ELIXIR-ITALY use case within H2020 INDIGO-Datacloud: developing a Galaxy "on demand" platform through cloud technologies.

M.A. TANGARO[1], G. DONVITO[2], G. PESOLE[1.4], F. ZAMBELLI[1,3]

(1) Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari.
(2) National Institute for Nuclear Physics, Bari.
(3) Department of Biosciences, University of Milano, Milano.
(4) Department of Biosciences, Biotechnologies and Pharmacological Sciences, University of Bari "Aldo Moro", Bari.

Next Generation Sequencing (NGS) and other high throughput technologies produce an unprecedented amount of biological data that require powerful computational resources to be analysed. Cloud computing technologies and infrastructures can be a powerful and scalable tool to exploit these kind of data to individual scientists, small research group, research facilities, etc... However, straightforward access to these resources can still be out of reach to many life scientists. To capitalize on the cloud infrastructure and expertise already present in our Node, ELIXIR-IIB is developing a fully customizable Galaxy "on demand" platform service. The service is developed as a case study within the H2020 INDIGO-DataCloud project, exploiting the INDIGO software framework to automate the creation of ready to use Galaxy instances, tailored to user's needs, on the cloud. Once deployed, each Galaxy instance is fully customizable with tools and reference data by its administrator(s). Moreover, each instance is deployed in an insulated environment in order to provide a suitable platform for research and clinical scenarios involving sensible human data.

INTRODUCTION

Galaxy is a workflow manager adopted in many life science research environments in order to facilitate the interaction with bioinformatics tools and the handling of large quantities of biological data. Through a coherent work environment and an user-friendly web interface it organizes data, tools and workflows providing reproducibility, transparency and data sharing functionalities to users.

Currently, Galaxy instances can be deployed in three ways, each one with pros and cons: public servers, local servers and commercial cloud solutions. In particular, the demand for cloud solutions is rapidly growing (over 2400 Galaxy cloud servers launched in 2015, since they allow the creation of a ready-to-use galaxy production environment avoiding initial configuration issues, requiring less technical expertise and outsourcing the hardware needs. Nevertheless relying on commercial cloud providers is quite costly and can pose ethical and legal drawbacks in terms of data privacy.

ELIXIR-IIB in the framework of the INDIGO-DataCloud project is developing a cloud Galaxy instance provider, allowing to fully customize each virtual instance through a user-friendly web interface, overcoming the limitations of others galaxy deployment solutions. In particular, our goal is to develop a PaaS architecture to automate the creation of Galaxy-based virtualized environments exploiting the software catalogue provided by the INDIGO-DataCloud community (www.indigo-datacloud.eu/service-component).

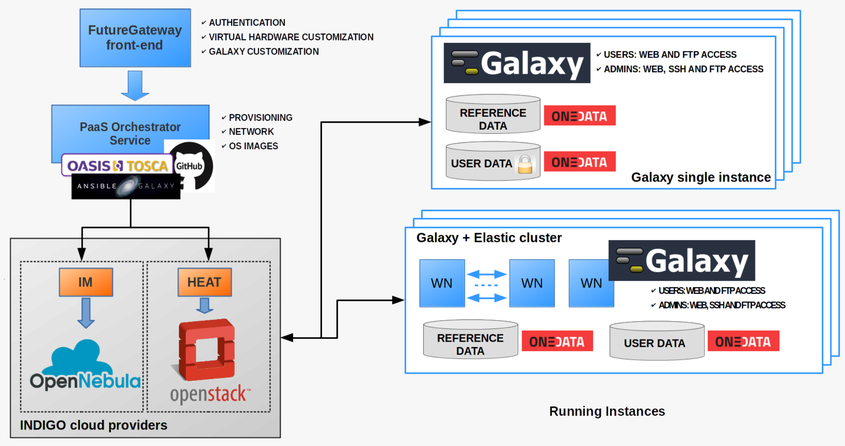
SERVICE ARCHITECTURE

The web front-end is designed to grant user friendly access to this service, allowing to easily configure and launch each Galaxy instance through the INDIGO **FutureGateway** portal component.

All the required components to automatically setup Galaxy instances (Galaxy and all its companion software) are deployed using the INDIGO **Orchestrator** service, based on the TOSCA orchestration language. The service is compatible with both OpenNebula and OpenStack, its deployment on different e-infrastructures. Moreover, it supports both VMs and Docker containers, leaving the selection of the virtual environment to the service providers. This effectively removes the need to depend on particular configurations (e.g. OpenStack, OpenNebula or other private cloud solution like Amazon or Google).

Persistent storage is provided to store users and reference data and to install and run new (custom) tools and workflows. Data security and privacy are granted through the INDIGO **Onedata** component which, at the same time, allows for transparent access to the storage resources through token management. Data encryption implemented at file system level protects user's data from any unauthorized access.

Automatic elasticity, provided using the **CLUES** INDIGO service component, enables dynamic cluster resources scaling, deploying and powering-on new working nodes depending on the workload of the cluster and powering-off them when no longer needed. This provides an efficient use of the resources, making them available only when really needed.

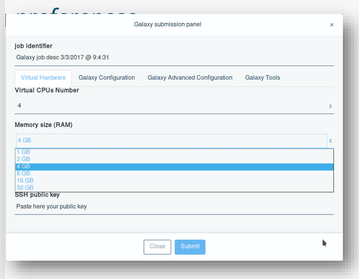


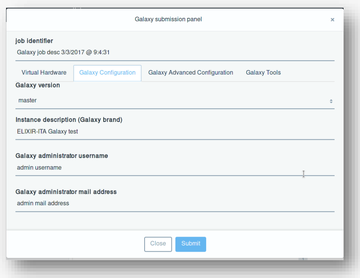
Galaxy cloud service architecture – *The prototype is based on the coordination of separated components, provided by the INDIGO e-infrastructures.*

INSTANCE CUSTOMIZATION

The web front-end provides different tabs for each configuration task: virtual hardware configuration, Galaxy configuration, tools configuration and elastic cluster support. The instance administrator credentials are provided by the users during the configuration phase.

Any instance will then be automatically configured according to the virtual machine hardware specifications according to the user preferences.





GALAXY PRODUCTION ENVIRONMENT

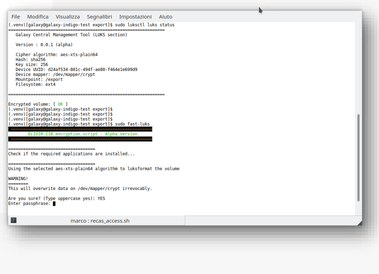
The system allows to setup and launch a virtual machine (or Docker container) configured with the Operative System (CentOS 7 or Ubuntu 14.04) and the auxiliary applications needed to support a Galaxy production environment such as PostgreSQL, Nginx, uWSGI and Proftpd and to deploy the Galaxy platform itself. Once deployed each Galaxy instance can be further customized with tools and reference data.

Each configuration provide also an external volume for reference data and one for users data. Instances are accessible via http, ftp and ssh.

INSTANCE ISOLATION

Users' data access rights will be controlled, by default, through the OneData INDIGO component.

Alternatively, each Galaxy instance can be deployed as an insulated environment, i.e. data are isolated from any other instance on the same platform and from the cloud service administrators, opening to the adoption of Galaxy based cloud solutions even within clinical environments.

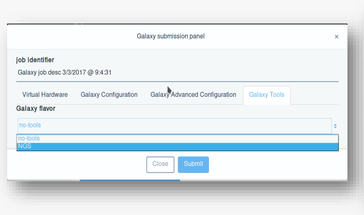


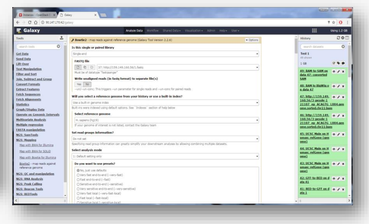
Data privacy is granted through LUKS file system encryption: users will be required to insert a password to encrypt/decrypt data directly on the virtual instance during its deployment, avoiding any interaction with the cloud administrator(s).

TOOLS AND REFERENCE DATA

Each Galaxy instance is customizable, through the web front-end, with different sets of pre installed tools (e.g. SAMtools, BamTools, Bowtie, MACS, RSEM, etc...), exploiting CONDA as default dependency resolver.

Each instance comes with reference data (e.g. genomic sequences) already available for many species, shared among all the instances through the Onedata INDIGO technology, thus avoiding unnecessary and costly data duplication.

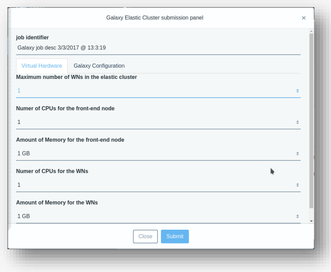




AUTOMATIC ELASTICITY

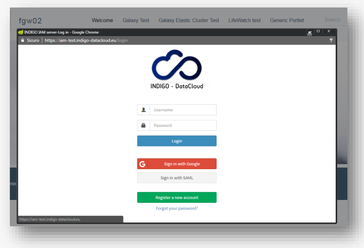
The service is scalable and both users and service providers can chose among a full range of different computational capabilities: from limited ones to serve e.g. small research groups, Galaxy developers or for didactic and training purposes, to instances with elasticity cluster support to deliver enough computational power for any kind of bioinformatic analysis and number of users, opening the route for the migration of public Galaxy instances to this service.

The service provides support for virtual clusters through a dedicated section of the web front-end and allows to instantiate Galaxy with SLURM as Data Resource Manager and to customize the number of virtual nodes, nodes and master virtual hardware.



AUTHENTICATION

The authentication system relies on the IAM INDIGO component. It's integration is complete and it is provided through the service web portal. Integration with the user authentication and authorization infrastructure provided by the ELIXIR AAI is already foreseen.



AUTOMATIC VALIDATION

An array of more than 30 automatic tests on the instance and tools functionalities are executed to validate and test the service when launched . Installed tools are and results compared to reference results.

A video demo showing the front-end functionalities has been presented to the INDIGO First Periodic Review 7-8 November 2016.

CONCLUSIONS

Our service aims to provide the Galaxy workflow manager to end users ranging from small research groups to institutions or SMEs, on suitable computation resources, removing the need to maintain their own hardware and software infrastructure and using resources in a more efficient way, ensuring improved reliability, better performances and the capability to handle larger research activities exploiting the features of the INDIGO-Datacloud components.

When production ready the service will enter the ELIXIR-IIB bioinformatic resources portfolio and will be sponsored by ELIXIR-IIB to become part of the ELIXIR Core Resources. Public Galaxy instances created and running thanks to this service will be part of the public Galaxy resources registry.

INDIGO - DataCloud receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement RIA 653549