Galaxy for AnswerALS on Microsoft Azure and Kubernetes

RC Carter, Microsoft (rocarter@microsoft.com)
Abhik Ghosh, Applied Information Systems
Alex Lenail, MIT
Enis Afgan, Johns Hopkins University
Gaurav Hind, Microsoft
Nuwan Goonasekera, University of Melbourne

Temporary Website URL:  http://galaxy.answerals.net
Temporary code repository:  https://github.com/rc-ms/galaxy-azure-k8s-helm-htcondor

We discuss a Galaxy configuration to execute the Neurolincs ATAC-Seq and RNA-Seq data pipelines in support of the AnswerALS Foundation research plan seeking a cure for Amyotrophic Lateral Sclerosis (ALS). We set out to build a system capable of supporting a variety of workloads and opted for a Kubernetes-based implementation using Helm Charts that would dynamically scale the available compute infrastructure. The developed system relies on current Galaxy capabilities, and is further consistent with long-term support of Galaxy CloudMan 2.0.

The system was designed to support the entire corpus of data and compute artifacts from 1,000 ALS patients (over 60 TB of data). This sizable volume of data is being ingested through a combination of upload methods, including FTP, HTTP upload via the Galaxy web user interface, as well as external bulk upload into an NFS volume via tools such as AzCopy. Currently, the system is deployed using Azure Container Service (AKS) and a dedicated NFS server as a cluster-wide file system.

In addition to the current system design, we will present learnings and considerations for alternative cloud configurations, including some of the performance impacts of Persistent Volume Claims (PVCs), and external cloud storage options.