Genome Sequencing Simulation & Synthetic Data Generation

AstraZeneca
Hariprasad Radhakrishnan

Introduction

This project/Experimentation involved simulating and generating synthetic genomic data (Using Varsim) in a highly scalable fashion to test Genomic pipeline tools used in Next generation Sequencing (NGS). It helps newly establishing projects in this space to test their NGS platforms with synthetic genomic data. The project would also discuss the architecture in achieving the big data scale of operations and the tools used and experimented (Docker, Kubernetes, Container engine & Registry, Cloud functions, storage, Google cloud) with to achieve the end results.

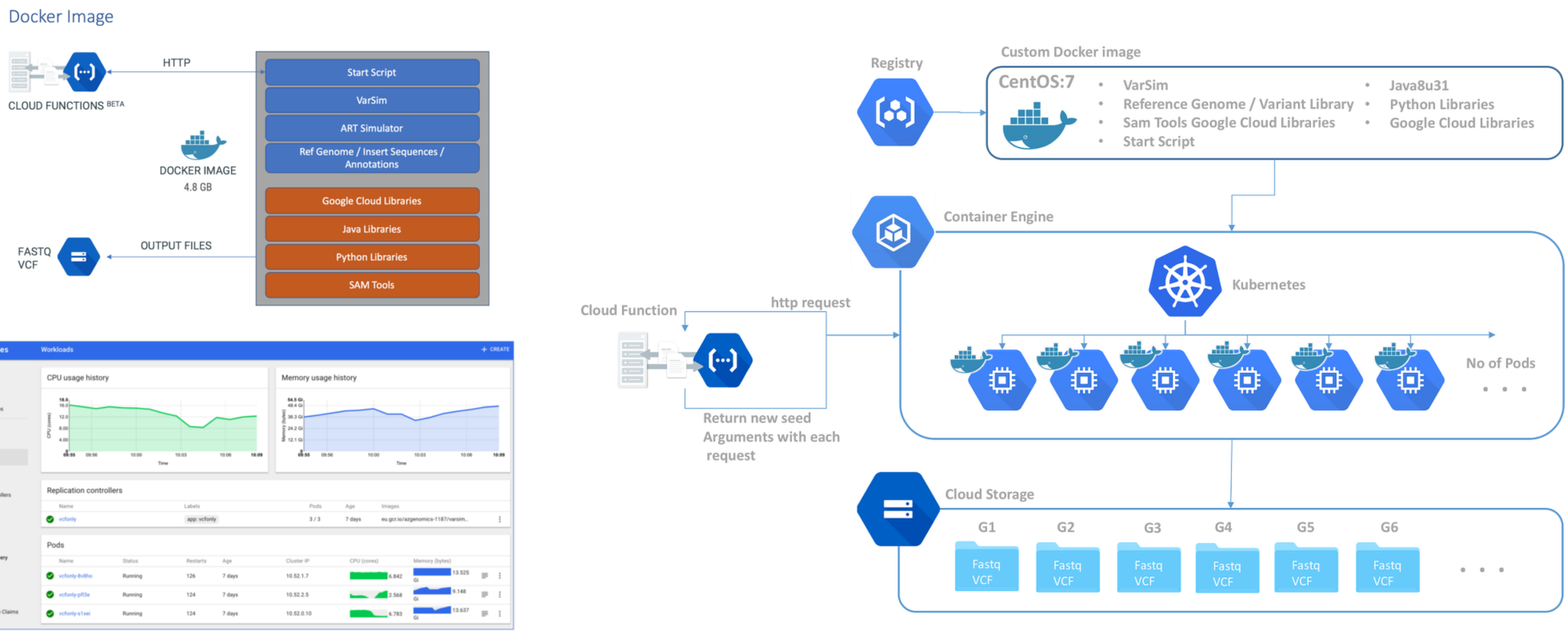
Google Cloud Platform

Purpose

Ability to simulate Genome Sequencing
Generated Genomic Data can be used to test pipelines and infrastructure
If run in a distributed mode can generate sufficient data to create near production line scenarios
Test various analytics tools
Synthetic so no issues with privacy, patient de-identification or transfer across regions

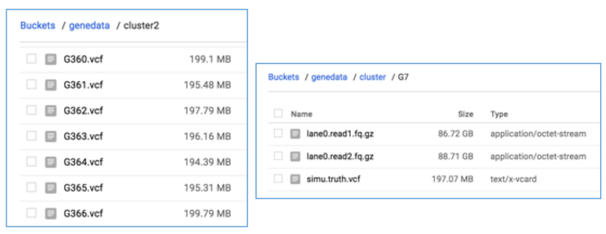
Methods

VarSim is a Python/Java based tool that would simulate one genome per run.
We looked into ways how we could parallelize the execution to generate more genomes.
Build Docker container/Image for the tool.
Execution on Google Cloud – Container engine, Kubernetes.
Parameters like Coverage, Unique ID & Seed value were externalized in a Lambda function that the Docker images could talk to and receive arguments before execution.
Output FASTQ files and VCF's would then be stored in Cloud storage.
Ability to choose to generate FASTQ & VCF or just VCF.



Results

We have around 1000 unique VCF files generated so far on the trial run.
We have the ability to generate the required number of Genome FASTQ and VCF's
The Architecture can be re-used for other distributed compute applications.



Varsim was pickedup as the tool for Genome Simulation, as it provided ability to insert variations into the Genome Simulation and output FASTQ and VCF files

THANKS TO :
John C. Mu, Marghoob Mohiyuddin, Jian Li, Narges Bani Asadi, Mark B. Gerstein, Alexej Abyzov, Wing H. Wong, and Hugo Y.K. Lam
VarSim: A high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications
Bioinformatics first published online December 17, 2014 doi:10.1093/bioinformatics/btu828

http://bioinform.github.io/varsim/