Bringing Hadoop into Bioinformatics with Cloudgene and CloudMan

Sebastian Schönherr[1], Lukas Forer[1], Davor Davidović[3], Hansi Weissensteiner[1], Florian Kronenberg[1], Enis Afgan[2,3]

[1] Division of Genetic Epidemiology; Department of Medical Genetics, Molecular and Clinical Pharmacology; Innsbruck Medical University, Innsbruck, Austria. *Email*: sebastian.schoenherr@i-med.ac.at

[2] Department of Biology, Johns Hopkins University, Baltimore, MD, USA

[3] Centre for Informatics and Computing, Ruđer Bošković Institute (RBI), Zagreb, Croatia

Project Website: http://cloudman.irb.hr; http://cloudgene.uibk.ac.at
Source Code: https://github.com/galaxyproject/cloudman; https://github.com/genepi/cloudgene
License: MIT (CloudMan); GPL (Cloudgene)
Acknowledgment: EU FP7 project (grant agreement 602133); Michigan Imputation Server Team

Despite the evident potential of the MapReduce model and existence of bioinformatic algorithms and applications, those are still to become widely adopted in the bioinformatics data analysis. The Hadoop MapReduce model offers a simple framework for data parallelism by providing automated runtime recovery (for both task runtime and hardware failures), implicit scalability (tasks automatically run in parallel batch mode), as well as data replication and locality (reduce data movement, hence increase processing capacity). We identify two prerequisites for wider adoption and higher utilization of MapReduce tools: (1) abstract the technical details of how multiple existing MapReduce tools are composed, and (2) provide easy access to the necessary compute infrastructure and the appropriate environment. Satisfying these requirements would allow bioinformatics domain experts to focus on the analysis while the required technical details are hidden.

At BOSC 2012, two platforms were presented: Cloudgene - a MapReduce tool execution platform leveraging Hadoop, and CloudMan - a cloud resource manager. Since then, we have combined and extended these two platforms to provide a readily available and an accessible Hadoop-based bioinformatics environment for the Cloud. Cloudgene, other than allowing arbitrary MapReduce tools to be integrated and used to craft an analysis, has been extended as a job execution engine for currently two dedicated services: an imputation service developed in cooperation with the Center for Statistical Genetics, University of Michigan (available at *imputationserver.sph.umich.edu*) and a mtDNA analysis service (available at *mtdna-server.uibk.ac.at*). Thus far, the "Michigan Imputation Server" has shown remarkable popularity and scalability with over 690,000 human genomes being imputed within one year. These services have been deployed on dedicated hardware and offer a simple interface for the specific tasks while the jobs are being executed in the MapReduce fashion. This demonstrates a positive disposition towards wider adoption of MapReduce paradigm in the bioinformatics data analysis space given accessible and effective solutions.

To facilitate easy access to such MapReduce solutions for bioinformatics and broaden the availability of these services, we have extended CloudMan to provide a Hadoop-based environment with pre-configured Cloudgene. CloudMan handles the tasks of procuring required cloud resources and configuring the appropriate environment, thus insulating the user from the low-level technical details otherwise required. Because CloudMan is compatible with multiple cloud technologies, it is now feasible to deploy this environment on a range of private and public clouds. This makes it possible for anyone to obtain a scalable Hadoop-based cluster with Cloudgene pre-installed and readily execute MapReduce tools.

This talk will present the motivation for supporting greater adoption of MapReduce-based applications in the bioinformatics data analysis space followed by the details of the described services and their functionality.