Scientific Workflow and Data Management with the Arvados Platform

Peter Amstutz

Tom Clegg, Lucas Di Pentima, Stephen Smith, Ward Vandewege, Alexander (Sasha) Wait Zaranek, Sarah Wait Zaranek

Curii Corporation, info@curii.com

Summary

Reproducibility benefits largely from robust workflow management. The open-source platform Arvados integrates a data management system called “Keep” and the compute management system called “Crunch”, creating a unified environment to store and organize data, and run Common Workflow Language workflows on that data. Arvados is multi-user and multi-platform, running on various cloud and high performance computing environments.

Arvados management features including the ability to (1) identify the origin and verify the content of every dataset, track every workflow run, and reliably reproduce any output (2) organize and search for datasets using metadata (3) securely and selectively share your data and workflows (3) efficiently manage data (minimizing storage costs) and (4) efficiently rerun workflows (minimizing time and compute costs).

Why Reproducibility?

Reproducible workflows are necessary to:

- Further study or to support scientific claims
- Answer questions from collaborators or regulators
- Fulfill regulatory requirements to retain data

To confidently reproduce results, need a *complete record* of computational analysis you have done.

Learn More

Website
arvados.org

Documentation
doc.arvados.org

Try at No Cost
playground.arvados.org

Enterprise Support
info@curii.com

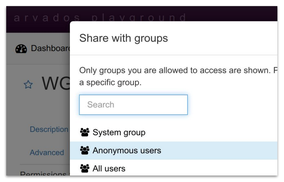
Security and Sharing

Comply with data protection regulations

- Arvados provides authentication, access and audit controls, data integrity, and transmission security

Collaborate with others by selectively and securely sharing data and workflows

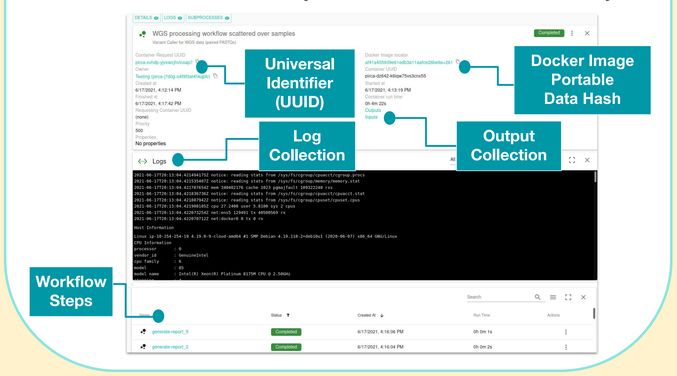
- Private by default
- Read-only, read/write, or manage (to grant permission to others)

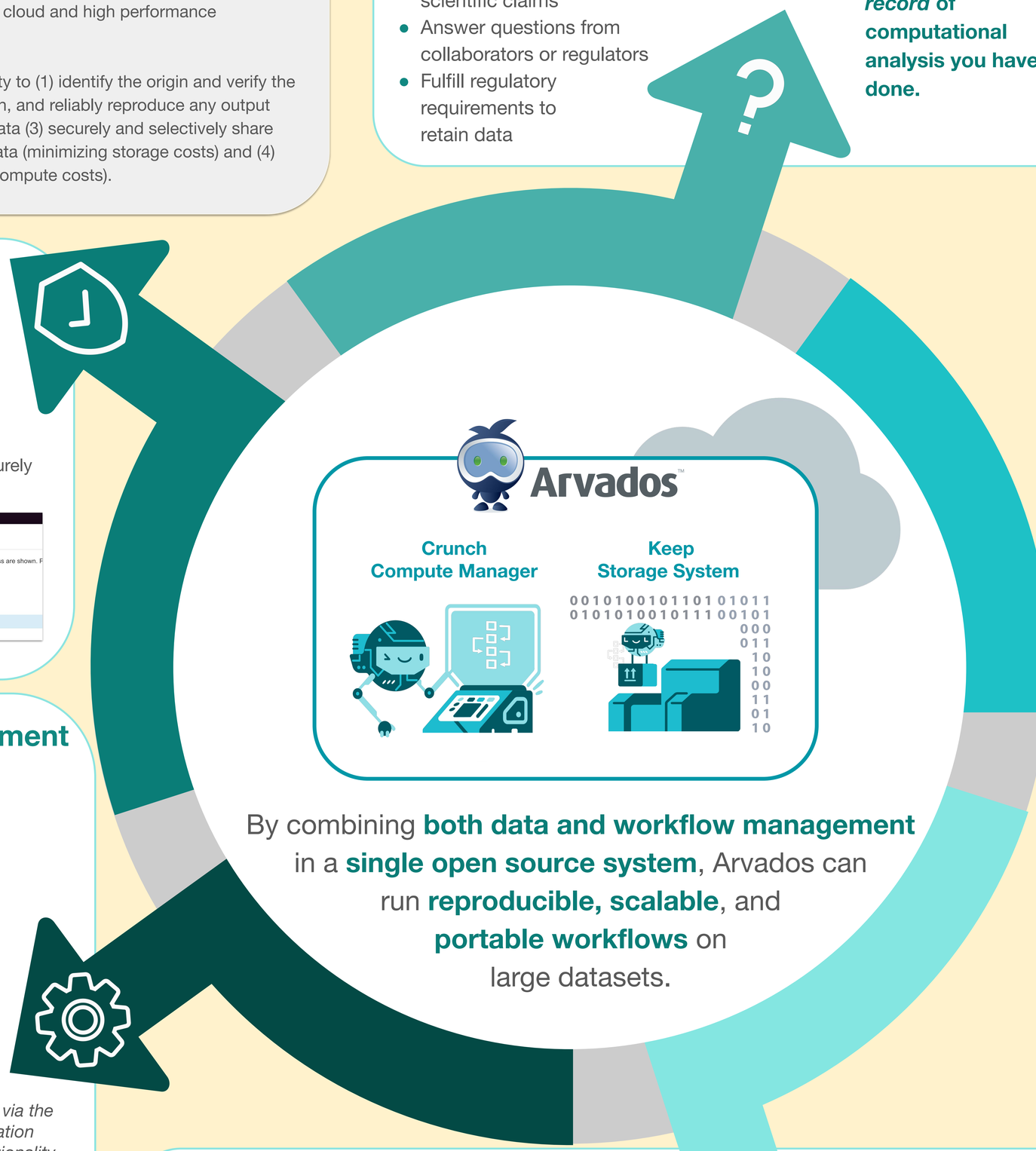


Arvados Workflow Management

- Crunch Workflow Manager:
  - Scalable container orchestration system for running CWL workflows
  - Designed to maintain data provenance and workflow reproducibility
- Automatically stores complete record of workflow execution in collections
  - Inputs/outputs, docker image, logs
  - Referenced by content address (portable data hash)
  - Reorganization *does not* break references

(Below) An executed workflow in Arvados viewed via the Arvados Workbench. The Workbench web application allows users to interactively access Arvados functionality





By combining **both data and workflow management** in a **single open source system**, Arvados can run **reproducible, scalable**, and **portable workflows** on large datasets.

Workflow Management *Includes* Data Management

- For a given result want to track how it was produced (*provenance)*
- Storing of this information in a reliable, findable, accessible, interoperable, and reusable (i.e. FAIR) way **requires a data management system**

Workflow Management requires keeping track of:

- Workflow execution
- Input, output, and intermediate datasets
- Software (e.g. Docker images) used to produce results

This data should be:

- **Identifiable** at a specific point in time and/or by content
- **Findable** through naming conventions and **searchable** attached metadata
- **Associated with robust identifiers** that don’t change if data is reorganized
- **Versioned** to keep track of all data changes
- **Secure** and **shareable**

Arvados Data Management

Keep Storage System combines:

- Content addressing and distributed storage architecture

Collections contain set of files (dataset)

- Organized into shareable “Projects”
- Add and query metadata
- Keep history of changes
- Associated with multiple identifiers: content address, database UUID, name

(Right) A collection in Arvados as viewed via the Arvados Workbench. The Workbench web application allows users to interactively access Arvados functionality