Supporting NGS Pipelines in the cloud

Ignacio Blanquer Blanquer[1], Goetz Brasche[2] Jacek Cala[3], Fabrizio Gagliardi[2], Dennis Gannon[2], Hugo Hiden[3], Hakan Soncu[2], Kenji Takeda[2], Andrés Tomás[1], Simon Woodman[3] ✉

[1]Insitute of Instrumentation for Molecular Imaging – I3M, Universitat Politècnica de València, València, Spain
[2]Microsoft Research, Aachen, Germany
[3]Newcastle University, Newcastle Upon Tyne, United Kingdom

Motivation and Objectives

The availability of workflow management systems (Romano, 2007, Smedley *et al.*, 2008) and public cloud computing infrastructures (Schatz *et al.*, 2010) have become a major breakthrough in the usage of computing resources for scientists / the science community. However, the combination of both approaches has shortcomings (Magellan Final Report, 2011, Blanquer *et al.*, 2012), such as the need to reduce administration effort to user, or the need for simple programming models for the transition from previous more conventional computing approaches and the support of legacy software. Although projects such as GATK, galaxy, 1000genomes (http://www.1000genomes.org/) use cloud, end users must have sophisticated knowledge of IT to deploy and use such resources. With this in mind, Microsoft Research (Microsoft Research– Cloud Research Engagement, 2013) has started several initiatives to improve the use of clouds in science. The "cloud4science" initiative, see http://www.cloud4science.eu/, considers next generation sequencing (NGS) as an excellence reference use case. This initiative builds on the results of the VENUS-C and e-Science Central projects, see http://www.venus-c.eu/ and http://www.escicencecentral.co.uk/, in which two different scientific workflow engines, namely the *Generic Worker and e-Science Central* were applied to solve specific bioinformatics problems requiring intensive computing. We propose an integration and enhancement of these two workflow engines with a set of selected bioinformatics tools to provide an easy-to-use framework for a cloud-enabled NGS pipeline for mutation analysis.

The resulting framework and components will simplify the deployment of processing services, the access to data and the sharing of the results.

Methods

The development of the framework focuses on three different categories: Computing resources; workload management and orchestration software; and bioinformatics legacy software.

As public cloud computing resources we selec-ted Microsoft's Windows Azure, see http://www.windowsazure.com. Windows Azure provides both users and application developers with different abstraction levels (PaaS and IaaS), including the support of Windows or Linux Virtual Machines (which is a requirement for many Bioinformatics tools). This makes Windows Azure an attractive platform to build upon.

The workload management and orchestration software need to deal with the two main types of NGS workflows: Coarse-grained data flows and fine-grained complex workflows. To tackle the first problem, the Generic Worker has been selected as it has been proven to be efficient in large-scale alignment problems (Carrión *et al.*, 2012). To deal with finer-grain complex workflows, eScience Central is used. E-science Central can scale very well in drug discovery problems (Cala *et al.*, 2012). Reference data for alignment is obtained from the UCSC Genome Bioinformatics repository (http://hgdownload.cse.ucsc.edu), and it is replicated in the Azure storage for convenience.

Finally, for the implementation of the pipeline, several tools have been selected for the initial prototype, covering sequence conversion (seqtk – https://github.com/lh3/seqtk and samtools - http://samtools.sourceforge.net/), Quality Control Analysis (fastqc - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), alignment (bowtie2 - http://bowtie-bio.sourceforge.net/bowtie2 and HPG aligner - http://docs.bioinfo.cipf.es/projects/hpg-aligner), Variant call file generation (GATK - http://www.broadinstitute.org/gatk/) and visualization (GenomeMaps - http://www.genomemaps.org/ and JBrowse - http://jbrowse.org/).

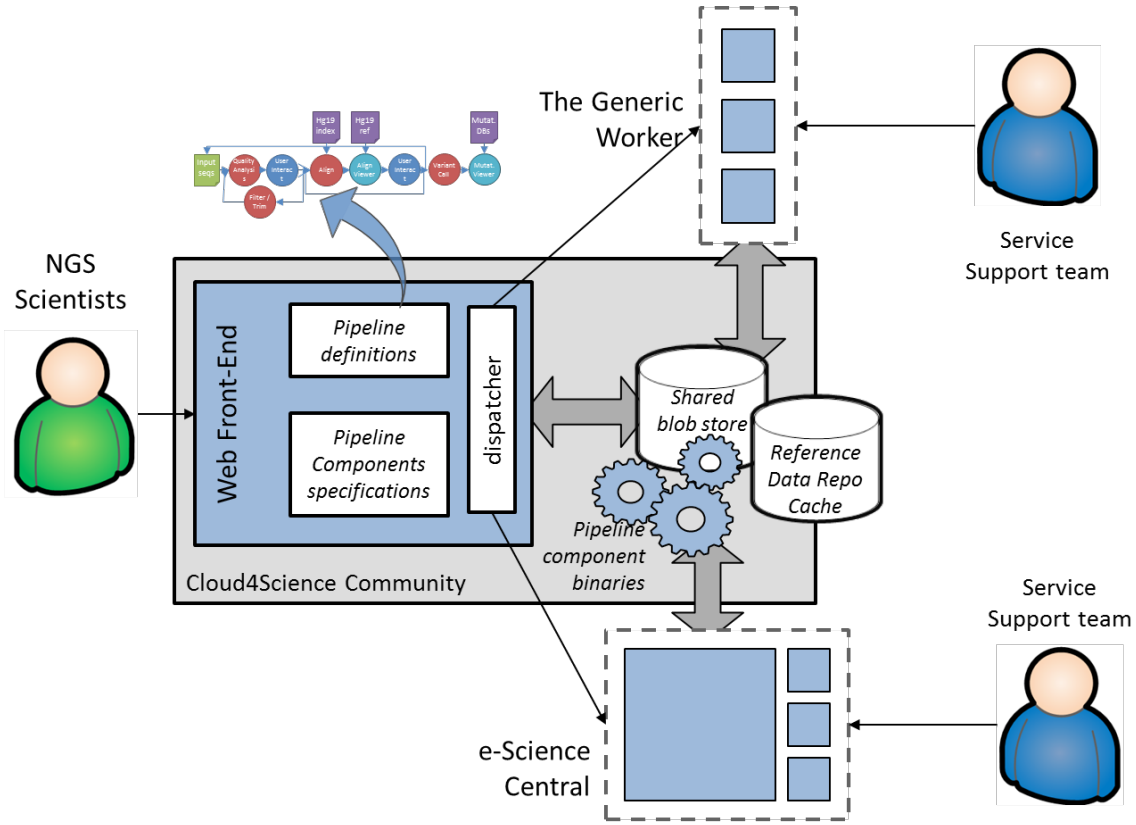

Figure 1. Simplified Architecture of the C4S platform-NGS framework

Results and discussion

The components are integrated through a shared blob store that hosts user-specific and shared data. As shown in Figure 1, the two different deployments of the enactment services are orchestrated through a web front-end, which exposes the interface to the users. The user interface is customized to guide the user through the different steps: uploading, conversion, quality control, trimming and filtering (repeatedly), alignment, visualization, quality control, variant call and final visualization. Interaction is managed through the use of server-based visualization tools, such as GenomeMaps. Early results of the performance achieved by the individual components demonstrated speed-up factors in the order of 63 times faster with 100 Azure cores (efficiency of 63%) using Generic Worker (Carrión *et al.*, 2012) and 176 times faster with 200 Windows Azure cores (efficiency of 88.2%) using e-Science Central (Cala *et al.*, 2012), . Improved results will be obtained by optimizing the data transfer among the processing nodes.

This first prototype will be improved in the first half of 2013 providing the users with more flexibility to alter the workflow, repeating or skipping individual or grouped steps and enhanced management of data to foster sharing and collaboration.

Eventually, it is important to note that the software in development in this project aims to becoming self-supported in the future by the establishment of an open end-user community, by the adoption of an open source software license scheme. Generic Worker and e-Science Central are released through Open Source licenses and the entire Cloud4Science environment will be released in the form of an open and collaborative project.

Acknowledgements

The authors want to thank Microsoft and the cloud4Science project for funding this research activity.

References

Blanquer I, Brasche G, Lezzi D: Requirements of Scientific Applications in Cloud Offerings. Proceedings of IBERGRID 2012, September 2012, in press.

Cala J, Hiden H, Woodman S, Watson P: Fast Exploration of the QSAR Model Space with e-Science Central and Windows Azure. Microsoft Cloud Futures, Berkeley, May 2012.

Carrión A, Blanquer I, Hernández V (2012) "A service-based BLAST command tool supported by cloud infrastructures", *Stud Health Technol Inform.* **175**, 69-77.

Magellan Final Report, December 2011, http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_final_report.pdf, last visited Jan 2013.

Microsoft Research – Cloud Research Engagement, http://research.microsoft.com/en-us/projects/azure/, last visited February 2013.

Romano P (2007) Automation of in-silico data analysis processes through workflow management systems. *Briefings in Bioinformatics* **9**(1), 57–68.

Schatz MC, Langmead B, Salzberg SL (2010) Cloud computing and the DNA data race, *Nature Biotechnology* **28**, 691–693. doi:10.1038/nbt0710-691.

Smedley D, Swertz MA, Wolstencroft K, Proctor G, Zouberakis M, *et.* al. (2008) Solutions for data integration in functional genomics: a critical assessment and case study. *Briefings in Bioinformatics* **9**(6), 532–544