

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



MEDIZINISCHE UNIVERSITÄT
INNSBRUCK



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!

Page 3

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering - www.mipro.hr



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”

Page 4

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering - www.mipro.hr



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

Page 5

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



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

Page 6

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Cloudfene: Overview

- Cloudfene
 - 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. Cloudfene: a graphical execution platform for MapReduce programs on private and public clouds. *BMC Bioinformatics*, 13(1):200, Jan. 2012.

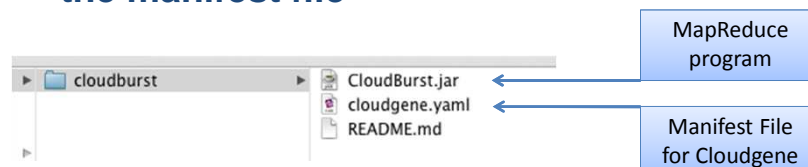
Page 7

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Cloudfene: 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



Page 8

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Cloudfuge: 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: 80,56930,56970

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

```



Submit Job

Set all parameters

General

Job Name: cloudburst-20140408-114029

Input Parameters

Reference Genome: Browse...

Reads: Browse...

min length of reads: 36

max length of reads: 36

mismatches: 3

Allow Differences: mismatches only

Filter Alignments: only report unambiguous best alignment

< Back Finish Cancel

Page 9

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Cloudfuge: Interface

The job was executed successfully. The results can be downloaded here.

Details

Job-Id: greyhound/cloudgeneAlign-20131211-143124

Started At: Wed Dec 11 2013 14:52:50

Finished At: Wed Dec 11 2013 15:16:06

Execution Time: 23 min 16 sec

Logs: [View](#) | [Statistics](#)

Results

BWA-MEM out: [bwaMemOut.sam.txt](#) (0 bytes)

Arguments

files (fastq): hdfs:///user/hadoop/workspace/seb/ExomeData/1Exome

reference: hdfs:///user/hadoop/workspace/seb/Greyhound/ref/genome-ref.tar.gz

trim x last pos: 0

BWA-MEM read batch: 99010

last chunks combined to batch: false

Download links to result files

Used Parameters

Page 10

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Problem 2: Missing Infrastructure

- Cloudbase 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

Page 11

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



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;

Page 12

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Configure the Cluster

Initial CloudMan Platform Configuration

Welcome to CloudMan. This application will allow you to manage this cluster platform and the services provided within. To get started, choose the type of platform you'd like to work with and provide the associated value, if any.

☒ **Galaxy Cluster:** Galaxy application, available tools, reference datasets, SGE job manager, and a data volume. Specify the initial storage type:

☒ Volume - Default (10 GB) ☐ Volume - Custom: GB

☐ Transient Storage

☐ **Share-an-Instance Cluster:** derive your cluster from someone else's cluster. Note that this form field works only for instances that were shared after July 1, 2013! For instances shared before that date, please use [CloudLaunch](#) and provide the share string there. Specify the provided cluster share-string (for example, cm-0011923649e9271f1704f83ba5846db0/shared/2013-07-01-21-00):

☐ **Data Cluster:** a persistent data volume and SGE. Specify the initial storage size (in Gigabytes): GB

☐ **Test Cluster:** SGE only. No persistent storage is created.

[Hide extra options](#)

[Choose platform type](#)

Page 13

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering - www.mipro.hr



Manage the Cluster

CloudMan from Galaxy [Admin](#) | [Report bugs](#) | [Wiki](#) | [Screencast](#)

CloudMan Console

Welcome to CloudMan. This application allows you to manage this instance cloud cluster and the services provided within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to manage services provided by the application.

[Terminate cluster](#) [Add nodes ▼](#) [Remove nodes ▼](#) [Access Galaxy](#)

Status

Cluster name: ghem

Disk status: 0 / 0 (0%)

Worker status: Idle: 4 Available: 2 Requested: 5

Service status: Applications ● Data ●

Autoscaling is off. Turn on?

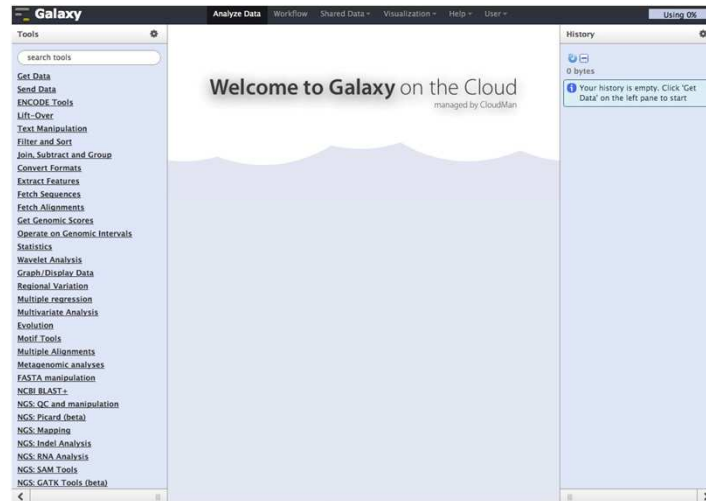
[Cluster status log](#)

Page 14

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering - www.mipro.hr



Use the Cluster



Page 15

21 May, 2014 DC-VIS - Distributed Computing, Visualization and Biomedical Engineering - www.mipro.hr



CloudMan-as-a-Platform

- CloudMan implements a service dependency management framework



- Enables creation of user-specific cloud platforms

Idea: Implementing Cloudfgen as a CloudMan service

Page 16

21 May, 2014 DC-VIS - Distributed Computing, Visualization and Biomedical Engineering - www.mipro.hr



Cloudfene meets CloudMan

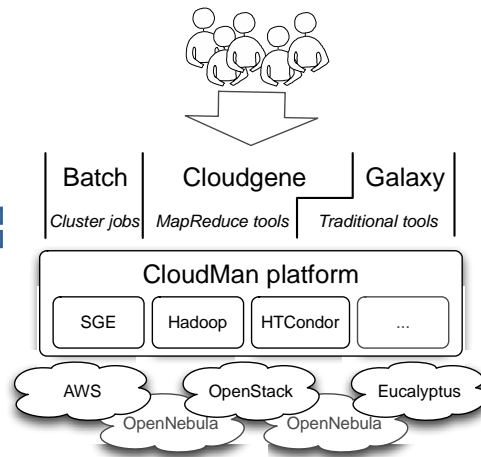
CloudMan

A cloud manager for
delivering application
execution environments



Cloudfene

A Bioinformatics
MapReduce workflow
framework



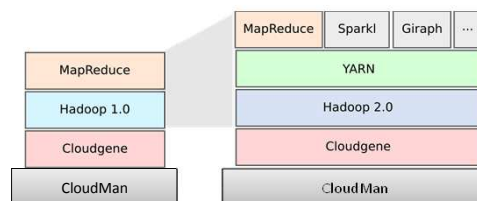
Page 17

21 May, 2014 DC-VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



New Opportunities

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



Page 18

21 May, 2014 DC-VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Conclusion

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

Page 19

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr



Acknowledge

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



MEDIZINISCHE UNIVERSITÄT
INNSBRUCK

Page 20

21 May, 2014 DC VIS - Distributed Computing, Visualization and Biomedical Engineering www.mipro.hr

