GA4GH Standards in Gen3

Gen3 Community Forum
18 September 2024

THE UNIVERSITY OF CHICAGO
CENTER FOR TRANSLATIONAL DATA SCIENCE

OHSU

Australian BioCommons

NeSI
New Zealand eScience Infrastructure

Agenda

- Introduction
- GA4GH Background and Driver Projects
- Gen3 services employing GA4GH standards
- Future standards to be adopted or influenced by Gen3
- DRS statistics
- Gen3 community - using Gen3 with TES

GA4GH Background

- GA4GH unites an international community dedicated to advancing human health through genomic data. They build technical standards and policy frameworks and tools that will expand responsible, voluntary, and secure use of genomic and other related health data.
- Gen3 strives to follow GA4GH standards to enable interoperability with other systems and simplify the use of a Gen3 commons or mesh

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Driver Projects

- A GA4GH Driver Project is an initiative that shapes GA4GH products and applies them to real genomic data.
- Driver Projects range from national health services to rare disease networks to cancer data centres
- Four driver projects are built using Gen3 technology and allow us to both influence and more easily adopt new GA4GH standards.

Biomedical Research Hub
Powered by GEN3

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

NIH National Heart, Lung, and Blood Institute | BioData CATALYST®

NCPI

Gen3 services employing GA4GH standards

Data Repository Service (DRS)

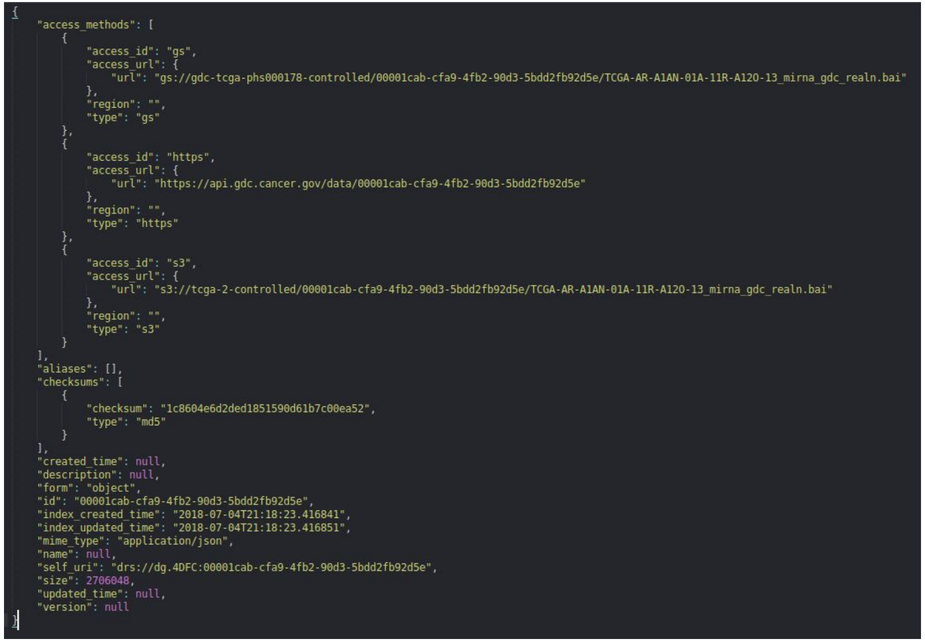
- The Data Repository Service is responsible for hosting data objects.
- Indexd is the Gen3 service that conforms to the DRS standard and returns data objects to the users.
- Currently this is the standard that Gen3 has most influenced. Contributions include prefixes, additional metadata for data objects, and statistics for DRS servers.
- DRS provides a vital service to allow researchers get the data they need as well as for systems to download data for further analysis

Data Repository Service (DRS)

DRS Request:

GET https://nci-crdc.datacommons.io/ga4gh/drs/v1/objects/dg.4DFC/00001cab-cfa9-4fb2-90d3-5bdd2fb92d5e

DRS Output:

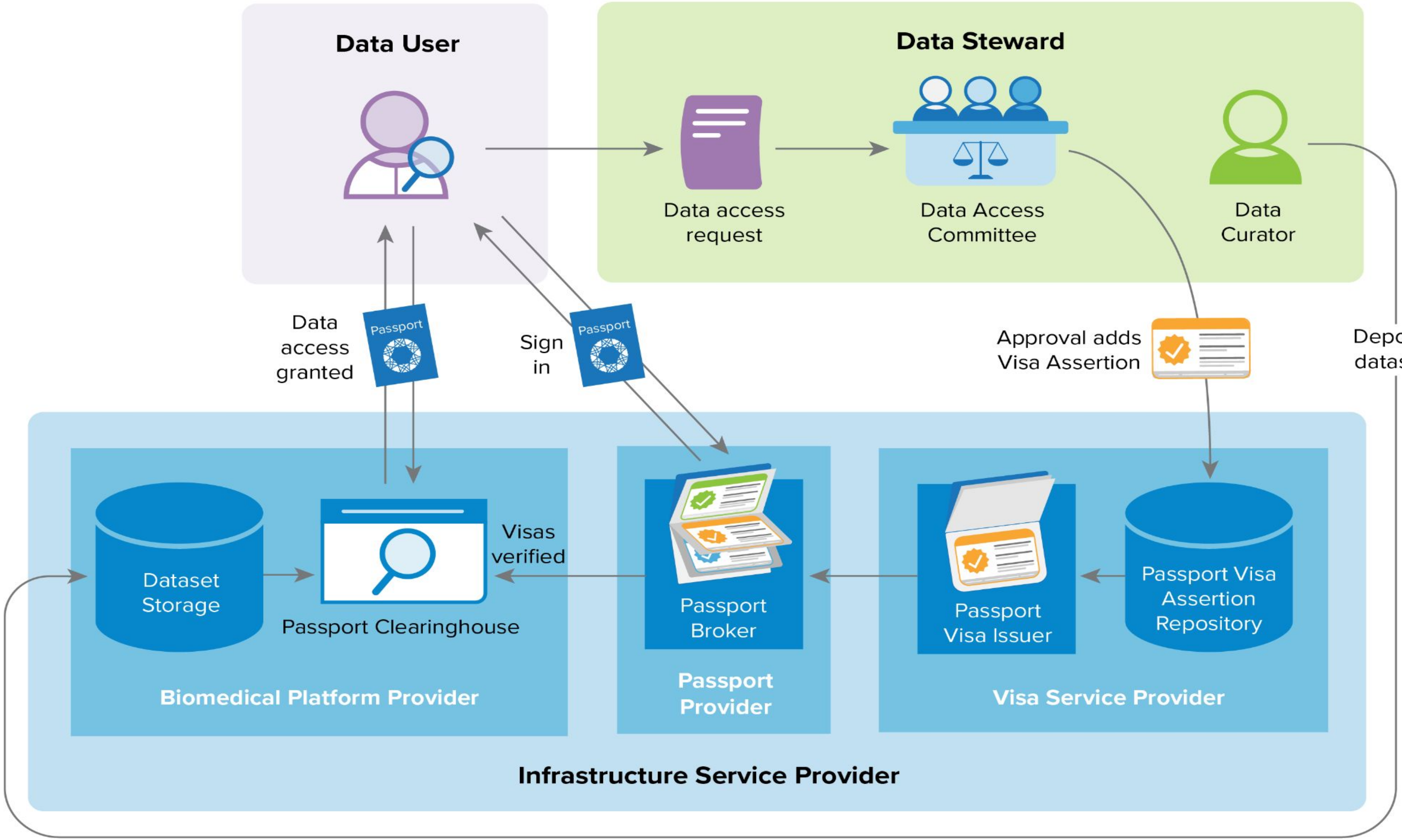


```
{
    "access_methods": [
        {
            "access_id": "gs",
            "access_url": {
                "url": "gs://gdc-tcga-phs000178-controlled/00001cab-cfa9-4fb2-90d3-5bdd2fb92d5e/TCGA-AR-A1AN-01A-11R-A120-13_mirna_gdc_realn.bai"
            },
            "region": "",
            "type": "gs"
        },
        {
            "access_id": "https",
            "access_url": {
                "url": "https://api.gdc.cancer.gov/data/00001cab-cfa9-4fb2-90d3-5bdd2fb92d5e"
            },
            "region": "",
            "type": "https"
        },
        {
            "access_id": "s3",
            "access_url": {
                "url": "s3://tcga-2-controlled/00001cab-cfa9-4fb2-90d3-5bdd2fb92d5e/TCGA-AR-A1AN-01A-11R-A120-13_mirna_gdc_realn.bai"
            },
            "region": "",
            "type": "s3"
        }
    ],
    "aliases": [],
    "checksums": [
        {
            "checksum": "1c8604e6d2ded1851590d61b7c00ea52",
            "type": "md5"
        }
    ],
    "created_time": null,
    "description": null,
    "form": "object",
    "id": "00001cab-cfa9-4fb2-90d3-5bdd2fb92d5e",
    "index_created_time": "2018-07-04T21:18:23.416841",
    "index_updated_time": "2018-07-04T21:18:23.416851",
    "mime_type": "application/json",
    "name": null,
    "self_uri": "drs://dg.4DFC:00001cab-cfa9-4fb2-90d3-5bdd2fb92d5e",
    "size": 2706048,
    "updated_time": null,
    "version": null
}
```

Passports and Visas

- The Passports and Visas standards handles authentication and authorization of data across platforms. Each user has a passport and platforms can grant visas to a user's passport (via a clearing house). A user can then visit different platforms and use their visas to essentially bring data access with them.
- Gen3 conforms to NIH RAS (Researcher Auth Service) which is a modified version of Passports and Visas.
- Currently NIH RAS and Passports are different standards. There is work within GA4GH and the NIH to bring these standards to a unified standard.

Gen3 services employing GA4GH standards



Data Connect

- Data Connect is a standard for discovery and search of biomedical data. It provides mechanisms for describing data and its data model along with searching the data with the given data model.
- Gen3 has a similar service, the Metadata Service (MDS) that handles arbitrary metadata for data objects. This means that for any object in indexd (DRS) you can add additional metadata.
- Work is being done to add the additional features of Data Connect into MDS to make the Gen3 Metadata Service Data Connect compliant.

Data Connect

- Data connect uses the concept of tables where each table has its own schema to which all data within the table conforms.
- Gen3 will modify the MDS so each data object will have its own table and thus each data object will have arbitrary metadata associated with it.

Future standards to be adopted or influenced by Gen3

Portable Format for Bioinformatics (PFB)

- PFB is a file format developed for Gen3 System interoperation. It is contains a self describing schema with serialized data. We have been coming up with a plan to introduce PFB to GA4GH as a new standard. We still need to identify a workstream that it fits into and when to introduce it.

DRS Statistics

- Indexd has had a stats endpoint for many years. This endpoint will give you total file size (bytes) and file count of all objects in an Indexd server.
- The cloud workstream noticed this endpoint and decided that it would be a beneficial addition to the standard. We made a PR to the DRS standard and it will be voted on in the current plenary in Melbourne.

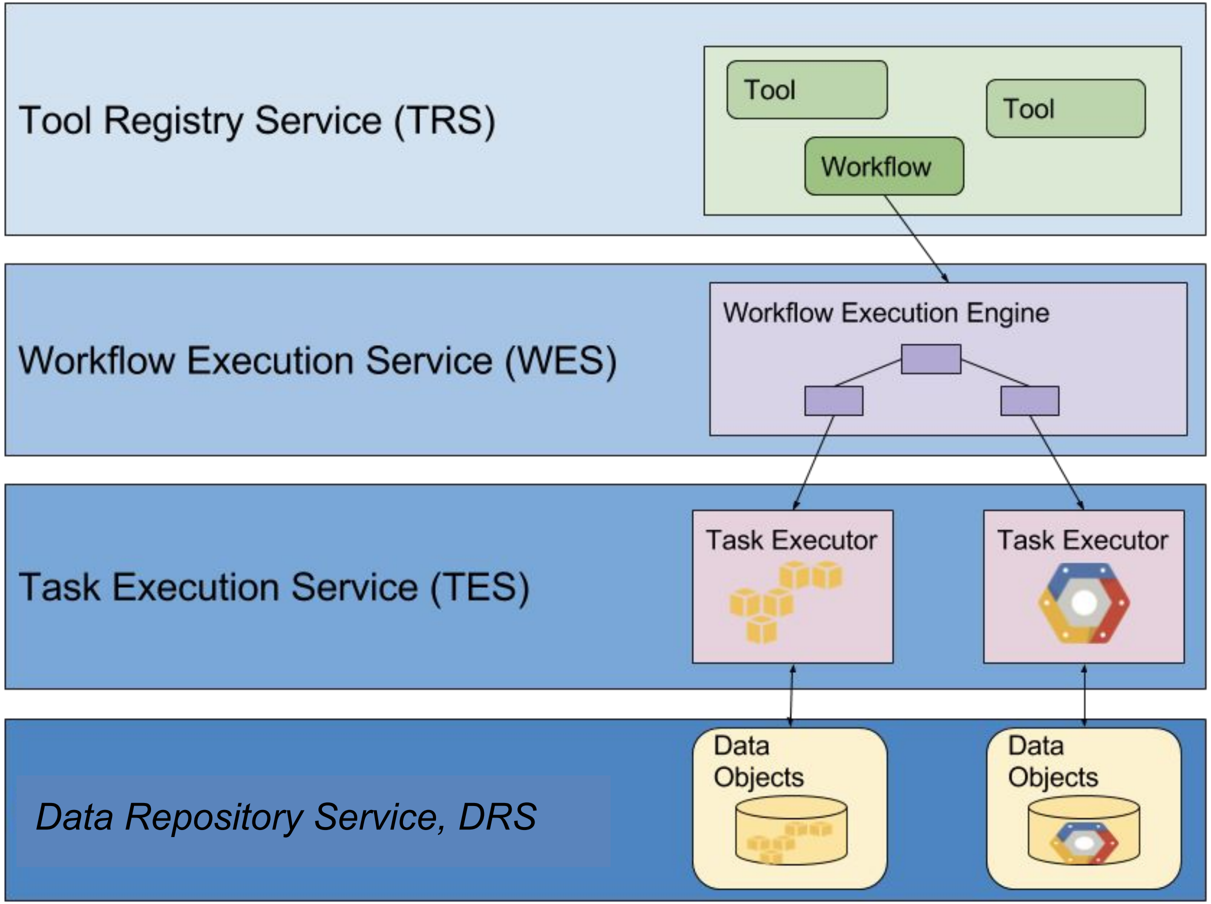
Task Execution Service

- The Task Execution Service (TES) is a standard for defining a schema and API for describing batch execution tasks
- Each Task can be thought of a computational process with input files and commands that are run on a docker container.
- TES goes hand in hand with WES (workflow execution service) where each task of a workflow can be executed by TES.
- Future benefits to users
    - Allow TES-based workflows to be run on Gen3
    - Provides visibility into progress of workflows
    - Enables Gen3 to offer dashboard control and monitoring of workflows

Gen3 Community Presentation
Using Gen3 with TES

Kyle Ellrott
Oregon Health & Science University

GA4GH Cloud Work Stream APIs



Tool Registry Service (TRS)

Tool

Tool

Workflow

Workflow Execution Service (WES)

Workflow Execution Engine

Task Execution Service (TES)

Task Executor

Task Executor

Data Repository Service, DRS

Data Objects

Data Objects

Sharing Tools and Workflows

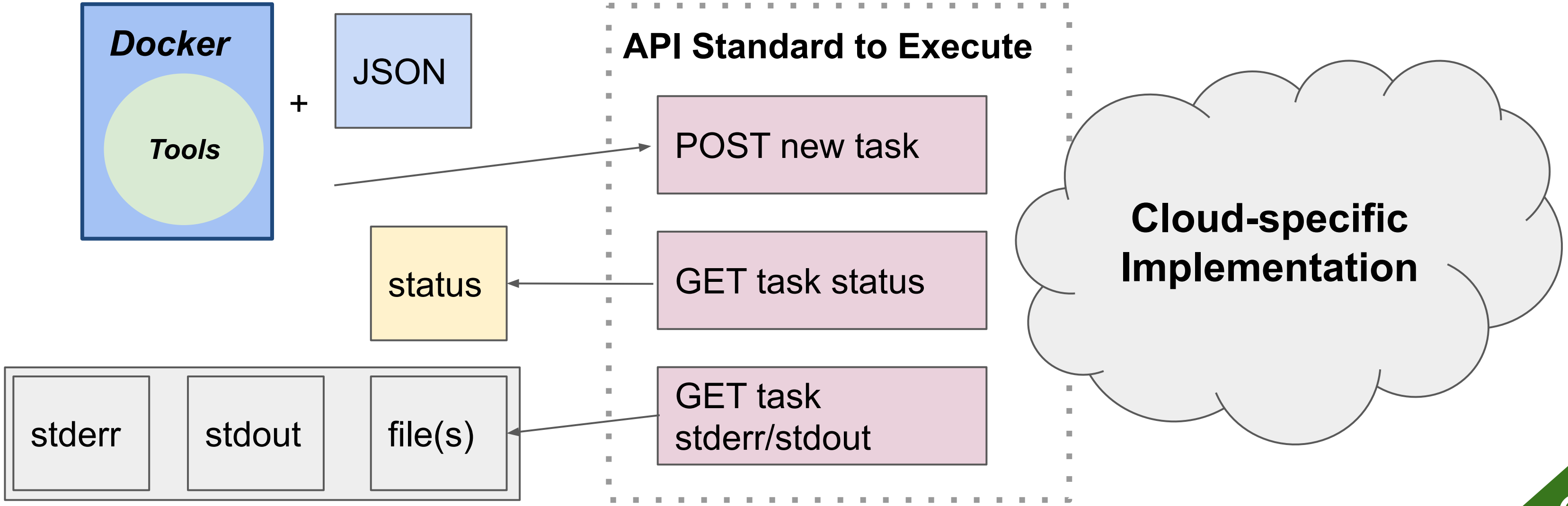
Executing Workflows

Executing Individual Tasks

Accessing Data

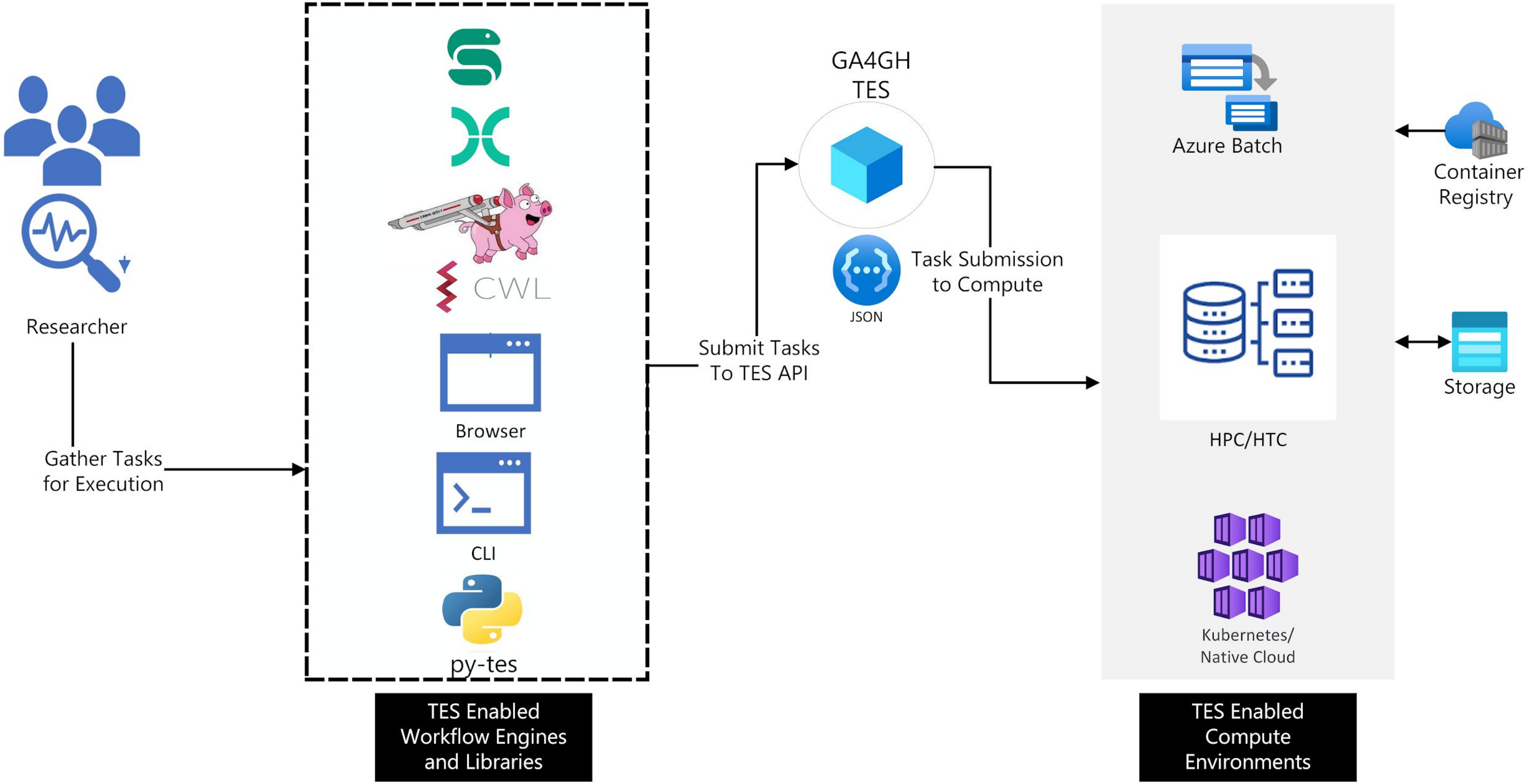
Task Execution Service (TES) API

A way to send a request to run a Docker-based tool in a remote environment, monitor progress, and retrieve the result. *(TES does tasks, WES does workflows)*



- **GitHub page**: https://github.com/ga4gh/task-execution-schemas
- **Latest release**: 1.1.0

Official GA4GH Standard!



Researcher

Gather Tasks for Execution

Browser

CLI

py-tes

TES Enabled Workflow Engines and Libraries

Submit Tasks To TES API

GA4GH TES

JSON

Task Submission to Compute

Azure Batch

HPC/HTC

Kubernetes/ Native Cloud

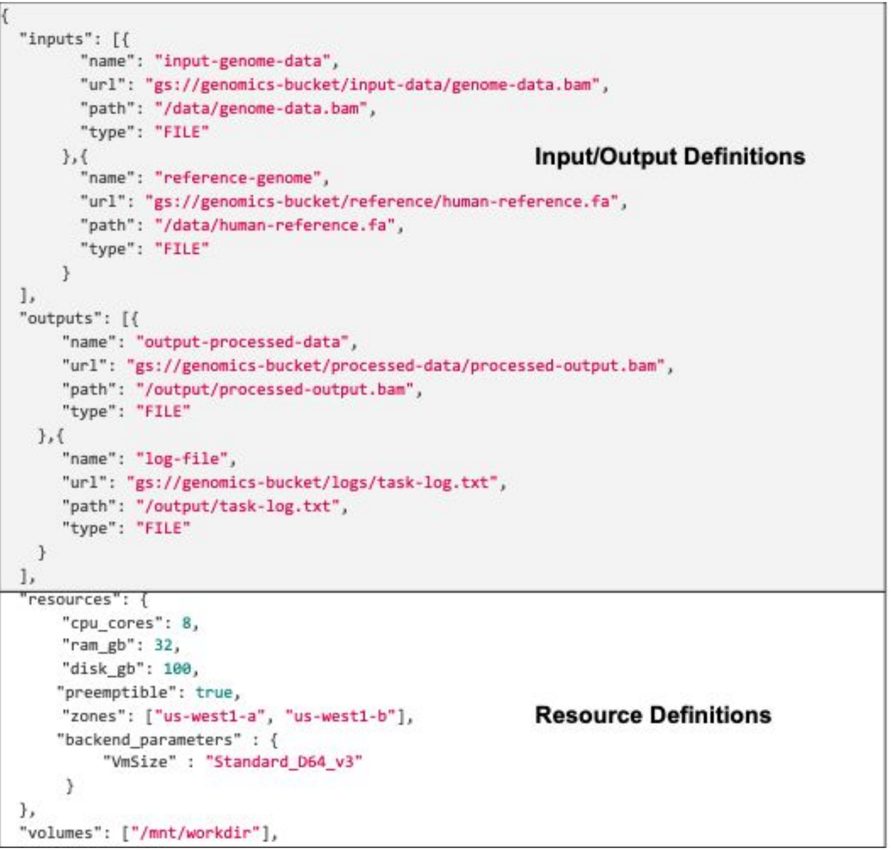
TES Enabled Compute Environments

Container Registry

Storage

A TES Request

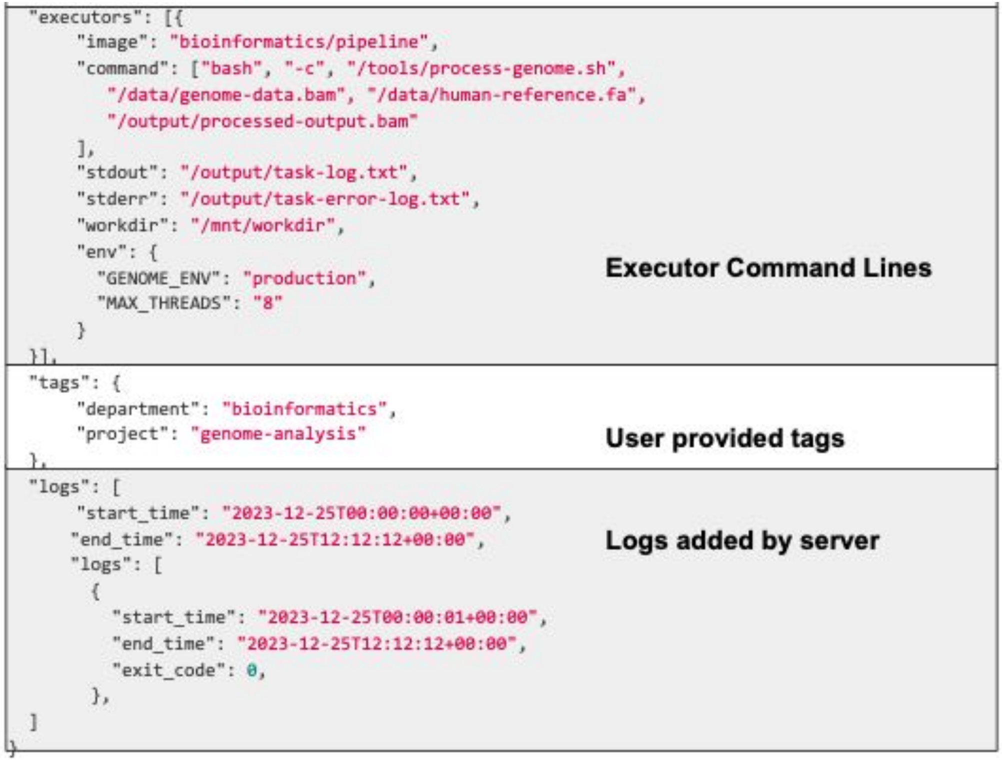
GEN3 DATA COMMONS



```
{
  "inputs": [{
        "name": "input-genome-data",
        "url": "gs://genomics-bucket/input-data/genome-data.bam",
        "path": "/data/genome-data.bam",
        "type": "FILE"
      },{
        "name": "reference-genome",
        "url": "gs://genomics-bucket/reference/human-reference.fa",
        "path": "/data/human-reference.fa",
        "type": "FILE"
      }
  ],
  "outputs": [{
      "name": "output-processed-data",
      "url": "gs://genomics-bucket/processed-data/processed-output.bam",
      "path": "/output/processed-output.bam",
      "type": "FILE"
    },{
      "name": "log-file",
      "url": "gs://genomics-bucket/logs/task-log.txt",
      "path": "/output/task-log.txt",
      "type": "FILE"
    }
  ],
  "resources": {
      "cpu_cores": 8,
      "ram_gb": 32,
      "disk_gb": 100,
      "preemptible": true,
      "zones": ["us-west1-a", "us-west1-b"],
      "backend_parameters" : {
          "VmSize" : "Standard_D64_v3"
      }
  },
  "volumes": ["/mnt/workdir"],
```

Input/Output Definitions

Resource Definitions



```
  "executors": [{
      "image": "bioinformatics/pipeline",
      "command": ["bash", "-c", "/tools/process-genome.sh",
         "/data/genome-data.bam", "/data/human-reference.fa",
         "/output/processed-output.bam"
      ],
      "stdout": "/output/task-log.txt",
      "stderr": "/output/task-error-log.txt",
      "workdir": "/mnt/workdir",
      "env": {
        "GENOME_ENV": "production",
        "MAX_THREADS": "8"
      }
  }],
  "tags": {
      "department": "bioinformatics",
      "project": "genome-analysis"
  },
  "logs": [
      "start_time": "2023-12-25T00:00:00+00:00",
     "end_time": "2023-12-25T12:12:12+00:00",
     "logs": [
       {
         "start_time": "2023-12-25T00:00:01+00:00",
         "end_time": "2023-12-25T12:12:12+00:00",
         "exit_code": 0,
       },
  ]
}
```

Executor Command Lines

User provided tags

Logs added by server

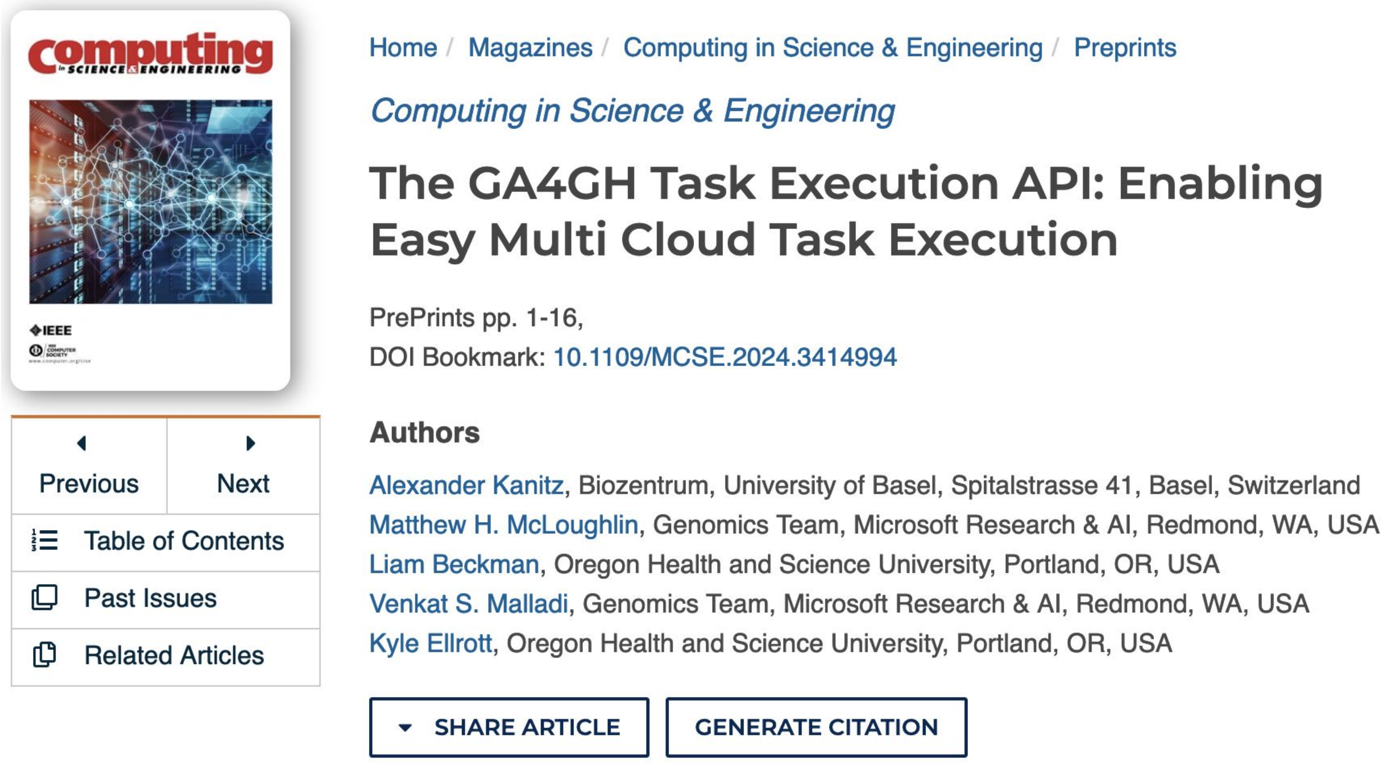
TES Ecosystem - Servers

Funnel	TES server implementation for HPC/HTS systems including AWS Batch, Google Cloud, Kubernetes, Slurm, GridEngine and HTCondor	github.com/ohsu-comp-bio/funnel
Pulsar	TES server implementation for the Galaxy/Pulsar federated distributed network	pulsar.readthedocs.io
TES-Azure	TES server implementation for the Microsoft Azure	github.com/microsoft/ga4gh-tes
TESK	TES server implementation for Kubernetes/ Native Cloud systems	github.com/elixir-cloud-aai/TESK
proTES	Proxy service for injecting middleware into GA4GH TES requests	github.com/elixir-cloud-aai/proTES

TES Ecosystem - Clients

Project	Description	Source
Cromwell	Workflow management system for executing composed in the Workflow Definition Language (WDL) domain-specific language (DSL)	cromwell.readthedocs.io
cwl-tes	Workflow management system for executing workflows in the Common Workflow Language (CWL) DSL	github.com/ohsu-comp-bio/cwl-tes
ELIXIR Cloud Components	Web Component library for interacting with TES services (and other GA4GH APIs)	elixir-cloud-components.vercel.app
Nextflow	Workflow management system for executing workflows composed in the Nextflow DSL	nextflow.io
py-tes	Python client library for interacting with TES services	github.com/ohsu-comp-bio/py-tes
Snakemake	Workflow management system for executing workflows composed in the Snakemake DSL	snakemake.github.io
Toil	Workflow management system for executing workflows composed in the Toil and CWL DSLs	toil.readthedocs.io

TES Paper in Pre-print



Home / Magazines / Computing in Science & Engineering / Preprints

Computing in Science & Engineering

The GA4GH Task Execution API: Enabling Easy Multi Cloud Task Execution

PrePrints pp. 1-16,
DOI Bookmark: 10.1109/MCSE.2024.3414994

Authors

Alexander Kanitz, Biozentrum, University of Basel, Spitalstrasse 41, Basel, Switzerland
Matthew H. McLoughlin, Genomics Team, Microsoft Research & AI, Redmond, WA, USA
Liam Beckman, Oregon Health and Science University, Portland, OR, USA
Venkat S. Malladi, Genomics Team, Microsoft Research & AI, Redmond, WA, USA
Kyle Ellrott, Oregon Health and Science University, Portland, OR, USA

Gaps

Authentication and authorization

- Current implementations provide basic auth
    - Funnel not connected to user database
- Credential handoff
    - How do things like S3 bucket credentials get handed off?
        - TESK and Funnel use pre-configured 'global configs' for storage access

Questions and Discussion

Acknowledgements

- Speakers
  - Robert Grossman - Center for Translational Data Science, University of Chicago
  - Michael Lukowski - Center for Translational Data Science, University of Chicago
  - Kyle Ellrott - Oregon Health and Science University
- Forum Support
  - Sara Volk de Garcia - Center for Translational Data Science, University of Chicago
  - Fay Booker - Center for Translational Data Science, University of Chicago
- Gen3 Forum Steering Committee
  - Robert Grossman - Center for Translational Data Science, University of Chicago
  - Steven Manos - Australian BioCommons
  - Claire Rye - New Zealand eScience Infrastructure
  - Plamen Martinov - Open Commons Consortium
  - Michael Fitzsimons - Center for Translational Data Science, University of Chicago