Delivering Bioinformatics MapReduce Applications in the Cloud

Lukas Forer*, Tomislav Lipić**, Sebastian Schönherr*, Hansi Weißensteiner*, Davor Davidović**, Florian Kronenberg* and Enis Afgan**

* Division of Genetic Epidemiology, Medical University of Innsbruck, Austria
** Center for Informatics and Computing, Ruđer Bošković Institute, Zagreb, Croatia

Introduction

- Cooperation between
  - Division of Genetic Epidemiology, Innsbruck, Austria
  - Ruđer Bošković Institute (RBI), Zagreb, Croatia
- Project
  - Aim: Developing a Bioinformatics Platform for the Cloud

Motivation: Bioinformatics

- More and more data is produced
- Next-Generation Sequencing (NGS)
  - Allows us to sequence the complete human genome (3.2 billion positions)
  - Decreasing Costs → Increasing Data Production

Bottleneck is no longer the data production in the laboratory, but the analysis!

Motivation: Next Generation Sequencing

- NGS Data Characteristics
  - Data size in terabyte scale
  - Independent text data rows
  - Batch processing required for analysis files
- MapReduce
  - One possible candidate for batch processing
  - scalable approach to manage and process large data efficiently

“High potential for MapReduce in Genomics”

MapReduce in Bioinformatics

Hadoop MapReduce libraries for Bioinformatics	Hadoop BAM	Manipulation of aligned next-generation sequencing data (supports BAM, SAM, FASTQ, FASTA, QSEQ, BCF, and VCF)
	SeqPig	Processing NGS data with Apache Pig; Presenting UDFs for frequent tasks; using Hadoop-BAM
	BioPig	Processing NGS data with Apache Pig; Presenting UDFs
	Biodoop	MapReduce suite for sequence alignments / manipulation of aligned records; written in Python
DNA - Alignment algorithms based on Hadoop	CloudBurst	Based on RMAP (seed-and-extend algorithm) Map: Extracting k-mers of reference, non-overlapping k-mers of reads (as keys) Reduce: End-to-end alignments of seeds
	Seal	Based on BWA (version 0.5.9) Map: Alignment using BWA (on a previously created internal file format) Reduce: Remove duplicates (optional)
	Crossbow	Based on Bowtie / SOAPsnp Map: Executing Bowtie on chunks Reduce: SNP calling using SOAPsnp
RNA - Analysis based on Hadoop	MyRNA	Pipeline for calculating differential gene expression in RNA; including Bowtie
	FX	RNA-Seq analysis tool
	Eoulsan	RNA-Seq analysis tool
Non-Hadoop based Approaches	GATK	MapReduce-like framework including a rich set of tools for quality assurance, alignment and variant calling; not based on Hadoop MapReduce

Problem 1: Complex Analysis Pipelines

- Bioinformatics MapReduce Applications
  - available only on a per-tool basis
  - cover one aspect of a larger data analysis pipeline
  - Hard to use for scientists without background in Computer Science

Needed: System which enables building MapReduce workflows

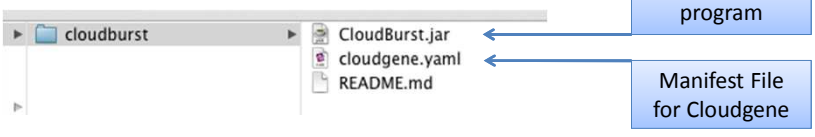
Cloudgene: Overview

- Cloudgene
  - MapReduce based Workflow System
  - assists scientists in executing and monitoring workflows graphically
  - data management (import/export)
  - workflow parameter track→ reproducibility
- Supports Apache's Big-Data stack

S. Schönherr, L. Forer, H. Weißensteiner, F. Kronenberg, G. Specht, and A. Kloss-Brandstätter. Cloudgene: a graphical execution platform for MapReduce programs on private and public clouds. *BMC Bioinformatics*, 13(1):200, Jan. 2012.

Cloudgene: Workflow Composition

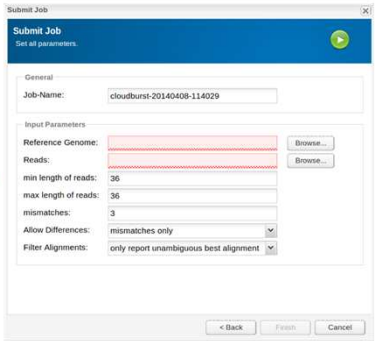
- Reuse existing Apache Hadoop programs
- No adaptations in source code needed
- All meta data about tasks and workflows are defined in one single file
  - **the manifest file**



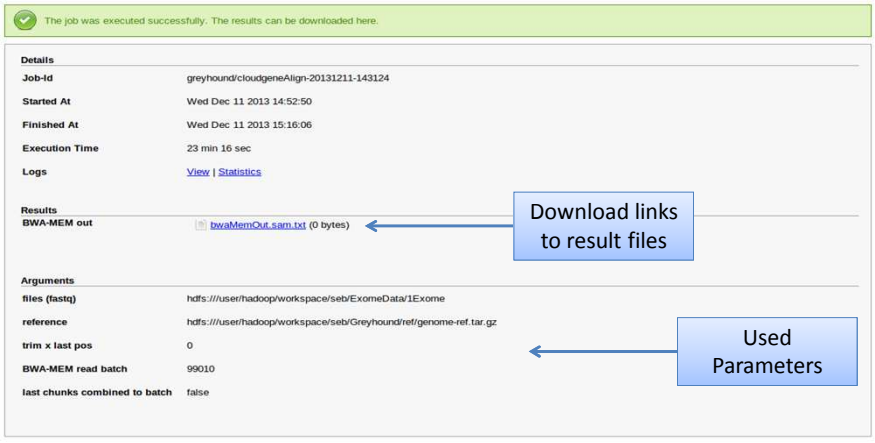
Cloudgene: Workflow Composition

- Manifest file defines the workflow

```
name: CloudBurst
description: Highly Sensitive
  Short Read Mapping with MapReduce.
version: 1.1.0
author: Michael Schatz
category: Genetics

cluster:
  image: us-east-1/ami-da0cf8b3
  type: m1.large,m1.xlarge
  ports: 80,50030,50070

mapred:

  jar: CloudBurst.jar
  params: $reference $reads $output
    $min_read_len  $max_read_len $k
    $allowdifferences $filteralignment
     240 48 24 24 128 16

  inputs:
    - id: reference
      description: Reference Genome
      type: hdfs-file
      makeAbsolute: false

    - id: reads
      description: Reads
      type: hdfs-file
      makeAbsolute: false
```



Cloudgene: Interface



Problem 2: Missing Infrastructure

- Cloudgene requires a functional compatible cluster
  - Small/Medium sized research institutes can hardly afford own clusters
- Possible Solution: Cloud Computing
  - Rent computer hardware from different providers (e.g. AWS, HP)
  - Use resources and services on demand

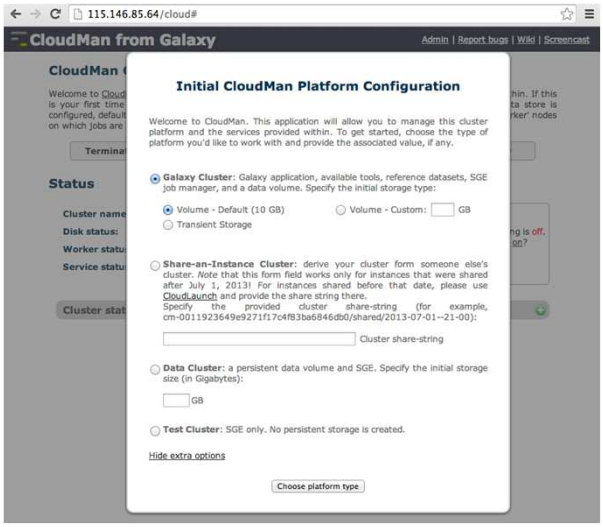
Needed: System which enables delivering MapReduce clusters in the cloud

CloudMan: Overview

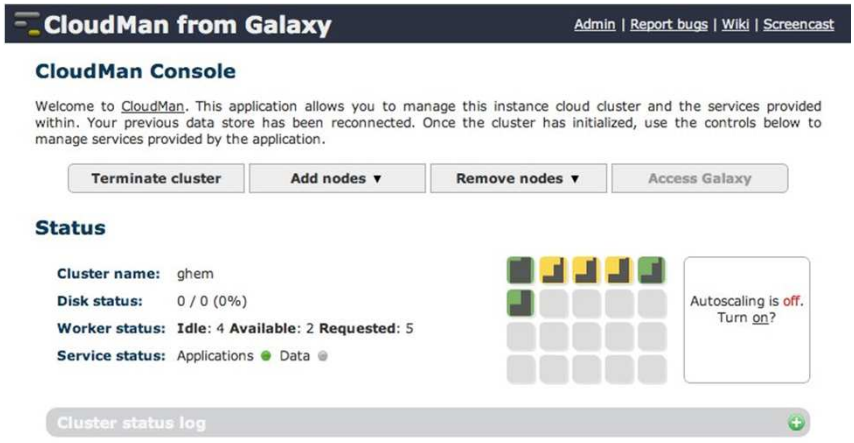
- Enables launching/managing a analysis platform on a cloud infrastructure via a web browser
  - Delivers a scalable cluster-in-the-cloud
- Preconfigure a number of applications
  - Workflow System: Galaxy
  - Batch scheduler: Sun Grid Engine (SGE)
  - MapReduce using Hadoop and SGE integration

Afgan E, Chapman B, Taylor J. CloudMan as a platform for tool, data, and analysis distribution. BMC Bioinformatics 2012;

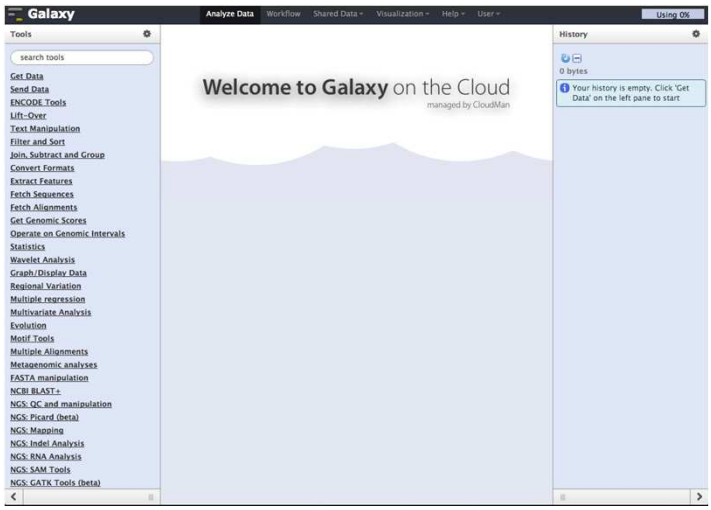
Configure the Cluster



Manage the Cluster



Use the Cluster



CloudMan-as-a-Platform

- CloudMan implements a service dependency management framework

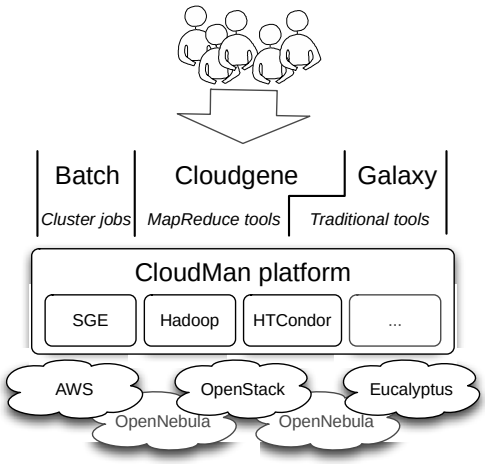
- Enables creation of user-specific cloud platforms

Idea: Implementing Cloudgene as a CloudMan service

Cloudgene meets CloudMan

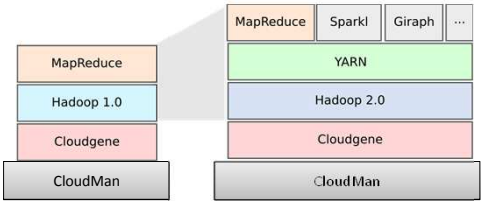
CloudMan
A cloud manager for delivering application execution environments

Cloudgene
A Bioinformatics MapReduce workflow framework



New Opportunities

- One single data analysis cloud platform
  - MapReduce and SGE integrated
  - Additionally, more big data computational models can be supported in future



Conclusion

- MapReduce
  - simple but effective programming model
  - well suited for the Cloud
  - still limited to a small number of highly qualified domain experts
- **Cloudgene**
  - as a possible MapReduce Workflow Manager
- **CloudMan**
  - as a possible Cloud Manager
- Implementing Cloudgene as a CloudMan service
  - Enables domain experts to incorporate and make use of additional big data models

Acknowledge

- CloudMan
  - Enis Afgan
  - Tomislav Lipić
  - Davor Davidović
- Cloudgene
  - Lukas Forer
  - Sebastian Schönherr
  - Hansi Weißensteiner
  - Florian Kronenberg