Bringing Hadoop into Bioinformatics with Cloudgene and CloudMan

Sebastian Schönherr, Lukas Forer, Davor Davidovic, Hansi Weissensteiner, Florian Kronenberg, Enis Afgan
Dublin, BOSC 2015
All started at BOSC 2012
BOSC 2012 - CloudMan

- “Cluster on the Cloud” for everyone
- Configures Galaxy automatically
- Features
  - Private/public cloud support, Instance sharing, dynamic cluster scaling, Persistent storage, re-launch your cluster

Enis Afgan, Johns Hopkins University & RBI
CloudMan 2015

• Cloud manager in several cloud infrastructures
  – Amazon AWS: Since 2010
  – Nectar: Since 2012
  – Jetstream: Coming late 2015
  – EGI ENGAGE H2020 project

• Deploy your own version of Galaxy on the Cloud
  – Using Ansible playbook + Packer
BOSC 2012
BOSC 2012 - Cloudgene

• Improve usability of Hadoop in Bioinformatics
• A graphical execution platform for Hadoop programs
  – Interface to integrate programs (YAML)
  – Combine several programs into a workflow
• Setting up a Hadoop cluster on the cloud

Lukas Forer  Sebastion Schönherr - Medical University of Innsbruck
Cloudgene 2015

• From a general workflow system to a **Software-as-A-Service** platform
  – Dedicated service for a given workflow
  – Already 2 services up and running
• Supports Hadoop YARN Stack
  – MRv2, Apache Spark
• Combine Hadoop + Pig + Command Line Programs + R (RMarkdown) programs into one workflow
  – Automatic file staging
BOSC 2012 - Cloudgene + CloudMan

• Similar ideas, different context
BOSC 2012 - Cloudgene + CloudMan
Project started in 2014

- Platform for Big Data Bioinformatics Analysis
- Combine the projects
  - CloudMan for Hadoop cluster provisioning
  - Cloudgene for Hadoop execution
- Find a suitable use case
MapReduce in Bioinformatics

Question: Why is Hadoop not used a lot in bio-informatics?

You are correct in noting most of Hadoop for bioinformatics papers are proofs of concept and real-world use of Hadoop in bioinformatics is quite low.

https://www.biostars.org/p/115260/
A Real World Use case

- Michigan Imputation Server
  - Cloudgene as the underlying framework
  - Our workflow includes QC + Phasing + Imputation
  - Cooperation with Center of Statistical Genetics, University of Michigan
  - https://imputationserver.sph.umich.edu

Gonçalo Abecasis  Michael Boehnke  Christian Fuchsberger
Overall Workflow

Reference Panels: 1000 Genomes / Hapmap / HRC

Genotypes

Scientist

Imputation Server

Reference Panel

Imputed Genotypes
QC-Report

Allele-Frequency Correlation

Uploaded Samples vs. Reference Panel

$r^2 = 0.963$

Potential Frequency Mismatches

Markers where $\chi^2$ is greater than 300.
Benefits

• Why CloudMan?
  – Provide our services on private & public clouds
    – Data sensitivity
  – Provide “best practices” pipeline to everyone
  – Reach a wide user community (Nectar, Jetstream)
Benefits

• Why Cloudgene?
  – Well-tested platform for running (Hadoop) services
    • Provides user management, admin dashboards, ...
  – Focus on the service implementation itself, not on the infrastructure
  – Service 1: Michigan Imputation Server
  – Service 2: mtDNA-Server
    • Detecting heteroplasmy and contamination in mtDNA NGS data [http://mtdna-server.uibk.ac.at](http://mtdna-server.uibk.ac.at)
  – Service 3: ? (Maybe after this meeting)
Software Stack

- Imputation Server
- Cloudgene
  MapReduce Platform
- CloudMan
  Infrastructure Manager
Current Project Status

• Hadoop + Cloudgene running on CloudMan
  – Fully distributed mode
  – Run a WordCount YARN example with Cloudgene
• Current work
  – Install services as apps (Cloudgene), scaling of cluster (CloudMan)
• Updates / Screenshots
  https://wiki.galaxyproject.org/CloudMan/Services
Codefest 2015

• Build a Docker Image for Hadoop + Cloudgene
  – We integrated mtDNA-Server
    `docker pull seppinho/cdh5-pseudo-mtdnaserver`

• Hadoop Galaxy Adapter (CRS4)
  – Perfect fit
  – Export our workflow and integrate it into Galaxy (tbd)
Acknowledgement

• CloudMan
  – Enis Afgan and Davor Davidovic
  – wiki.galaxyproject.org/CloudMan

• Cloudgene
  – Lukas Forer and Sebastian Schönherr
  – cloudgene.uibk.ac.at

• Michigan Imputation Server
  – Gonçalo Abecasis; Michael Boehnke; Christian Fuchsberger
  – imputationserver.sph.umich.edu
Thanks to BOSC!