential and protected workloads	✔	*	*	*
WashU IT managed workloads	Servers, Databases, Web tools	Servers, Databases, Web tools		
Discounts	Competitive University pricing	6% across the board, Hybrid Use Benefit	5% across the board	25% on compute/storage, 19% on other services, Sustained use discounts
Data egress waiver	No egress data charges	No data egress charges if < 15% of total charges	No data egress charges if < 15% of total charges	No data egress charges if < 15% of total charges
Ease of integration with:	On-campus technology resources	Microsoft/Windows technology stack	Linux/Unix Technology stack	Linux/Unix Technology stack

* WashU's agreements for Microsoft Azure, Amazon Web Services, and Google Cloud include HIPAA BAA and FERPA compliance guarantees. However, workloads (servers, databases, etc.) in these environments are not inherently compliant. Workloads must be built and configured following appropriate guidelines. When you will be using public cloud platforms to store and transmit protected data, It is necessary to seek guidance from Information Security and WashU IT.

If you still aren't certain which cloud offering would best suit your business needs, contact <washuit-publiccloud@wustl.edu> to discuss the options.

Source: https://it.wustl.edu/services/cloud-computing/cloud-computing-services-comparison/ (4/21/2022)

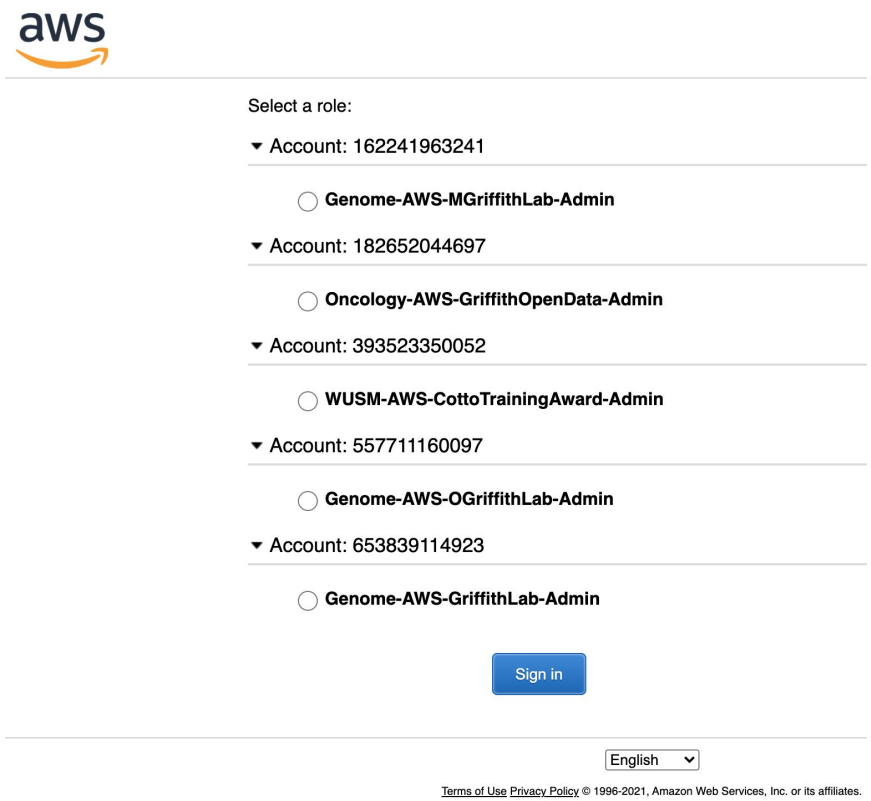
WUIT Public Cloud: Agreements

Normally, putting data into the cloud (ingress) is free. Taking data back out of the cloud (egress) is not typically free. If egress is less than 15% of charges, egress fees are waived.

WUIT has made agreements with the major public cloud providers to allow certain compute & storage discounts on WUIT-managed cloud accounts.

See WUIT's Cloud Computing Service Comparison for the latest details.

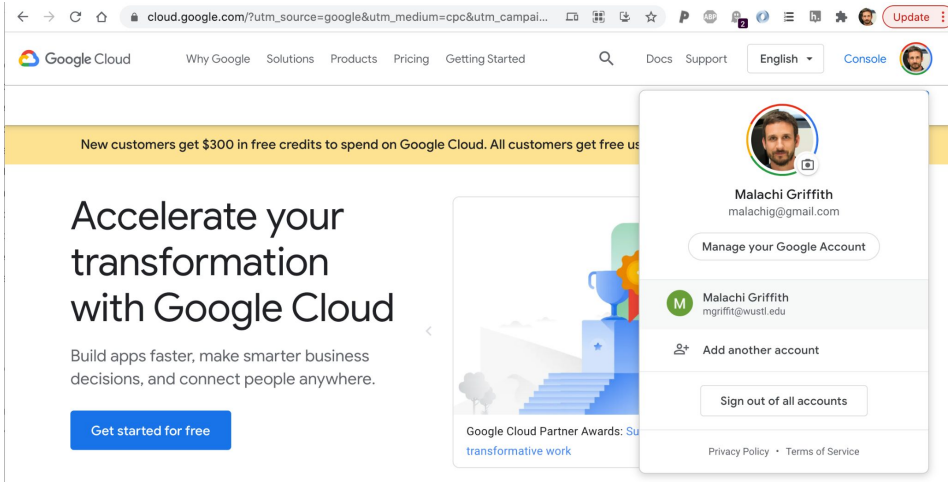
Logging in to AWS (http://connect.wustl.edu/awsconsole)

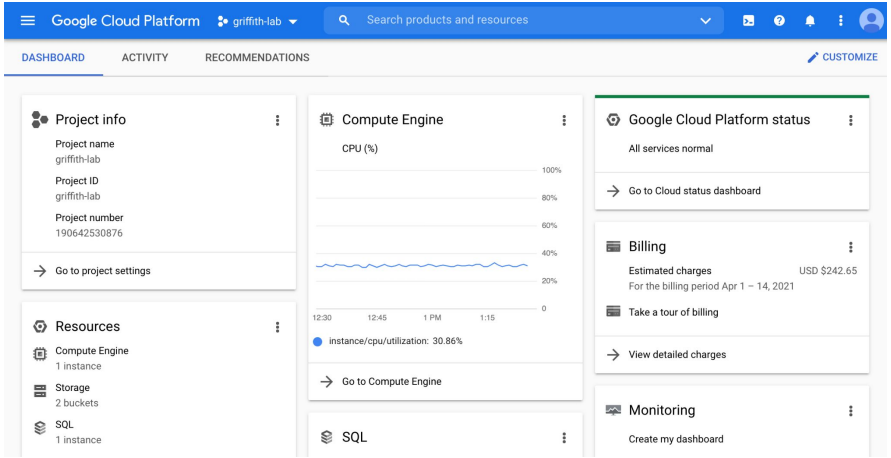


- Authentication uses your WUSTL Key
- You can have multiple accounts (optional)
  - Per project
  - Per billing source
- Or you have can one account and manage billing by paying one bill from multiple sources
- Multiple accounts can also be used to compartmentalize access to members of the lab working on different projects
- Each account will generate distinct billing, billing alerts, etc.

Logging into GCP (https://cloud.google.com/)

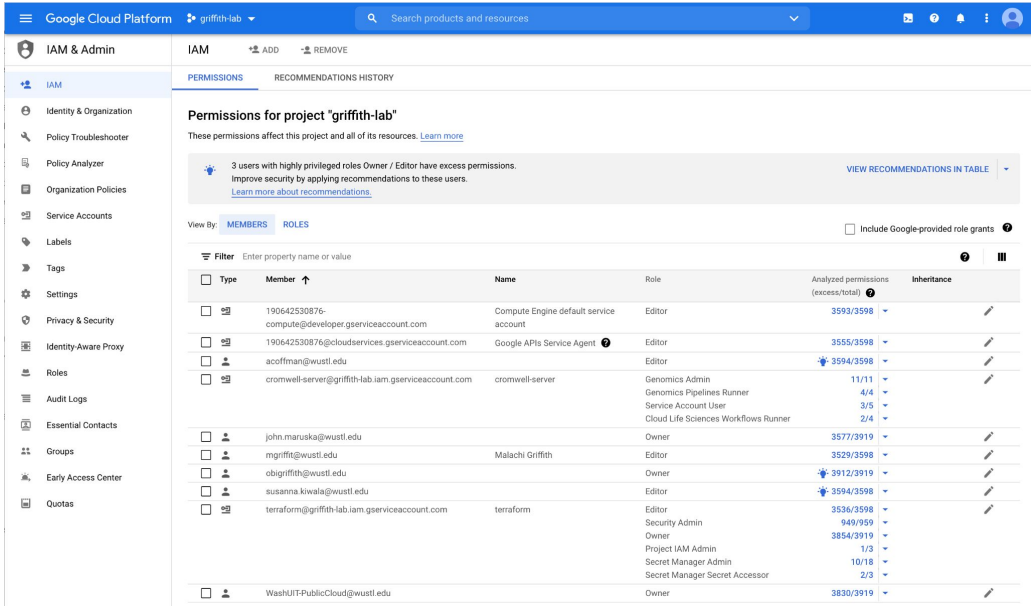
- WUSTL Key Google account
  - A very limited functionality Google account associated with your WASHU login
  - You now have two Google accounts. If you use the Google ecosystem a lot, this will be annoying.
  - Better functionality would happen with Google Workspace (formerly known as “G Suite”) integration; however, WUIT has provided G Suite access to some users, contact WUIT public cloud for info.



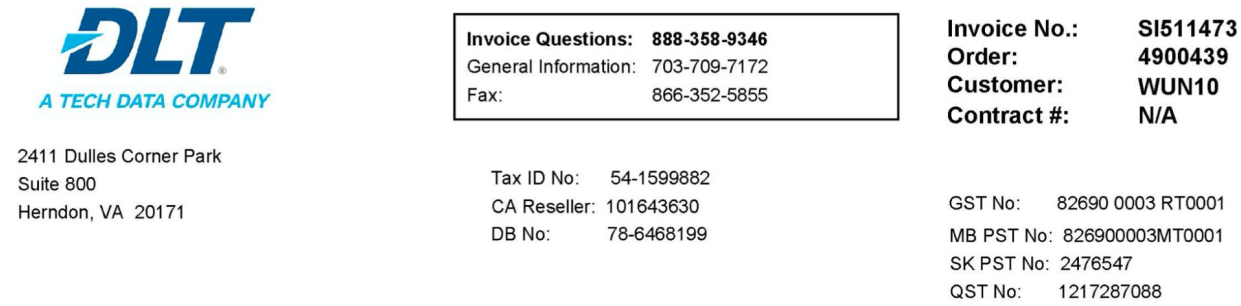


Adding users from a lab to a cloud account

- For AWS, the ability to add users from the lab to access AWS accounts is managed through WUIT (Eric Suiter) by email
  - No way for PI to see what users are currently authorized without asking WUIT
- For GCP, the ability to add users is also limited to WUIT but at least it is possible to view the current configuration



Invoice example



DLT
A TECH DATA COMPANY

2411 Dulles Corner Park
Suite 800
Herndon, VA 20171

Invoice Questions: **888-358-9346**
General Information: 703-709-7172
Fax: 866-352-5855

Tax ID No: 54-1599882
CA Reseller: 101643630
DB No: 78-6468199

Invoice No.: **SI511473**
Order: **4900439**
Customer: **WUN10**
Contract #: **N/A**

GST No: 82690 0003 RT0001
MB PST No: 826900003MT0001
SK PST No: 2476547
QST No: 1217287088

Bill To: Washington University in St. Louis
Accounts Payable
Campus Box 1056
700 Rosedale Ave
SAINT LOUIS, MO 63112-1408

Ship To: Washington University in St. Louis
Accounts Payable
ONCOLOGY / SCB
425 SOUTH EUCLID
SOUTHWEST TOWER
SAINT LOUIS, MO 63108

Date	Period of Performance	ACT # / PDN #		Terms		
02/26/21	01/01/21 - 01/31/21			Net 30 Days		
Purchase Order Number		**Order Date**	**Salesperson**	**Our Order Number**		
2940074K		02/26/21	Jennifer Triplett	4900439		
Quantity		**CLIN No.**	**Item Number**	**Tax**	**Unit Price**	**Amount**
Req.	**Ship**	**B.O.**	**Description**			
1	1		6538-3911-4923 **413500**	N	1,491.27	1,491.27

AWS 01-2021 - WUSTL McDonnell Genome Institute

Invoices will be emailed to you monthly for approval. Work with DLT to ensure the right people get emailed.

- Performance Period
- WASHU PO
- AWS Account #

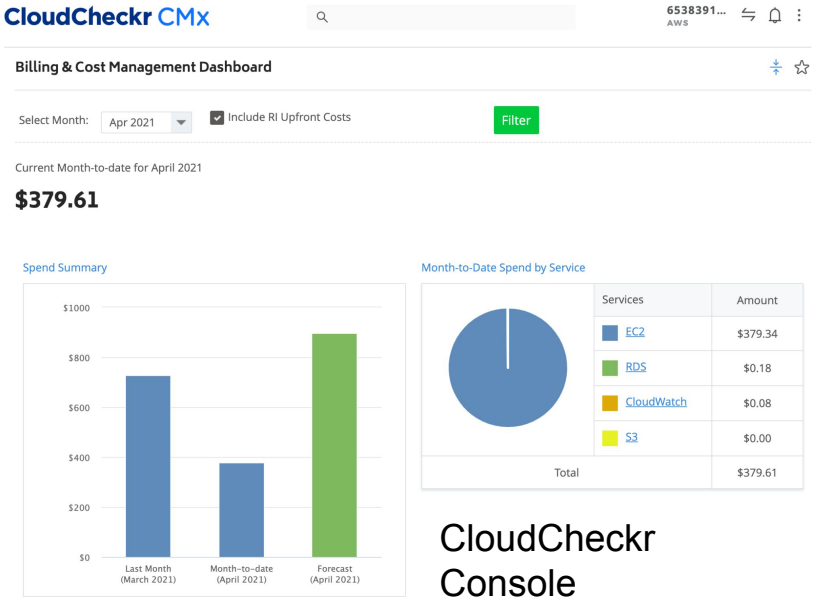
Invoice example

Description	Qty	Rate	Price
4900439 - 6538-3911-4923 - McDonnell Genome Institute - Genome Griffith Lab			
CloudWatch			
Alarm metric month (standard resolution) - US West (Oregon)	2	0.09500	0.19
GB-mo of log storage - US East (Northern Virginia)	0.01	0.00000	-
GB-mo of log storage - US West (Oregon)	0	0.00000	-
CloudWatch Total	**2.01**	**0.03167**	**0.19**
DynamoDB			
GB - first 10 TB / month data transfer out beyond the global free tier	0	0.00000	-
GB - next 40 TB / month data transfer out	0	0.00000	-
DynamoDB Total	**0**	**0.00000**	**-**
EC2			
Elastic IP address not attached to a running instance per hour (prorated)	2231	0.00475	10.60
EUN1-AWS-Out-Bytes in EU (Stockholm)	0	0.00000	-
GB - first 10 TB / month data transfer out beyond the global free tier	1738.81	0.08550	148.67
GB - regional data transfer - in/out/between EC2 AZs or using elastic IPs or ELB	26.48	0.00944	0.25
GB - US West (Oregon) data transfer to Asia Pacific (Mumbai)	0	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Seoul)	0.01	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Singapore)	0.23	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Sydney)	0	0.00000	-
GB - US West (Oregon) data transfer to Asia Pacific (Tokyo)	0.06	0.00000	-
GB - US West (Oregon) data transfer to AWS GovCloud (US)	0.02	0.00000	-
GB - US West (Oregon) data transfer to Canada (Central)	0.23	0.00000	-
GB - US West (Oregon) data transfer to EU (Germany)	0.01	0.00000	-
GB - US West (Oregon) data transfer to EU (Ireland)	1.3	0.01538	0.02
GB - US West (Oregon) data transfer to EU (London)	0	0.00000	-
GB - US West (Oregon) data transfer to EU (Paris)	0	0.00000	-
GB - US West (Oregon) data transfer to South America (Sao Paulo)	0	0.00000	-
GB - US West (Oregon) data transfer to US East (Northern Virginia)	19.4	0.01907	0.37
GB - US West (Oregon) data transfer to US East (Ohio)	0.27	0.03704	0.01
GB - US West (Oregon) data transfer to US West (Northern California)	0	0.00000	-
GB-month of General Purpose SSD (gp2) provisioned storage - US West (Oregon)	2560	0.09500	243.20
GB-Month of snapshot data stored - US West (Oregon)	1824.65	0.04750	86.67
In Asia Pacific (Hong Kong)	0.02	0.00000	-
In Middle East (Bahrain)	0	0.00000	-
On Demand Linux m5.2xlarge Instance Hour	744	0.36480	271.41
On Demand Linux m5a.2xlarge Instance Hour	744	0.32680	243.14
On Demand Linux t2.large Instance Hour	744	0.08816	65.59
On Demand Linux t2.medium Instance Hour	744	0.04409	32.80

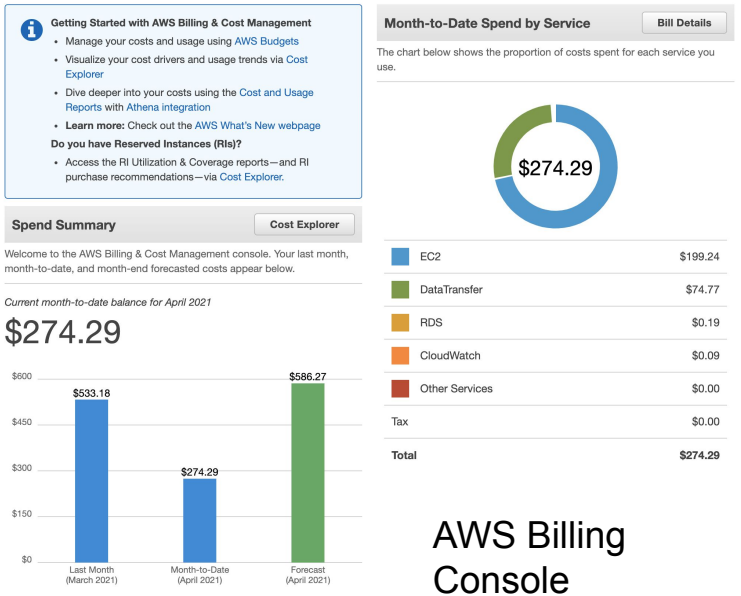
- DLT gives a detailed breakdown of cloud services you are being charged for
- This only loosely corresponds to the billing information you see in AWS console
  - The usage should be the same, but *how* you are billed for it can be very different
  - We have seen DLT bills >2x what is shown in AWS
  - Related to reserved instances. DLT calls this “RI arbitrage”. Basically DLT is reserving instances, keeping the savings from Amazon and billing you the full agreed rate.

CloudCheckr CMx (https://app-us.cloudcheckr.com/)

- More detailed information on what DLT will actually charge you for AWS usage
- Must request that DLT set this up for you



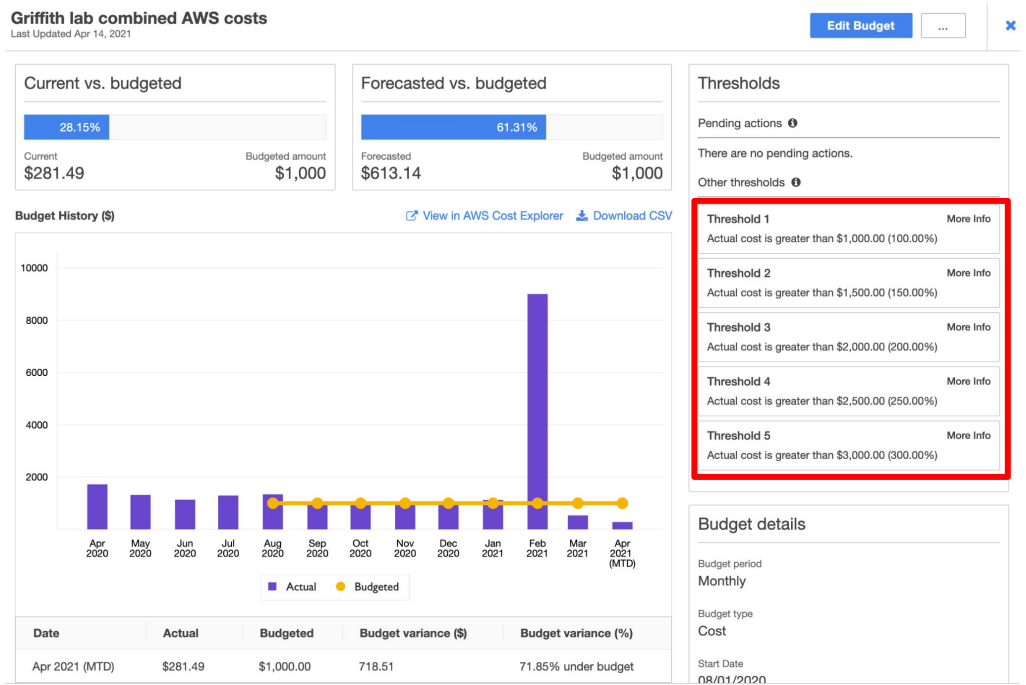
CloudCheckr Console



AWS Billing Console

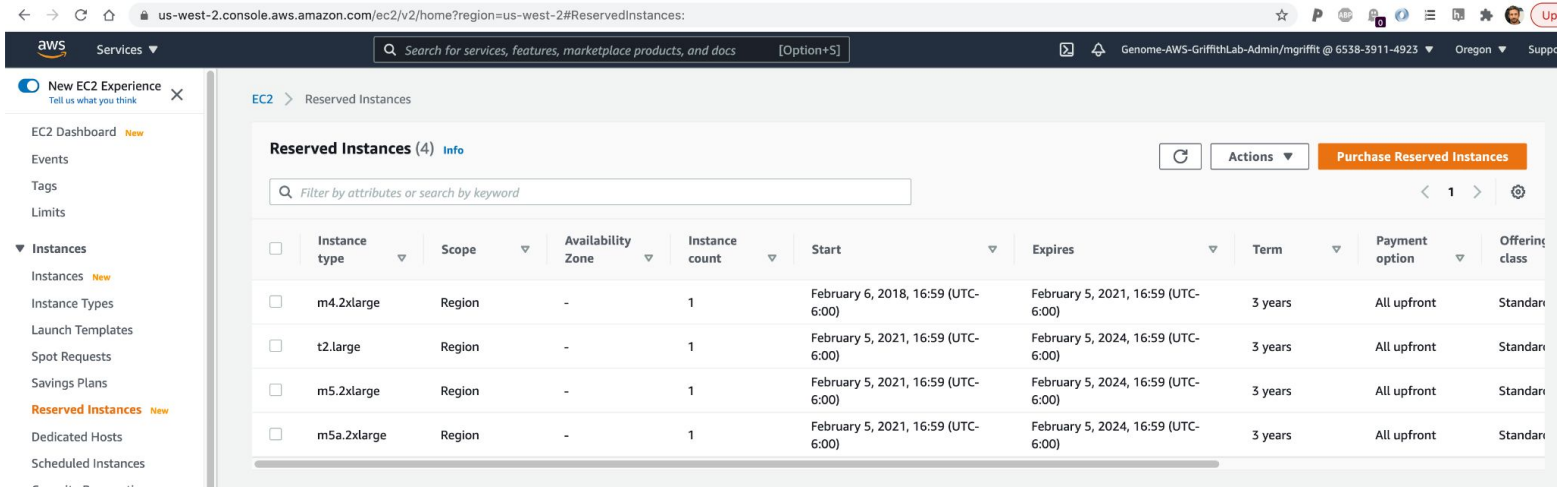
“Budgets” and Billing Alerts

- Creating a “budget” in AWS or GCP allows you to set up email/text alerts when you hit certain amounts each month



Long-Term Compute: “Reserved Instances” (AWS) and “Commitments” (GCP)

- If you have persistent compute services running use of Reserved Instances or Commitments can result in huge savings
  - E.g. by reserving instances for 3 years at a cost of ~$8,000 we project savings of ~$15,000 compared to if we just used on demand resources for that whole time



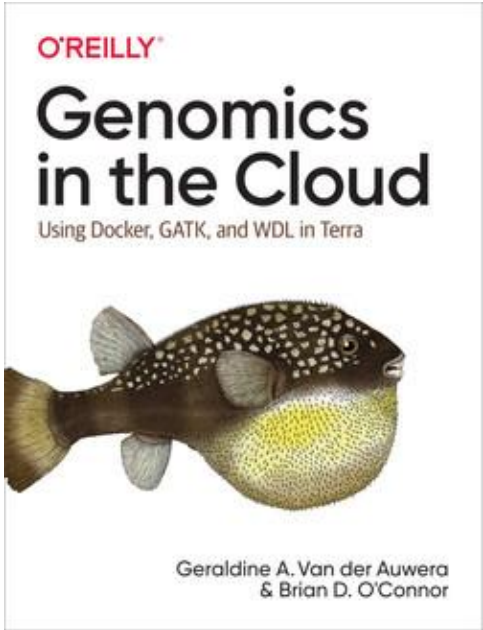
RIS CONFLUENCE LINK (RESTRICTED ACCESS)

Short-Term Compute:
“Spot” (AWS) or “Preemptible” (GCP) Instances

- Meant for “short-term” fault-tolerant jobs
- Much cheaper pricing (up to 90% at times) than regular compute instances
- Drawback is that AWS or GCP reserves the right to take back the machine at any given time depending on the state of the cloud
- Google Preemptible Instances can run at most for 24 hours
- See GCP Preemptible VM Instances or Amazon EC2 Spot for more details
- Often useful for running many batch jobs at scale cheaply

Public Cloud Resources: Education & Training

- Introduction to AWS | Griffith Lab
- Getting Started with AnVIL
- Training | NIH STRIDES
- Google Cloud Online Courses | Coursera
- Learn AWS with Training and Certification | Cloud Skills Courses and Programs
- Genomics in the Cloud [Book]

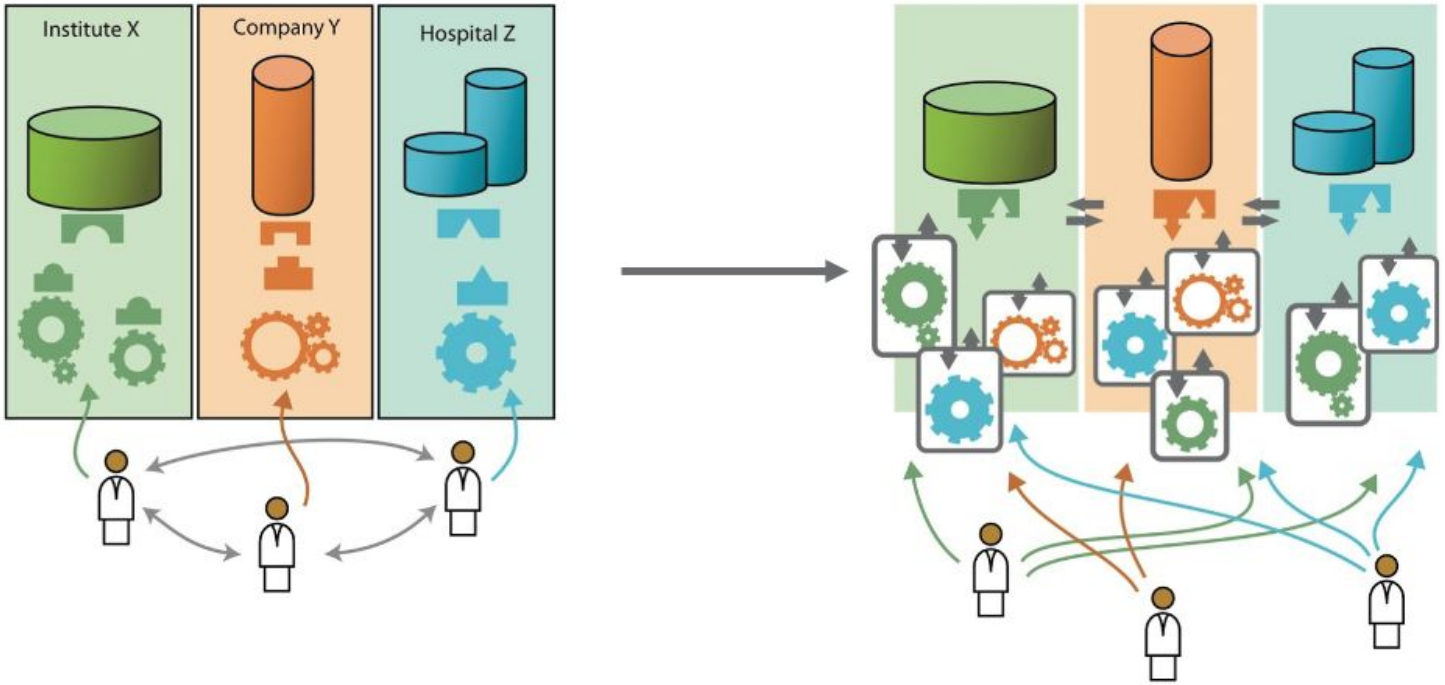


Public Cloud Resources: Funding Opportunities

- Current
  - Enrollment | NIH STRIDES
  - AWS Programs for Research and Education
  - Get and redeem education credits | Google Cloud Billing
- Past
  - AnVIL Cloud Credits Continued Program (AC3) (12/17/2021)

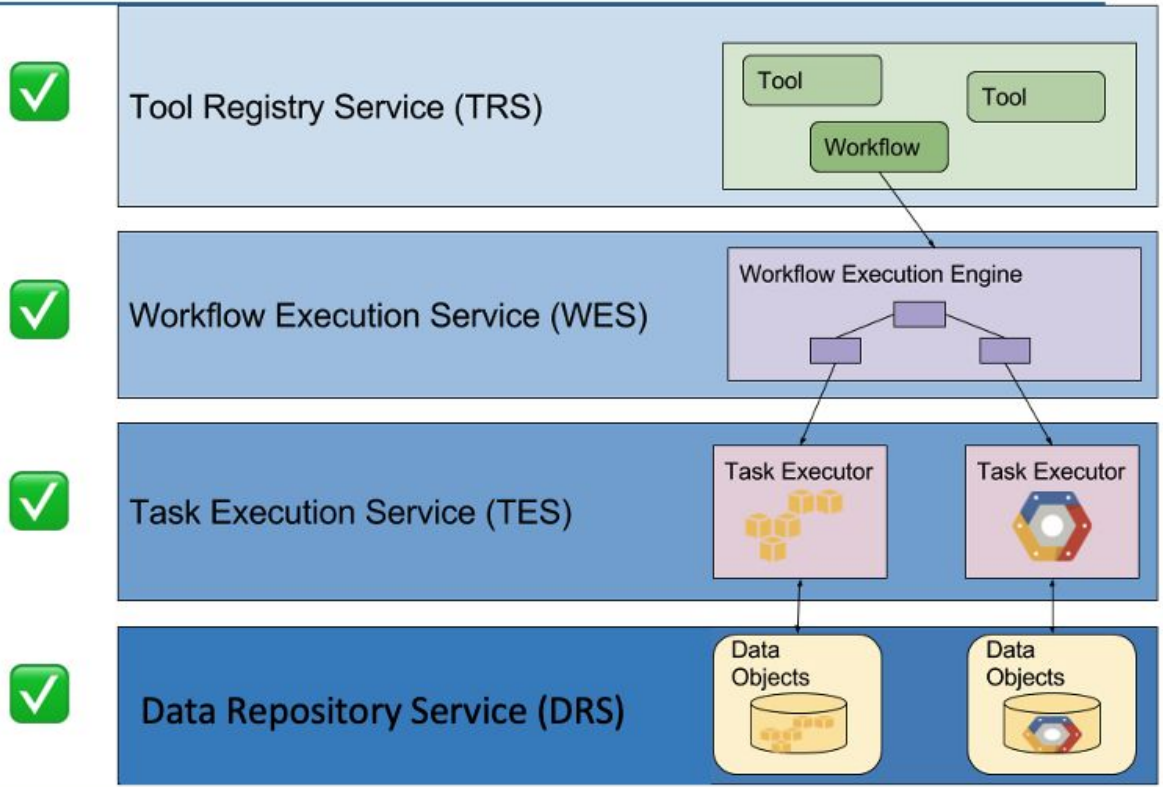
Cloud Work Stream Vision

Global Alliance for Genomics & Health



Cloud Work Stream APIs

Global Alliance for Genomics & Health



Sharing Tools and Workflows

Executing Workflows

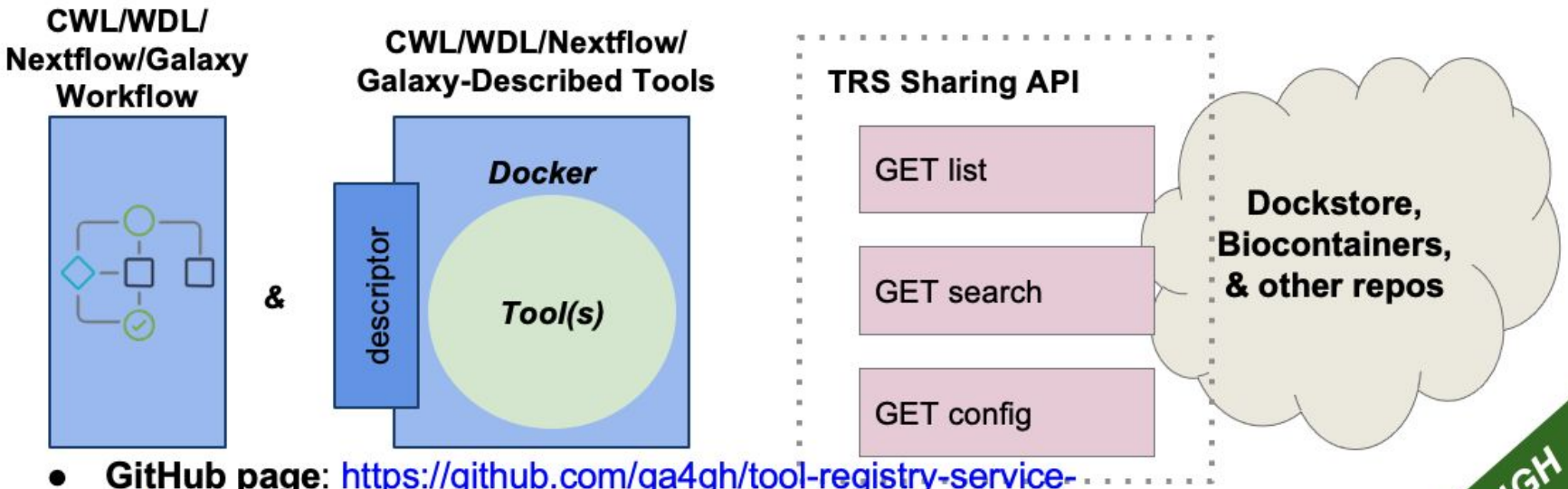
Executing Individual Tasks

Accessing Data

Tool Registry Service (TRS) API

Global Alliance for Genomics & Health

List, search, & config CWL/WDL/Nextflow/Galaxy-described Docker Tools and Workflows.



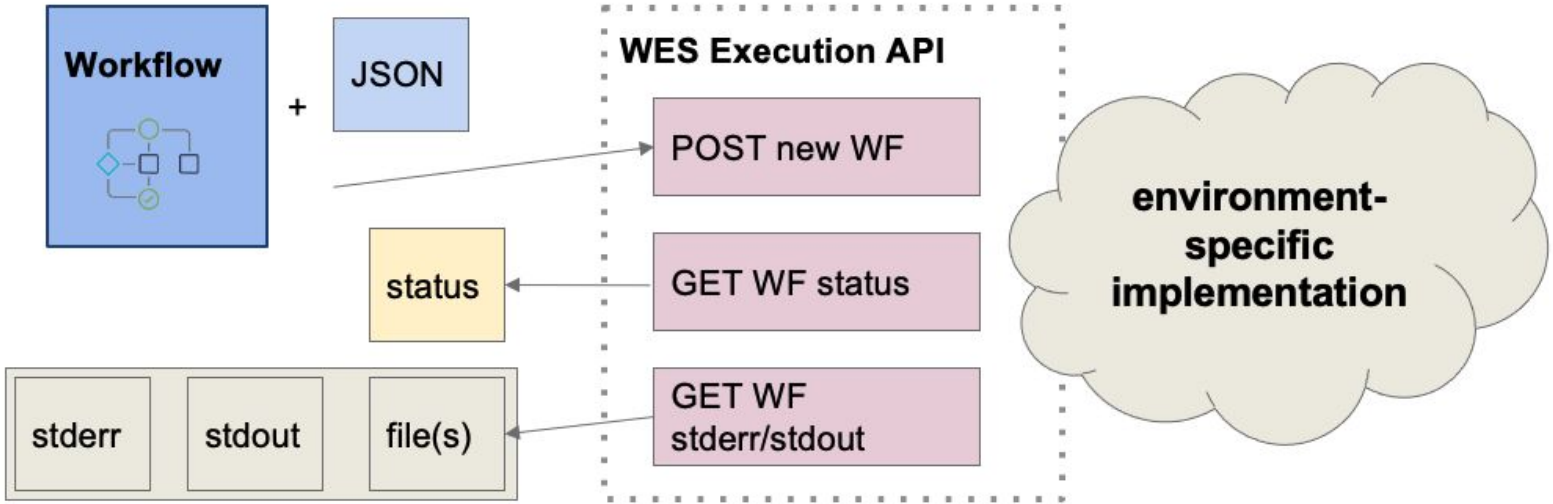
- **GitHub page**: https://github.com/ga4gh/tool-registry-service-schemas
- **Latest release**: 2.0.1-beta.0
- **Champion**: Denis Yuen

Official GA4GH Standard!

Workflow Execution Service (WES) API

Global Alliance for Genomics & Health

Execute Workflows in a cloud and platform-agnostic way. *(TES is very similar!)*



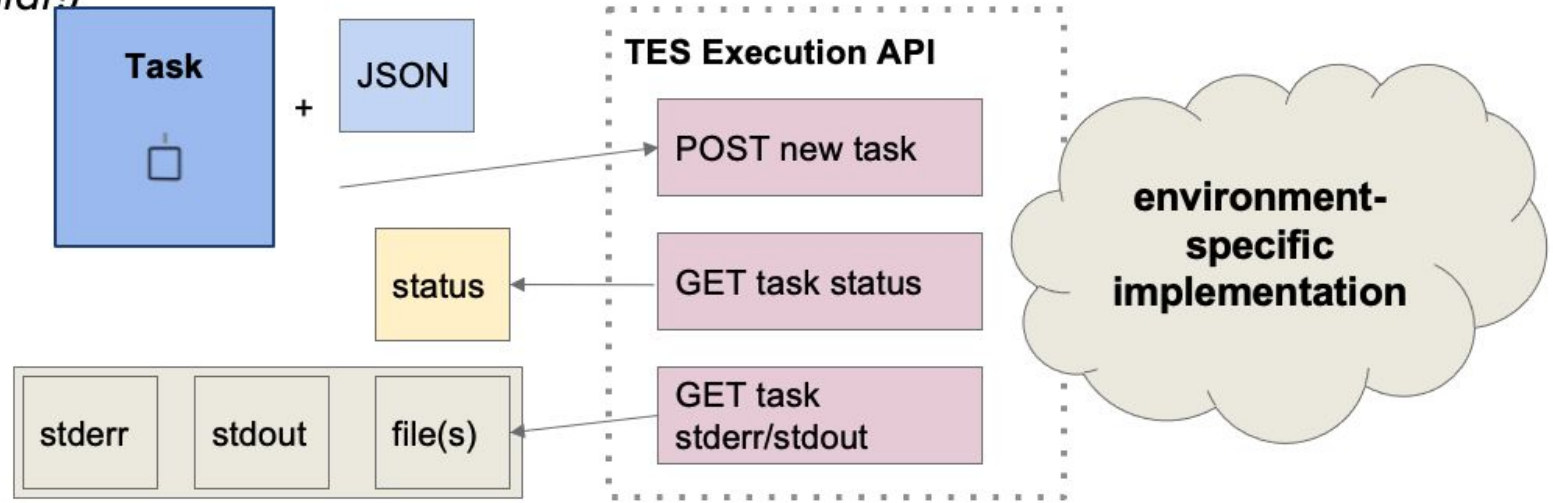
- **GitHub page**: https://github.com/ga4gh/workflow-execution-service-schemas
- **Latest release**: 1.0.0
- **Champions**: James Eddy, Ruchi Munshi (former), and Walt Shands

Official GA4GH Standard!

Task Execution Service (TES) API

Global Alliance for Genomics & Health

Execute individual Tasks in a cloud and platform-agnostic way. *(WES is very similar!)*



- **GitHub page**: https://github.com/ga4gh/wiki/wiki/Task-Execution-Service
- **Latest release**: 1.0.0
- **Champions**: Ania Niewielska and Kyle Ellrott

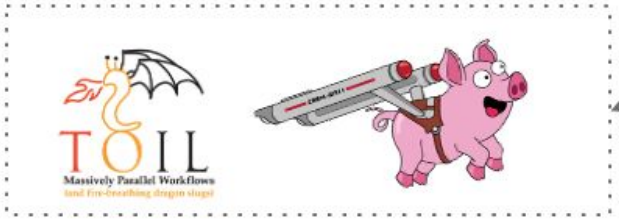
Official GA4GH Standard!

Data Repository Service (DRS) API

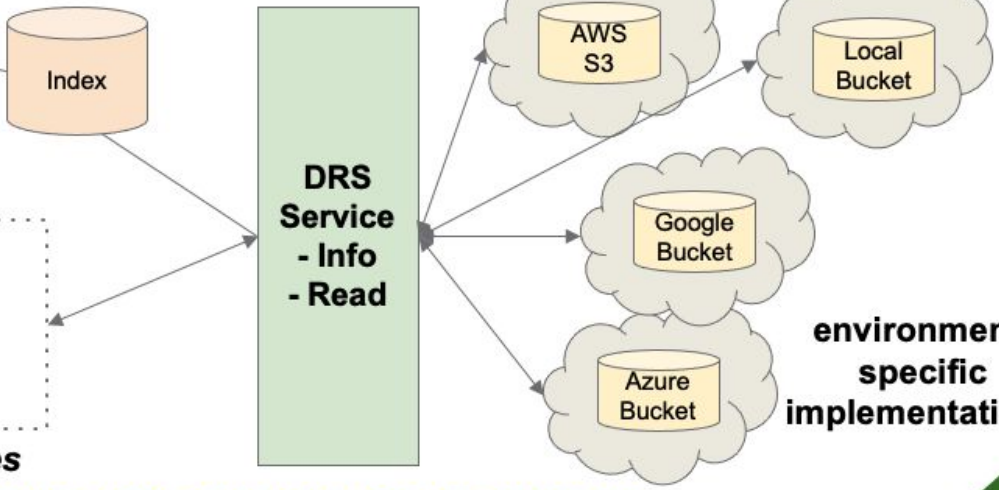
Global Alliance for Genomics & Health

A cloud agnostic way to map IDs to data objects and get info/read via signed URLs or other protocols

Data Browsing Portal



Workflow Engines



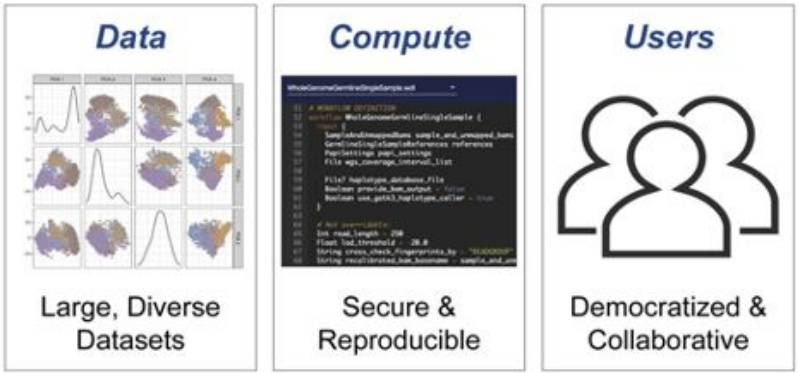
- **GitHub page**: https://github.com/ga4gh/data-repository-service-schemas
- **Latest release**: 1.2.0
- **Champions**: David Glazer and Brian O'Connor

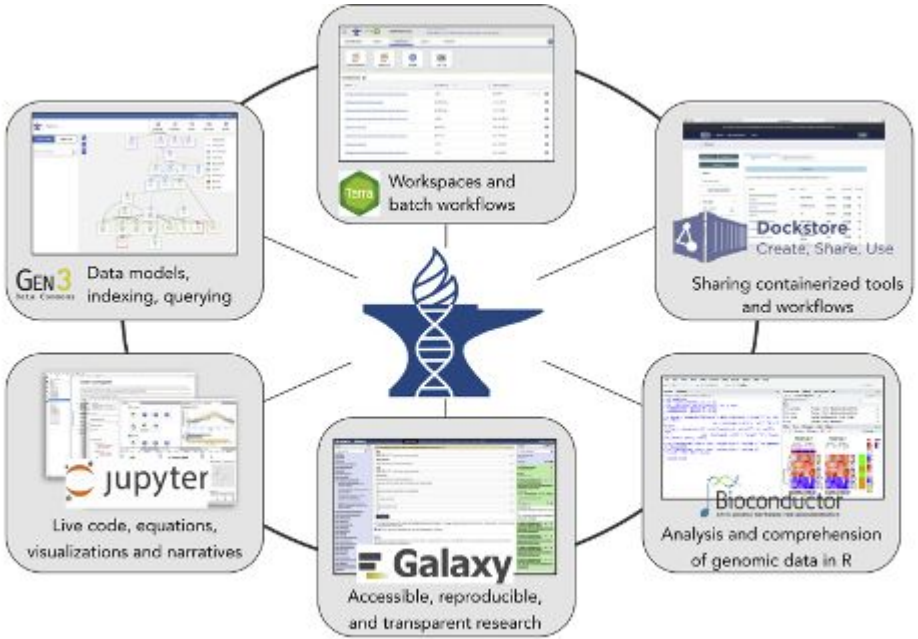
Official GA4GH Standard!

Analysis Visualization and Informatics Lab-space (AnVIL)

Inverting the model of genomics data sharing with the NHGRI Analysis Visualization and Informatics Lab-space (AnVIL)

AnVIL empowers users to collaboratively analyze data in a secure cloud environment





Source: "Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space": https://doi.org/10.1016/j.xgen.2021.100085

NIH Cloud Platform Interoperability (NCPI) Effort

RAS: Researcher Auth Service

Goal: Unified identity/authentication

RAS is an effort by the NIH's Center for Information Technology (CIT) to provide a common mechanism by which researchers can establish their identity and access data they are authorized to use.

DRS: Data Repository Service

Goal: Unified data access across storage infrastructures

Data Repository Service API, are a standardized set of cloud data access methods. The primary functionality is to map a logical ID to a means for physically retrieving the data represented by the `drs://URI` scheme.

FHIR: Fast Healthcare Interoperability Resources

Goal: data harmonization and an API for exchange of electronic medical records.

FHIR facilitates interoperation between health care systems, to make it easy to provide health care information on a wide variety of devices.