Arvados Technical Overview

Last Update: March 2021 for Arvados 2.1
Latest Version: http://arvados.org/technology/Arvados_Technical_Overview.pdf

Curii Corporation

- Accelerating biomedical research and precision medicine through open standards, open data, and free and open source software
- Core Competencies
  - Arvados development and support
  - Fast querying and machine learning on large genomic datasets
  - Assisting projects that focus on open biomedical data and standards (e.g. Personal Genome Project, Common Workflow Language, Genome In A Bottle)
- Current customers
  - Large pharmaceutical organizations, small biotechs, and individual investigators
- Founded by George Church, Sarah Wait Zaranek, Alexander (Sasha) Wait Zaranek, Tom Clegg and Ward Vandewege

curii

The Arvados Platform

- Open source platform for managing, processing, and sharing genomic and other large scientific and biomedical data
- Runs anywhere
  - Run in the cloud (e.g. Azure, AWS) as well as on-premises and hybrid clusters
  - Run on distributed data using federated multi-cluster workflows
- Large scale
  - Manage petabytes of data and use thousands of cores simultaneously
- Provenance and reproducibility
  - Repeat computations, reproduce results, identify the origin and verify the content of every dataset

curii

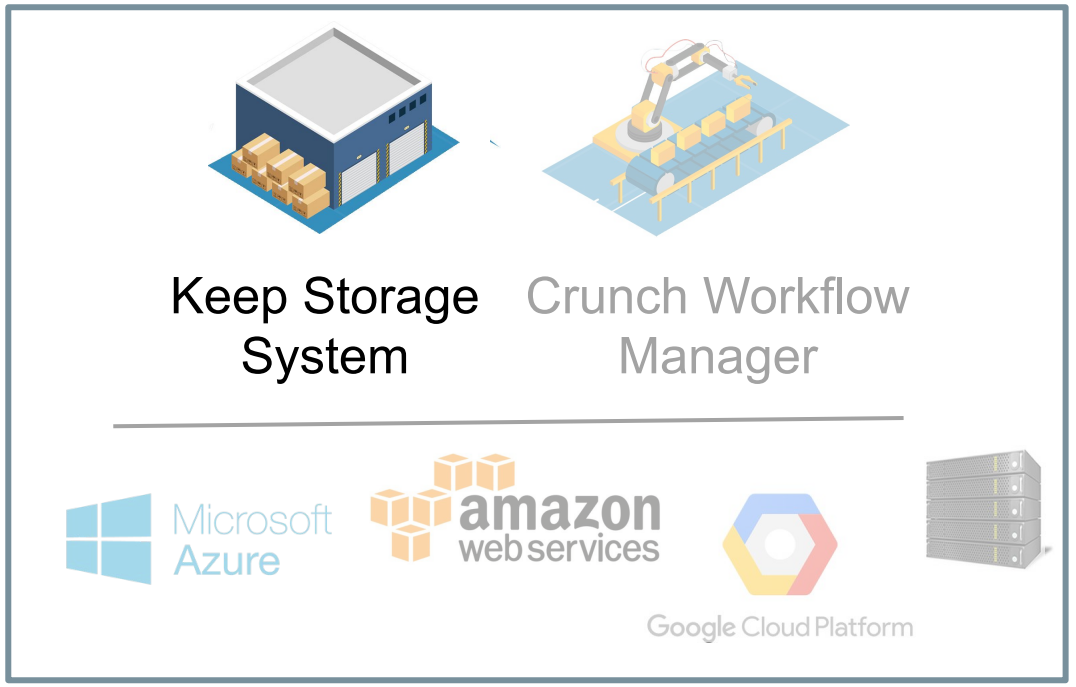
Expand Your Capabilities With Arvados

- Manage large datasets
  - Identify collections of files by name, uuid, or content address (hash)
  - Easily ascertain data origin and validate data content
  - Control access and share data between users and groups
- Run complex biomedical analysis workflows
  - Scale across multiple compute nodes
  - Repeat computations and reproduce results
- Build secure applications
  - Integrate with your existing infrastructure
  - Leverage APIs available for Arvados capabilities

curii

Arvados Core Components: Keep

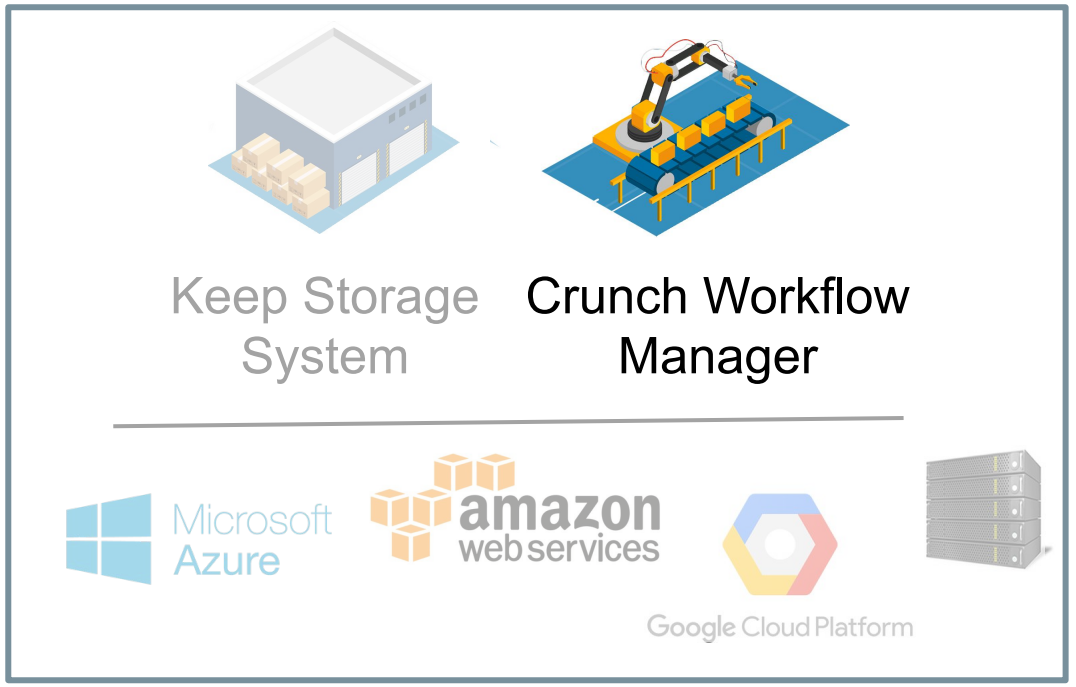
- Organize GB-PBs of data in virtual folders called collections
  - Add and query metadata
  - Track collection history
- Content addressable storage system
  - Validate and check data integrity
  - Automatic deduplication
  - Cheap to create, copy, edit, and delete collections
- Variety of access options
  - Arvados API, mounted as a filesystem, over HTTP(S), S3-API compatible endpoint



curii

Arvados Core Components: Crunch

- Manages reproducible complex computational workflows at scale
  - Dispatches to cloud or on-prem (e.g. Slurm)
  - Runs workflow step in containers (e.g. Docker)
  - Scales compute on demand in cloud
  - Optimizes compute costs by re-using past results whenever possible
  - Tracks resource usage (RAM, CPU, I/O) to reveal bottlenecks and underused resources
- Track input and output data through Keep
  - Input and output files, logs, and container images recorded with immutable hash-based identifiers



curii

Common Workflow Language (CWL)

- Community developed open standard for describing computational data-analysis workflows
- Native workflow language for Arvados
- Designed to make workflows portable and scalable across a variety of software and hardware environments
- Focused particularly on serving the data-intensive sciences

Participating Organizations

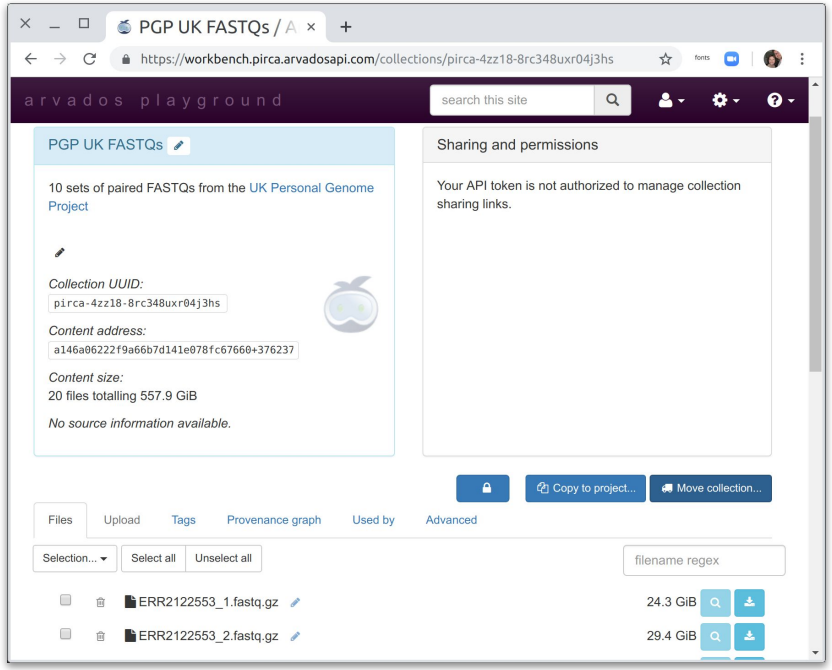
- Arvados Project
- Curii
- Seven Bridges Genomics
- Galaxy Project
- Apache Taverna
- Institut Pasteur
- Wellcome Trust Sanger Institute
- University of California Santa Cruz
- Harvard T.H. Chan School of Public Health
- Cincinnati Children's Hospital Medical Center
- Broad Institute
- University of Melbourne Center for Cancer Research
- Netherlands eScience Center
- Texas Advanced Computing Center Life Science Computing Group / Agave Platform
- CyVerse
- Institute for Systems Biology
- ELIXIR Europe
- BioExcel CoE
- BD2K
- EMBL Australia Bioinformatics Resource
- IBM Spectrum Computing
- DNAnexus
- CERN

curii

Arvados Working Environment

- Workbench web application to interactively access functionality
  - Query and browse data
  - Visualize provenance
  - Track progress of workflows
  - Run workflows

- Command line tools available for all interactive functionality
  - For power users, batch operations, and automation




```
$ arvados-cwl-runner wgs-processing-wf.cwl wgs-inputs.yml
```

curii

Working with Arvados: SDK and APIs

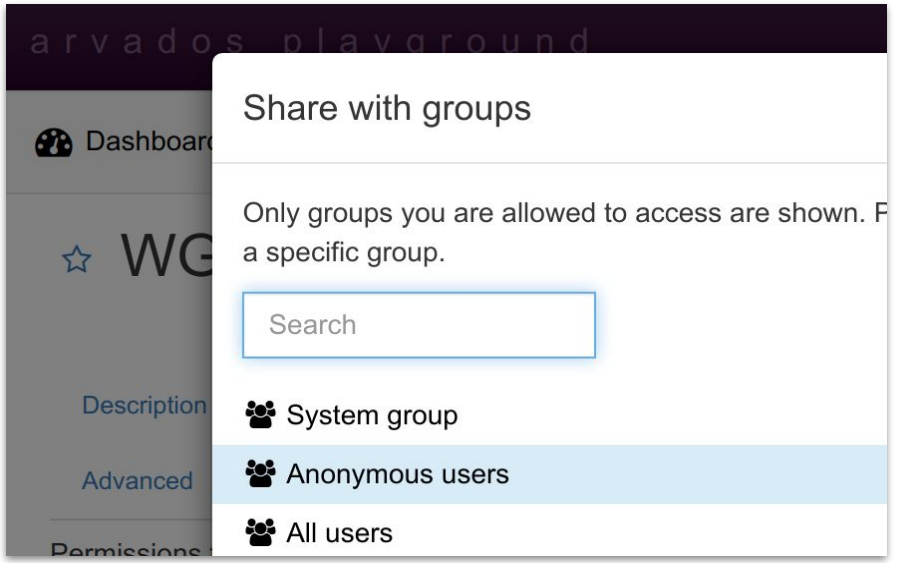
- Integrate or develop applications on top of Arvados using SDKs and APIs
- All services in Arvados are accessible through RESTful APIs
- SDKs available for Python, Go, R, Ruby, and Java
- WebDAV and S3 API provides read/write access to files stored in Keep

```
import arvados
import arvados.collection
api = arvados.api()
container_request_uuid="su92l-xvhdp-lgxv1nk09d6j1bl"
container_request = api.container_requests().get(uuid=co
collection = arvados.collection.CollectionReader(contain
print(collection.open("cwl.output.json").read())
```

curii

Security and Access Control

- All endpoints secured by access tokens
- Encryption in transit (TLS)
- Encryption at rest (encrypted disks or buckets)
- Supports various user authentication methods
  - Lightweight Directory Access Protocol (LDAP)
  - OpenID Connect (OAuth2)
  - Portable Authentication Modules (PAM)
- Control sharing of projects with other users or groups
  - Private by default
  - Read-only, read/write, or manage (to grant permission to others)

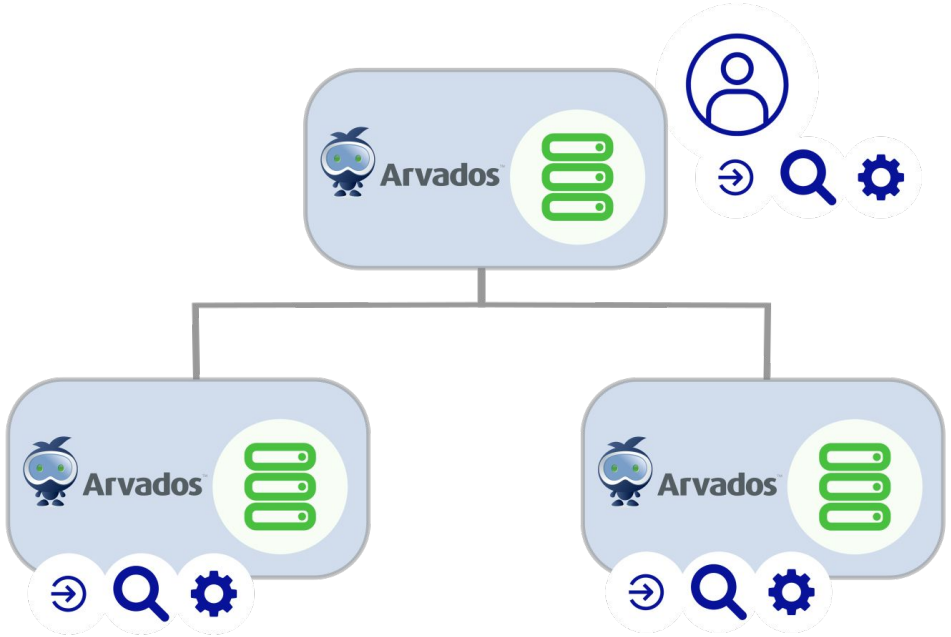


curii

Federated Computing with Arvados

- Federation is the ability to transparently access, create and manipulate data across Arvados clusters in different regions and organizations

- With federation, users can:
  - Access federated clusters with consistent identity and credentials
  - Search and access data hosted on federated clusters
  - Run workflows across clusters (minimize data transfer costs or meet regulatory compliance)

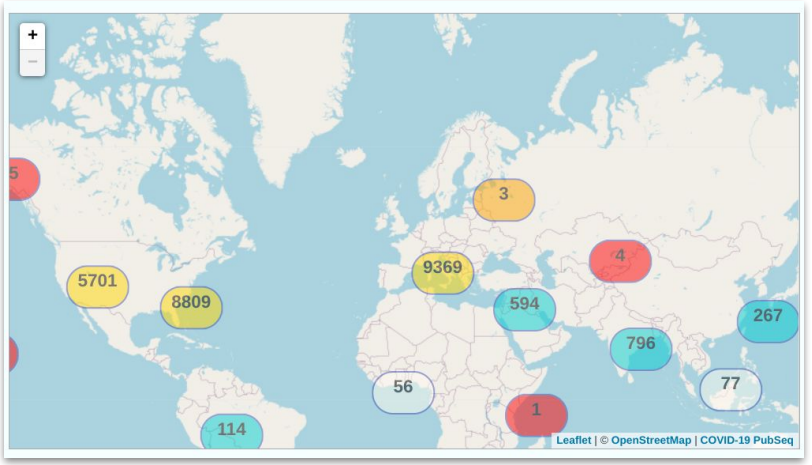


curii

Use Case: Public SARS-CoV-2 Sequence Resource

- Free and open online bioinformatics public sequence resource
  - Rapid analysis of sequenced SARS-CoV-2 samples allowing for a quick turnaround in identification of new virus strains
  - Standardised workflows triggered on upload of raw data from a sequencer
  - Results immediately available with metadata (30,000+ public sequences available)
  - http://covid19.genenetwork.org/
- Created at the Covid19 2020 BioHackathon
  - Arvados (on AWS) instance provided for participants
- Using Arvados:

  Idea → Working prototype in 5 days



curii

The Arvados Platform

- **Flexibility:** Run anywhere — in the cloud, as well as on premise and hybrid clusters
- **Freedom:** 100% free and open software that you can control
- **Scale:** Manage PBs of data and run workflows that use thousands of cores simultaneously
- **Confidence:** Verify the content and origin of every dataset, reliably reproduce any output
- **Security:** Selectively and securely share your data and workflows
- **Decentralized:** Search, access, and run on distributed data with federated clusters

Try it out: https://playground.arvados.org

curii