Laniakea: a Galaxy on-demand platform for data analysis in life science

M.A. Tangaro, M. Antonacci, V. Spinoso, S. Nicotri, M. Perniola, G. Pesole, F. Zambelli, G. Maggi, G. Donvito

Lifewatch Scientific Community Meeting
27-29 May 2019 - Rome

Outline

Motivation

Laniakea

Service architecture

Main features

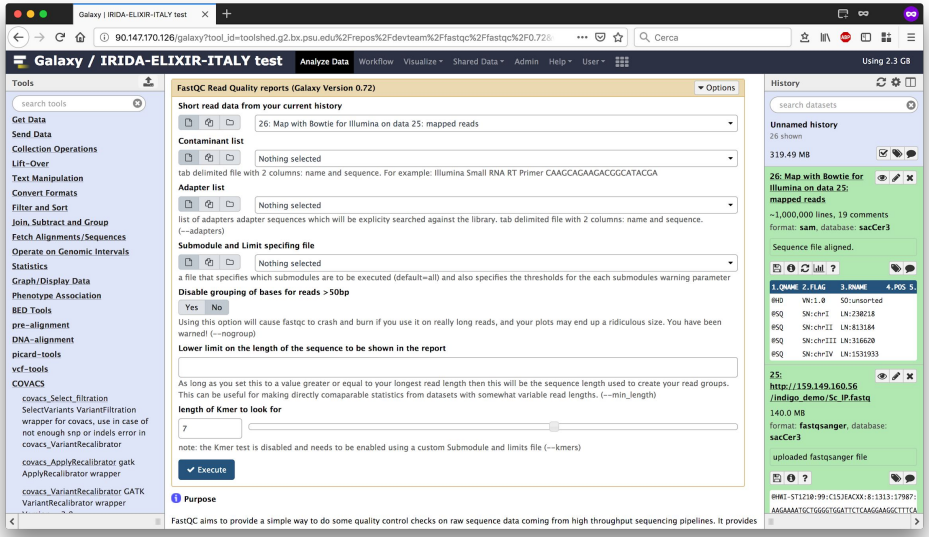
Laniakea for Lifewatch users

Conclusions and outlook

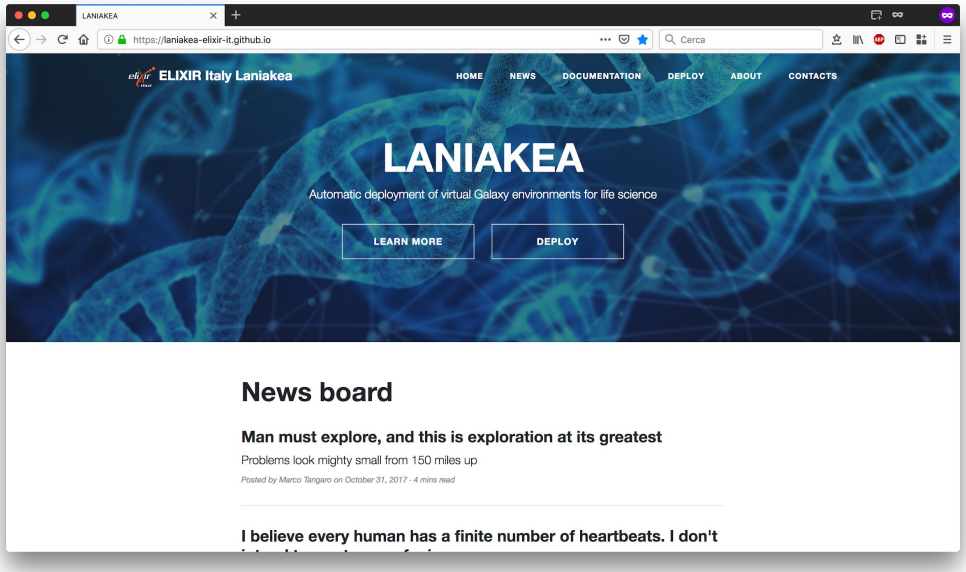
Motivation

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.



LANIAKEA



ELIXIR-Italy in the framework of the H2020 INDIGO-DataCloud project has developed a cloud Galaxy instance provider platform, named **LANIAKEA***, allowing to fully customize each virtual instance through web interface.

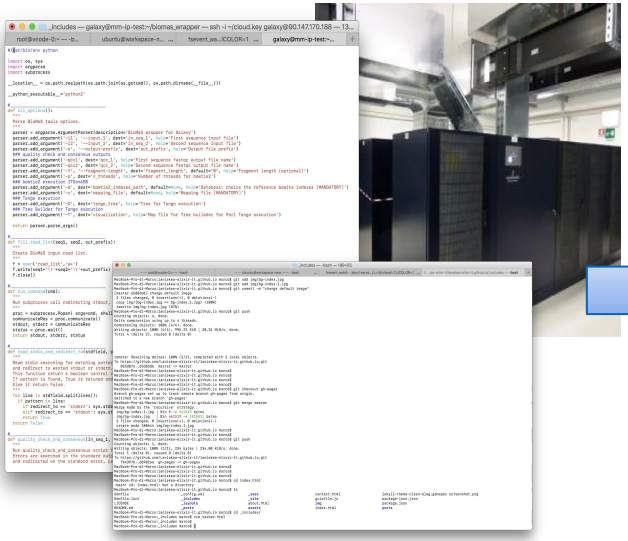
No need for the end user to know the underlying infrastructure.

No need for maintenance of the hardware and software infrastructure.

(*)The Laniakea Supercluster (Laniakea; also called Local Supercluster or Local SCI or sometimes Lenakaeia) is the galaxy supercluster that is home to the Milky Way and approximately 100,000 other nearby galaxies [Source: Wikipedia].

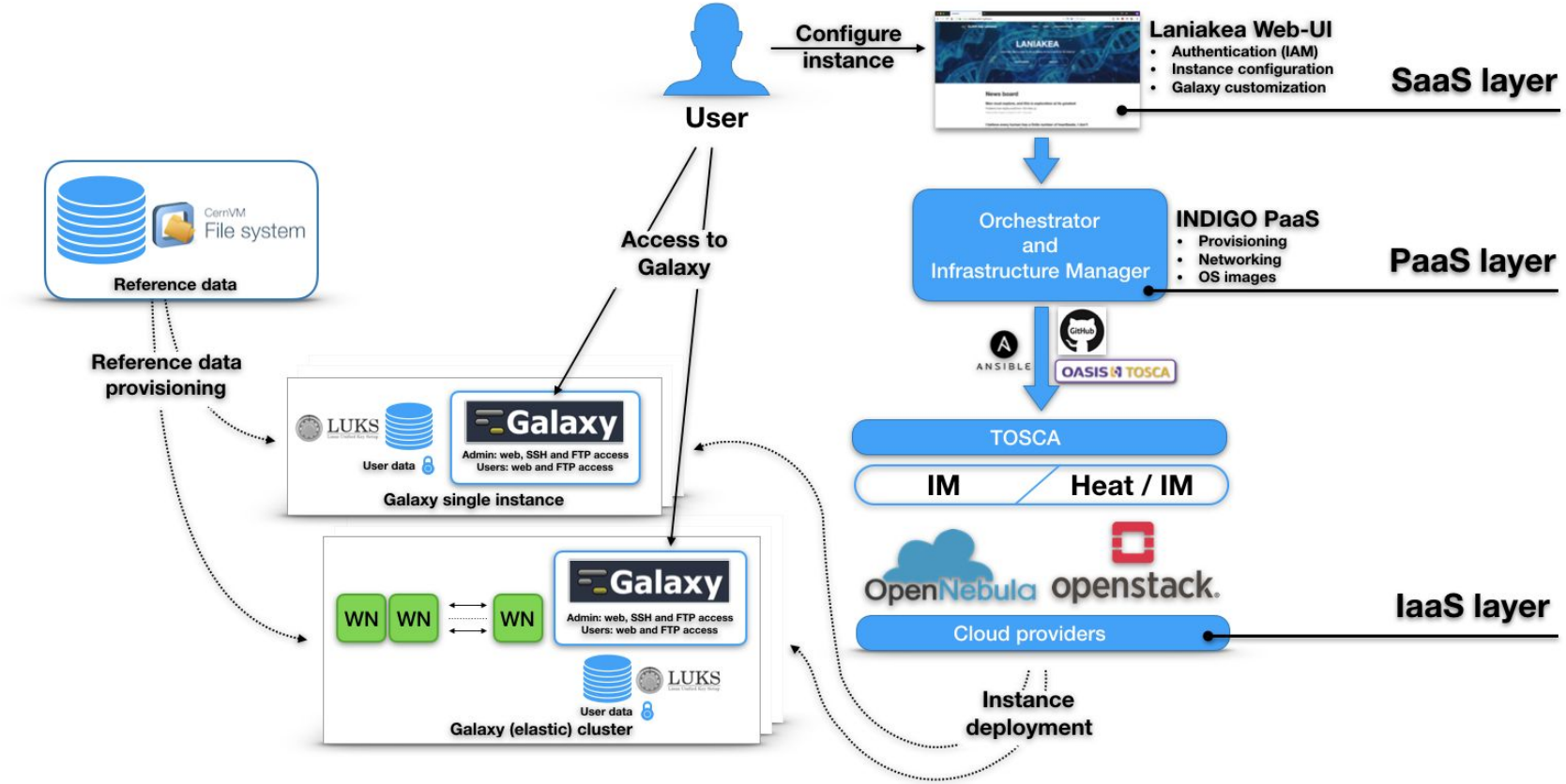
LANIAKEA

Allowing the community to move from command line tools to web user interfaces (on clouds).



Allowing multiple users to exploits same tools, enhancing data FAIRness (Findable, Accessible, Interoperable, Reproducible).

Service architecture



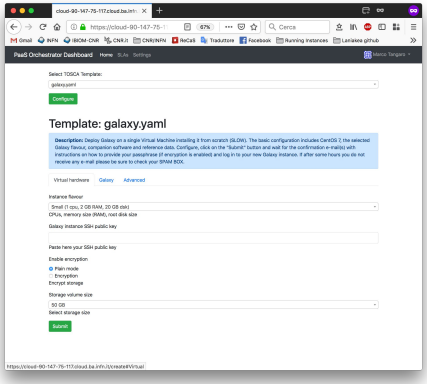
Main features

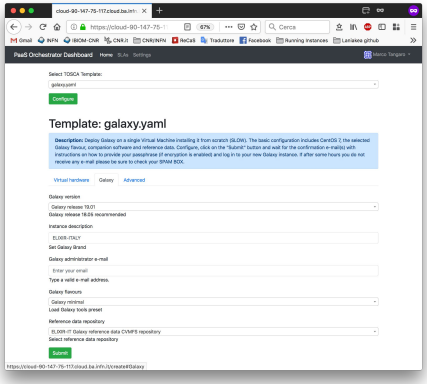
Instance customization

The web front-end provides two different tabs:

- Virtual hardware configuration;
- Galaxy configuration (e.g admin credential);

Dedicated section for cluster deployment.





Galaxy production environment

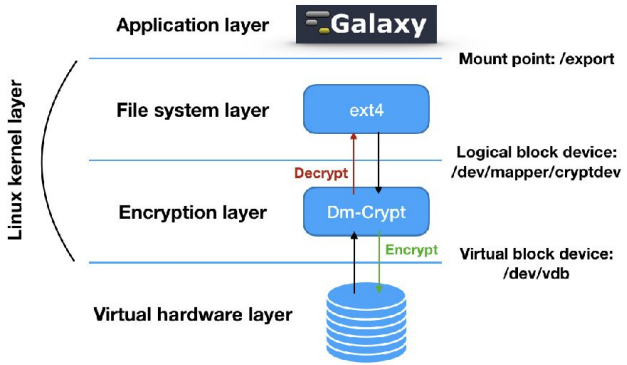
Galaxy is deployed for a multi-user production environment, i.e. there are some additional auxiliary application needed for the best performance (the basic Galaxy installation is suitable for development by a single user):

- PostgreSQL as database
- NGINX as web server (+ upload module)
- uWSGI link between the service and the web server
- Proftpd as FTP server

Main features

Storage encryption

Data privacy is granted through LUKS storage encryption as a service: users are 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 availability

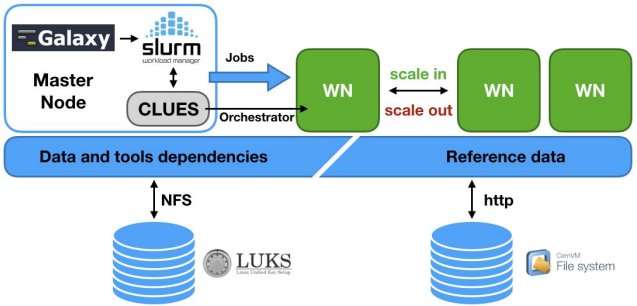
- Galaxy flavors available: each Galaxy instance is customizable, with different sets of pre installed tools.
- Reference data available: each instance comes with reference data (e.g. genomic sequences) already available for many species, shared among all the instances through the CERN-VM FileSystem (cernvm.cern.ch) technology, thus avoiding unnecessary and costly data duplication. Galaxy automatically is configured to properly use them.

Main features

Cluster support

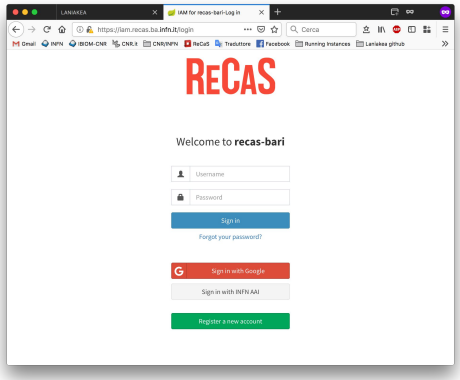
Virtual clusters support through a dedicated section of the web front-end, allowing to instantiate Galaxy with SLURM as resource manager and to customize the cluster virtual hardware.

Dynamic cluster resources scaling: deploying and powering-on new working nodes depending on the cluster workload and powering-off them when no longer needed, depending on the real user requests.



Authentication and Authorization

Robust Authentication and Authorization Infrastructure, supporting different auth mechanisms (e.g. SAML and OpenID Connect).



Laniakea for Lifewatch users

LANIAKEA available for Lifewatch users at ReCaS datacenter soon.

Basic instance:
4 CPUs
8 GB RAM
200 GB external (encrypted) storage
(different requirements will be discussed and possibly supported)
Feedback, suggestion, also for new Galaxy flavours, are more than welcome.

CONTACTS US:
support@recas-bari.it

Conclusions and outlook

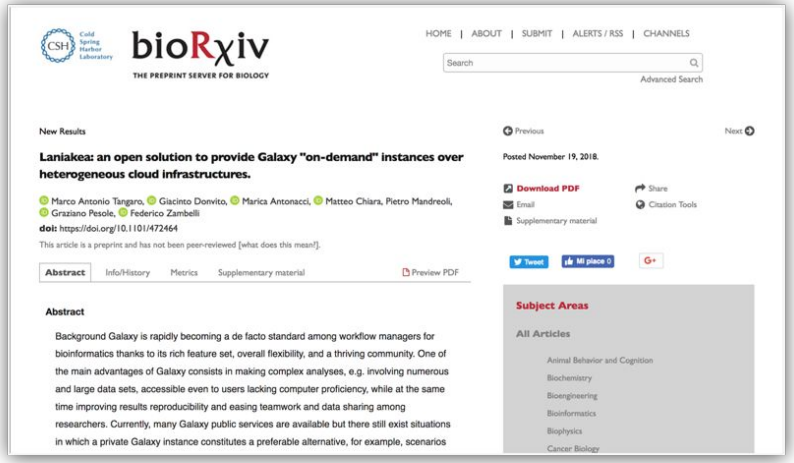
Paper: **Laniakea: an open solution to provide Galaxy "on-demand" instances over heterogeneous cloud infrastructures.**
url: https://www.biorxiv.org/content/early/2018/11/19/472464
doi: https://doi.org/10.1101/472464

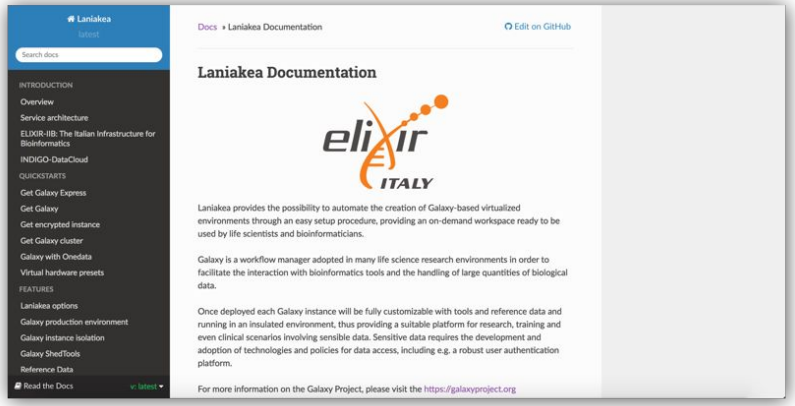
Documentation: http://laniakea.readthedocs.io

GitHub: https://github.com/Laniakea-elixir-it

Demo video: https://www.youtube.com/watch?v=rub3skcs84Q

Future improvements: deployment of dockerized Galaxy and tools.





Thank you!

CONTACTS:

- **support@recas-bari.it**
- Marco Antonio Tangaro (CNR-IBIOM) ma.tangaro@ibiom.cnr.it
- Giacinto Donvito (INFN - Bari Section) giacinto.donvito@ba.infn.it

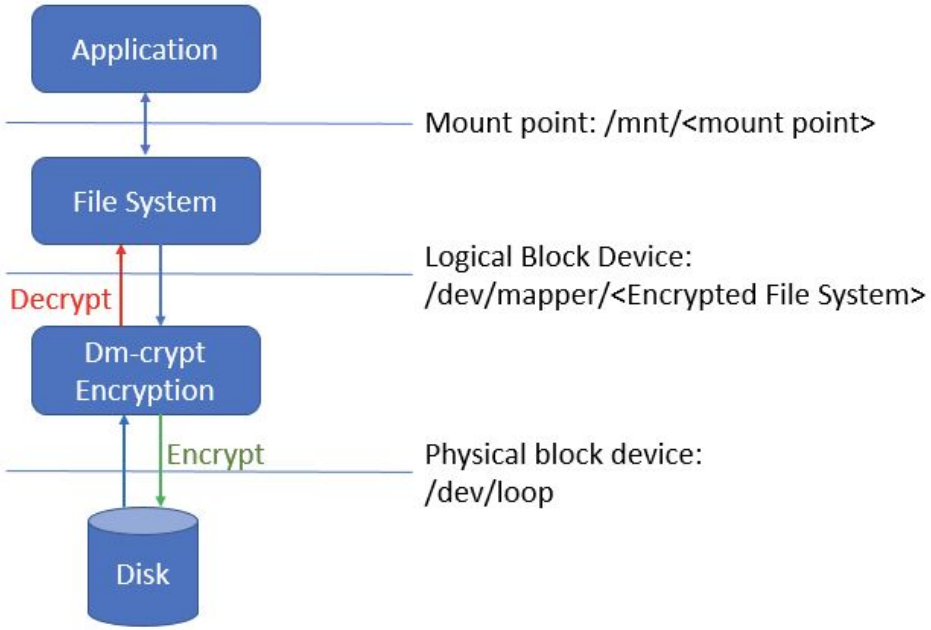
Backup

Motivation

	Ready to use	Quota	Galaxy Custom.	Maintenance	Costs	Data Privacy
Public Servers	✓	✗ Strongly Limited	✗	✓ Up to service provider	✓ No costs (usually)	✗
Local Install	✗	✓	✓	✗ Required	✗ Costly	✓
Cloud* (e.g. Amazon)	✓	✓ Costs Dependent	✓	✓ Only Galaxy Maintenance	✗ Costly	✗ ✓

(*) **Over 2400 Galaxy cloud servers launched in 2015** (Nucleic Acids Research (2016) doi: 10.1093/nar/gkw343)

Block Storage Encryption



Block Storage Encryption

Bash scripting + Ansible + INDIGO PaaS Orchestrator:

- Storage Encryption as a Service
- Dependency resolution
- Script instance lock, i.e. is not possible to run two instances of the encryption script.
- Configurable (encryption algorithm, key size, hash algorithm, mountpoint, filesystem).
- Automatic configuration file creation to open/close the volume with one command.

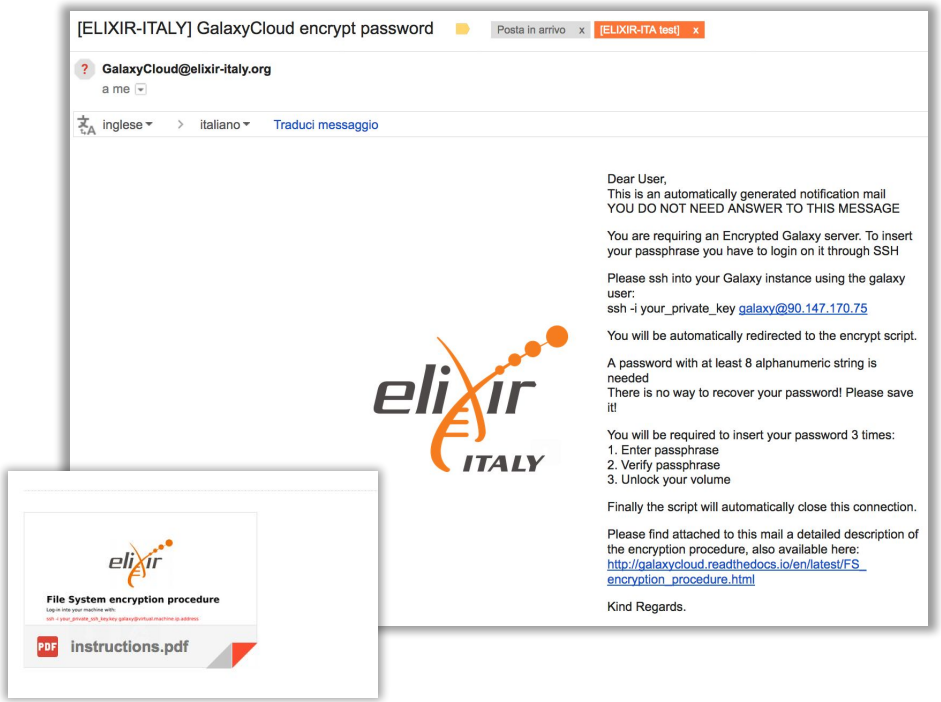
Block Storage Encryption

Ansible automates the encryption procedure, installing the scripts, informing, by mail, the user once the system is ready to accept the password.

The encryption procedure summary is reported by mail, while a detailed step-by-step how-to is sent attached.

Script to easily manage the LUKS volume is added to each virtual instance:

- check if the volume is correctly mounted,
- Mount and open LUKS volumes.
- Close and umount LUKS volumes.



Block Storage Encryption

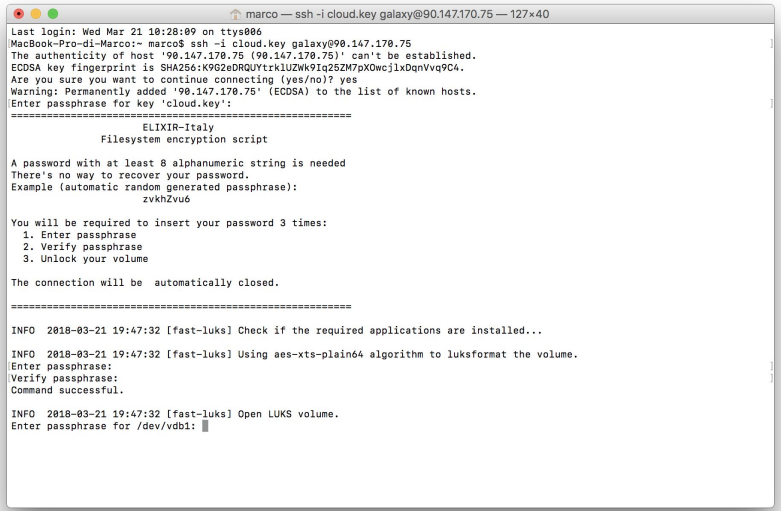
Automatic logout after password injection: the encryption procedure continues in background.

Default encryption algorithm:

- aes-xts-plain64 encryption
- 256 bit key
- sha256 as hash algorithm used for key derivation.

Script to easily manage the LUKS volume is added to each virtual instance:

- check if the volume is correctly mounted,
- Mount and open LUKS volumes.
- Close and umount LUKS volumes.



Block storage encryption

- Test on unmounted encrypted devices:
  - Create two volumes, one encrypted
  - Put inside the same file
  - Umount volumes
  - Create volume binary images and HexDump the binary image with xdd
  - Grep non-zero bytes and search for the file content

  It is possible to see the file content only on the un-encrypted volume.

- Try to open the volume when active (LUKS volume opened and mounted, Galaxy running) in the Virtual Machine.

  Test executed on the cloud controller as administrator.

  It is not possible to mount the volume without the user password.

Automatic elasticity

ELIXIR-IIB: Galaxy as a Cloud Service