Contribution ID: **66** Type: **Hackathon**

Galaxy jobs processed on the grid: what about the cloud?

Wednesday, 21 May 2014 16:00 (1h 30m)

As part of the labex CEBA (Center for the study of biodiversity in Amazonia), we set up a Galaxy based e-VirtualBiodiversityLab to make easy the use of biodiversity analysis tools to researchers. The increase of data production as a consequence of the progress in sequencing technologies (NGS) requires high computing resources like clusters, the EGI grid and the cloud.
We started to write tools for Galaxy to distribute the processing of such big NGS files (phylogeny, global alignment, taxonomic annotation, ...) to the Avakas cluster (Mesocentre Aquitaine, 1000 cores, Torque) and to the EGI grid (France-Grilles) using the DIRAC interware. This is completely transparent for the user of Galaxy who doesn't need to have an expertise on such technologies and can focus on analyzing and interpreting the results.
Software requirements (OS,version, dependencies, ...) could be a problem when using the grid. We propose for the hackaton to tackle this problem by distributing jobs on cloud resources from a Galaxy platform. This could also be a way to run jobs with parallelized softwares (MPI, ...) easily.
The hackaton challenge would be to write a Galaxy tool to launch jobs on virtual machine (VM) instances on Academic cloud.

Wider impact and conclusions

Diversity of tools used by researchers is currently impeded (for biologists, but not only ...) on efficient HPC infrastructure by technicalities of how to distribute and launch, and related dependencies problems. It can be circumvented using dedicated virtual machines instantiated on the cloud. One can take advantage of scaling capabilities of a cloud infrastructure as well (cloud burst). Our current intent to connect the Galaxy platform to the grid and the cloud, brings directly the high computing capabilities of EGI to the final user : the researcher. The e-VirtualBiodiversityLab will be able to access its own computing elements, the cluster of the Mesocentre Aquitaine, the EGI grid and cloud resources. All elements will be present to offer a scalability pattern for software and tools in data analysis for biodiversity. It will foster and enable the development of new tools and pipelines for a better exploration of unknown biodiversity.

URL(s) for further info

https://galaxy-pgtp.pierroton.inra.fr
http://diracgrid.org

Description of work

We use the Galaxy instance kindly provided by the Genomic Transcriptomic facility of Bordeaux. We already wrote a Galaxy tool to launch a pipeline producing a taxonomic inventory from NGS sequences on the Avakas cluster (Mesocentre Aquitaine). A second tool is developed to distribute the calculation repetitions (Structure software, PhyML phylogeny) to the grid using the DIRAC interware.

Development of a galaxy tool to launch jobs on the cloud could be accomplished performing the following steps:
- set up a Galaxy server for this hackaton (or use a prebuilt VM)
- configure authentication process
- write a generic script using DIRAC interware to:
. connect to a cloud resource
. instantiate virtual machine(s) (VM)
. send files and parameters specified in Galaxy interface
. run specific software on the VM with specified parameters
. get back the result files to Galaxy
- generate the xml file to declare the tool inside Galaxy

We propose to test the implementation by producing 3 Galaxy tools launching respectively:
- Phyml phylogeny, with bootstraps
- Structure bootstrapping calculation,
- Readsyst pipeline taxonomy inventory.

Primary author: LAIZET, Yec'han (INRA)

Co-authors: FRANC, Alain (CNRS); FRIGERIO, Jean-Marc (INRA); CHAUMEIL, Philippe (INRA)

Presenter: LAIZET, Yec'han (INRA)

Session Classification: Hackathon: Galaxy jobs