Procedia Computer Science

Volume 51, 2015, Pages 2864–2868

ICCS 2015 International Conference On Computational Science

Pairwise genome comparison workflow in the Cloud using Galaxy

Oscar Torreno[1]*, Michael T. Krieger[1]†, Paul Heinzlreiter[1]‡, and Oswaldo Trelles[2]§

[1] RISC Software GmbH, Hagenberg, Austria
oscar.torreno@risc-software.at,michael.krieger@risc-software.at,paul.heinzlreiter@risc-software.at
[2] Department of Computer Architecture, Malaga, Spain
ortrelles@uma.es

Abstract

Workflows are becoming the new paradigm in bioinformatics. In general, bioinformatics problems are solved by interconnecting several small software pieces to perform complex analyses. This demands certain expertise to create, enact and monitor such tools compositions. In addition bioinformatics is immersed in the big-data territory, facing huge problems to analyse such amount of data. We have addressed these problems by integrating a tools management platform (Galaxy) and a Cloud infrastructure, which prevents moving the big datasets between different locations and allows the dynamic scaling of the computing resources depending on the user needs. The result is a user-friendly platform that facilitates the work of the end-users while performing their experiments, installed in a Cloud environment that includes authentication, security and big-data transfer mechanisms. To demonstrate the suitability of our approach we have integrated in the infrastructure an existing pairwise and multiple genome comparison tool which comprises the management of huge datasets and high computational demands.

Keywords:

1 Introduction

Nowadays, the cost of obtaining genome-scale molecular and biomedical data continues to drop rapidly. Managing and analysing such ever-growing datasets is becoming a crucial issue, since most academic labs do not have the required computing and storage facilities.

Cloud Computing appears as a good solution due to its flexibility in providing computational resources at an attractive price. The price is affordable because users can adapt the resources

*Implemented and ported the workflow to Galaxy, and main contributor in the manuscript preparation.
†Installed and configured Galaxy in the cloud.
‡Contributed in the manuscript preparation (cloud description).
§Work supervision and fruitful comments/contributions to the workflow's design as well as manuscript writing/reviewing.

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.449

to their demand (elasticity), paying for what they actually need, and with no need to maintain a large infrastructure.

To analyse the mentioned generated data, researchers need to link, combine, and query the data using several different tools, what is commonly referred as workflows or pipelines. Particularly, Comparative Genomics workflows have been one of the main focuses in bioinformatics in the recent years. The sequencing of new genomes makes fundamental to study their relation with the already studied ones. Current sequences comparison software does not always address handling big datasets, because data to be processed is retained in RAM during analysis (for example Gepard [4], MUMmer [5] and LASTZ [3]). We present a case study of a pairwise and multiple sequences comparison workflow tackling this problem.

Furthermore, at the moment running tools in the Cloud requires specific knowledge of the computational infrastructure and the command line invoking style. We believe that it is essential to facilitate the end-users with no computer experience to take profit of the advantages offered by the Cloud Computing model. Therefore, we are using Galaxy as a user-friendly tool to execute workflows.

This paper presents our developments to examine the potential and feasibility of integrating Galaxy in a Cloud environment, making the exploitation of the available resources simpler and more efficient, and allowing users to execute tasks requiring a significant amount of resources. Additional distinguishing characteristics of the presented work are the user authentication and transfer of big datasets through grid computing technologies.

2 Related work

There exist some workflow management systems in bioinformatics such as Taverna [7] but they have limitations as they only provide a subset of features including workflows creation, publication and execution; also, they consist in desktop applications using pretty old technologies being replaced at the moment; and their performance seriously degrades when the number of processes grows up. The Galaxy [2] workflow and data management Web platform, appears as a very good option possesing all the mentioned features, and tackling the mentioned issues. In [1] they study the integration of Taverna and Galaxy each one complementing the other.

3 Methods

3.1 Cloud Computing Infrastructure

The Cloud Computing infrastructure used is a small community Cloud installation running OpenStack, which is composed by 3 nodes interconnected through an Ethernet 10Gbps network. Its core functionality has been extended with some functionality provided by the Grid community.

3.2 Data storage and transfer

Ceph is the base distributed file system of our cloud setup, because it implements the interfaces defined by the OpenStack Swift storage services. As the majority of Cloud infrastructures, OpenStack offers object containers and persistent volumes. In our installation the object containers are securely accessible through GridFTP and the persistent volumes are mounted as external disks to the instances.

3.3 User Authentication

User credentials are stored using the Lightweight Directory Access Protocol (LDAP) thus allowing for an easy and central management. The OpenStack users are organized in groups, which contain users from different organizations, who are for example working on a shared project.

The access rights are assigned on groups level and represented within the central LDAP tree of the installation, which is subsequently queried during authentication.

3.4 Galaxy

Galaxy is a platform which enables publishing of scientific workflows. Although it was initially developed for genomics research it is largely domain agnostic and is now used as a general bioinformatics and biomedicine workflow management system.

Galaxy enables users with no programming experience to graphically construct their workflows by interconnecting basic tools. These workflows can subsequently be stored for reuse or shared with other researchers, addressing the issue of having reproducible experiments.

In this work we are using a customised Galaxy Cloud installation, which aims to facilitate the work of the researchers. We are working to achieve a better integration (more details in the Section 7).

3.4.1 Data transfer

Due to data size, Galaxy separates the data transfer (either upload or download) from workflow execution. Users request data transfers by selecting the "Get Data → Upload File" menu entry.

Galaxy provides a Web upload, where the actual content could be pasted, or a file could be selected; transfer via a URL endpoint, from where Galaxy will retrieve the content internally; and FTP upload, where Galaxy checks the user FTP folder. The transfer status and progress are displayed to the end-user in a following dialog.

Galaxy automatically determines the data type of the uploaded file for some specific types and also allow to select it when is not automatically deducted. Such metadata is useful to filter the files matching the available services.

3.4.2 Defining a workflow

Galaxy incorporates a section to graphically define an analysis workflow by interconnecting the involved tools. Once defined, the user can determine if it will be private or public.

To include tools, the users need to generate XML tool wrappers, where the parameters and the actual command invocation line are defined. Then, the administrator installs these wrappers to make them available to the general public. In order to decouple the tool installation from the administrator, these tools are stored in a central tool repository called Galaxy ToolShed from where users can directly install them.

3.4.3 Invoking an existing workflow

At this point, the users search for the specific workflow to be executed and then fill the required input parameters. Once Galaxy receives the job submission its internal scheduler launches all the tasks which do not have unsatisfied dependencies.

At the end of the job submission, Galaxy shows the list of generated tasks (including their statuses). An important point to emphasise here is that the users only need to download the results they are interested in, what contrasts with the former Web Services strategies.

4 Case study: Pairwise sequences comparison workflow

The pairwise sequences comparison workflow reported in [6] has a modular organisation (see Figure 1) removing repetitive actions when processing collections of data sets.

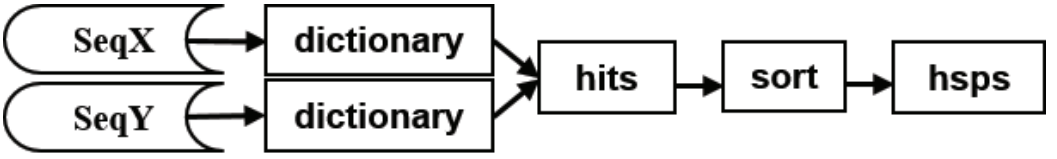

Figure 1: Workflow overview. This is a graphical representation of the steps to be described in the next paragraphs.

The workflow creates two dictionaries, one for each sequence. The dictionaries are composed by a sorted set of words of length K (K-mers) together with their occurrence positions along the sequence. Using the dictionaries, exact K-mers matches determine the starting points (hits) of the final alignments. These hits are sorted by diagonal (defined as position in query sequence minus position in reference sequence) and offset, optionally filtered by proximity and finally used to produce the alignments (High-scoring Segment Pairs, HSPs).

5 Results

The previous workflow has been executed in the Galaxy Cloud installation explained in this document. It is important to note that in the current deployment configuration the Galaxy web server and the tasks execution share the same machine (instance with 4 VCPU and 4 GB of RAM). The workflow was tested using 4 Mycoplasmas sequences of around 1Mbp each and two mammalian chromosomes of around 150Mbp each. The rationale of the experiments was not a speed-up study but the corroboration of the correct execution of the analyses.

The dictionary together with the seed points sorting are the most time-consuming steps. As improvements, in an all-versus-all comparison of a set of sequences, we calculate in parallel the sequence dictionaries since there are no dependencies; we avoid the re-calculation of the sequences dictionaries because they remain stored on disk; and we also perform several pairwise comparisons in parallel since each one is independent from the others.

To execute the test, we followed the data uploading and execution processes as described in Sections 3.4.1 and 3.4.3 respectively. The data uploading took about 2 seconds for the 4 Mycoplasma genomes and 6 minutes for the 2 mammalian chromosomes. The data was moved from the University of Malaga, Spain facilities to the Cloud Computing infrastructure in RISC-Hagenberg, Austria. The execution for all Mycoplasma comparisons (i.e. 6 pairwise comparisons) was of less than 2 minutes and less than 2 hours for the mammalian chromosomes pairwise comparison.

6 Conclusions

The reported use case workflow can be used to analyse sequences of different sizes. However, its distinguishing characteristics are noticed for larger sequences, specially for mammalian genomes in the range of Gbp but also for bacteria in the Mbp range. The state of the art methods mentioned in the introduction have two problems with them, the first problem is the memory consumption, and the second problem is the high execution time. The presented workflow works

also quite well with small sequences (e.g. genes whose average size is of 10-15Kbp in human genome), experiencing execution times comparable to the other methods.

We believe that Cloud infrastructures are the most suitable and flexible solutions for the computational needs of small bioinformatics research groups because of two main reasons, first because they require minimal computational know-how from the end-user in terms of installation, administration and scripting, and second because the resources can be scaled according to the user's needs what is good for all-against-all comparisons performed by the presented workflow.

Galaxy makes easy the way users execute their analysis workflows, because it provides an user-friendly web interface ubiquitously accessible over the internet, which abstracts the user from the specifics of the Cloud infrastructures. Obviously Galaxy introduces some overhead (around 1 minute compared to the command line execution of the use case workflow), but we think it will be negligible when we change the configuration to run tasks in a separate cluster.

7 Future work

We are currently working to integrate the LDAP authentication and GridFTP data transfer with Galaxy. In addition, we plan to separate the Galaxy web server from the execution cluster itself by using Galaxy DRMAA-based scheduler interface[1].

Acknowledgements

This publication is supported by the European Union's Seventh Framework Programme for research, technological development and demonstration through the IAPP project Mr. SymBioMath, grant agreement number 324554.

References

[1] Mohamed Abouelhoda, Shadi A Issa, and Moustafa Ghanem. Tavaxy: Integrating taverna and galaxy workflows with cloud computing support. *BMC bioinformatics*, 13(1):77, 2012.

[2] Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8), August 2010.

[3] RS Harris. Improved pairwise alignment of genomic dna. 2007. *PhD diss., The Pennsylvania State University*, 2007.

[4] Jan Krumsiek, Roland Arnold, and Thomas Rattei. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026–1028, 2007.

[5] Stefan Kurtz, Adam Phillipy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L Salzberg. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2), 2004.

[6] Andrés Rodríguez Moreno, Óscar Torreño Tirado, and Oswaldo Trelles Salazar. Out of core computation of hsps for large biological sequences. In *Advances in Computational Intelligence*, pages 189–199. Springer, 2013.

[7] Tom Oinn, Mark Greenwood, Matthew Addis, M Nedim Alpdemir, Justin Ferris, Kevin Glover, Carole Goble, Antoon Goderis, Duncan Hull, Darren Marvin, et al. Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience*, 18(10):1067–1100, 2006.

[1] https://wiki.galaxyproject.org/Admin/Config/Performance/Cluster#Runner_Configuration