Arvados: A Free Software Platform for Big Data Science

Peter Amstutz <peter.amstutz@curoverse.com>, Brett Smith <brett@curoverse.com>,
Ward Vandewege <ward@curoverse.com>, Tom Clegg <tom@curoverse.com>,
Radhika Chippada <radhika@curoverse.com>, Alexander Zaranek <awz@curoverse.com>
Curoverse, Inc.

http://arvados.org http://github.com/curoverse/arvados Affero GPL v3, Apache v2

Large-scale bioinformatics such as genomics requires the application of cluster computing, with many nodes working in parallel to produce results in a reasonable amount of time. When a compute job draws on terabytes of data, uses days compute time, and produces thousands of files, robust management of data sets and the analysis tools used on them is essential to avoid errors that may lead lead to wasted effort or invalid results. To best serve the needs of science, computing platforms should be designed from the ground up to achieve data integrity, provenance, and computational reproducibility.

This talk will introduce the Arvados (http://arvados.org) platform for data science. Arvados is a software system for managing compute clusters built around a scale-out content-addressed distributed file system (Arvados Keep) for storage, a cluster job queuing system designed for reproducibility (Arvados Crunch), and a user and group permission system for controlling and sharing access to those resources. Arvados provides web based and command line tools for transferring, managing, sharing, and computing on very large data sets.

Arvados is designed to scale from a single laptop to cluster and cloud based deployments with dozens of nodes. Arvados is also designed to federate with other Arvados instances, with easy transfer of data and computation between instances. For example, only a single command “arv-copy” is required to copy a complex computation pipeline from a laptop to a cluster or cloud instance (or between instances), where that computation can be run immediately with no additional provisioning or configuration on the target system. The Arvados project is also a founding member of the Common Workflow Language working group, and provides robust support for running computational workflows that are portable across multiple vendor platforms.

This talk will describe the Arvados architecture, describe how Arvados has been used successfully in research, and how interested participants can download and try Arvados for themselves and join the community. Arvados is free software, with services licensed under the GNU Affero General Public License version 3, with SDKs under the Apache License 2.0.