On-demand cloud-based secure environments for analysing personal and health data

Tangaro Marco Antonio (IBIOM-CNR)
Donvito G., Antonacci M., Foggetti N., Zambelli F.

CS3 2023 - Cloud Storage Synchronization and Sharing
March 06-08, 2023

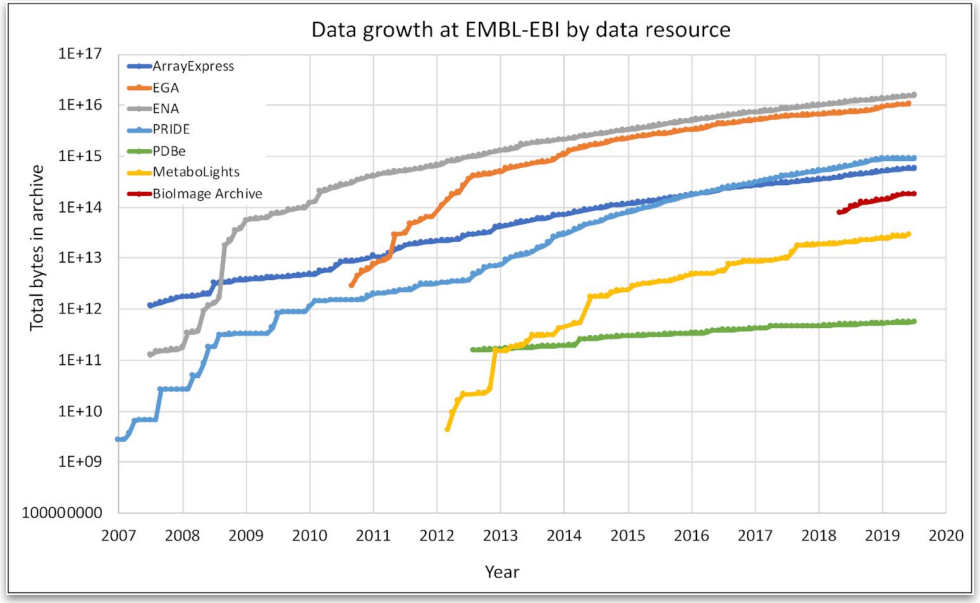
Outline

- Introduction: tools, data, compute and … GDPR
- Galaxy
- Laniakea
- Encryption
- VPN
- Conclusions

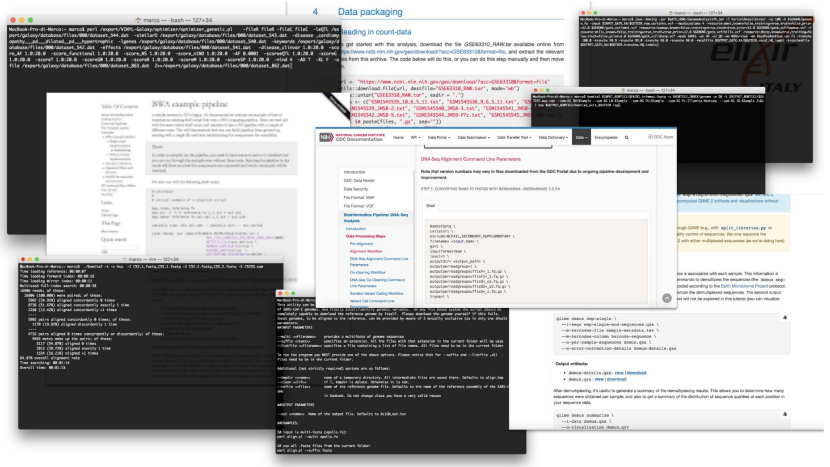
Motivation

DATA SOURCES

DATA STORAGE



DATA ANALYSIS TOOLS



DATA PROTECTION (GDPR)

Galaxy PROJECT

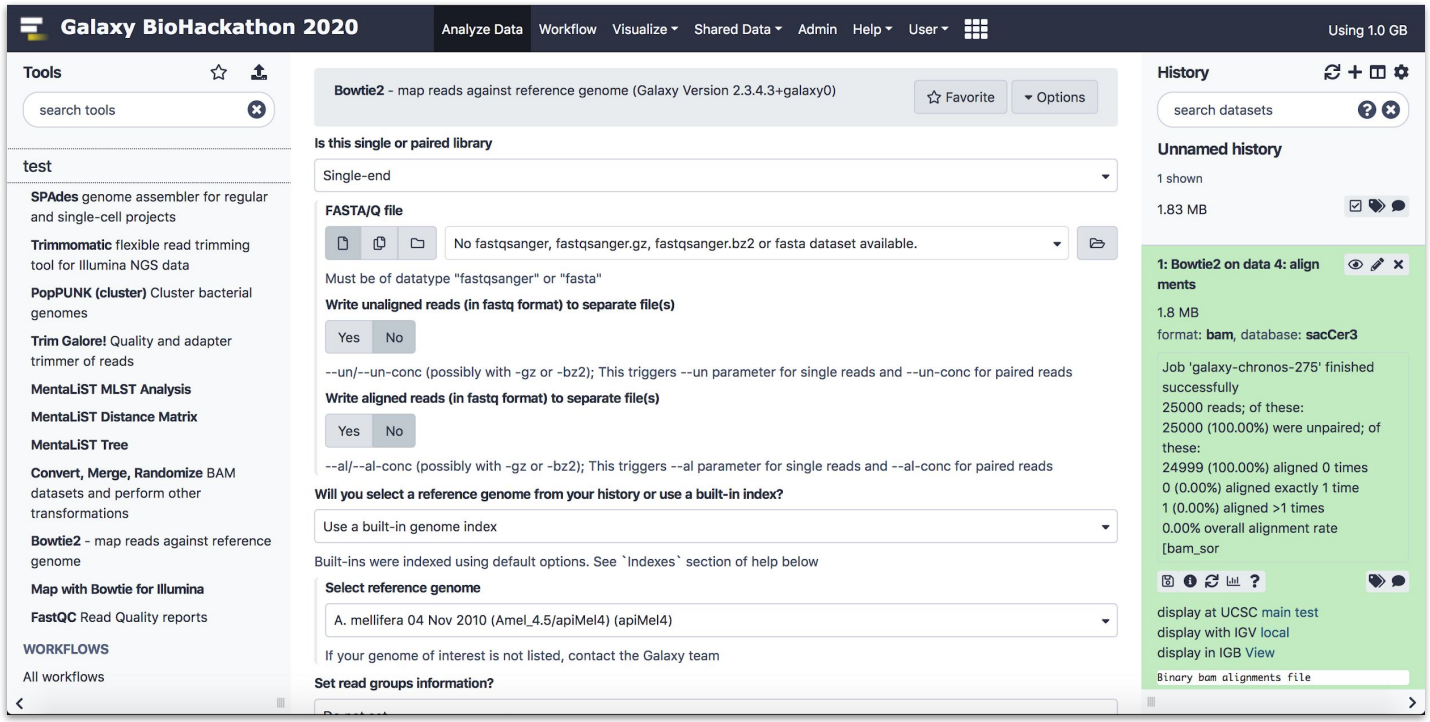
Galaxy is a workflow manager adopted in many life science research environments in order to facilitate the interaction with bioinformatics tools and the handling of large quantities of biological data.

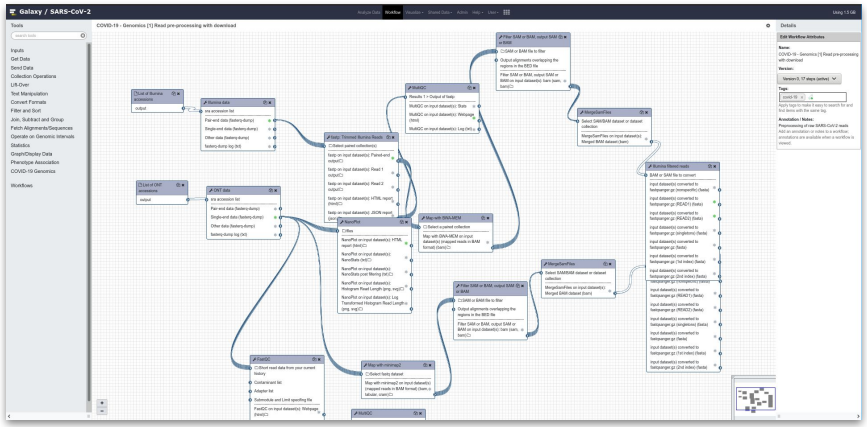
Through a coherent work environment and an **user-friendly web interface** it organizes data, tools and workflows providing **reproducibility, transparency** and **simple data sharing** functionalities to users.

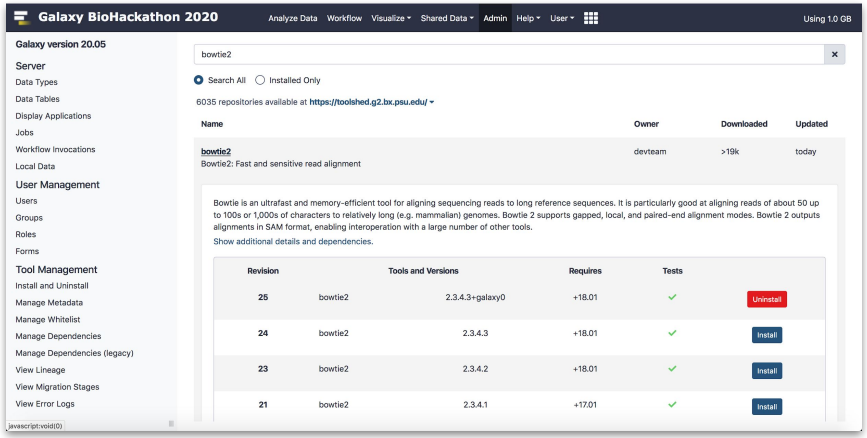
galaxyproject.org

Galaxy PROJECT

elixir ITALY







- Tools graphical user interface.
- Workflow graphical user interface
- App store" to all Galaxies worldwide
- Tools dependencies automatically solved

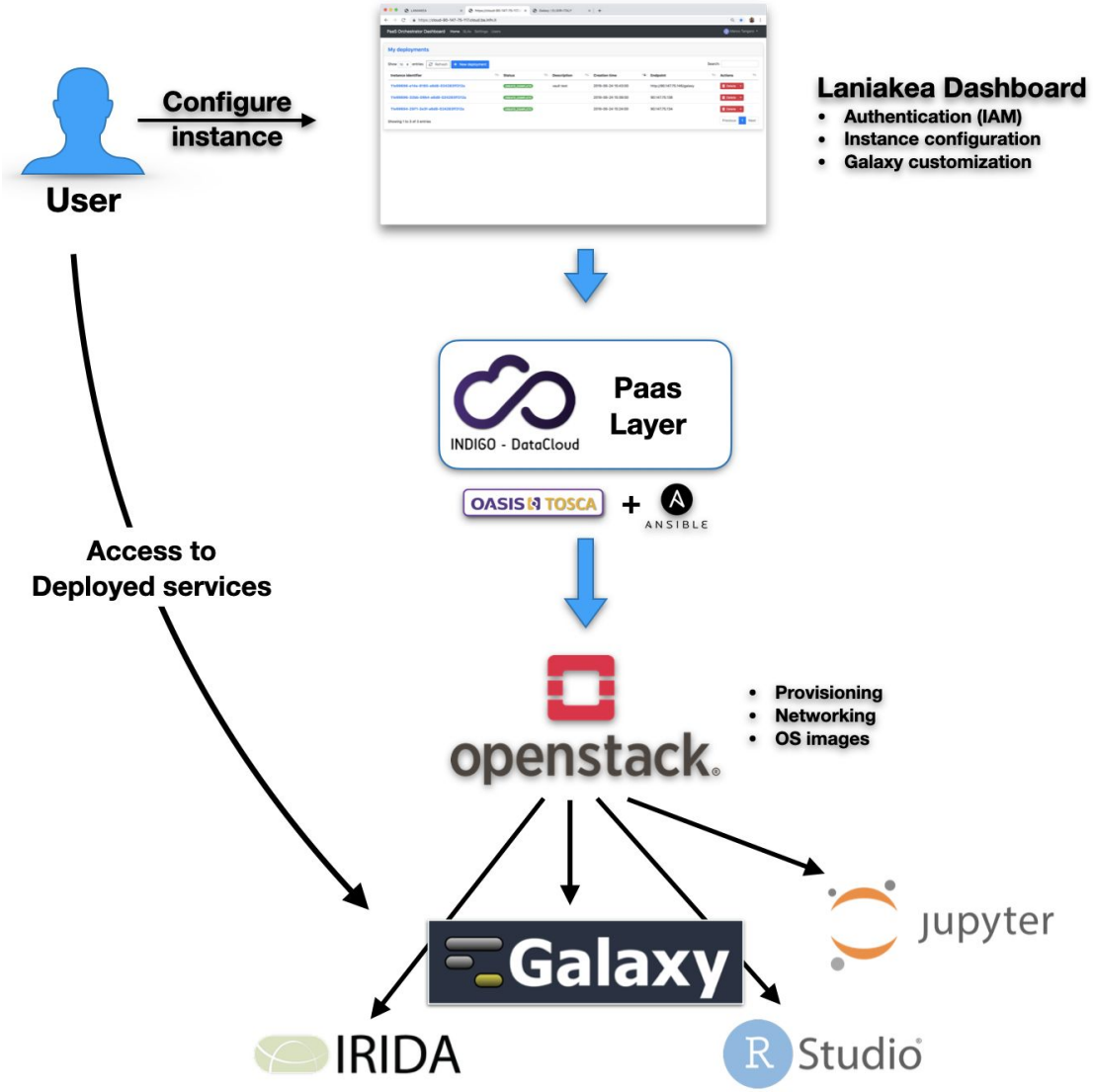
Laniakea

LANIAKEA IS A CLOUD BASED GALAXY INSTANCE PROVIDER.
https://laniakea-elixir-it.github.io/

- Laniakea relies on commonly used Life Science Open Source tools, e.g. Galaxy, RStudio, Jupyter, HashiCorp Vault, LUKS and SLURM.
- Laniakea is European Open Science Cloud service provider.

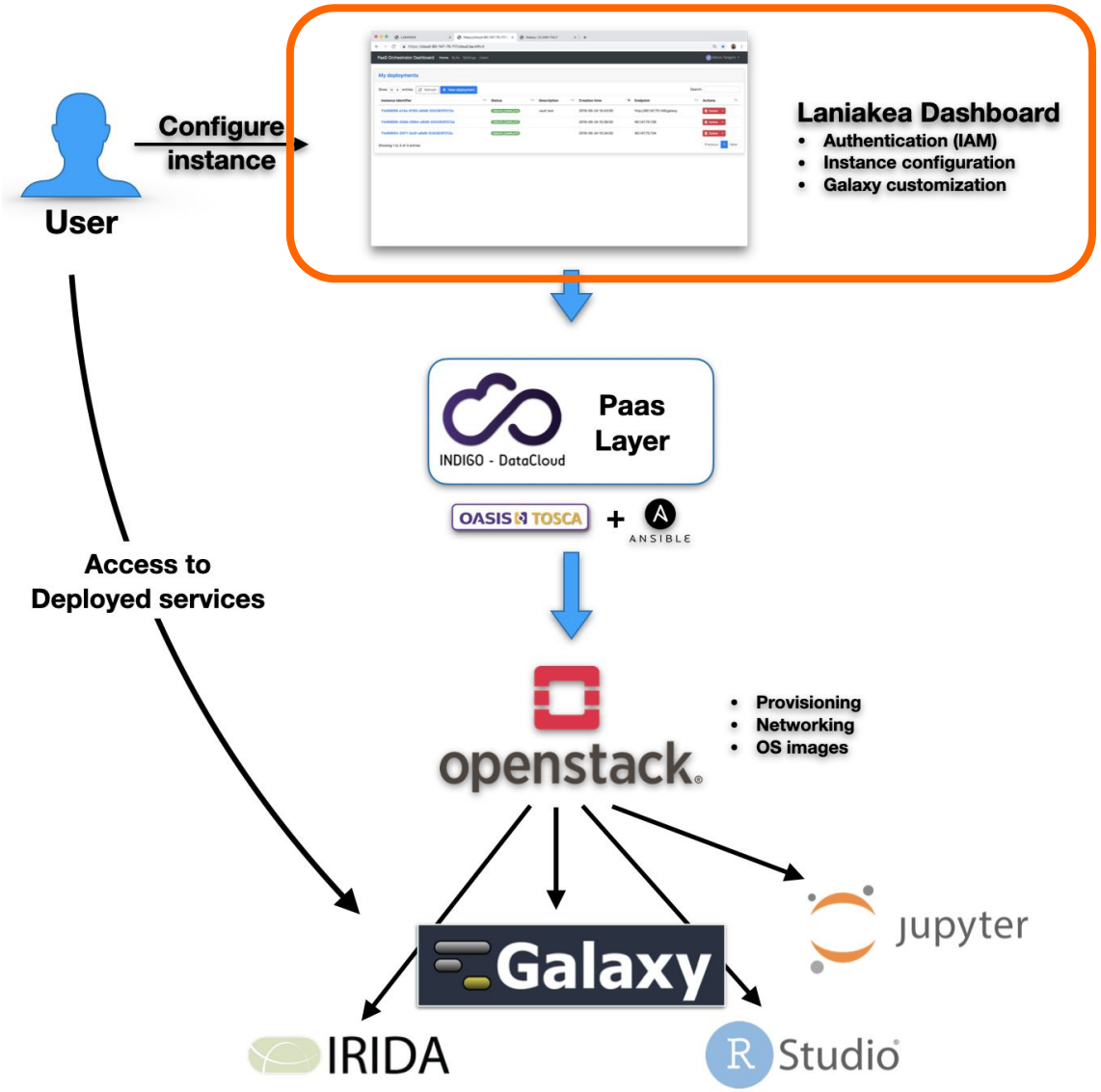
Recommended for scenarios where users need full administrative control over a private Galaxy instance.

Laniakea architecture (simplified view)



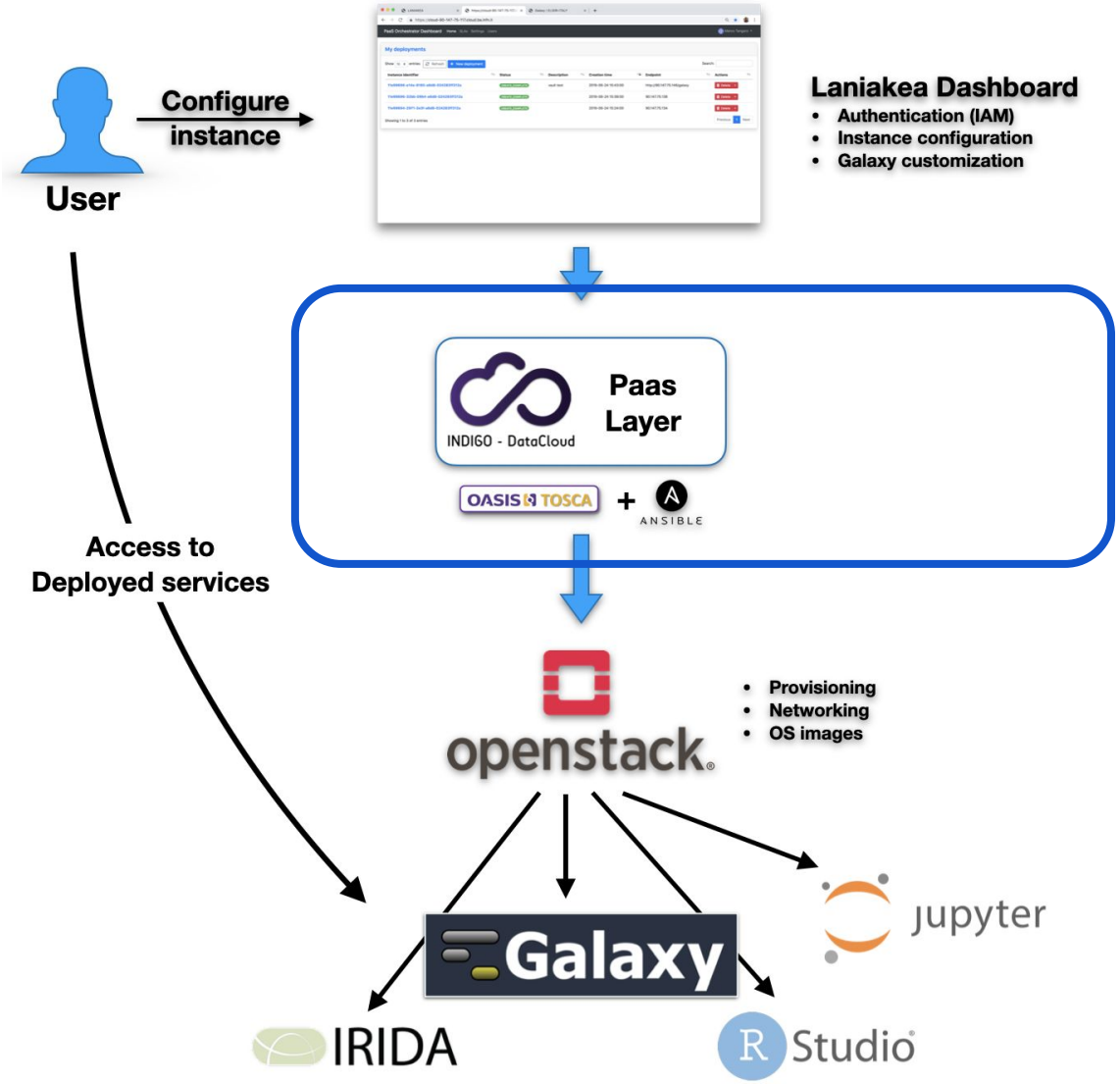
- **Dashboard -** User friendly access to configure and launch a Galaxy instance
- **INDIGO PaaS -** Galaxy automatic d
- **Cloud Providers -** ReCaS-Bari

Laniakea architecture



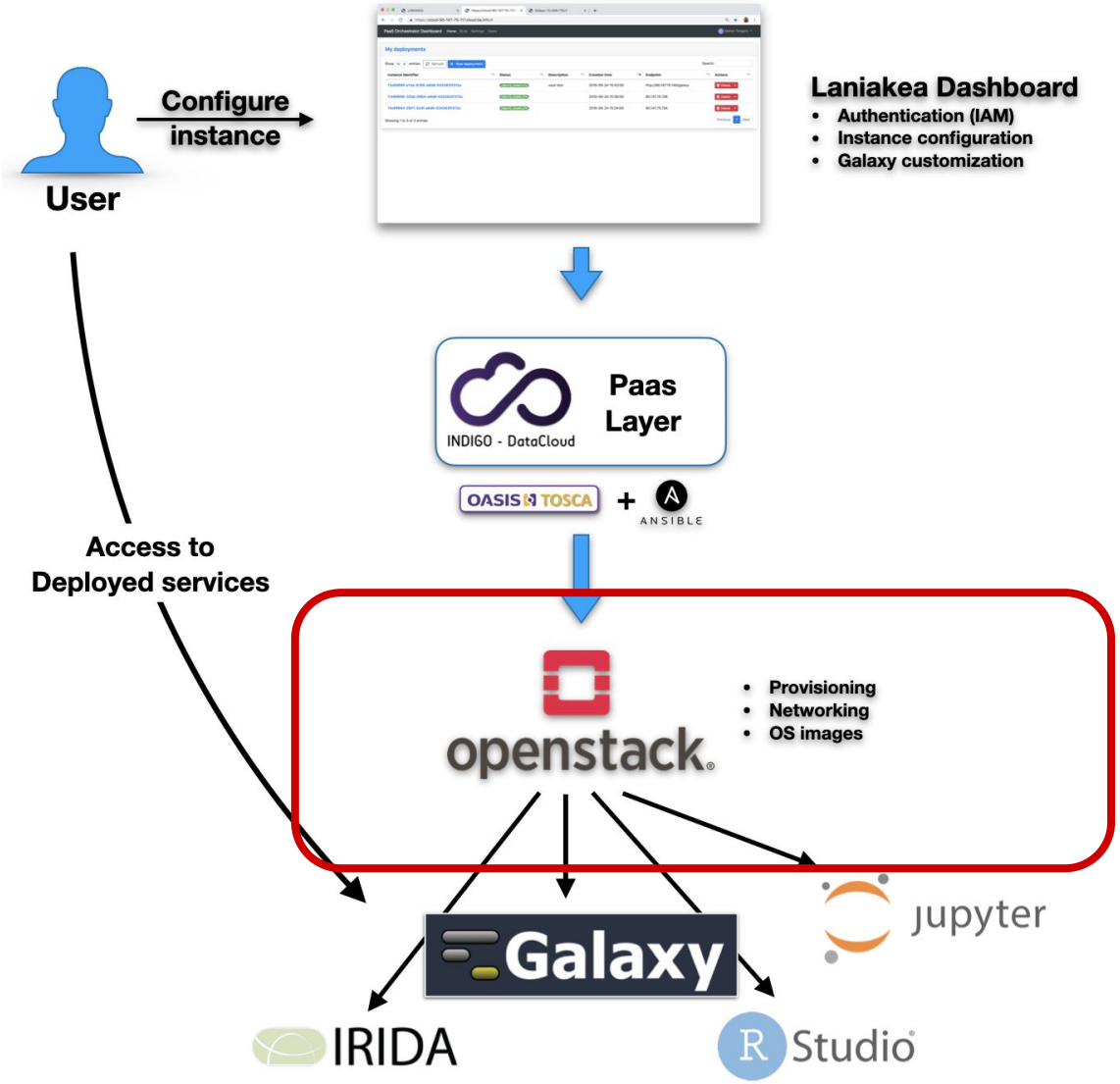
- **Dashboard -** User friendly access to configure and launch a Galaxy instance
- **INDIGO PaaS -** Galaxy automatic d
- **Cloud Providers -** ReCaS-Bari

Laniakea architecture



- **Dashboard -** User friendly access to configure and launch a Galaxy instance
- **INDIGO PaaS -** Galaxy automatic deployment
- **Cloud Providers -** ReCaS-Bari

Laniakea architecture



- **Dashboard -** User friendly access to configure and launch a Galaxy instance
- **INDIGO PaaS -** Galaxy automatic deployment
- **Cloud Providers -** ReCaS-Bari

Isolated environment features

Storage Encryption - Data privacy is provided through encryption "on-demand".

Deployments under Private Network - Automatic deployments of virtual environments on private networks.

Laniakea encryption

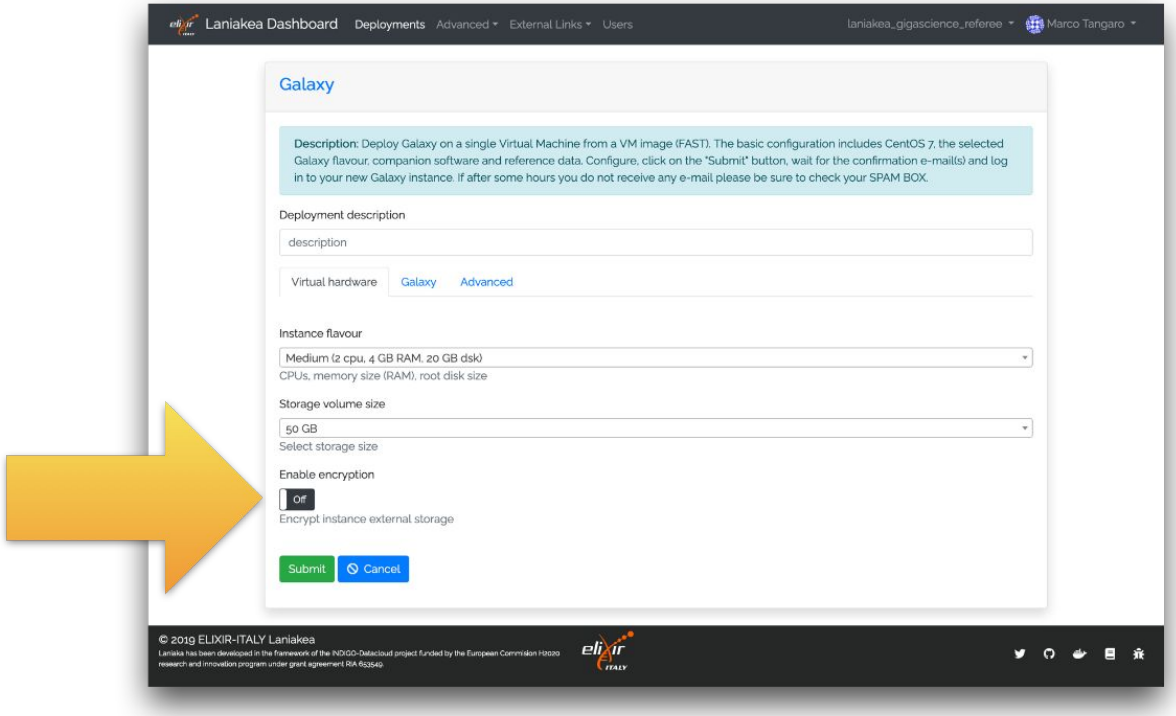
Laniakea encryption

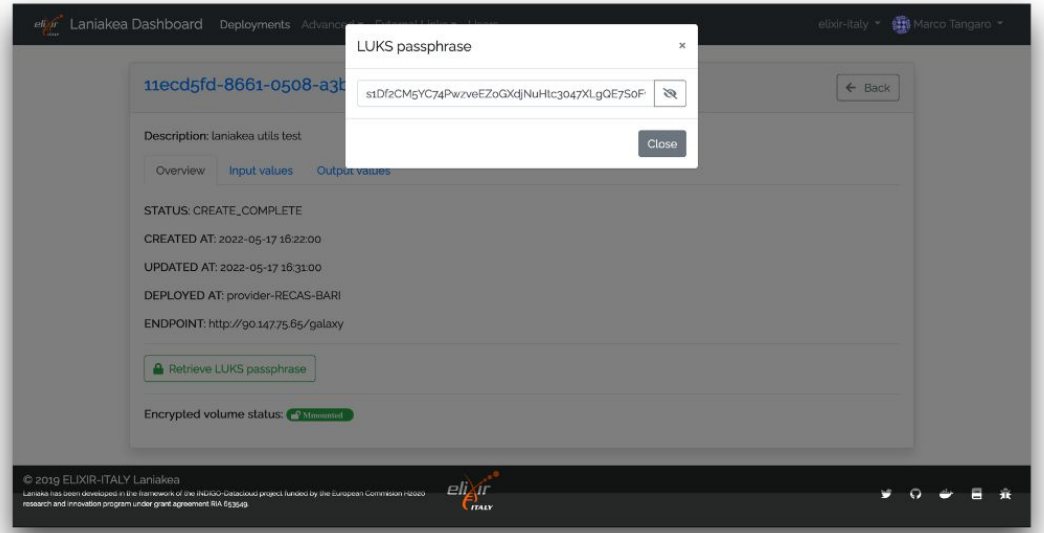
The user data privacy is granted through **LUKS** storage encryption as a service: the encryption procedure is automated in order to simplify the user experience, each user can encrypt storage on-demand, using a strong random alphanumerical passphrase.

This has been achieved integrating the Dashboard and the key management system **Hashicorp Vault** (vaultproject.io) to store encryption keys, which are shown in the Laniakea Dashboard only if explicitly requested by the user.

Vault is a tool for securely accessing "secrets". A secret is everything you want to tightly control access to, such as encryption passphrases and user creds.

User perspective





The user can enable the storage encryption using a switch toggle in the Instance "Virtual hardware" configuration tab.

The procedure is completely automated.

The storage is encrypted and the User can retrieve his random passphrase from the Instance overview page.

The underlying infrastructure

LUKS - Linux Unified Kernel Setup

A python package (pyLUKS) is used to encrypt the storage using a random passphrase and then store it on **Hashicorp Vault**.

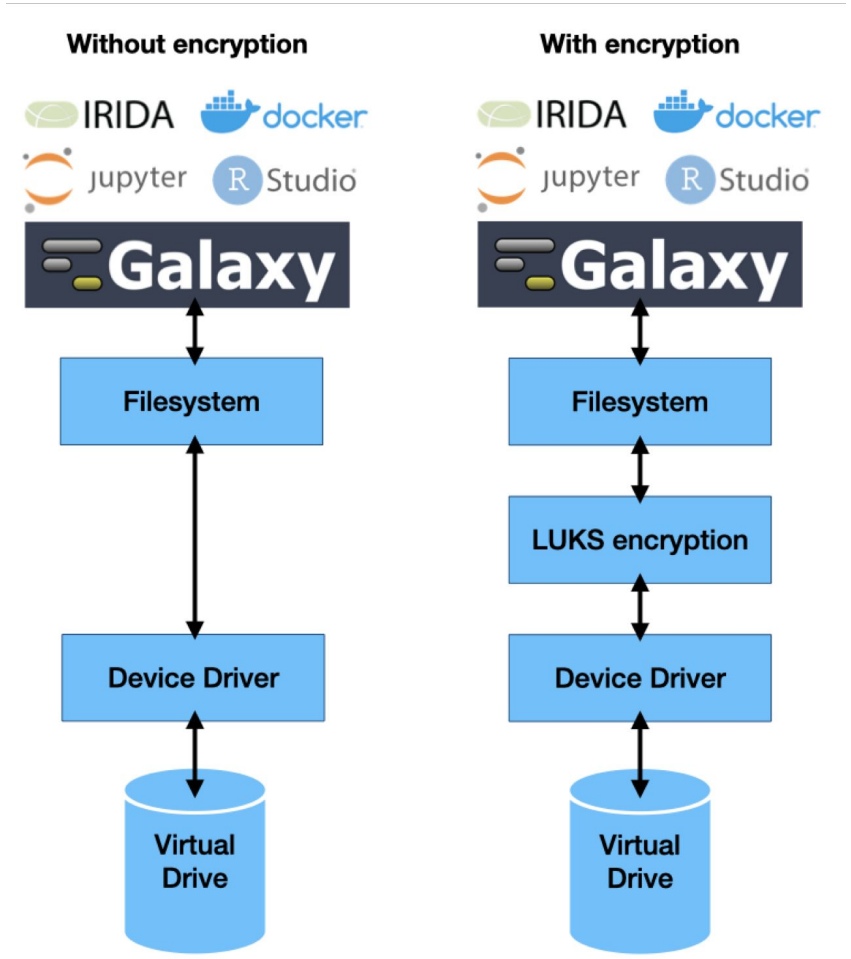
The encryption layer sits between the physical disk and the file system.

Galaxy, or any other application, is unaware of storage encryption.

Galaxy exploits a specific mount point in order to store and retrieve files. Files are encrypted when stored to disk and decrypted when read.

Default encryption algorithm:

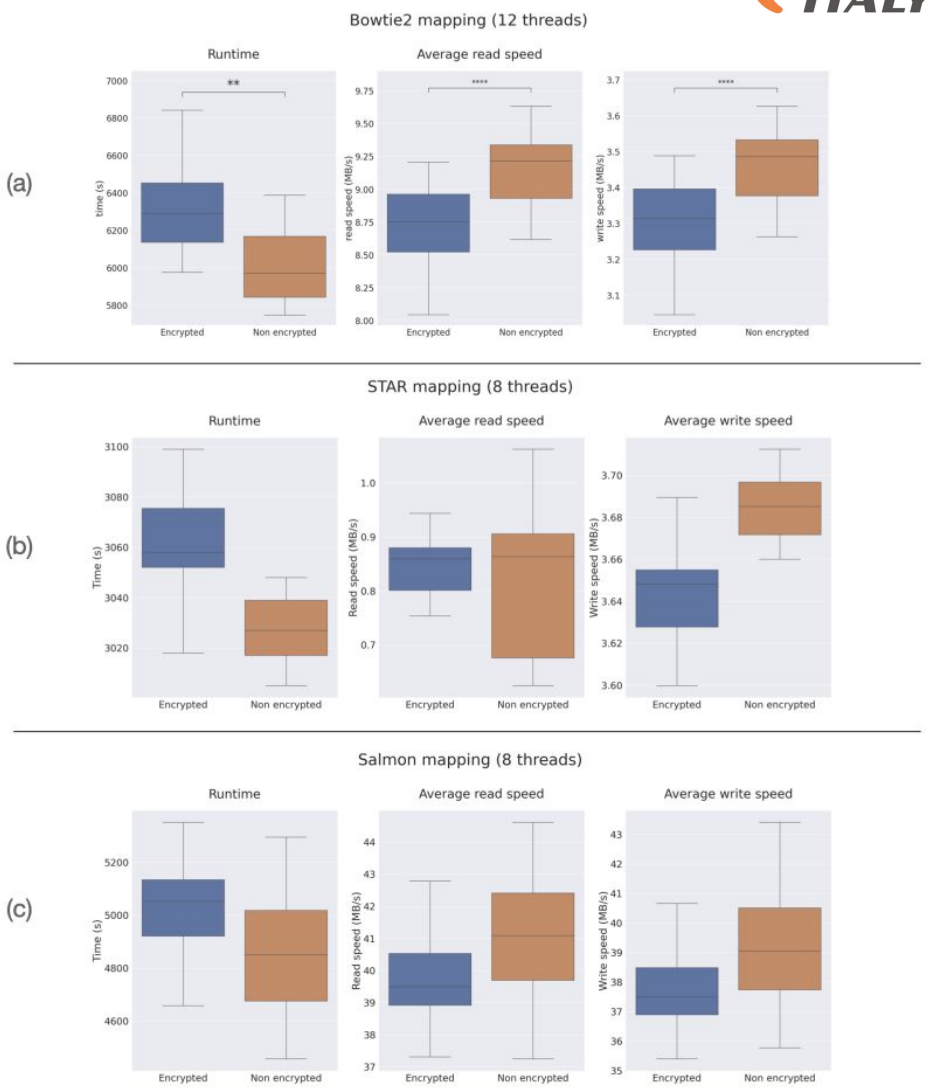
- aes-xts-plain64 encryption
- 256 bit key
- sha256 as hash algorithm used for key derivation.



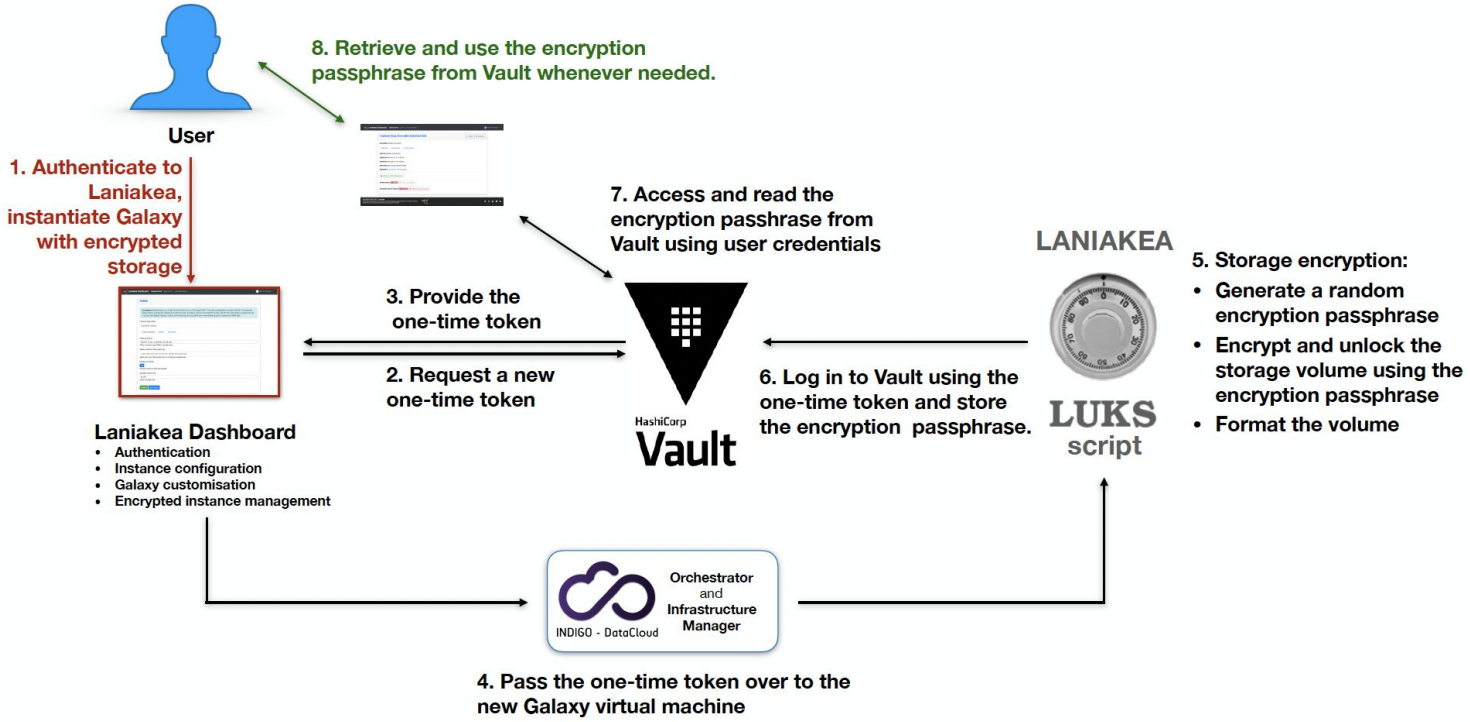
The underlying infrastructure

To evaluate the impact of the storage encryption layer on the performance of the main application supported by Laniakea, i.e., Galaxy, we measured jobs runtime and read/write speed on Virtual Machines generated by the Laniakea@ReCaS data center with and without storage encryption.

The impact on the performance of using the encryption layer, as measured in all our tests, is limited to ~5% or less across all the measured parameters and conditions.



The underlying infrastructure



(*) Tokens are the core method for authentication within Vault. After the authentication on the Laniakea Dashboard, tokens are dynamically generated based with a specific policy allows to write/read/update secrets.

1. User Authentication.
2. **A short lived, write only token, usable only once**, is delivered to the Laniakea encryption script on the VM. There's no update policy: this token can't overwrite other passphrases for security reasons.
3. The Storage volume is encrypted by Laniakea pyLUKS package.
4. The passphrase is sent to Vault by Laniakea pyLUKS.
5. After the instance has been successfully deployed the user can retrieve his password through the Dashboard.
6. The user reads the password on the Dashboard.

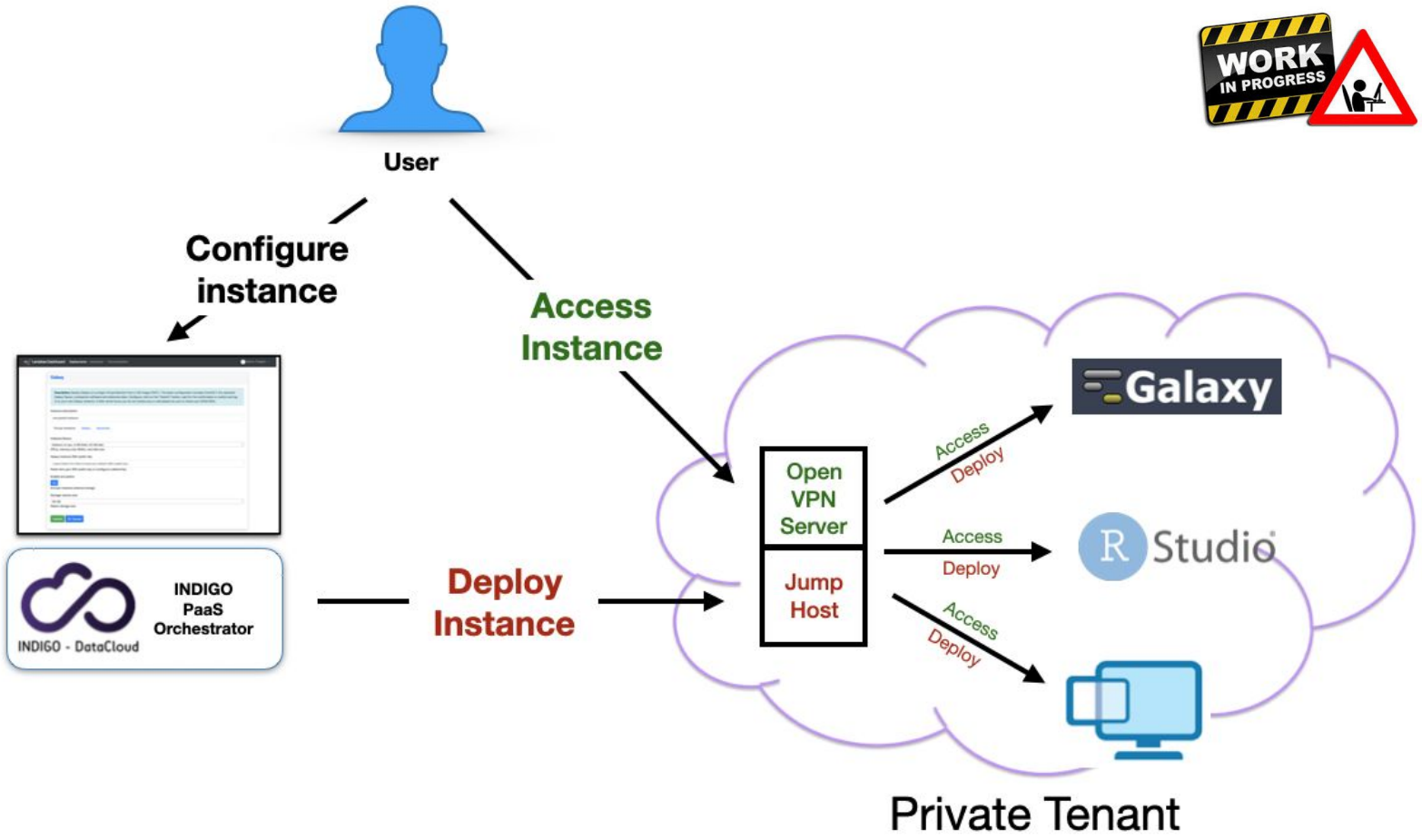
Deployments under VPN

Deployments under VPN

VPN isolated environments - Automatic deployments of virtual environments on private networks.

Isolation is reached using Tenant and security groups properties, granting the access only through VPN authentication.

User authentication to the VPN using the same Laniakea credentials.

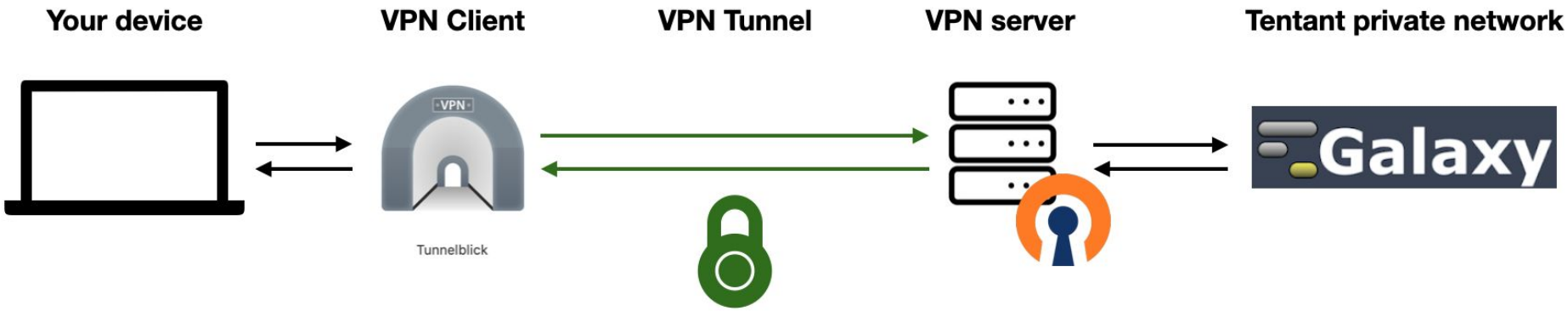


Deployments under VPN

The VPN is based on OpenVPN, with clients and server are configured to use TPC protocol.

We have developed a PAM plugin to enable authentication through OpenID Connect, exploiting Oauth2 device flow:

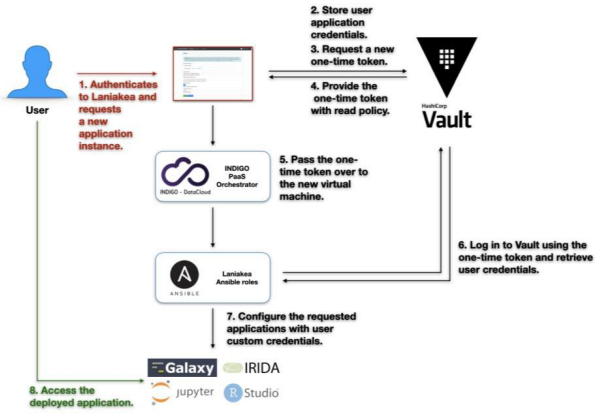
1. the user connects to the VPN server using an OpenVPN client
2. PAM is configured to send verification code by mail to the user.
3. the user can authenticate with its own Laniakea credentials.
4. the OIDC provider (INDIGO-IAM) sends the access token to the VPN server, that is now able to verify users identity and authorizations.
5. if the user owns the right tenant permissions, he is granted access to the private network and can finally interact with the deployed application



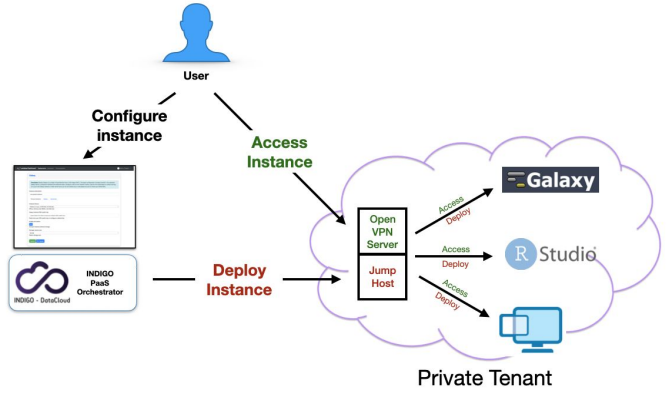
Conclusions

The storage encryption procedure has been extended to allow also users' credentials customisation for many applications.

Data are still potentially exposed to attacks against the VM itself, where Galaxy or other applications need to consume them. To tackle this, we are working to provide Laniakea's users with the possibility to hide deployed applications beyond a Virtual Private Network, achieving even more robust isolation of the research environment.



These approaches can help promoting the adoption of the on-demand model for Life Science and biomedical applications, making compute infrastructures more readily available to potential users even in the case of tight requirements for data protection.



Thanks for your attention

CONTACTS:

Graziano Pesole (ELIXIR-ITALY Head of Node) g.pesole@ibiom.cnr.it

Federico Zambelli (ELIXIR-ITALY technical coordinator) federico.zambelli@unimi.it

Giacinto Donvito (Compute platform ELIXIR-ITALY) giacinto.donvito@ba.infn.it

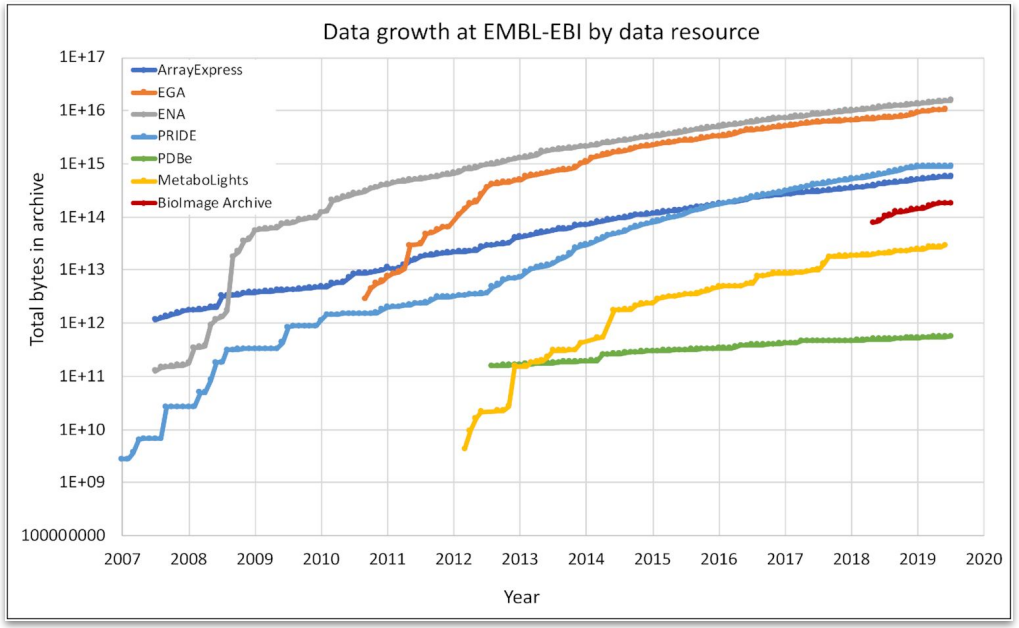
Nadina Foggetti (Legal expert) nadina.foggetti@ba.infn.it

Marica Antonacci (PaaS developer) marica.antonacci@ba.infn.it

Marco Antonio Tangaro (Laniakea chief developer) ma.tangaro@ibiom.cnr.it

Backup

Data



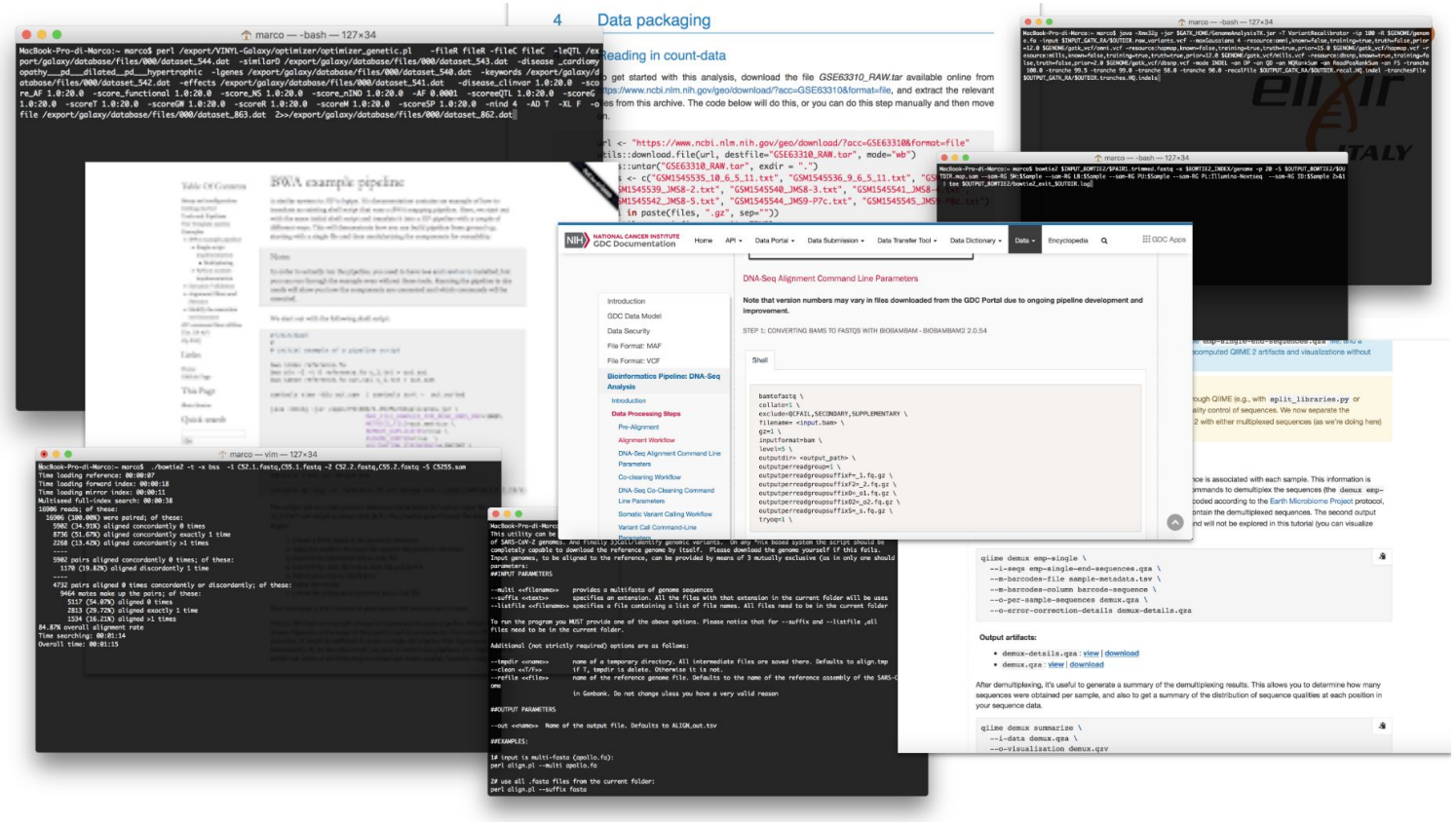
Genomic data are distributed across several sequencing centres and/or IT infrastructures

Data volume growing not only in quantity but also on variety!

Data growth at EMBL-EBI Source: Charles E. Cook et al. Nucl. Acids Res. 2020; Volume 48, Issue D1, Pages D17-D23

Discipline	Data size	# devices
HEP-LHC	15PB/year	1
Astronomy	15PB/year	several
Genomics	0.4TB/genome	>1000

Tools



The command line is the standard way to use most bioinformatics tools:

- Plenty of parameters
- Multiple input and output data and formats
- Reference data
- Need to run multiple times
- Need to change the parameters for each run

Workflows involve more than one tools!

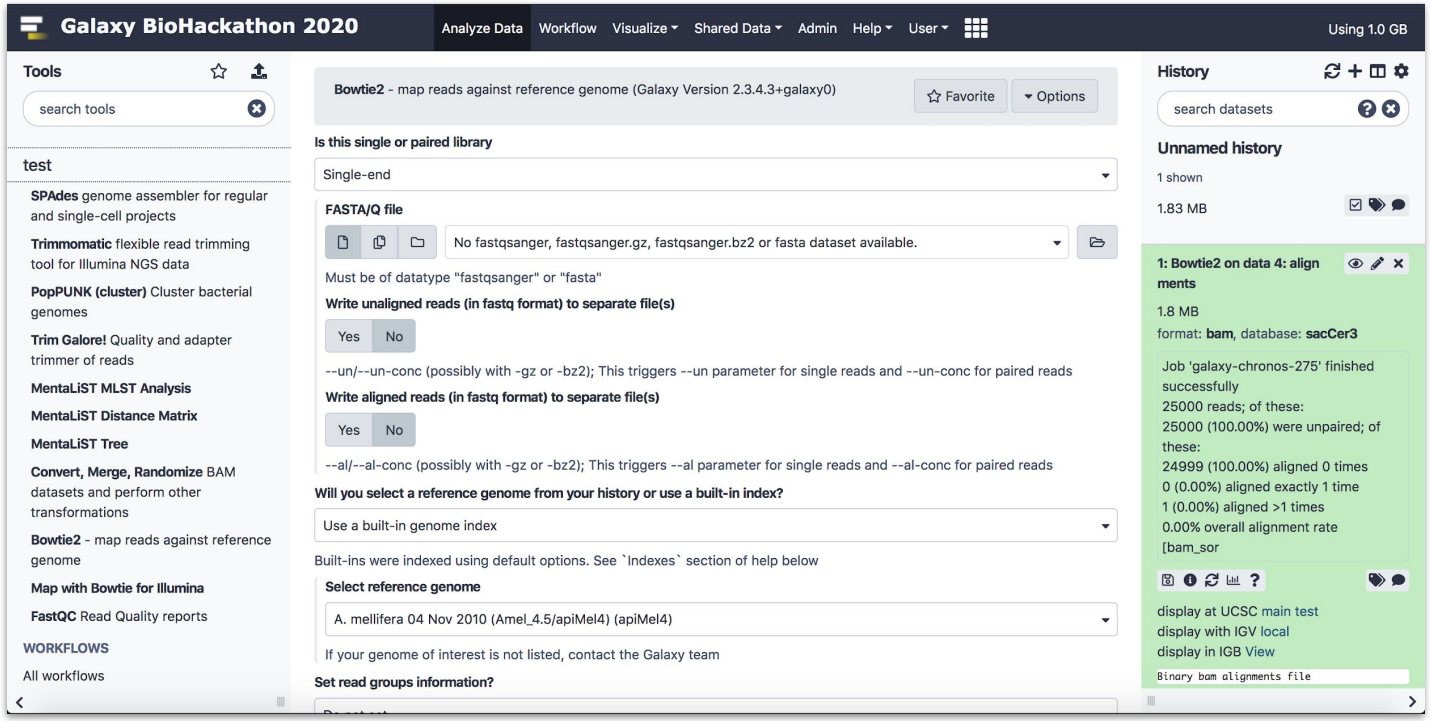
Tools are usually manually installed.

GDPR

The GDPR explicitly recognizes the sensitive nature of the collected genetic data (Article 9), but at the same time permits sensitive genetic data processing for scientific research purposes (Article 89(1)), provided this is allowed by EU or Member States law framework and appropriate safeguards measures are in place.

Galaxy PROJECT

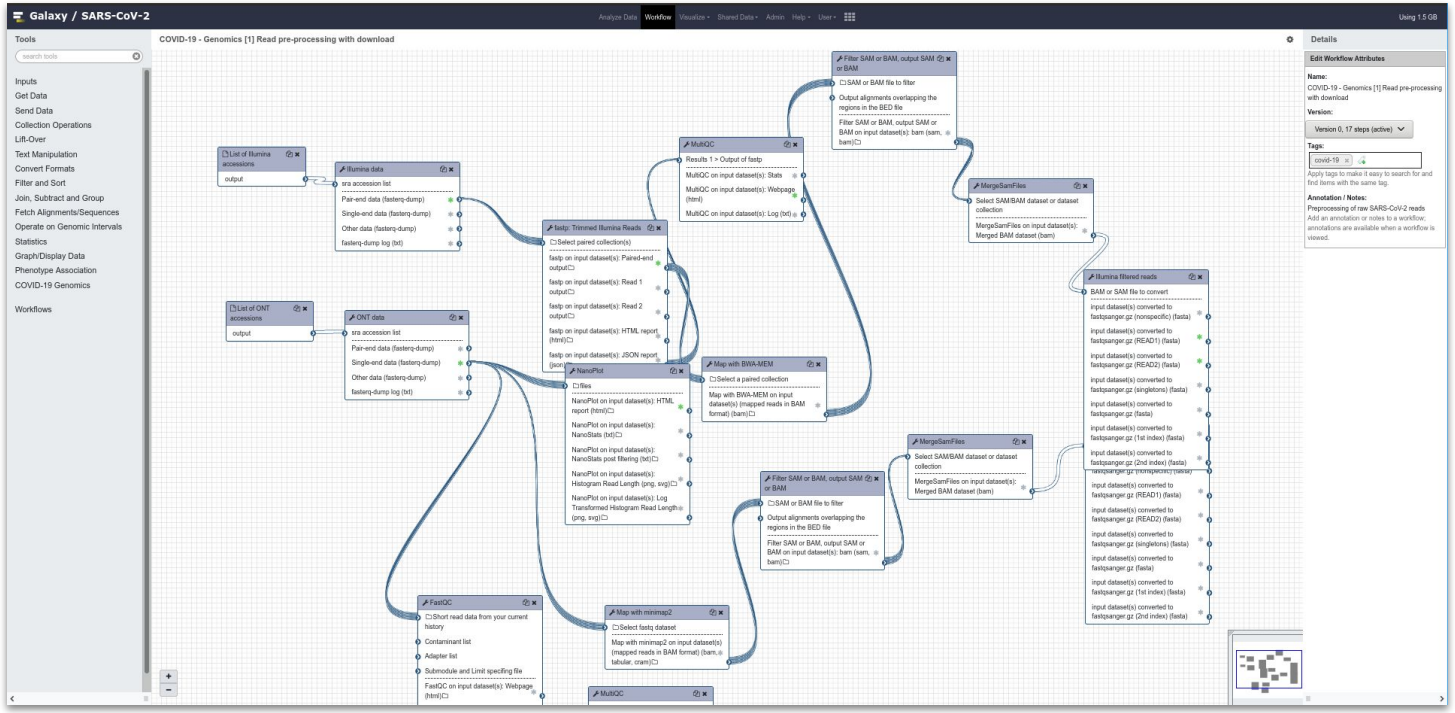
elixir ITALY



- Tools graphical user interface.
- Input and output data management.
- Output visualization.
- Data and analysis parameters sharing.
- Used tools and parameters configuration always available **-> analysis reproducibility**.
- Reference data already available for many tools.

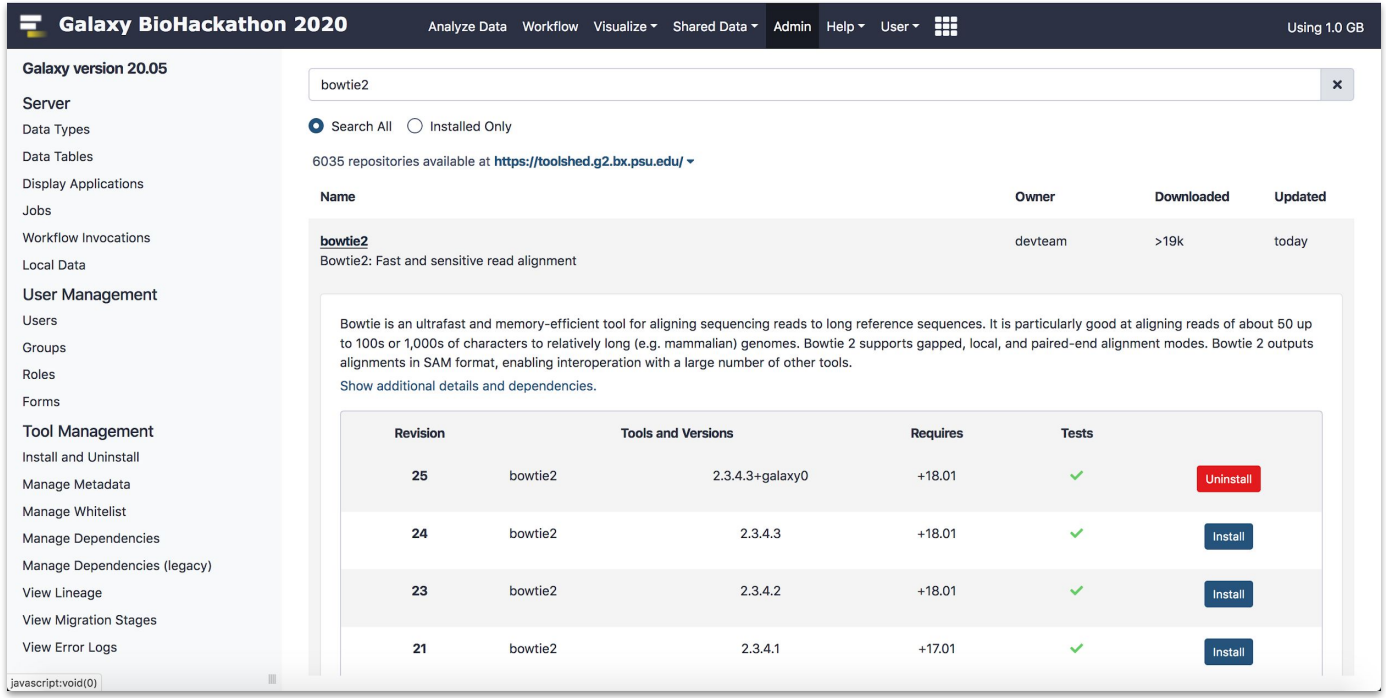
Galaxy PROJECT

elixir ITALY



Galaxy Workflow Editor

Graphical user interface to easily add, connect and configure tools for composing workflows.



Galaxy ToolShed

Serves as an "app store" to all Galaxies worldwide.

It is a **free service** Galaxy developers to share tools.

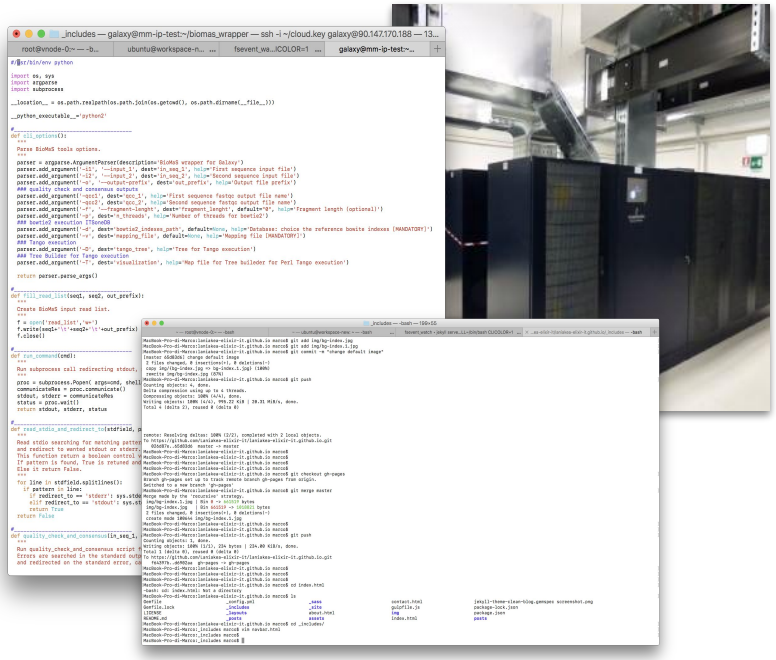
Galaxy Administrator can install tools on their instances.

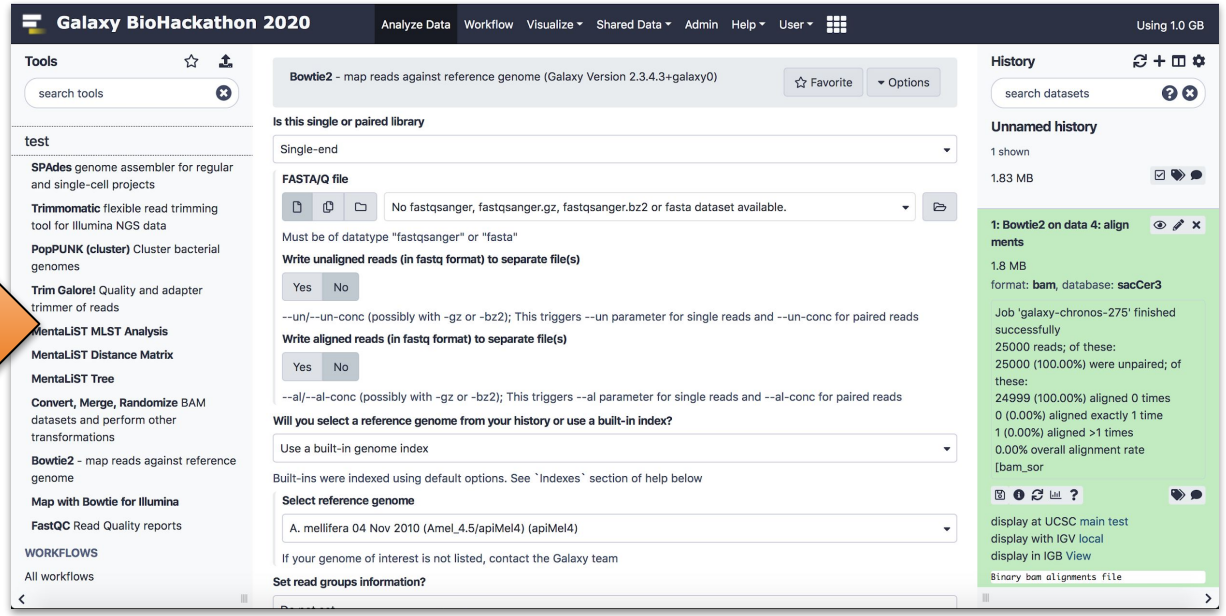
All Galaxy users can access to the tools available on a server.

Galaxy PROJECT

elixir ITALY

Allowing the community to move from command line tools to web user interfaces.





Allowing multiple users to exploit homogenous software environments, enhancing reproducibility.

Laniakea

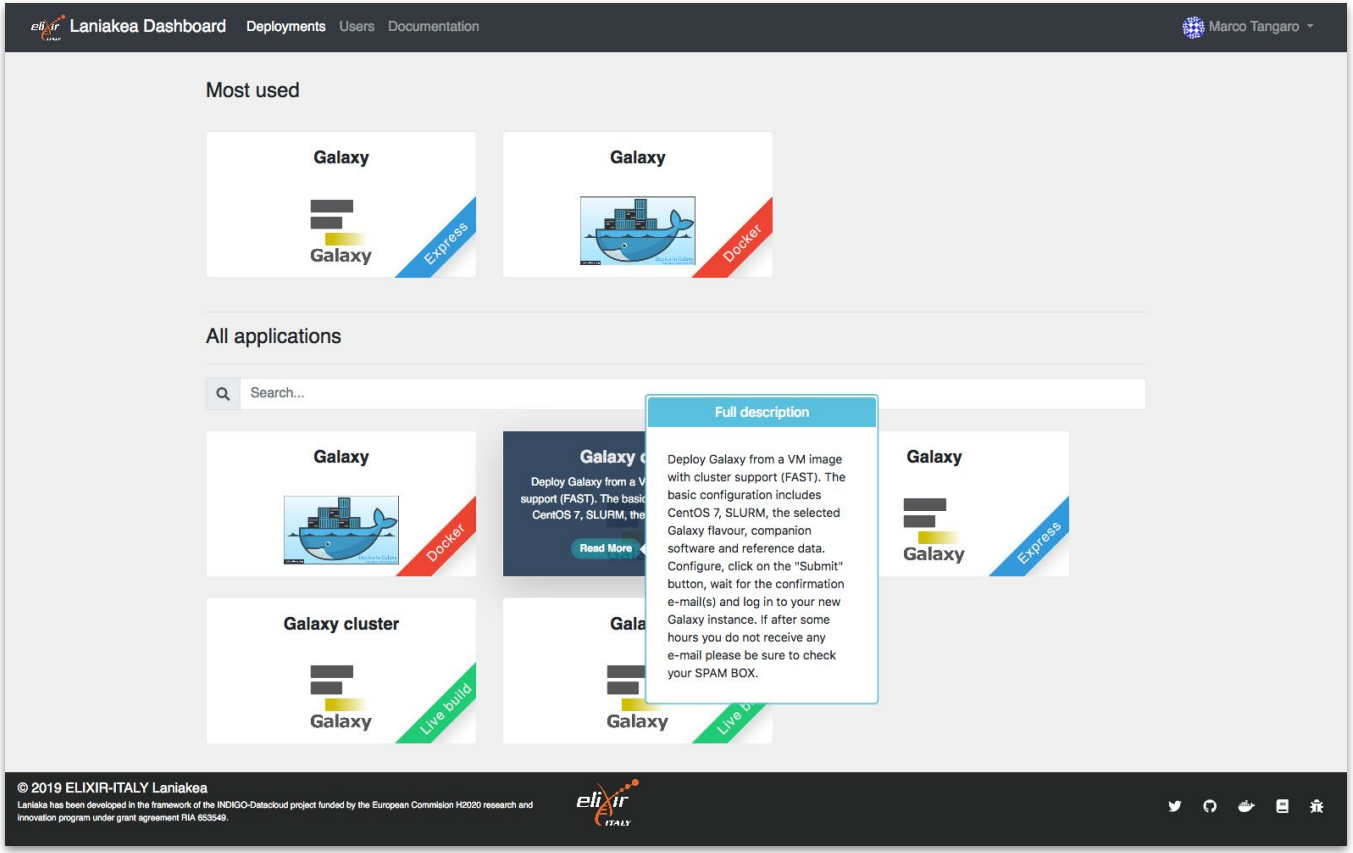
ELIXIR-Italy partners are actively involved in the service development and/or also contribute with cloud resources.

A Laniakea service is in production for ELIXIR-ITALY partner but also for ELIXIR and external users.

The ELIXIR-ITALY **Laniakea@ReCaS** Call offers access to Cloud resources to be used for the deployment of on-demand Galaxy instances.

https://laniakea-elixir-it.github.io/laniakea_at_recas

Laniakea

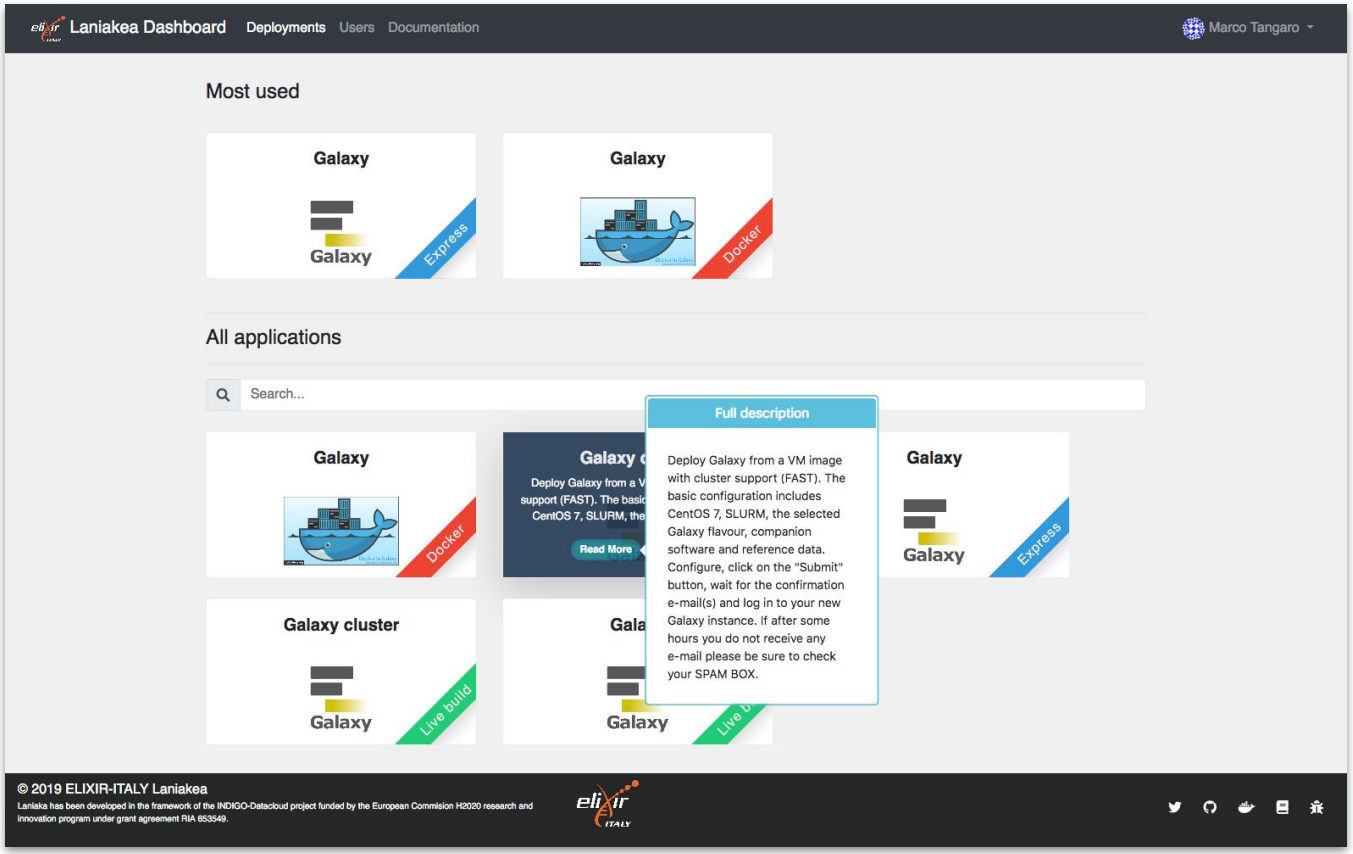


The Laniakea Dashboard home page.

Each tile provides a quick explanation of the application and links to the configuration and launch section.

Soon more applications available: Jupyter, RStudio, …

Laniakea



Different deployment strategies:

Live Build: build Galaxy from scratch -> always up-to-date (deployment time depending by the tools number).

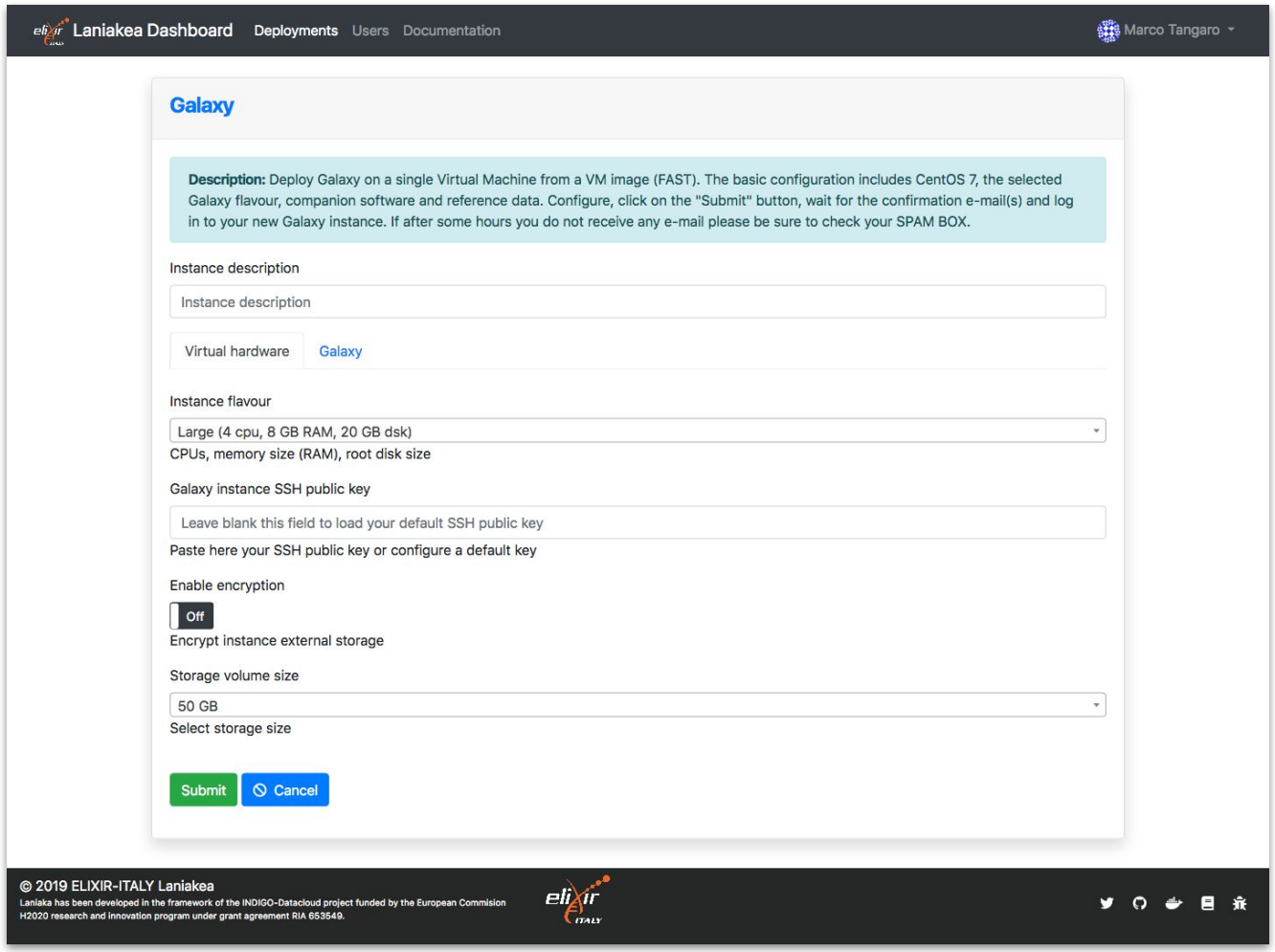
Express: pre-built Galaxy images -> fast deployment, but tools not always at the last available version.

Docker: fast deployment of new flavours.

Laniakea

The web front-end provides different tabs to configure your Galaxy.

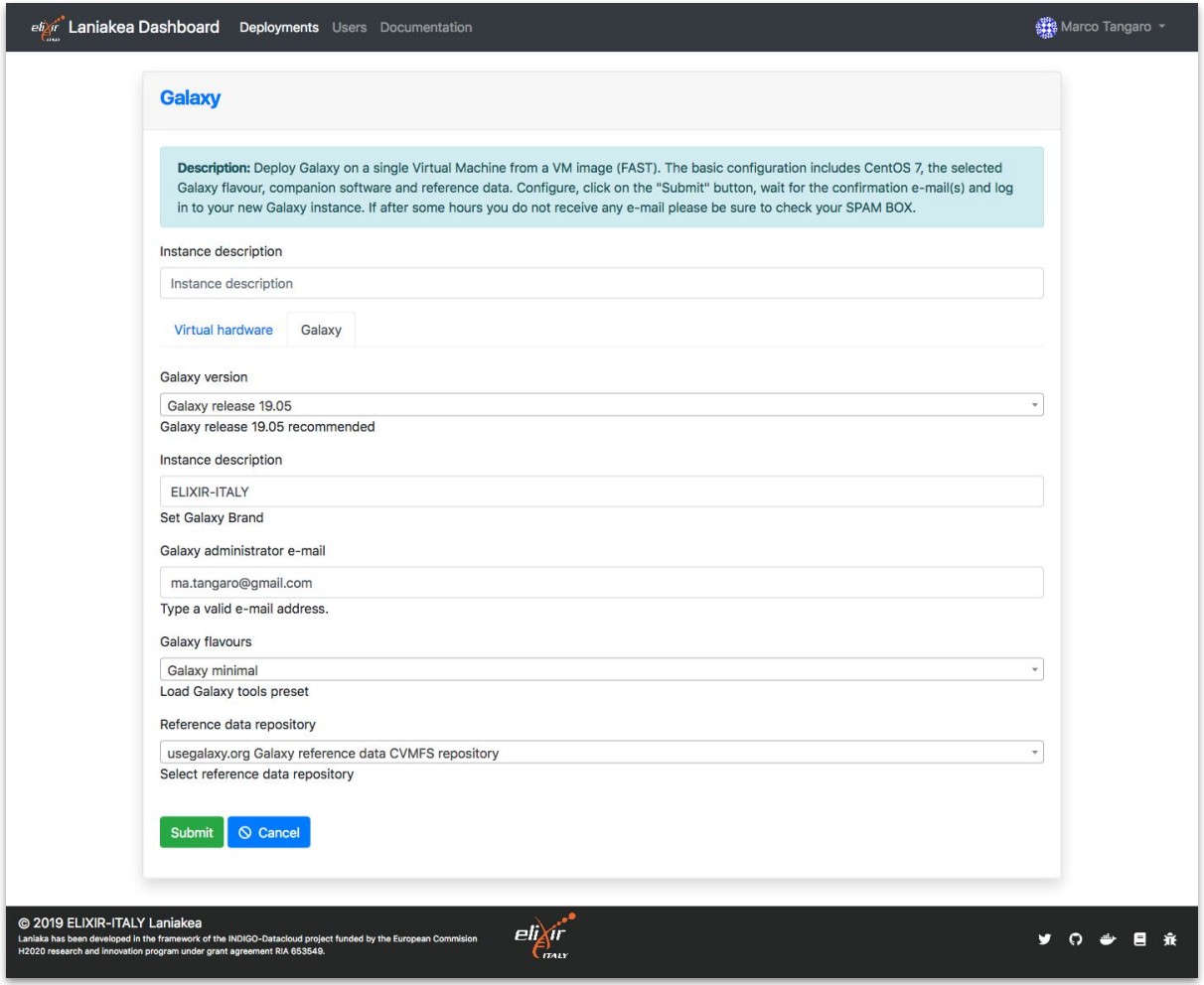
Virtual hardware: CPU, RAM and Storage



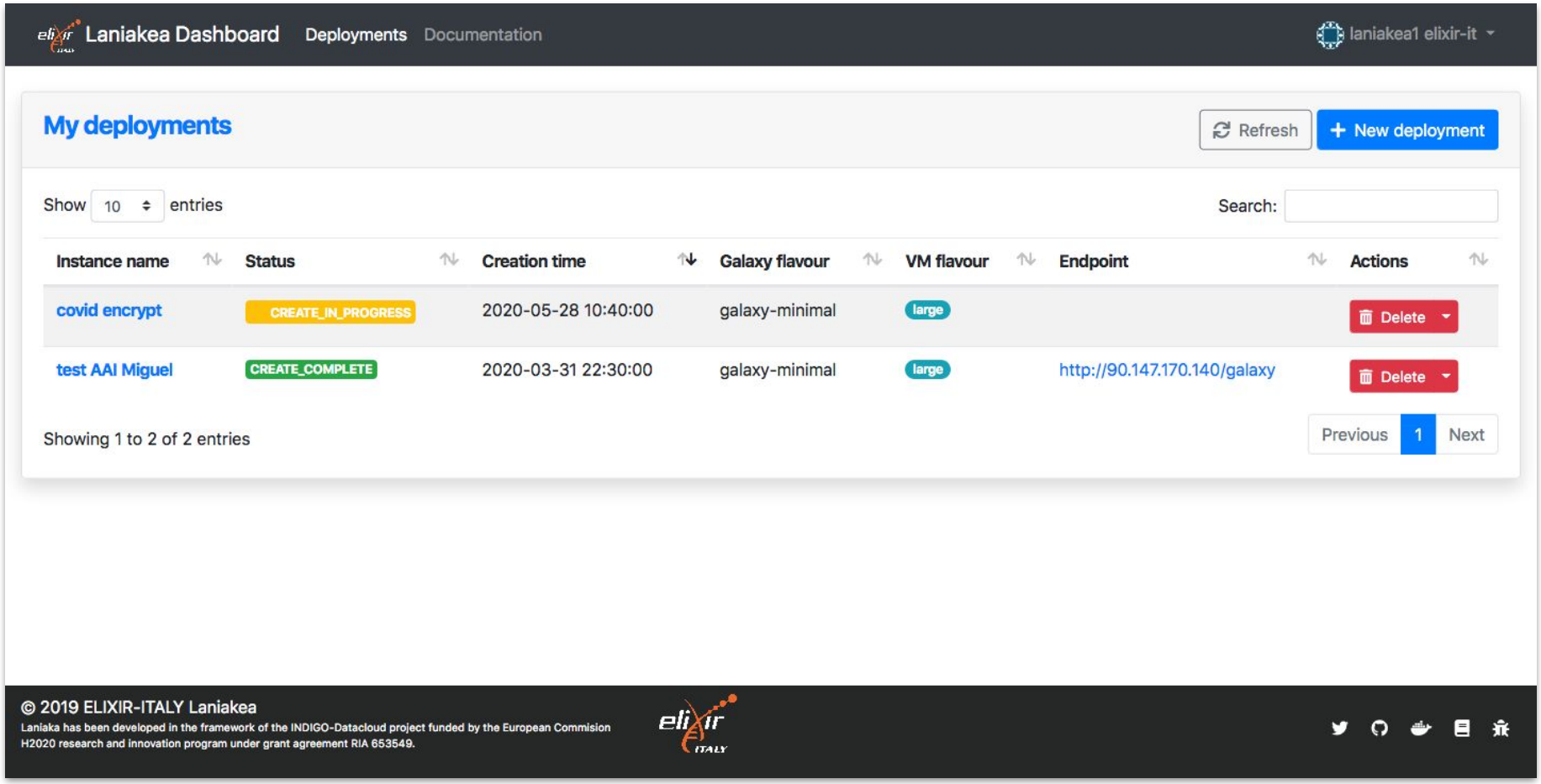
Laniakea

The web front-end provides different tabs to configure your Galaxy.

Galaxy software: version, credentials, flavor and reference data.



Laniakea



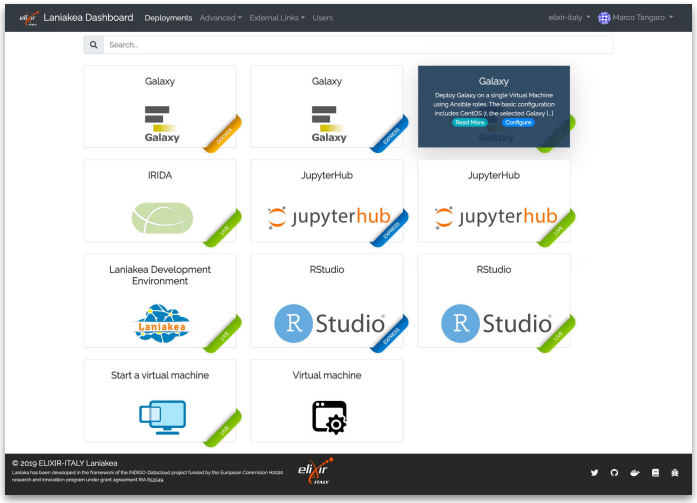
Galaxy instances management view

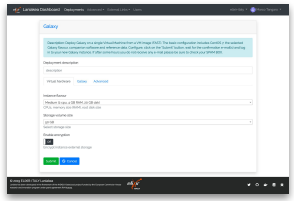
Laniakea main features

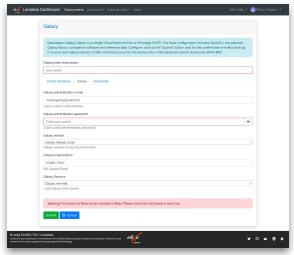
Dashboard - By hiding the technical complexity behind a user-friendly web front-end, Laniakea allows its users to configure and deploy "on-demand" Galaxy instances with a handful of clicks.

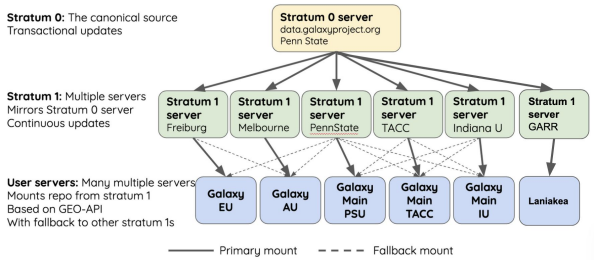
No need for the end user to know the underlying infrastructure.

No need for maintenance of the hardware and software infrastructure.









Shared reference data - 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 .

Galaxy with cluster - allowing to instantiate Galaxy with dedicated Resource Manager, allowing to customize the number of the virtual nodes to be created and their configuration in terms of number CPU and RAM.

Laniakea main features

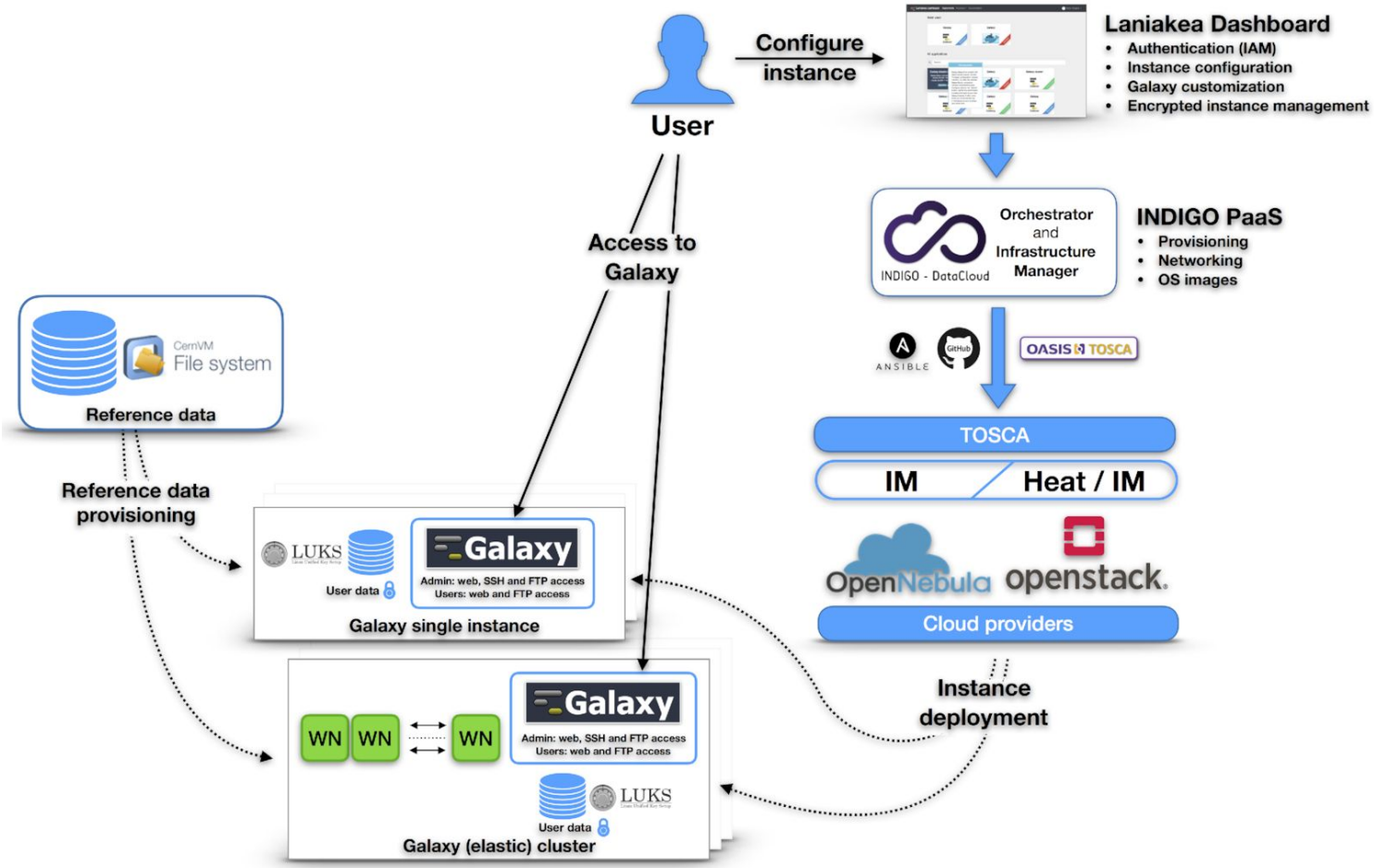
Galaxy flavors - Deploy Galaxy with sets of tested, validated and pre installed tools, named Galaxy flavors.

Current available tools presets: Galaxy Minimal, Galaxy CoVaCS, Galaxy GDC Somatic Variant, RNA Workbench, Galaxy Epigen, Covid-19.

More Applications - No more limited to Galaxy. Jupyter Notebooks, RStudio and IRIDA available.

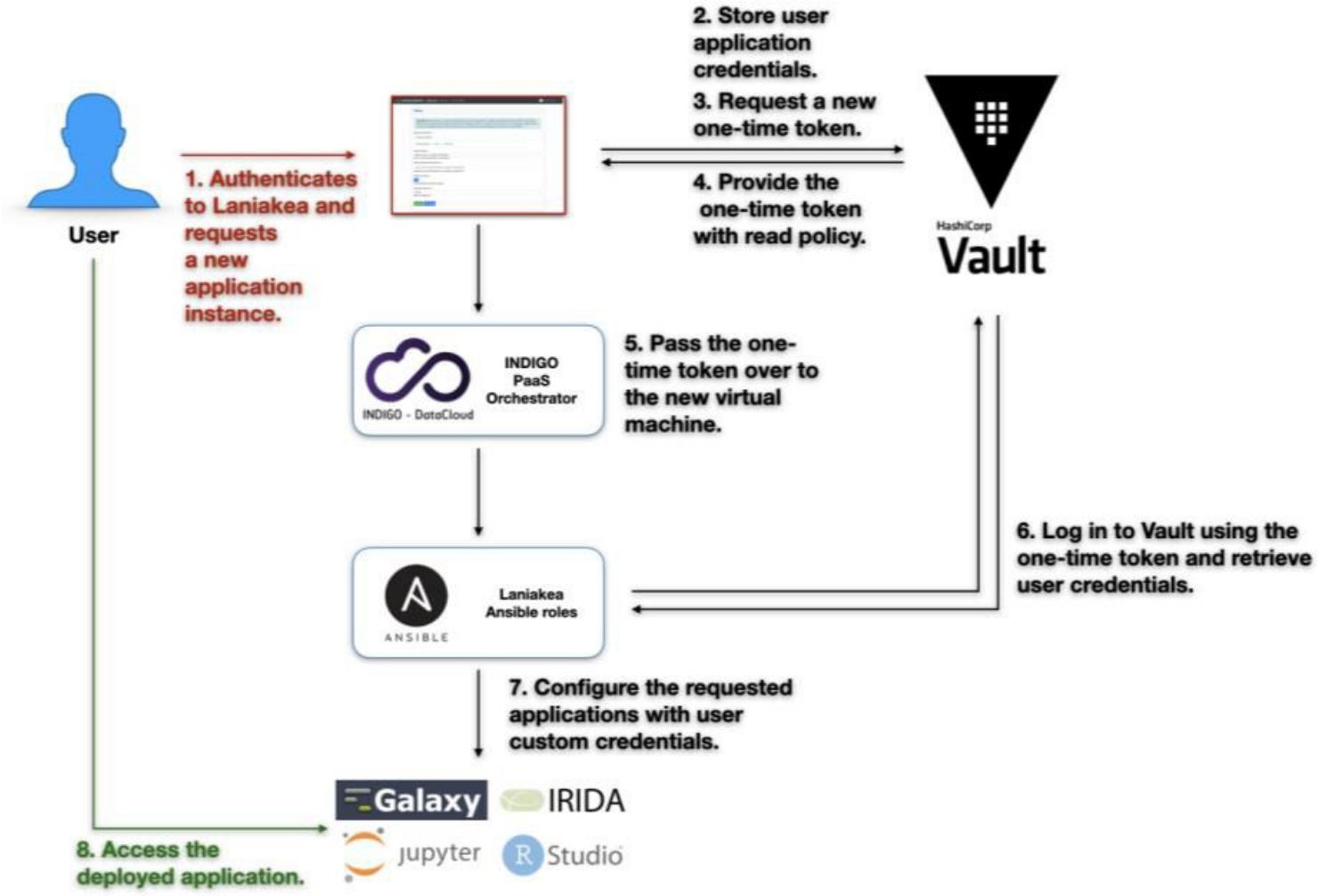
Environment with NextFlow, CWLtool and other development tools available.

Laniakea architecture



- **Dashboard** - User friendly access to configuration and and launch of a Galaxy instance
- **INDIGO-IAM** - Authentication and Authorization system
- **INDIGO-PaaS** - PaaS layer for Galaxy deployment
- Cloud Provider - ReCaS Bari
- **Persistent storage** with/without encryption
- **Hashicorp Vault** - secrets management
- **Reference data** availability with CERN-VM FS

Laniakea


User
1. Authenticates to Laniakea and requests a new application instance.
2. Store user application credentials.
3. Request a new one-time token.
4. Provide the one-time token with read policy.
HashiCorp
Vault
INDIGO PaaS Orchestrator
INDIGO - DataCloud
5. Pass the one-time token over to the new virtual machine.
Laniakea Ansible roles
ANSIBLE
6. Log in to Vault using the one-time token and retrieve user credentials.
7. Configure the requested applications with user custom credentials.
Galaxy
IRIDA
jupyter
Studio
8. Access the deployed application.

Laniakea@ReCaS

Currently, some important Italian Institutions are using Laniakea for their daily work:

- Istituto Ortopedico Rizzoli (2 internal Galaxy servers).
- Istituto Zooprofilattico Sperimentale della Puglia e della Basilicata (2 internal Galaxy servers and 1 IRIDA instance).
- Ospedale Pediatrico Giannina Gaslini (public server).
- University of Milan (public Galaxy server and tools development).
- IBIOM-CNR (public Galaxy server and tools development).
- University of Turin (training)

... and counting.