Arvados: Achieving Computational Reproducibility and Data Provenance in Large-Scale Genomic Analyses

Brett Smith[1], Adam Berrey[1], Alexander Wait Zaranek[1,2]
(1) Curoverse, Inc; (2) Harvard Medical School, Boston, USA

Project: https://arvados.org
Code: https://github.com/curoverse/arvados
License: GNU AGPL v3 and Apache 2 for SDKs

Arvados is an open source platform for storing large genomic and biomedical data sets, executing distributed computations, and federated data sharing between private clouds. The platform implements a series of computing strategies to support data provenance and computational reproducibility:

- Content addressing to provide canonical, globally-unique, cryptographically-verifiable references to data sets;
- Manifests that provide a means to describe very large data sets (e.g. exabyte scale) in a compact, canonical, durable form for long term referencing;
- Computational job management that uses Docker containers and virtualization to capture exact configurations for distributed computations that utilize multiple nodes for parallel computation with the map/reduce programming pattern;
- Generation of a metadata graph that records the provenance of individual data sets and the usage of those data sets by computational jobs within the system;
- Graphical visualization of provenance for data sets stored in the system.

We will describe how Arvados can be used in a public cloud service such as Amazon Web Services or a private cloud using a hypervisor such as XenServer to store and analyze genomic data. We will show how the system can run pipelines created with common bioinformatics tools such as GATK and languages such as Python that result in clear and verifiable records of data provenance and outputs that are consistently reproducible over extended periods of time.

Arvados is based on software originally developed at Harvard Medical School for the Harvard Personal Genome project and deployed in a multi-cluster federated system for storing and analyzing those data. The software is developed primarily in Ruby and Go, licensed as free/open source software, and maintained by the Arvados project and community.