AWS Cloud Computing of TOPMed Data

Analysis Pipeline on the Cloud
Presentation

● Review High Performance Computing (HPC) Support in Analysis Pipeline

● Overviews
  ○ AWS cfncluster
  ○ Docker
  ○ AWS Batch Service

● Examples Include:
  ○ Docker
  ○ Analysis Pipeline using AWS Batch Service
  ○ Monitoring Batch Service
High Performance Computing (HPC) Support in Analysis Pipeline

- **Currently Supported**
  - Local Linux cluster with SGE job scheduler
  - AWS cfncluster with SGE job scheduler (traditional)
  - AWS batch with docker images (new)

- **To Be Investigated**
  - Microsoft Azure using Batch/Docker
  - Google Cloud using Batch/Docker
Overview AWS cfncluer

- Encrypted S3 Storage
- EFS Storage
- Administrative Instance
- Install/Manage Software Tools
- Manage Compute Environments
- Shared Data

AWS Cloud

- 24 Node Homogeneous Cluster
- 400 Node Homogeneous Cluster
- Single Computer Environment
Overview Docker

docker commands:
- Run interactively
- Run command in image
- Optionally mount local volumes
- Build and deploy an image to the docker hub

docker commands:
- Run interactively
- Run command in image
- Optionally mount local volumes
- Build and deploy an image to the docker hub

docker hub on the web
- Ubuntu Image Repository
- UW-GAC Image Repository
- Other Image Repositories

docker container
- GENESIS
  - R
  - Linux

docker engine
- Linux or Windows Computer
- Local Storage Volume
- NAS Storage Volume
Docker Examples

1. Log into AWS docker instance

   $ ssh -i ~/.ssh/xxx.pem ubuntu@xx.xx.xx

2. List docker images and containers

   $ docker images  # list images
   . . .
   $ docker ps      # list containers
   . . .
3. Run the R TOPMed docker image

```r
$ docker run -it uwgac/r-topmed:dev /bin/bash
# R
library(SeqArray)
data.path <-
"https://github.com/smgogarten/analysis_pipeline/raw/devel/testdata"
vcffile <- "1KG_phase3_subset_chr1.vcf.gz"
if (!file.exists(vcffile)) download.file(file.path(data.path, vcffile),
vcffile)
gdsfile <- "1KG_phase3_subset_chr1.gds"
seqVCF2GDS(vcffile, gdsfile, fmt.import="GT", storage.option="LZMA_RA",
verbose=FALSE)
gds <- seqOpen(gdsfile)
gds
```
Overview AWS Batch Service

AWS Cloud Computing of TOPMed Data

August 2017
Analysis Pipeline and AWS Batch Examples

4. Preliminaries

$ # cd to working directory
$ cd /projects/topmed/analysts/kuraisa/tm-workshop
$ # vi the config file
$ vi assoc_window_burden.config
...

5. Print out commands

$ # use print option and cluster_type AWS_Batch
$ python /projects/topmed/dev_code . . .
Analysis Pipeline and AWS Batch Examples

6. Execute pipeline

```bash
$ # specify cluster_type AWS_Batch and cluster cfg
$ # file for using test data/environment
$ python /projects/topmed/dev_code/analysis_pipeline . . .
```

7. Monitor jobs
   a. Batch Console
      i. Dashboard (overview)
      ii. Job queue (Optimal_topmed_testdata)
      iii. Logs
   b. ec2 instances