On-demand Galaxy-based secure environments with Laniakea

Donvito Giacinto (INFN)
Marco Antonio Tangaro (IBIOM-CNR)
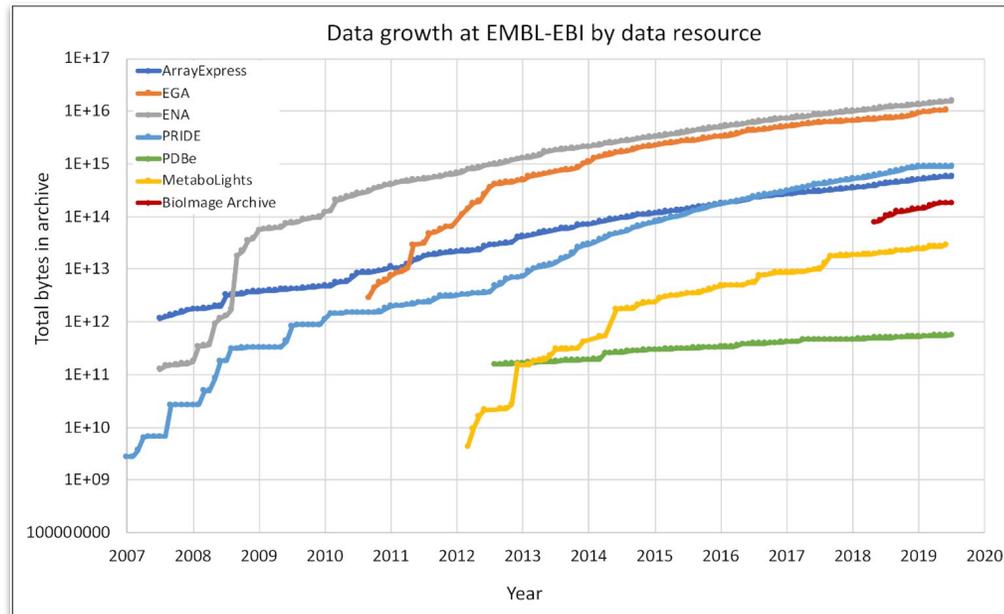
Datathon on Sensitive Data
30th May - 1st June, 2023

Outline

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

Data

Data growth at EMBL-EBI by data resource

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

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

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 genetic data as "Special Categories of Personal data".

Sensitive genetic data processing for scientific research purposes is possible, provided this is allowed by EU or Member States law framework and appropriate safeguards measures are in place.

Motivation

DATA SOURCES

DATA STORAGE
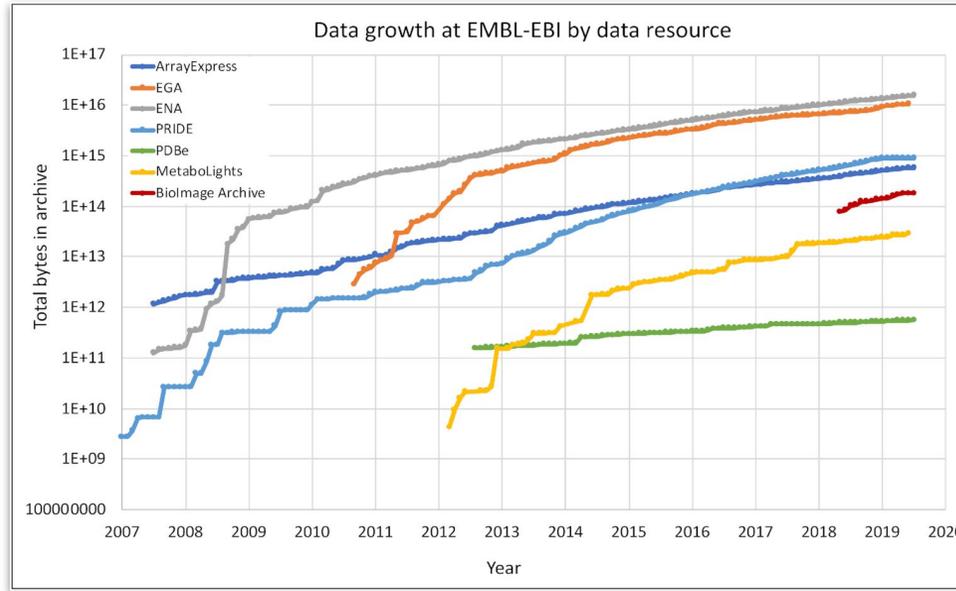
Data growth at EMBL-EBI by data resource
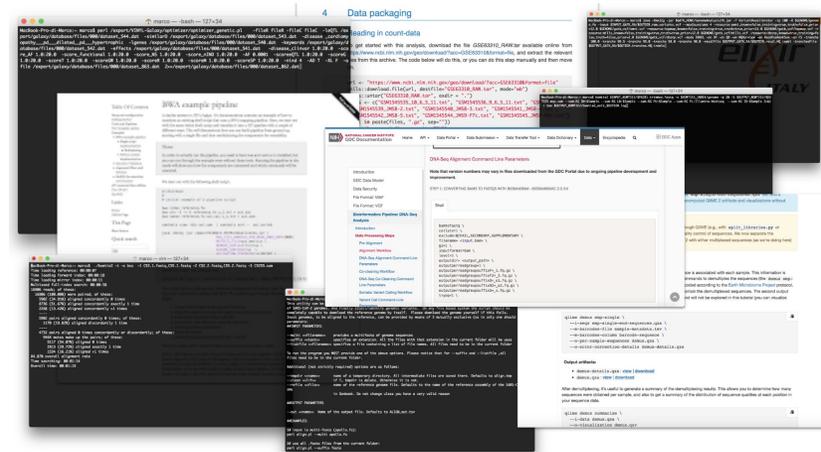
DATA ANALYSIS TOOLS

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

- 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.
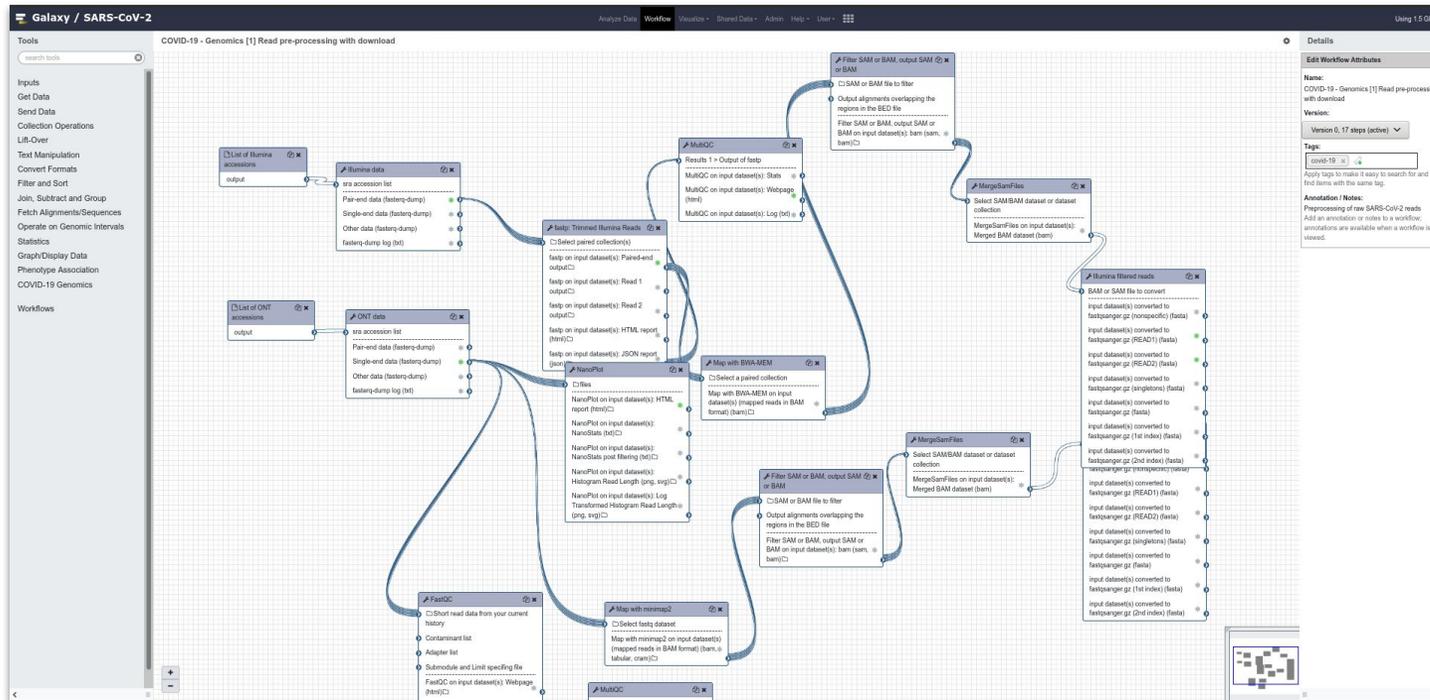
Galaxy Workflow Editor

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.

Tools dependencies automatically solved.

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

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.
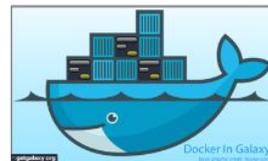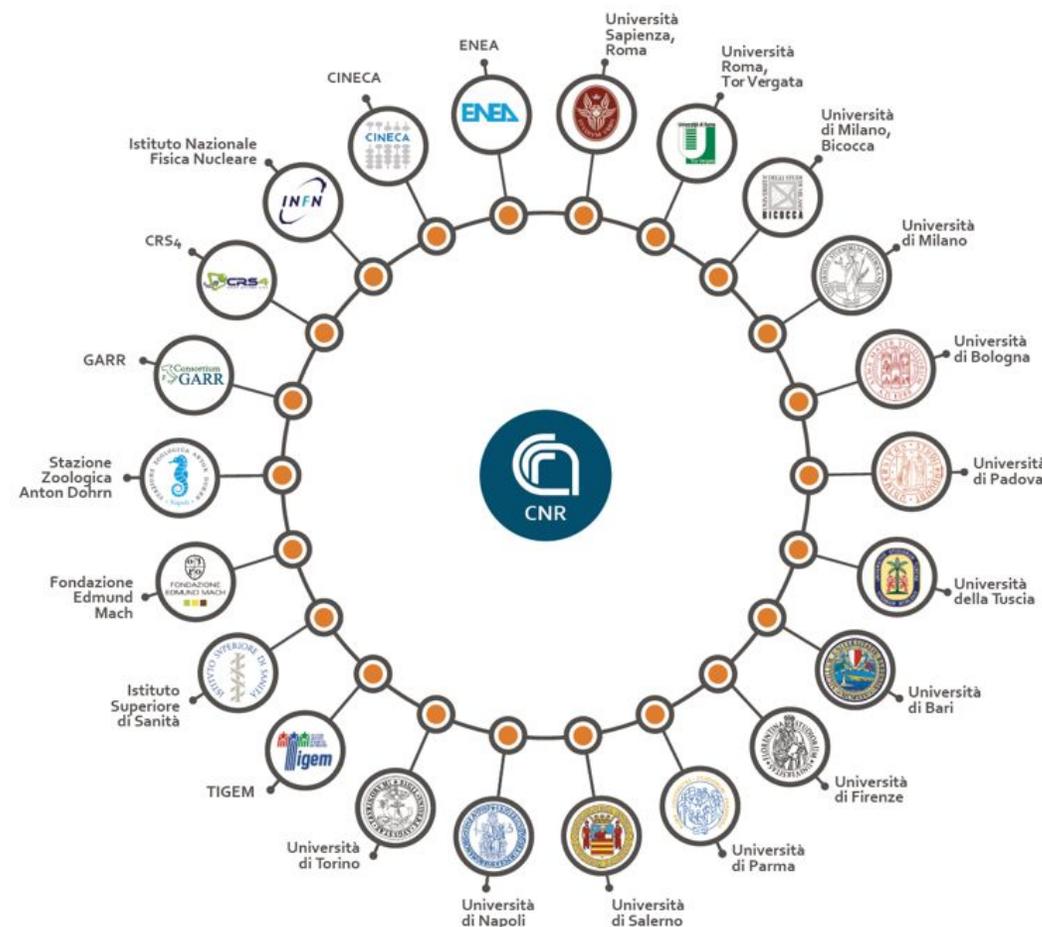
Laniakea

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 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.
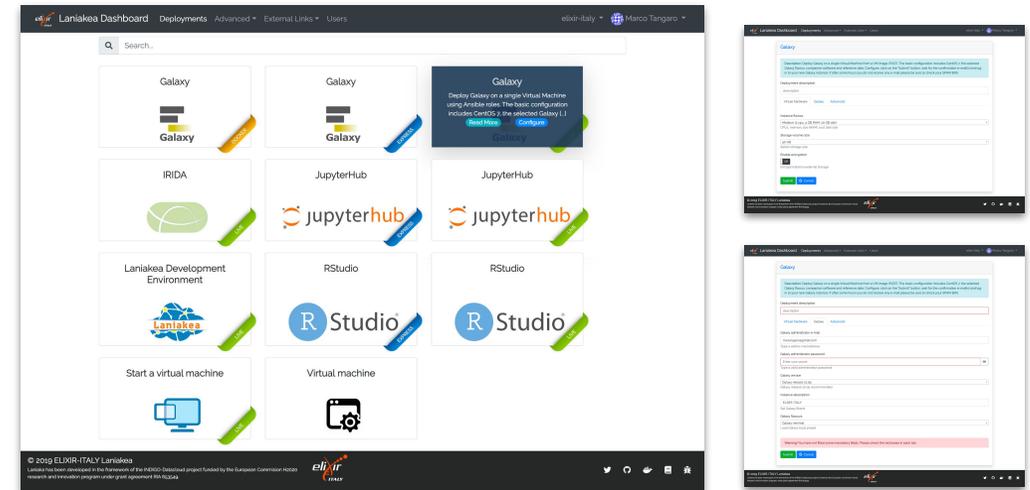
Different deployment strategies:

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 main features
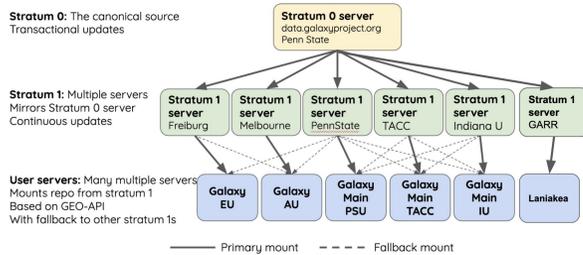
Shared reference data - Each instance comes with reference data (e.g. genomic sequences) already available for many species, shared among all the instances through the CERN-VM FileSystem .

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.
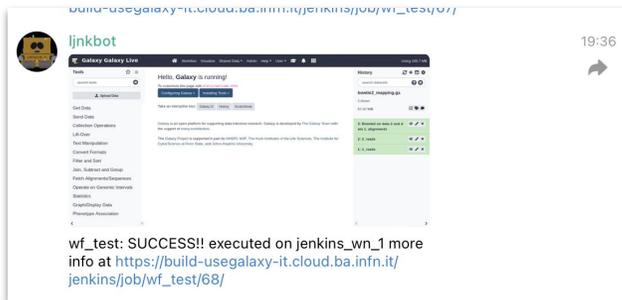
Continuous Integration - Cloud images creation and service testing are automated using Jenkins and Github.
Cloud images automatically updated when there are changes on github
Services are tested regularly every week.

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.

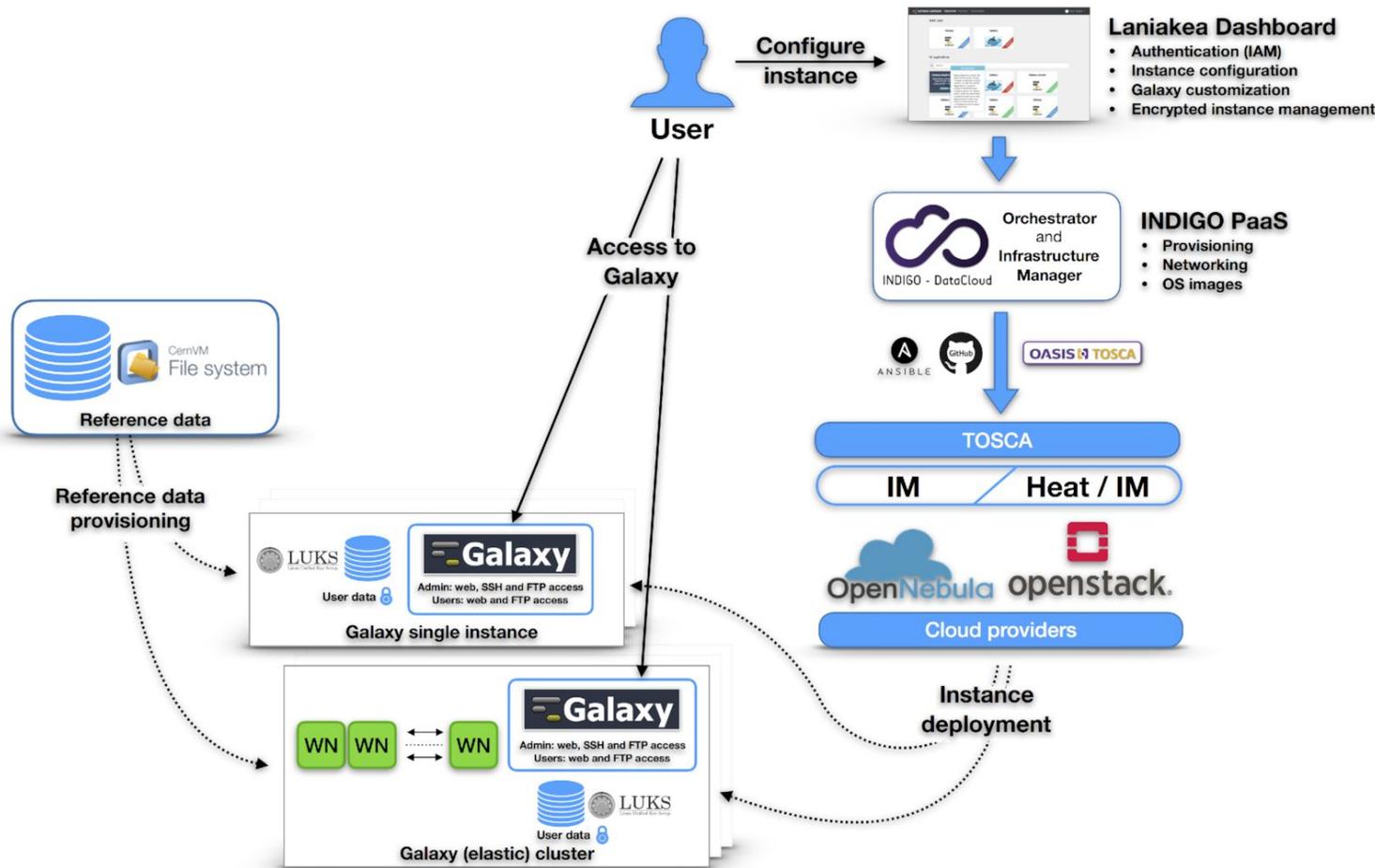Environment with NextFlow, CWLtool and other development tools available.

Secured environment features

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

OPENVPN

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

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 architecture
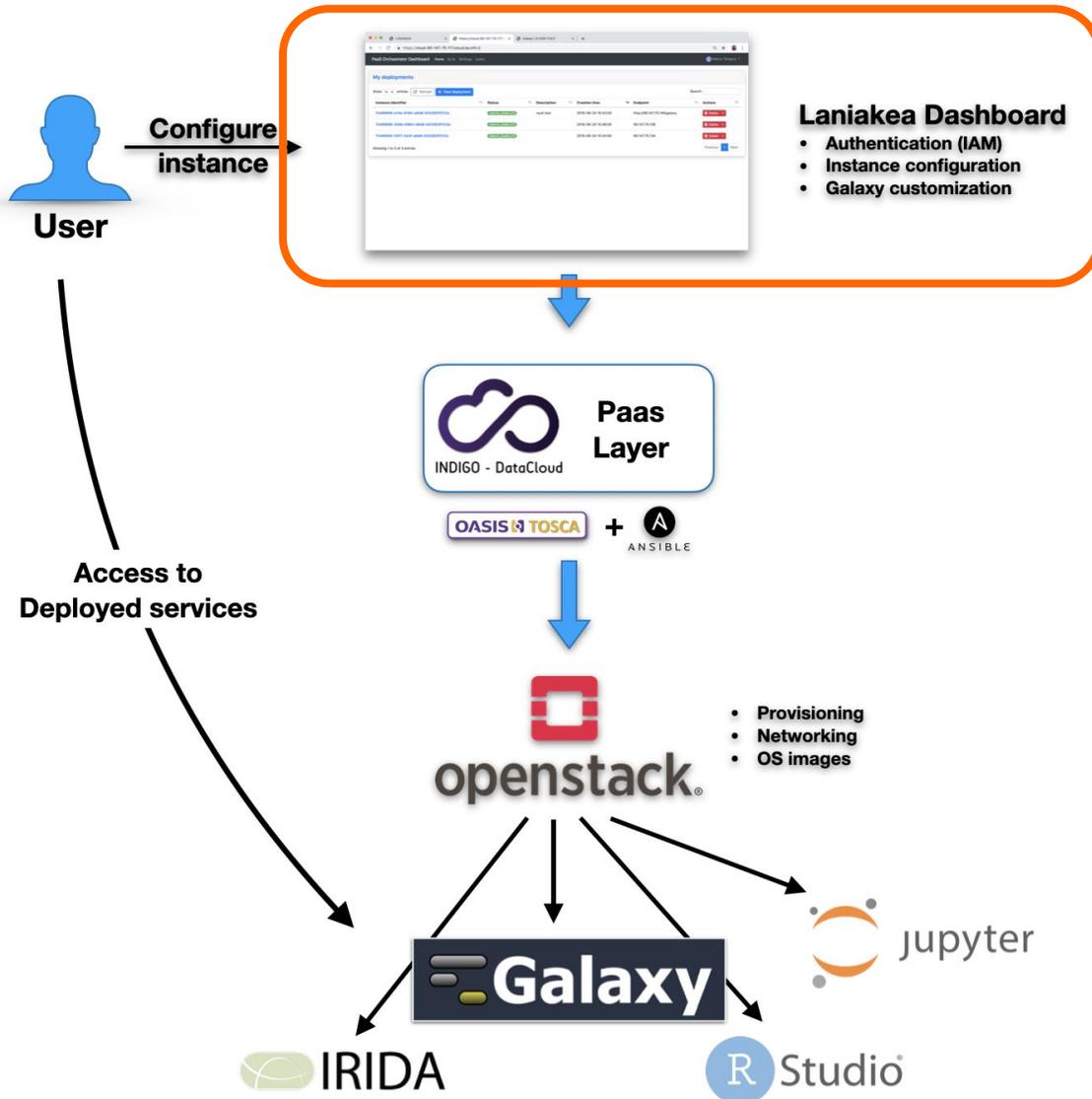
Laniakea Dashboard
- **Authentication (IAM)**
- **Instance configuration**
- **Galaxy customization**

User

Configure instance

Access to Deployed services

Paas Layer INDIGO - DataCloud

OASIS TOSCA + ANSIBLE

openstack
- **Provisioning**
- **Networking**
- **OS images**

Galaxy

Jupyter

IRIDA

R Studio

- **Dashboard -** User friendly access to configure and launch a Galaxy instance

- **INDIGO PaaS -** Galaxy automatic deployment

- **Cloud Providers -** ReCaS-Bari

18

Laniakea architecture

- **Dashboard -** User friendly access to configure and launch a Galaxy instance

- **INDIGO PaaS -** Galaxy automatic deployment

- **Cloud Providers -** ReCaS-Bari

19

Laniakea architecture

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

20
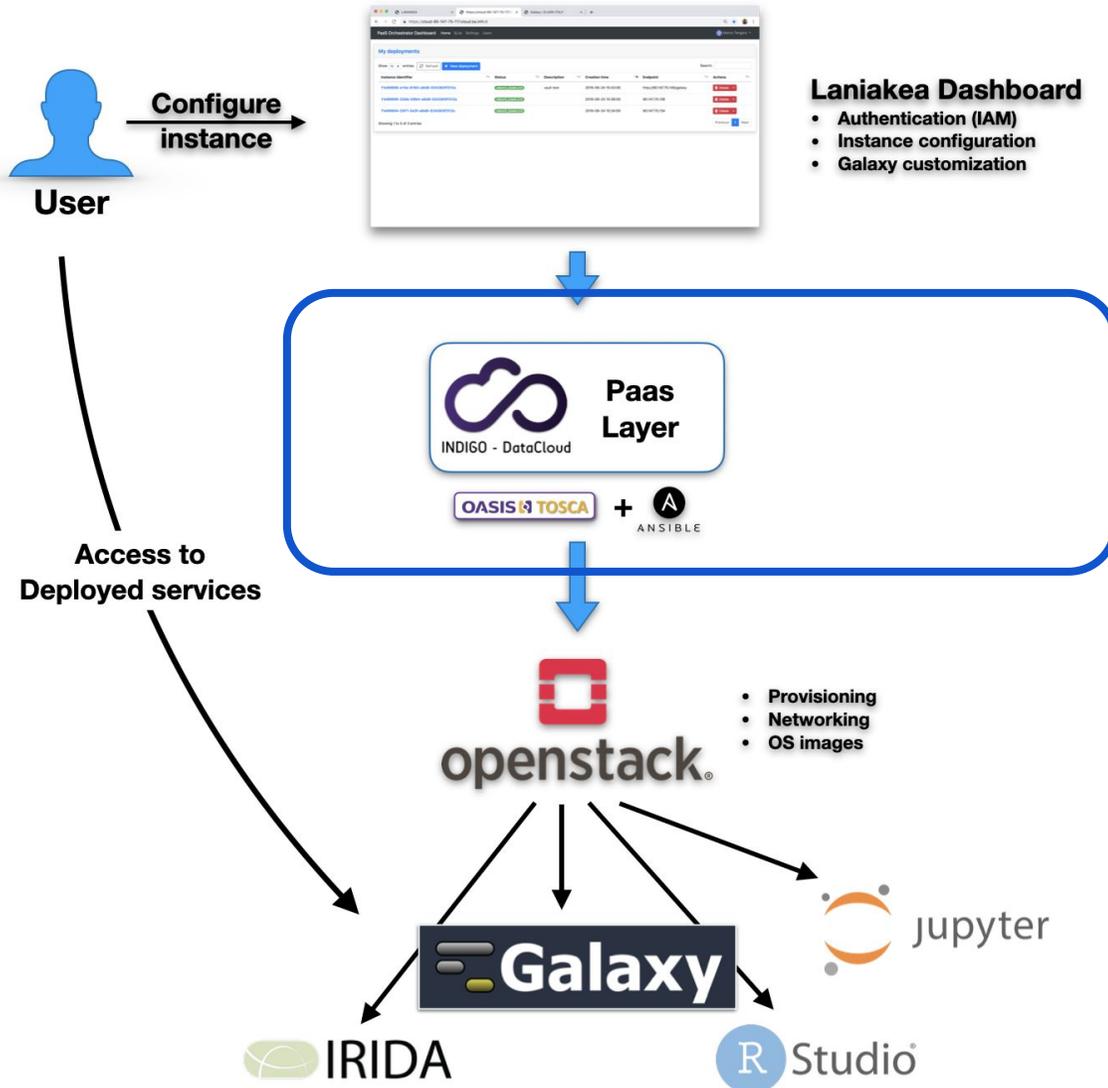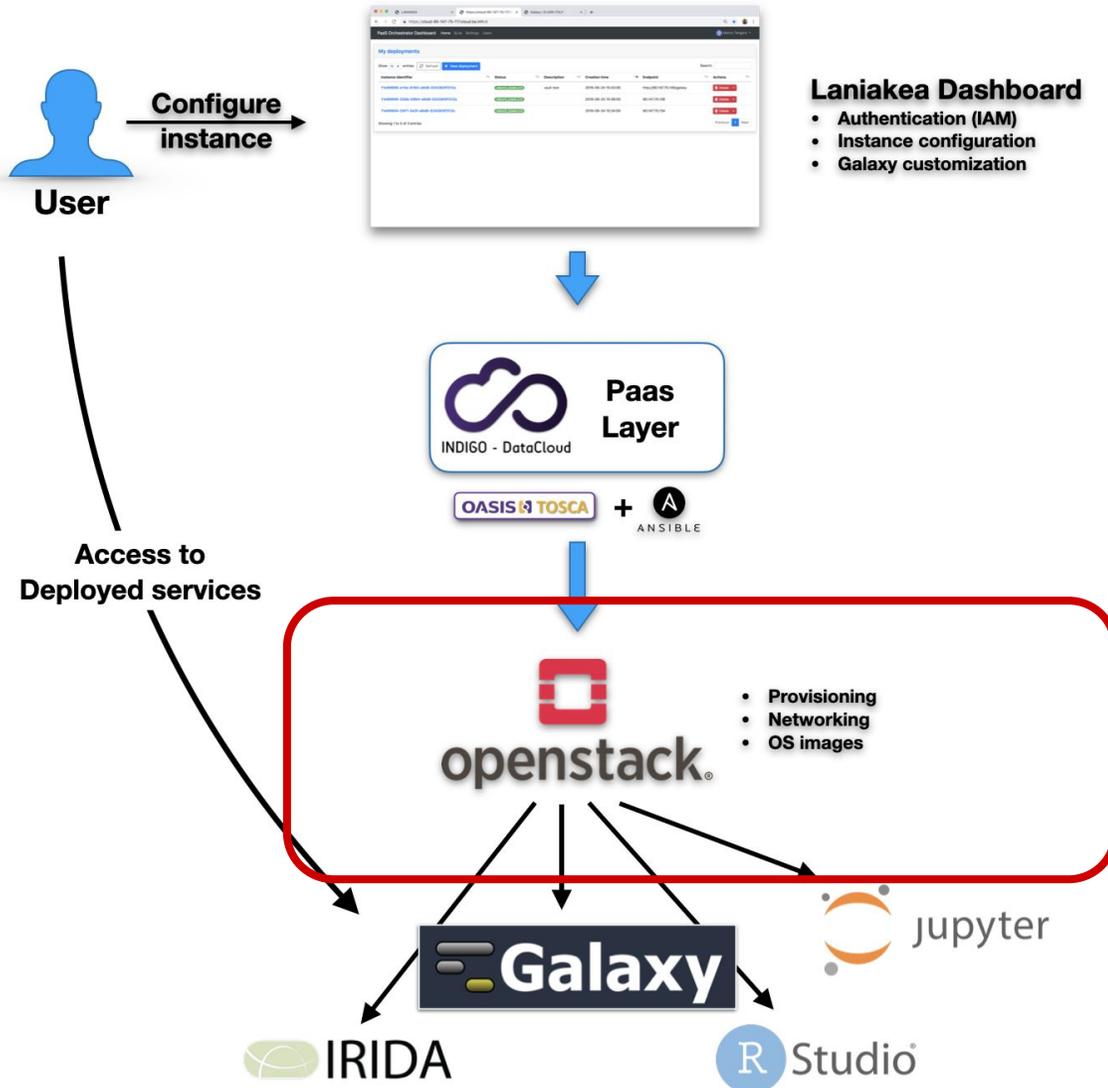
Laniakea Dashboard

Laniakea Dashboard
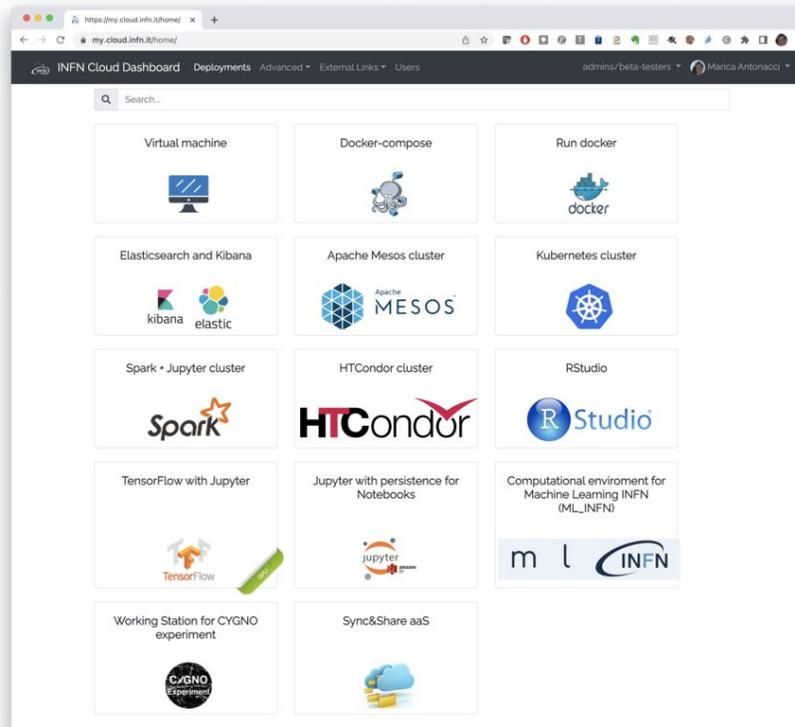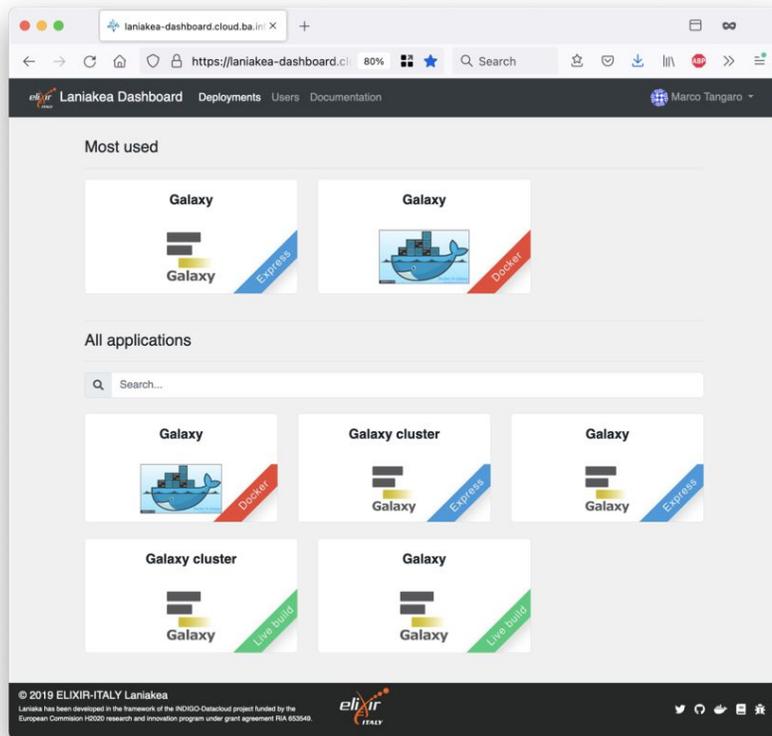
Flask web micro-framework (flask.pocoo.org/),

Jinja2 template engine (jinja.pocoo.org/)

Bootstrap 4 toolkit (getbootstrap.com/).

Integrated with Hashicorp Vault for user secrets management.

Laniakea Dashboard

The PaaS Layer accepts deployment requests in the form of TOSCA Templates: a document (YAML syntax) describing the infrastructure to deploy, e.g. the virtual hardware and the software to be installed and configured. The Dashboard parses the TOSCA document automatically and renders the user interface with user friendly forms. This allows to extend Laniakea functionalities just adding new templates without any code modification.

We extended the TOSCA templates inputs to create configurable forms. This creates a flexible web interface, allowing straightforward customisation of the user experience through human readable YAML configuration files. The dashboard template is modular and can be easily adapted adding new functionalities to the user interface (e.g. adding a dropdown menu, text fields, toggles…) based on the Laniakea administrator requirements.

Laniakea Dashboard

```
instance_flavor:                                    TOSCA template
    type: string
    description: instance flavor (num_cpu, memory, disk)
    default: small
```

User input field in the TOSCA template to select instance flavour in terms of vCPUs, RAM and root disk storage.

```
instance_flavor:                                    Parameters file
  display_name: "Instance flavour"
  tag_type: "select"
  description: "CPUs, memory size (RAM), root disk size"
  constraints:
    - { value: "small",  label: "Small (1 cpu, 2 GB RAM, 20 GB dsk)" }
    - { value: "medium", label: "Medium (2 cpu, 4 GB RAM, 20 GB dsk)" }
    - { value: "large",  label: "Large (4 cpu, 8 GB RAM, 20 GB dsk)" }
    - { value: "xlarge", label: "xLarge (8 cpu, 16 GB RAM, 20 GB dsk)" }
  tab: "Virtual hardware"
```

Dashboard render

Virtual hardware Galaxy Advanced

Instance flavour

Small (1 cpu, 2 GB RAM, 20 GB dsk)

CPUs, memory size (RAM), root disk size

Dashboard render

Virtual hardware Galaxy Advanced

Instance flavour

Small (1 cpu, 2 GB RAM, 20 GB dsk)

Small (1 cpu, 2 GB RAM, 20 GB dsk)

Medium (2 cpu, 4 GB RAM, 20 GB dsk)

Large (4 cpu, 8 GB RAM, 20 GB dsk)

xLarge (8 cpu, 16 GB RAM, 20 GB dsk)

24

Laniakea Dashboard

Laniakea Dashboard

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.

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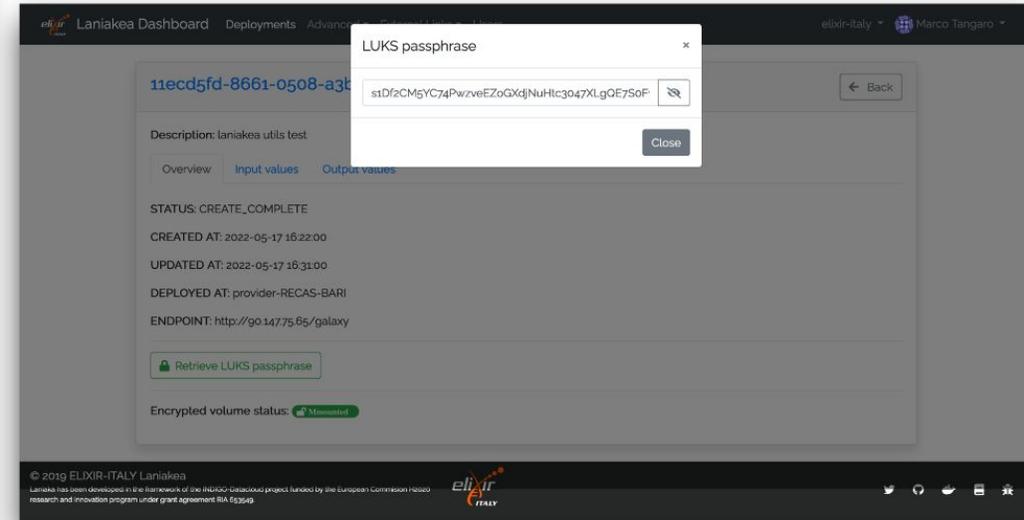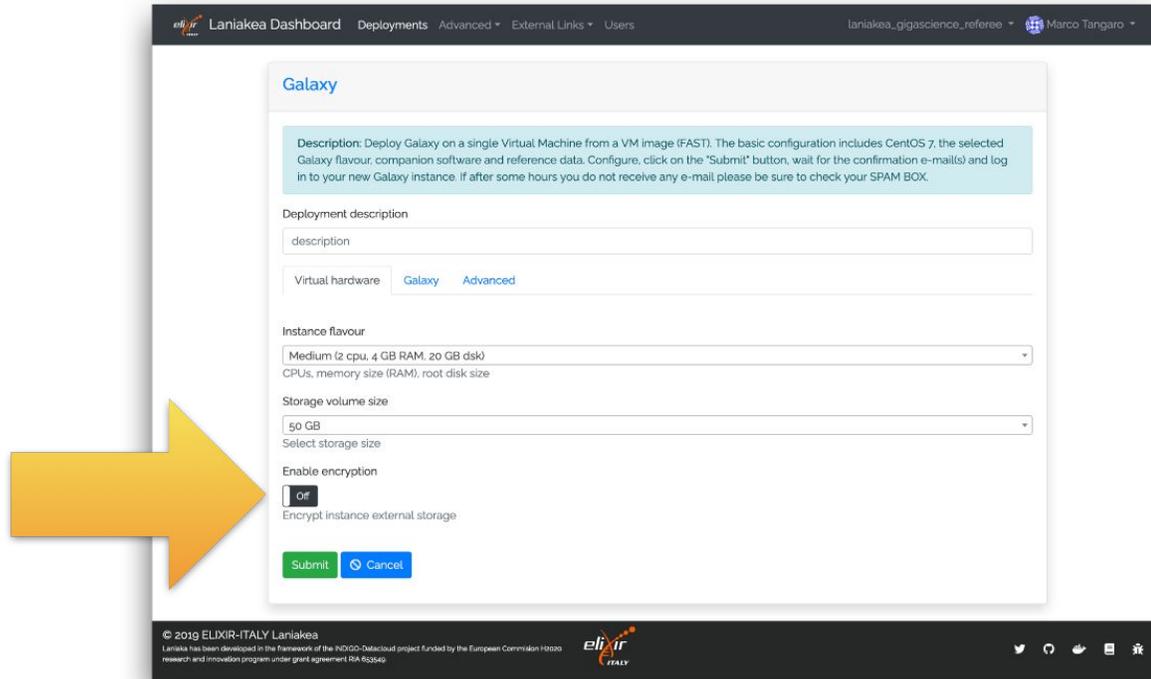
The user can enable the storage encryption using a switch toggle in the Instance "Virtual hardware" configuration tab.

The procedure is completely automated.

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

Laniakea encryption: performance

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.
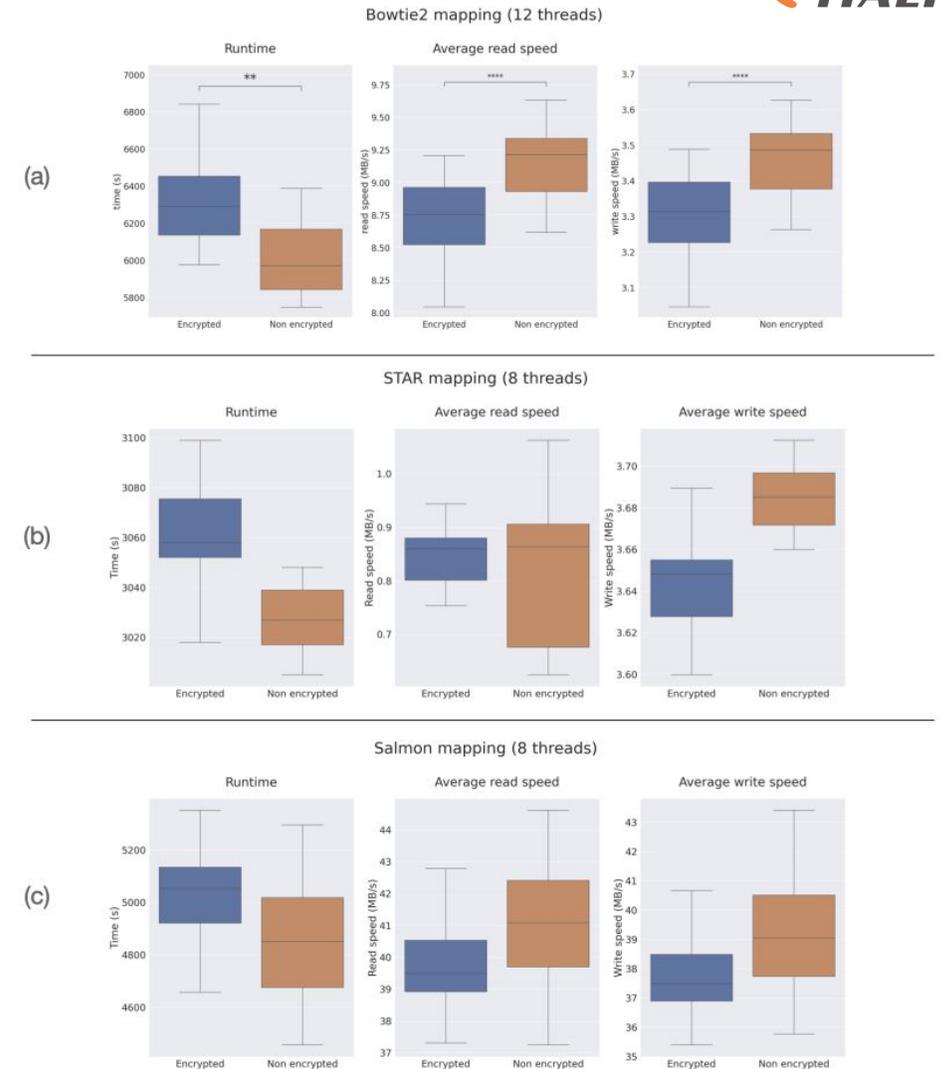
Deployments under VPN

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.
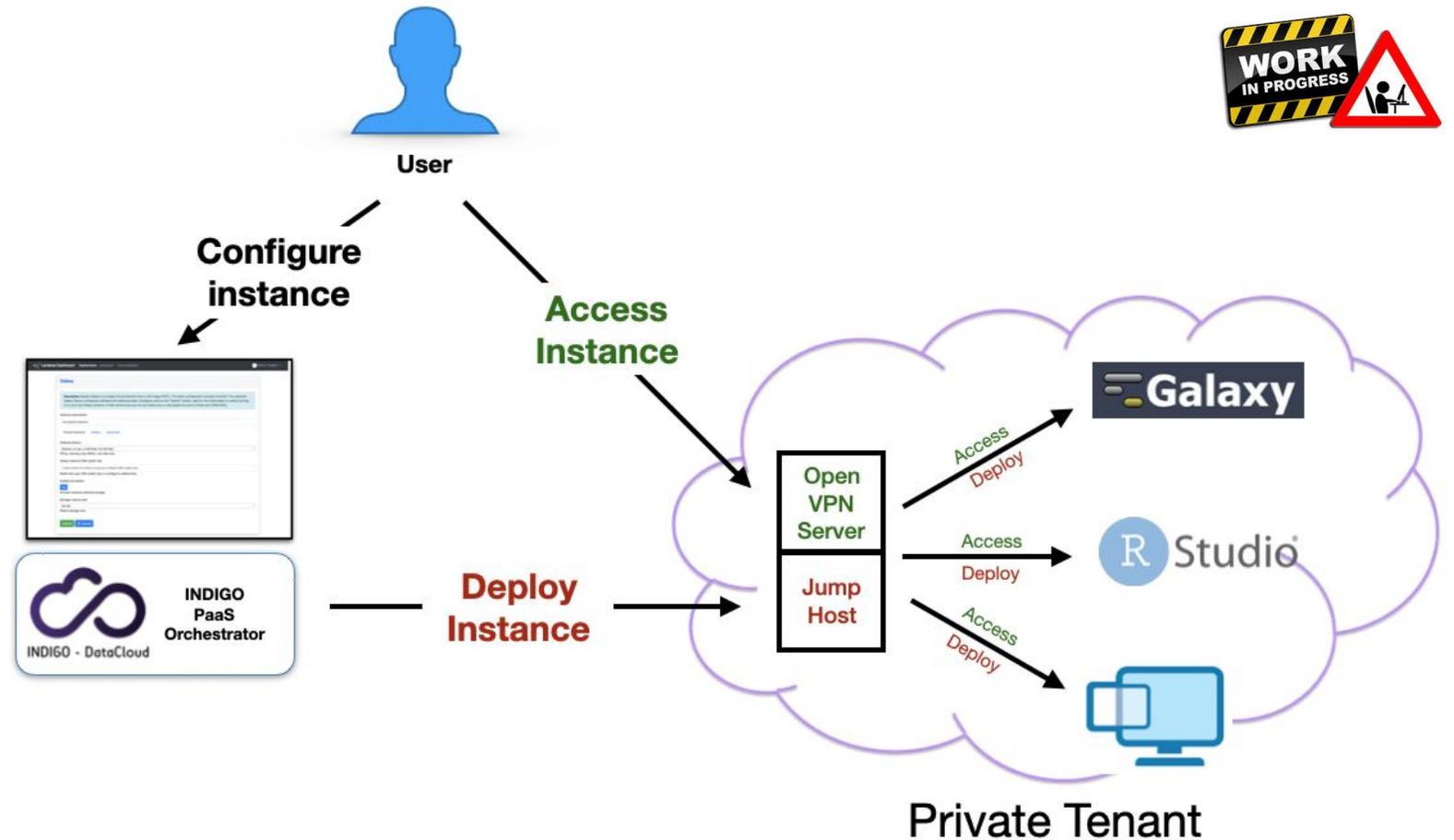
Thanks for your attention

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

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

Allowing the community to move from command line tools to web user interfaces.
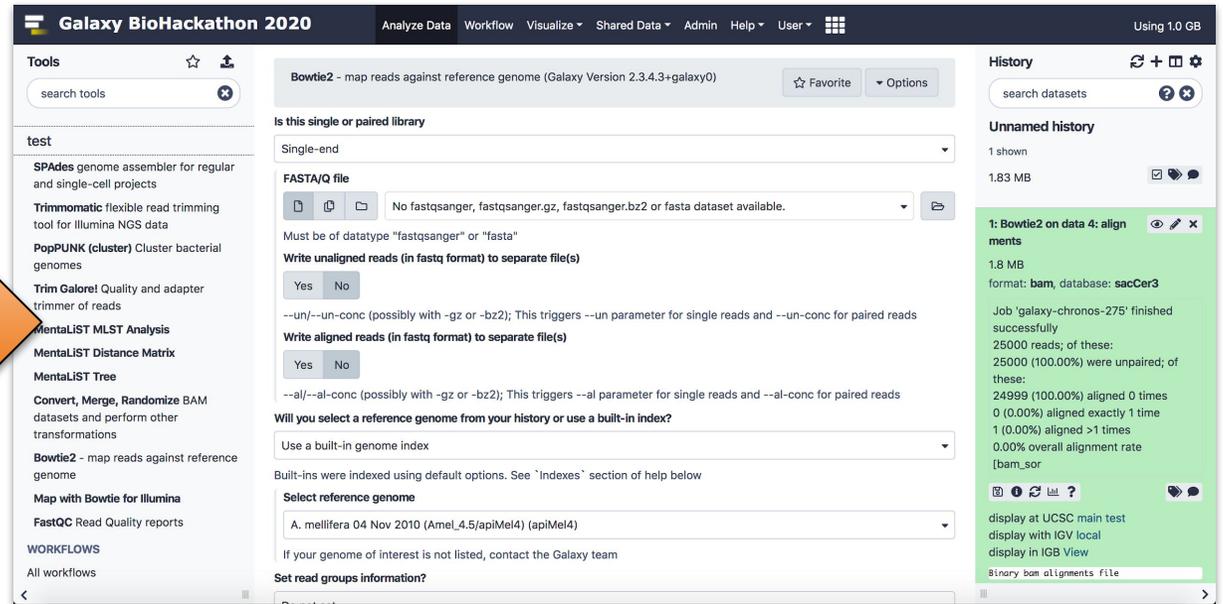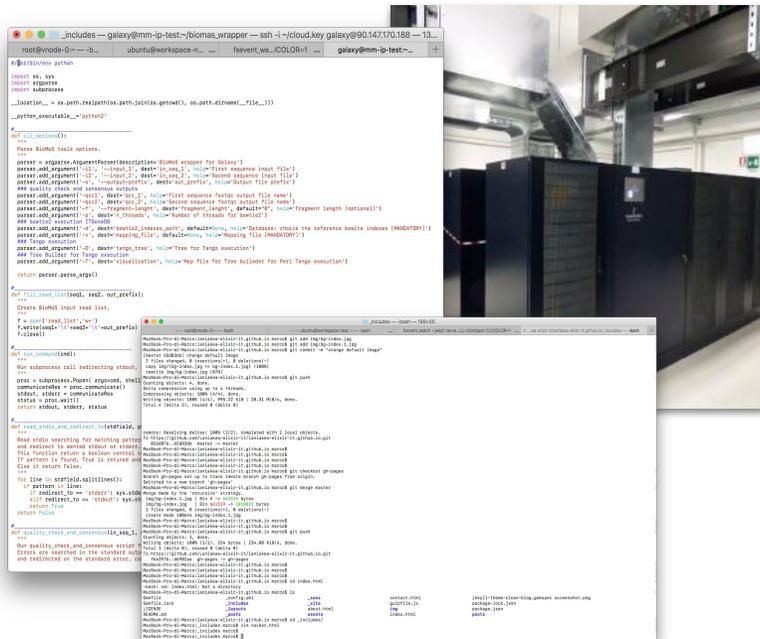
Allowing multiple users to exploit homogenous software environments, enhancing reproducibility.

Laniakea

The Laniakea Dashboard home page.

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

Also more applications available: Jupyter, RStudio, ...
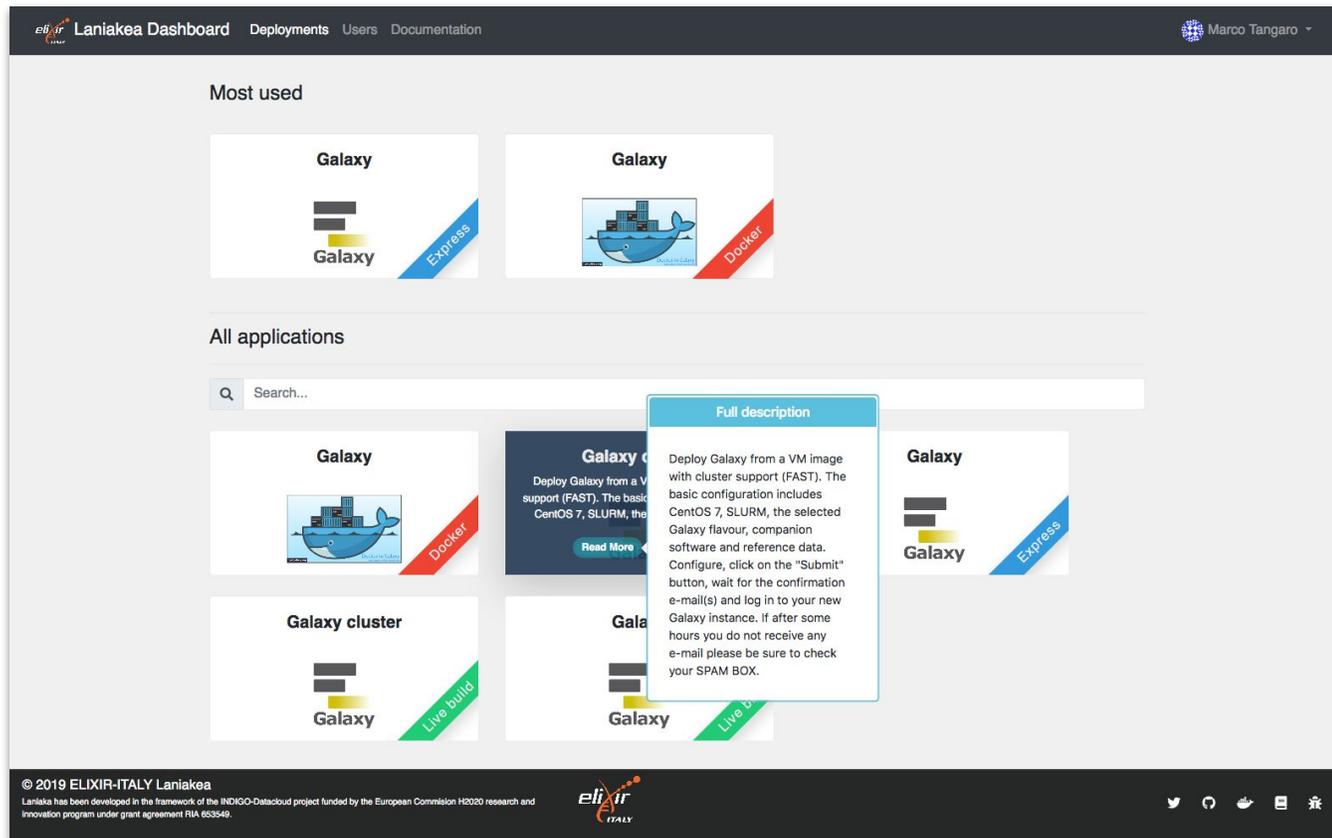
Laniakea
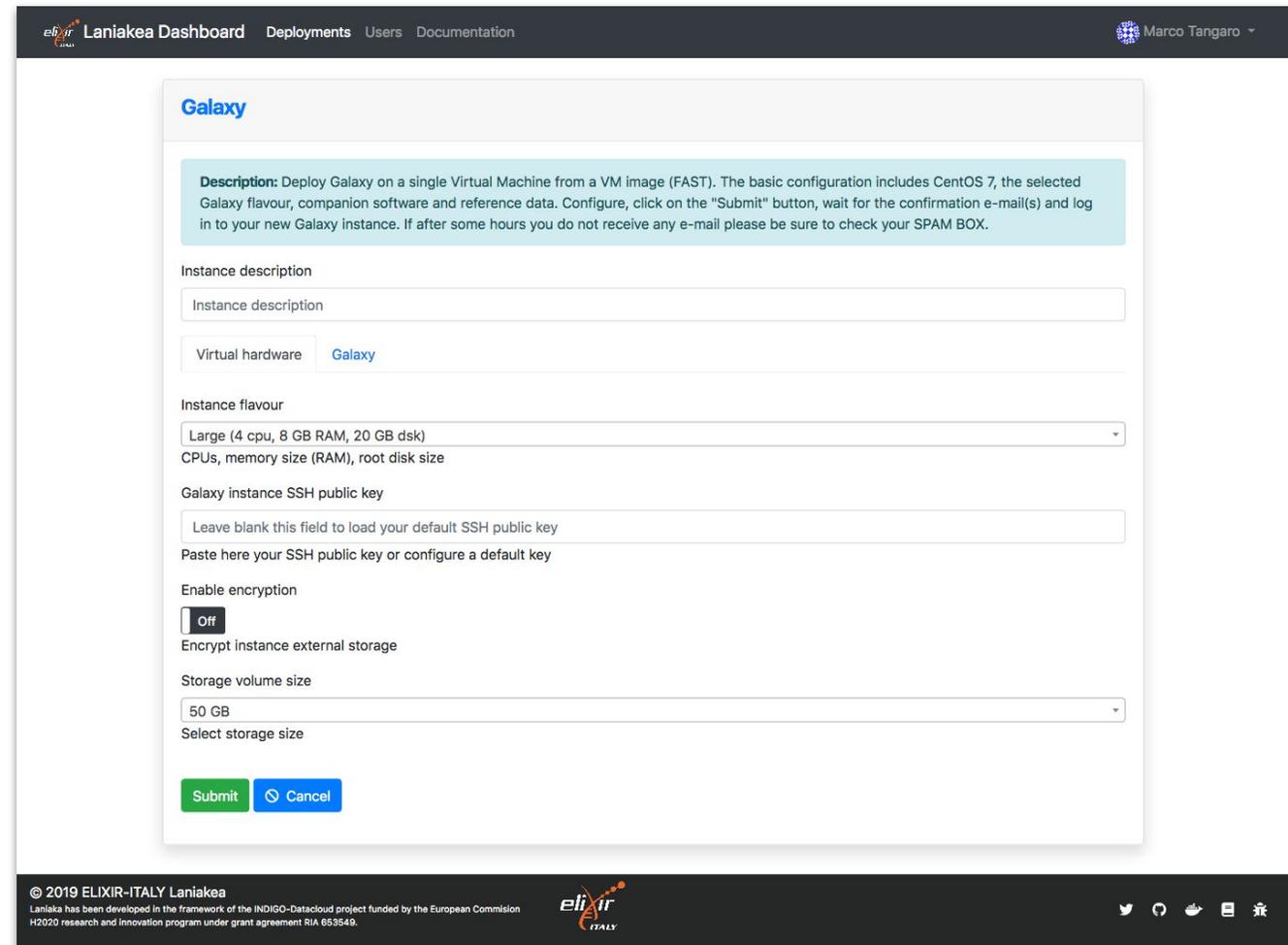

Different deployment strategies:

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.

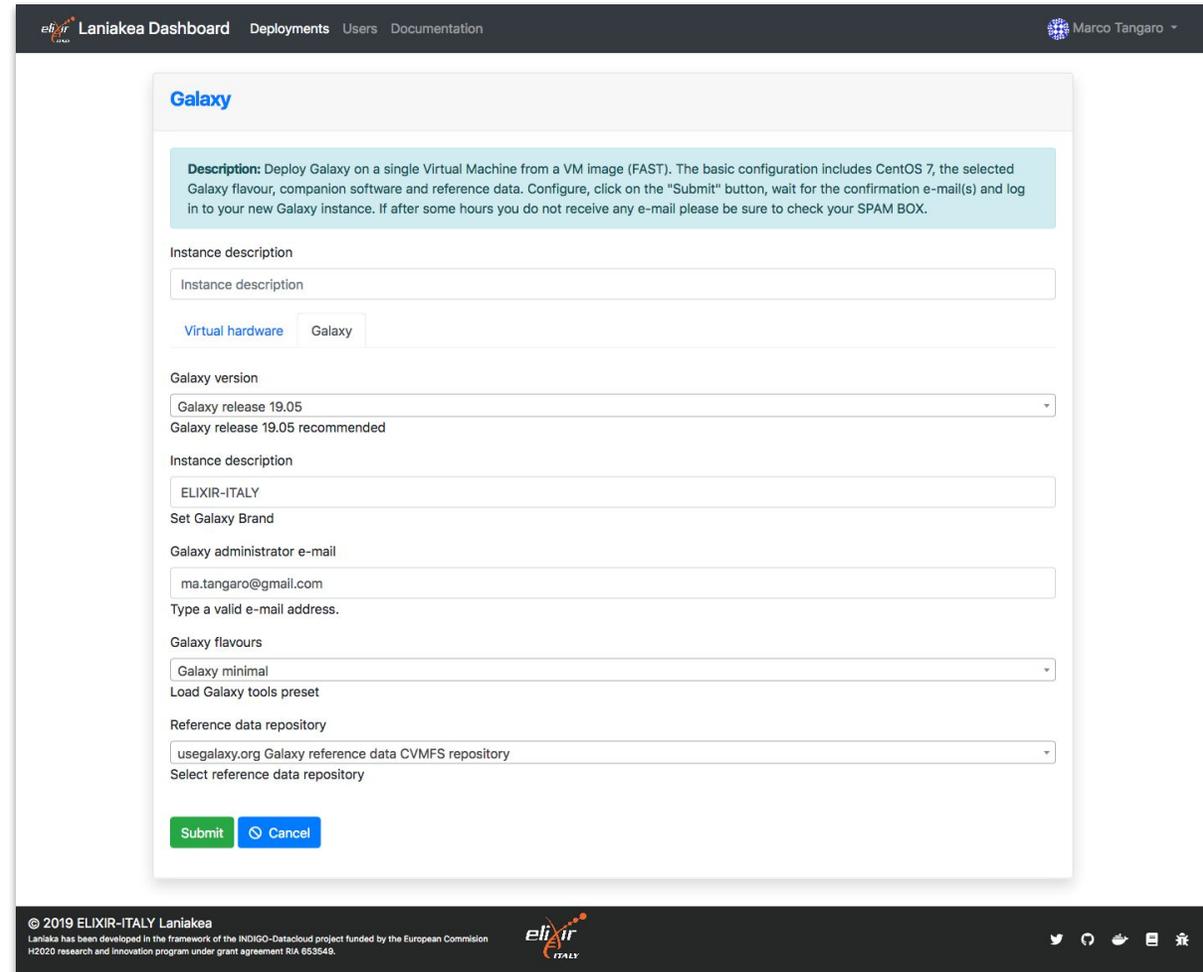**Docker:** fast deployment of new flavours.

Laniakea

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

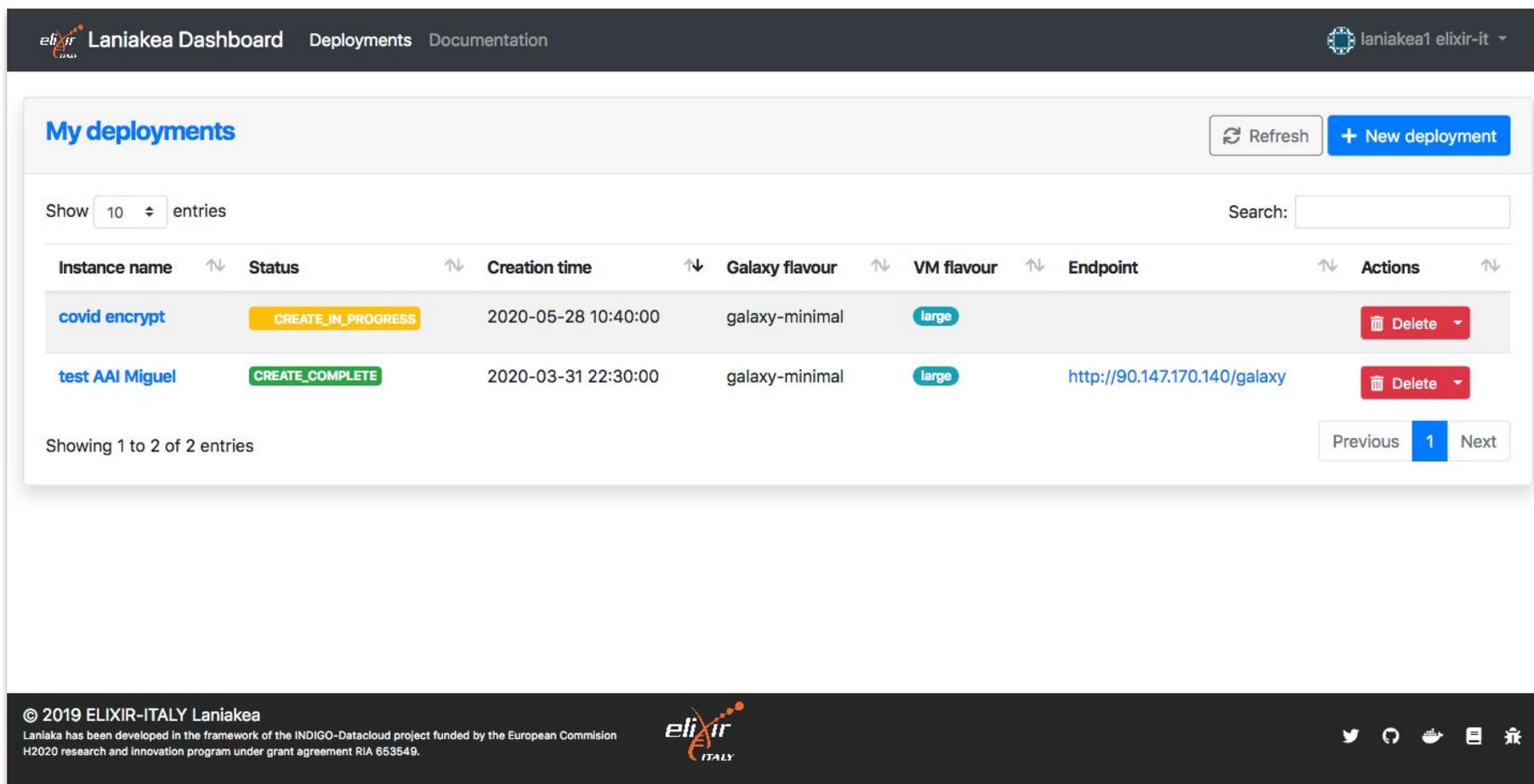Virtual hardware: CPU, RAM and Storage

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 architecture (simplified view)
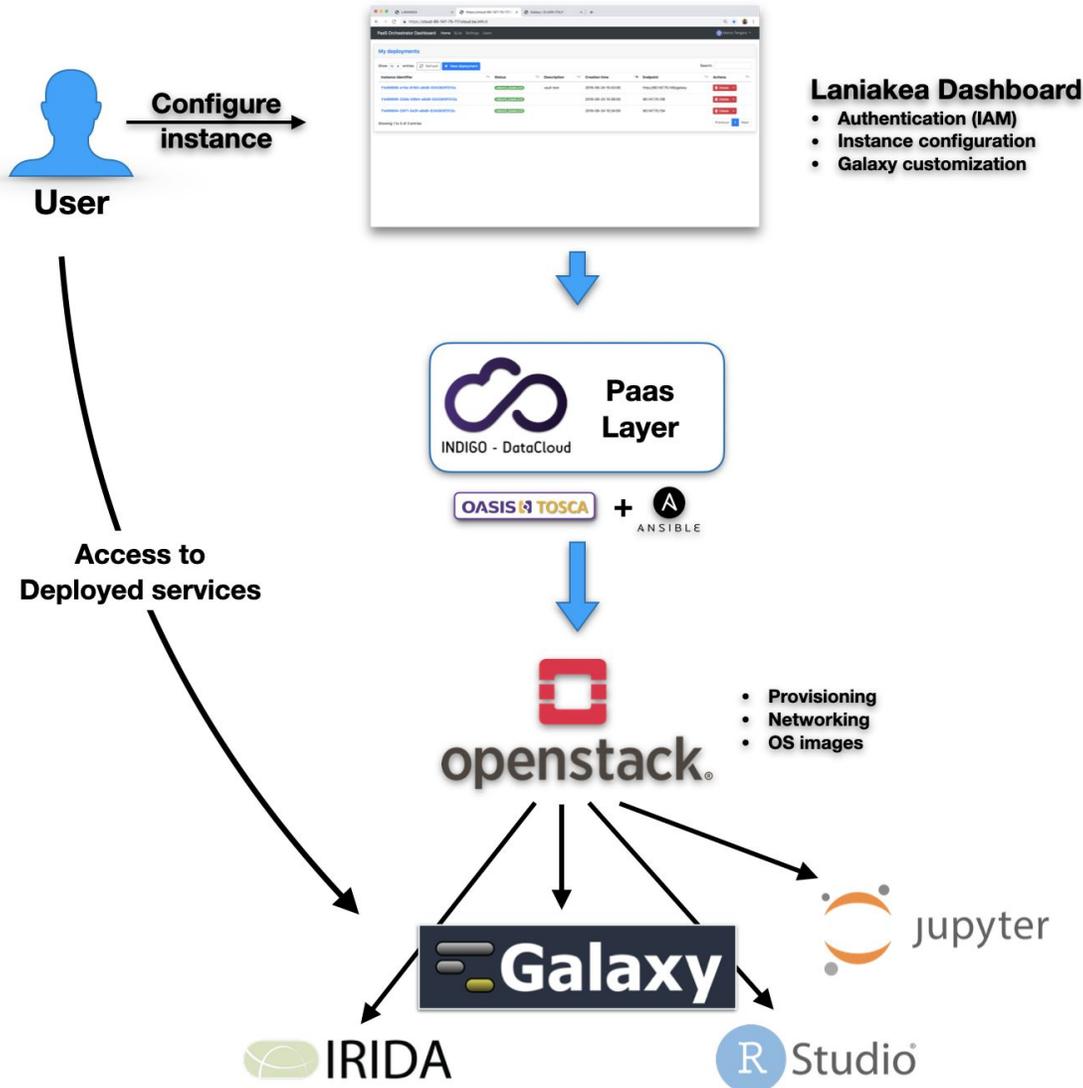
Laniakea Dashboard
- **Authentication (IAM)**
- **Instance configuration**
- **Galaxy customization**

Configure instance → User

Access to Deployed services

Paas Layer INDIGO - DataCloud

OASIS TOSCA + ANSIBLE

openstack
- **Provisioning**
- **Networking**
- **OS images**

Galaxy

Jupyter

IRIDA

R Studio

- **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
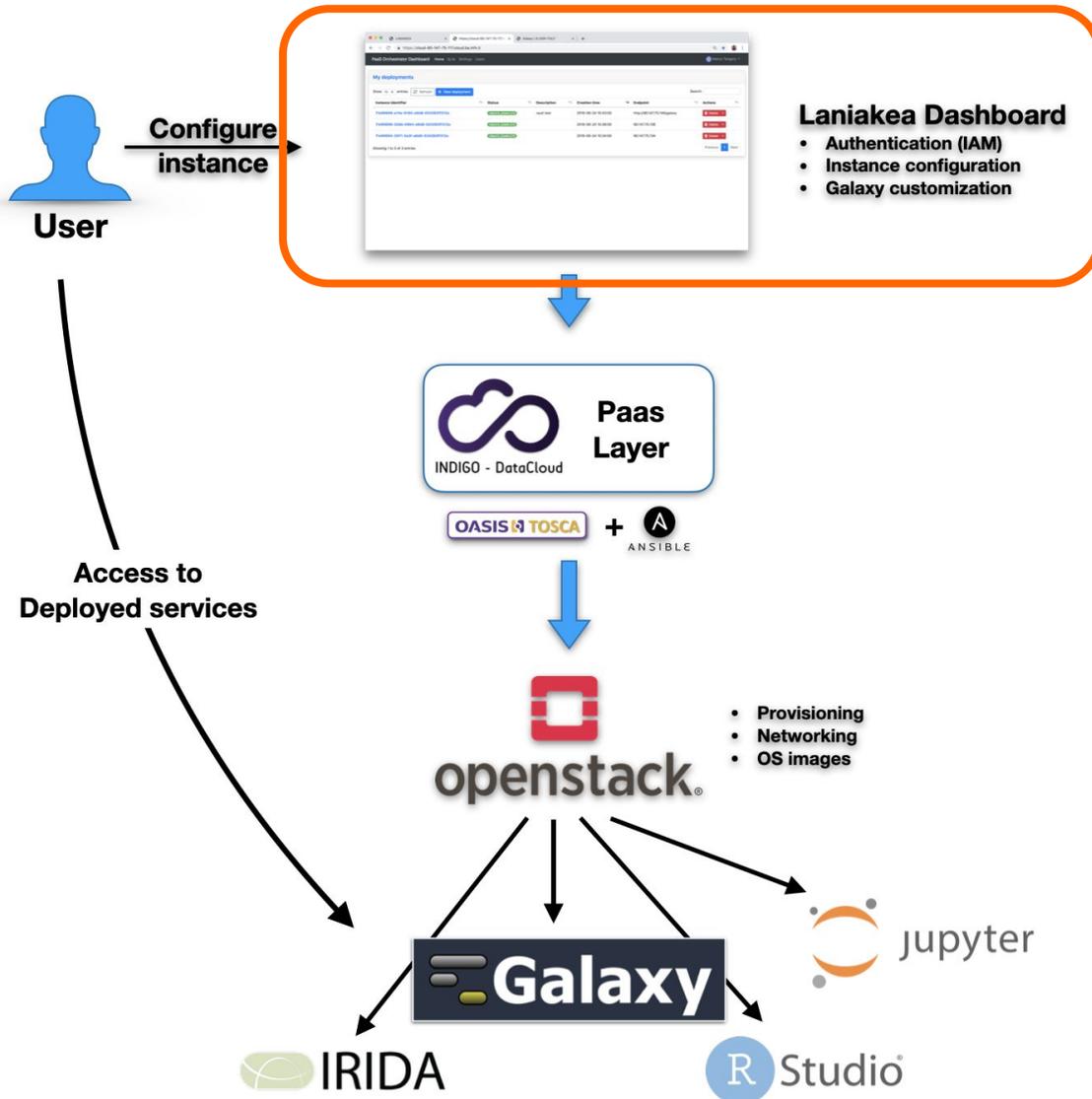
Laniakea architecture

Laniakea Dashboard
- **Authentication (IAM)**
- **Instance configuration**
- **Galaxy customization**

- **Dashboard -** User friendly access to configure and launch a Galaxy instance

- **INDIGO PaaS -** Galaxy automatic deployment

- **Cloud Providers -** ReCaS-Bari

44

Laniakea architecture

Laniakea Dashboard
- **Authentication (IAM)**
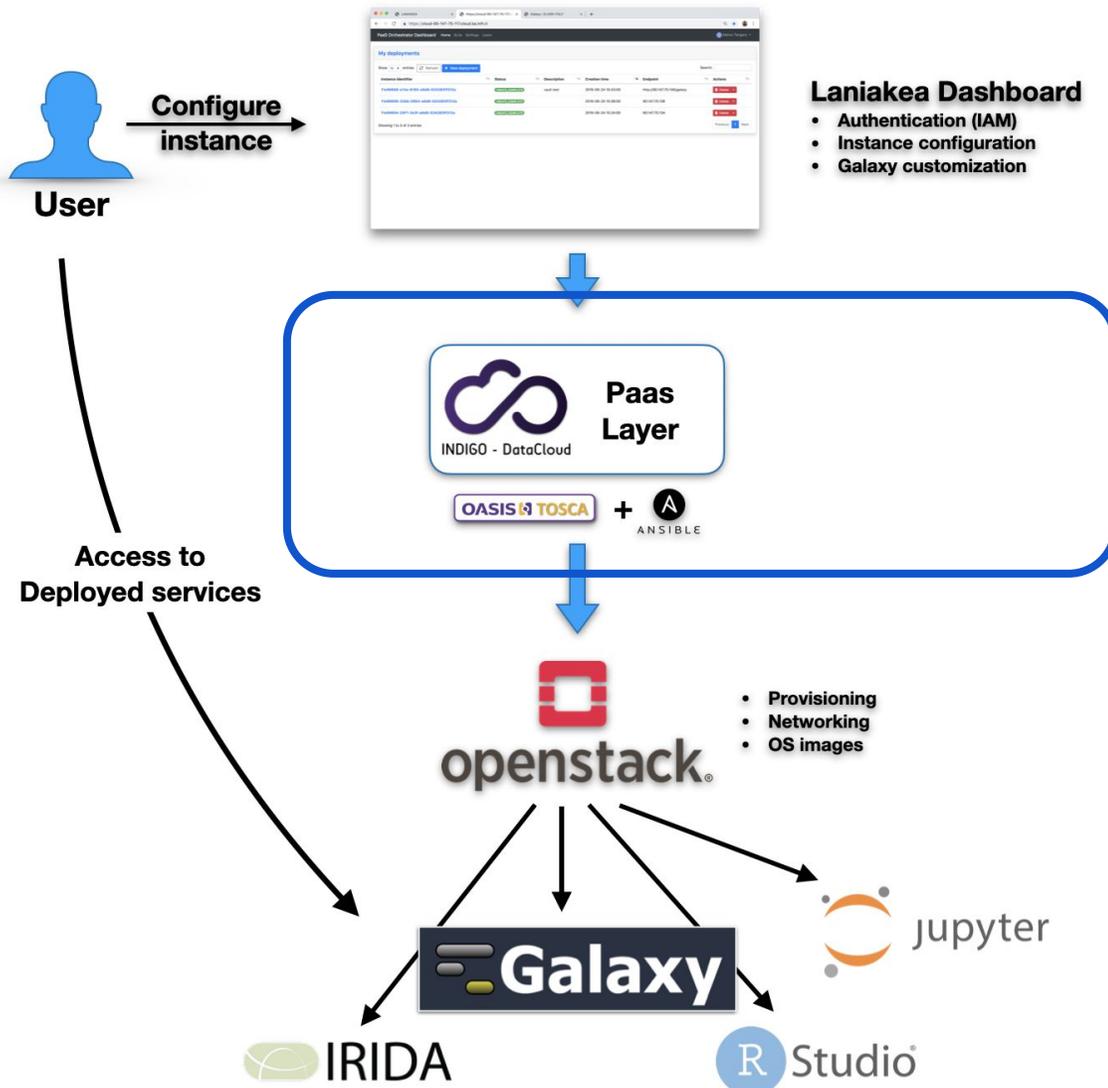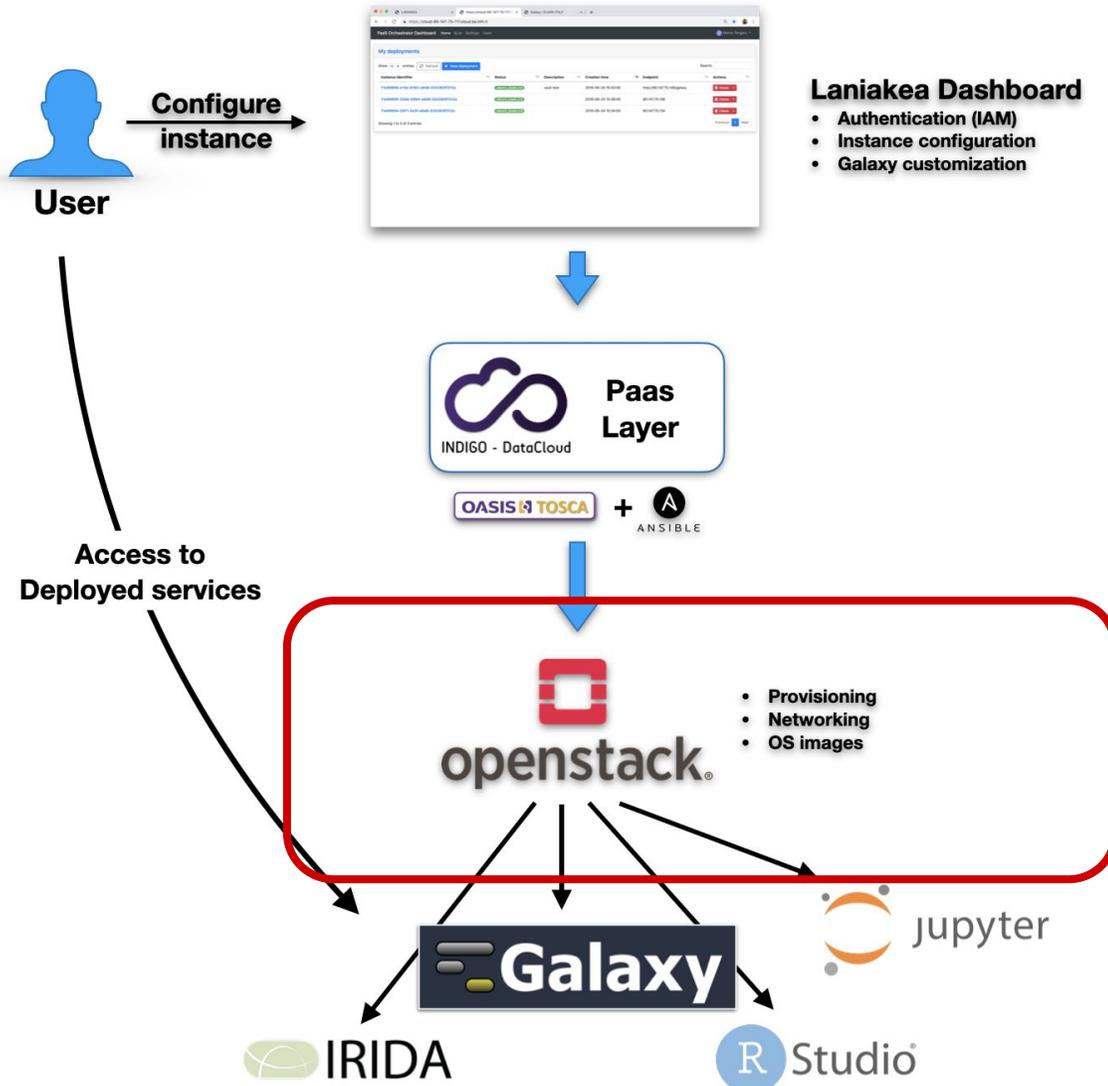- **Instance configuration**
- **Galaxy customization**

- **Dashboard -** User friendly access to configure and launch a Galaxy instance

- **INDIGO PaaS -** Galaxy automatic deployment

- **Cloud Providers -** ReCaS-Bari

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.