



*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

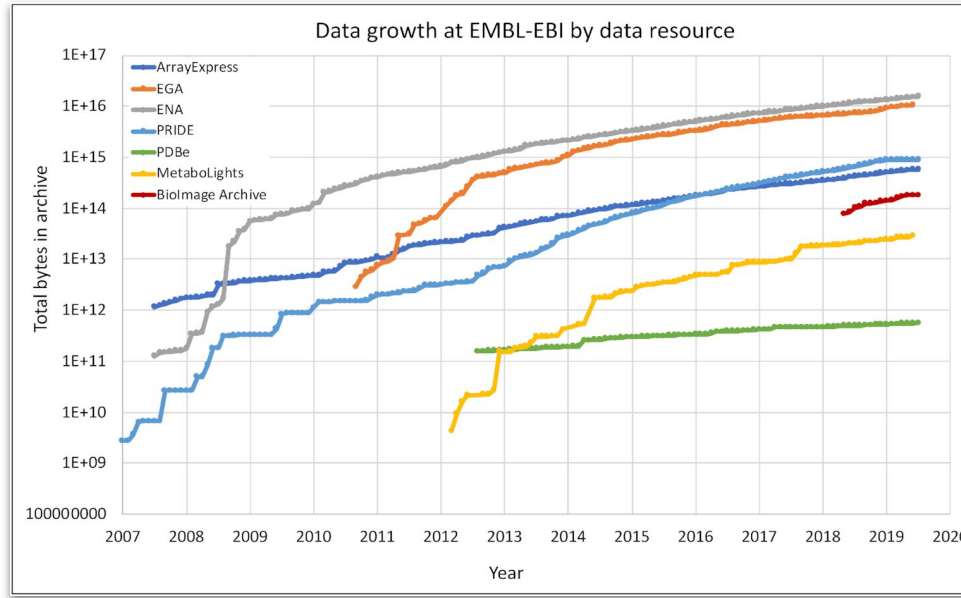
Outline



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

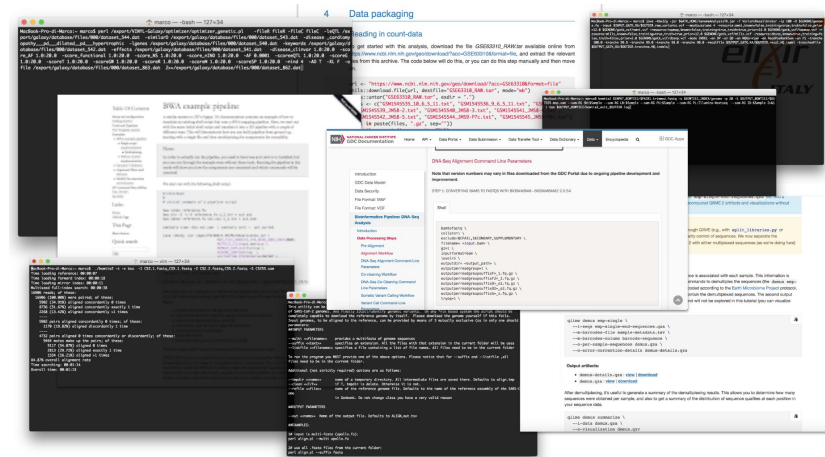
Motivation

DATA SOURCES



DATA STORAGE

DATA ANALYSIS TOOLS



DATA PROTECTION (GDPR)



The Galaxy Project logo, consisting of a stylized icon of three horizontal bars (two dark grey, one yellow) to the left of the word "Galaxy" in a large, bold, dark grey font, with the word "PROJECT" in a smaller, spaced-out, dark grey font below it.

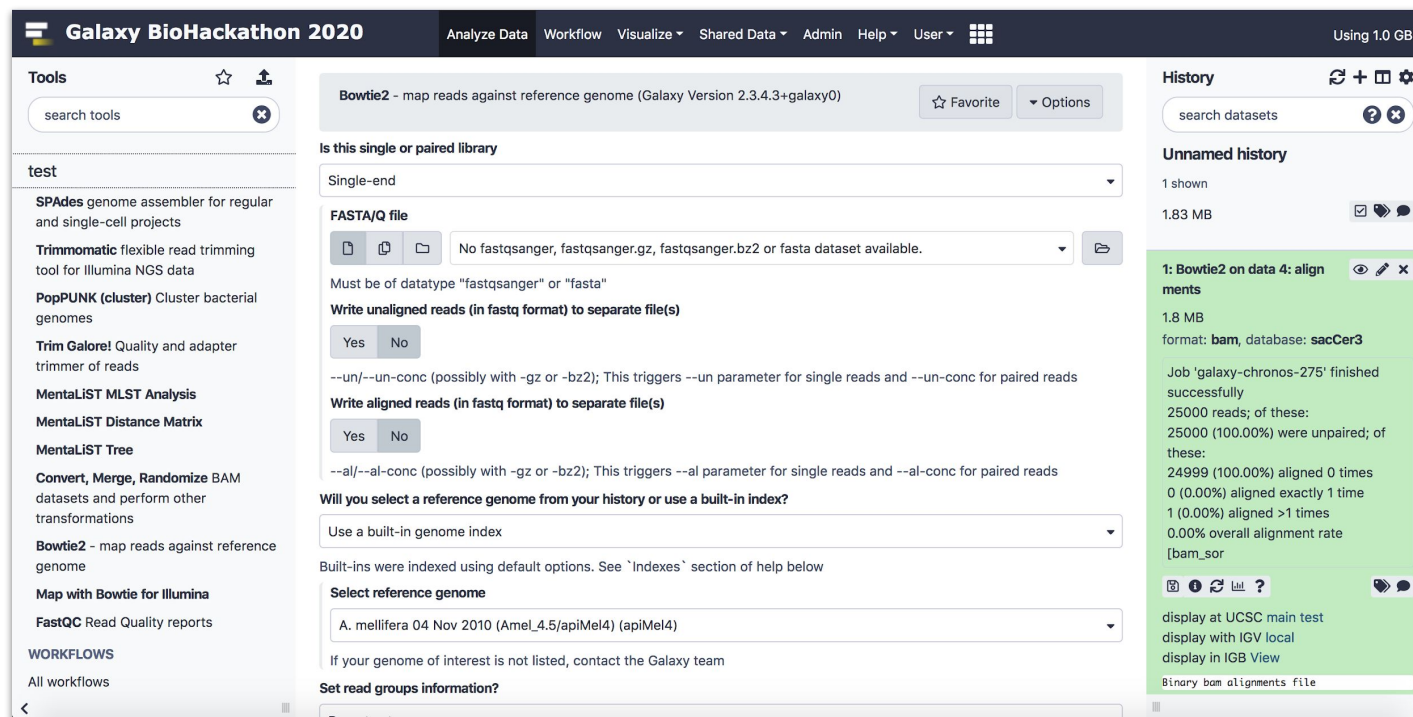
Galaxy

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 BioHackathon 2020 | Analyze Data | Workflow | Visualize | Shared Data | Admin | Help | User | Using 1.0 GB

Tools

search tools

test

- SPAdes genome assembler for regular and single-cell projects**
- Trimmomatic** flexible read trimming tool for Illumina NGS data
- PopPUNK (cluster)** Cluster bacterial genomes
- Trim Galore!** Quality and adapter trimmer of reads
- MentalIST MLST Analysis**
- MentalIST Distance Matrix**
- MentalIST Tree**
- Convert, Merge, Randomize BAM datasets and perform other transformations**
- Bowtie2** - map reads against reference genome
- Map with Bowtie for Illumina**
- FastQC** Read Quality reports

WORKFLOWS

All workflows

Bowtie2 - map reads against reference genome (Galaxy Version 2.3.4.3+galaxy0)

☆ Favorite | Options

Is this single or paired library

Single-end

FASTA/Q file

No fastqsanger, fastqsanger.gz, fastqsanger.bz2 or fasta dataset available.

Must be of datatype "fastqsanger" or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

Yes No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

Yes No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See `Indexes` section of help below

Select reference genome

A. mellifera 04 Nov 2010 (AmeL4.5/apiMeL4) (apiMeL4)

If your genome of interest is not listed, contact the Galaxy team

Set read groups information?

History | search datasets

1 shown | 1.83 MB

1: Bowtie2 on data 4: alignments | 1.8 MB

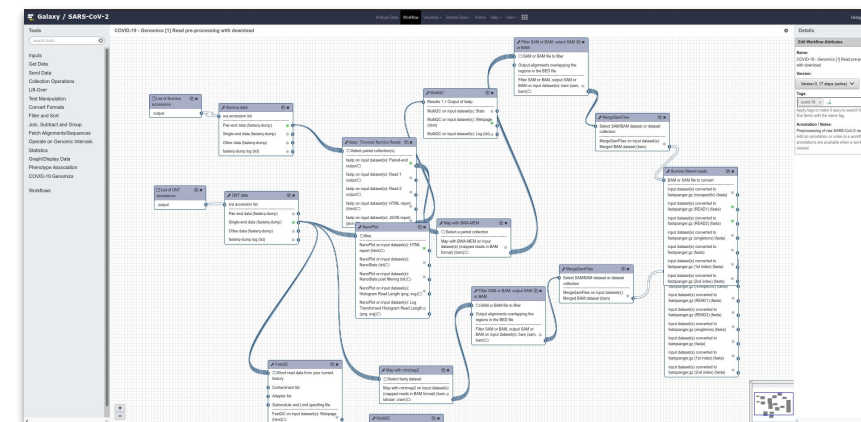
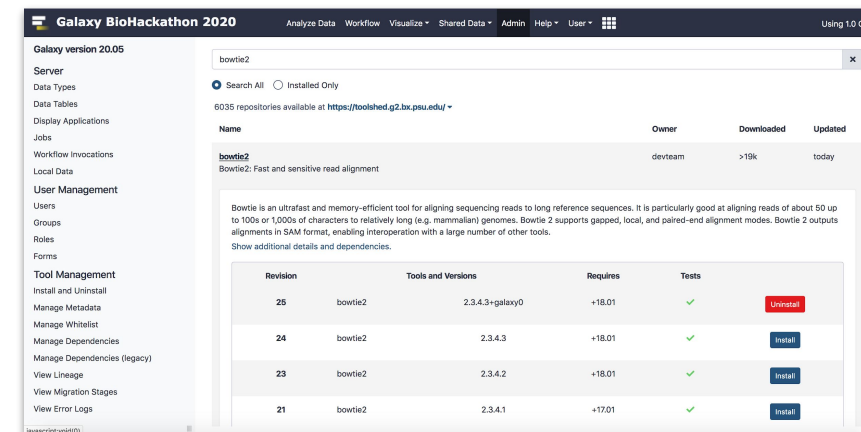
format: bam, database: sacCer3

Job 'galaxy-chronos-275' finished successfully
25000 reads; of these:
25000 (100.00%) were unpaired; of these:
24999 (100.00%) aligned 0 times
0 (0.00%) aligned exactly 1 time
1 (0.00%) aligned >1 times
0.00% overall alignment rate

[bam_sor]

display at UCSC main test
display with IGV local
display in IGB View

Binary bam alignments file

Galaxy BioHackathon 2020 | Analyze Data | Workflow | Visualize | Shared Data | Admin | Help | User | Using 1.0 GB

Galaxy version 20.05

Server

- Data Types
- Data Tables
- Display Applications
- Jobs
- Workflow Invocations
- Local Data
- User Management
- Users
- Groups
- Roles
- Forms

Tool Management

Install and Uninstall
Manage Metadata
Manage Dependencies
Manage Dependencies (legacy)
View Lineage
View Migration Stages
View Error Logs

bowtie2

Search All | Installed Only

6035 repositories available at <https://toolshed.g2.bx.psu.edu/>

Name	Owner	Downloaded	Updated
bowtie2	devteam	>19k	today

Bowtie2: Fast and sensitive read alignment

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 supports gapped, local, and paired-end alignment modes. Bowtie 2 outputs alignments in SAM format, enabling interoperability with a large number of other tools. Show additional details and dependencies.

Revision	Tools and Versions	Requires	Tests
25	bowtie2 2.3.4.3+galaxy0	+18.01	✓ Uninstall
24	bowtie2 2.3.4.3	+18.01	✓ Install
23	bowtie2 2.3.4.2	+18.01	✓ Install
21	bowtie2 2.3.4.1	+17.01	✓ Install

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

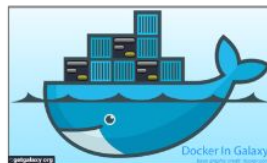
Laniakea

LANIAKEA IS A CLOUD BASED GALAXY INSTANCE PROVIDER.

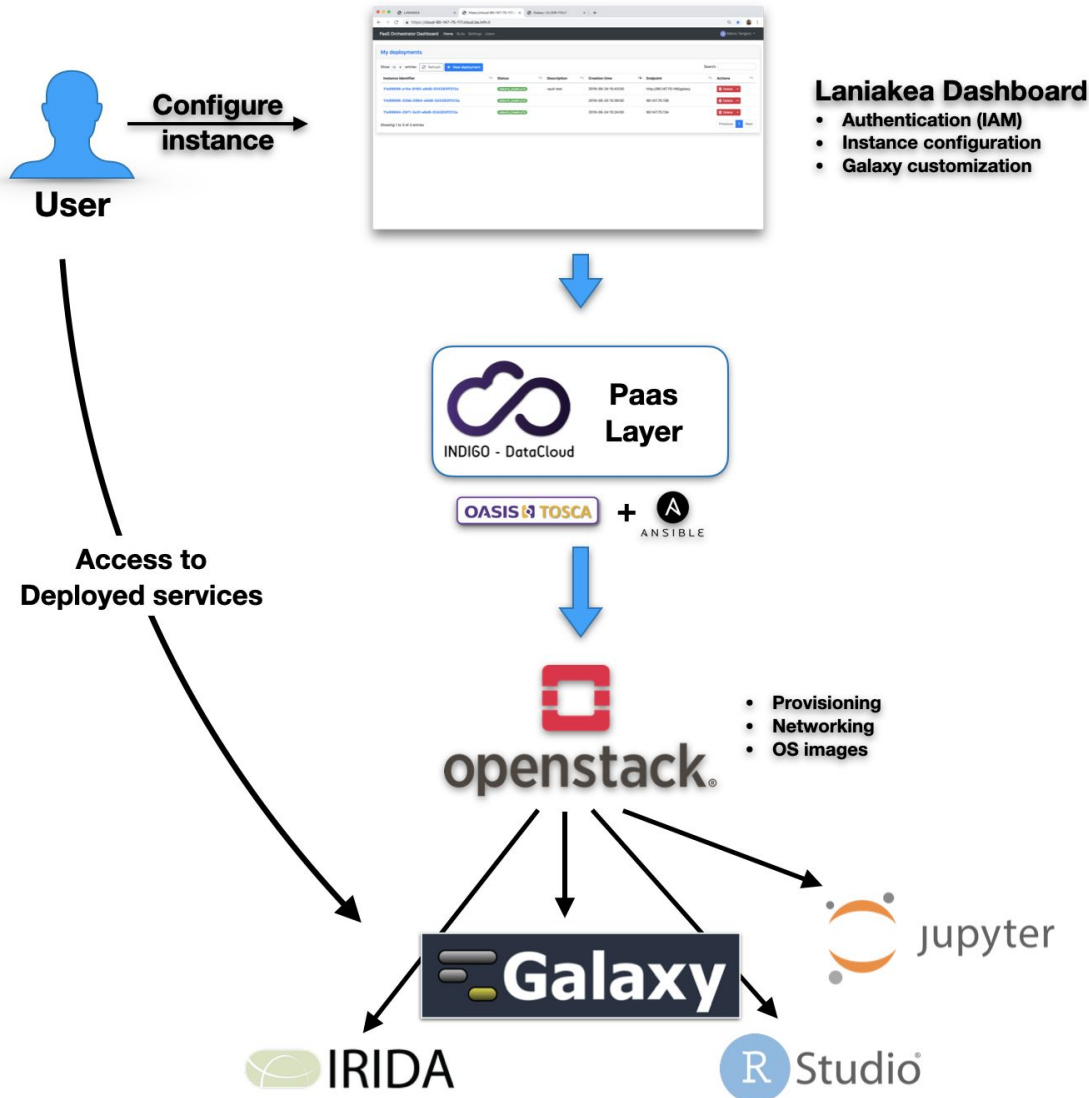
<https://laniakea-elixir-it.github.io/>

- 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)

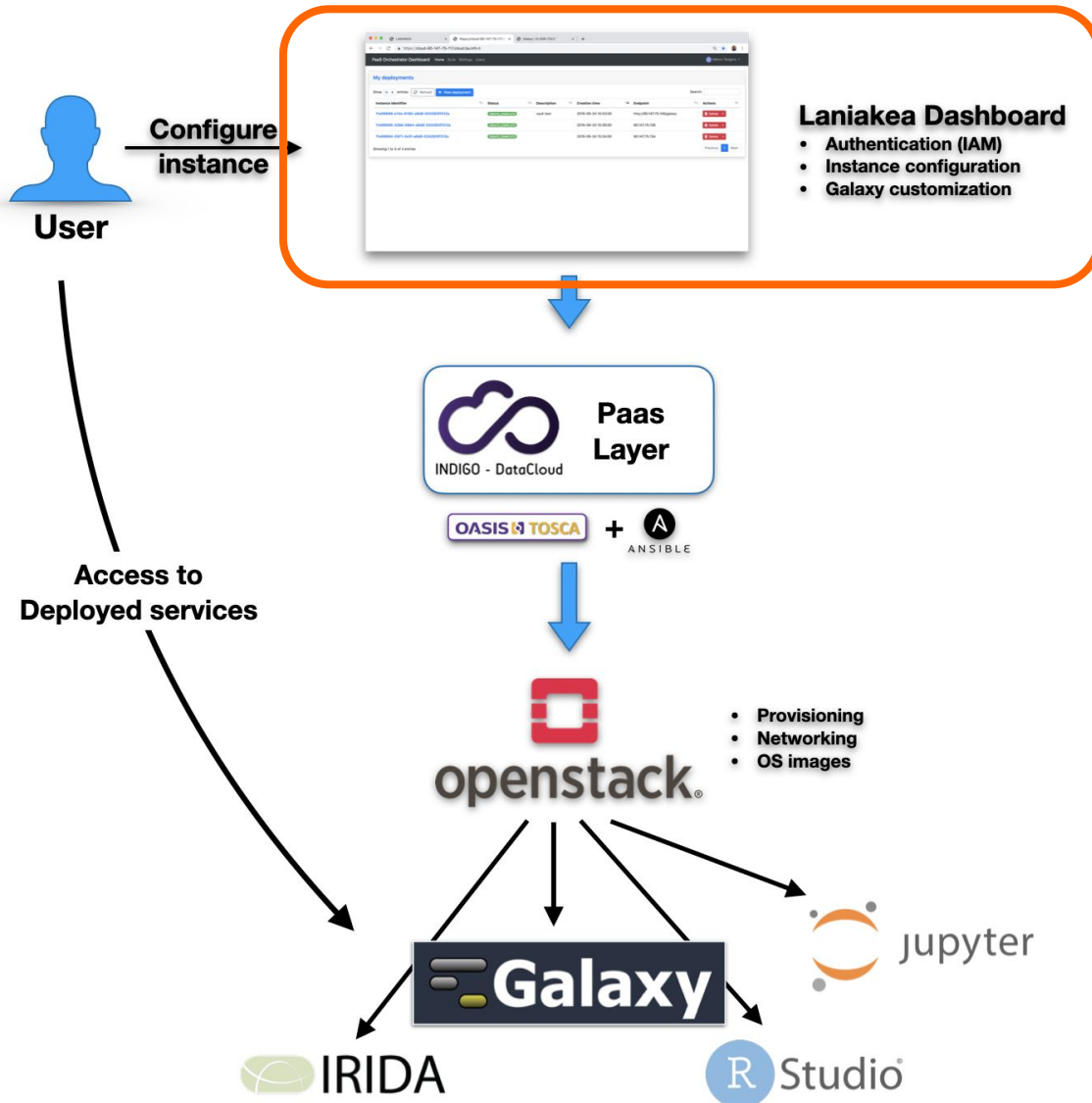


Laniakea Dashboard

- Authentication (IAM)
- Instance configuration
- Galaxy customization

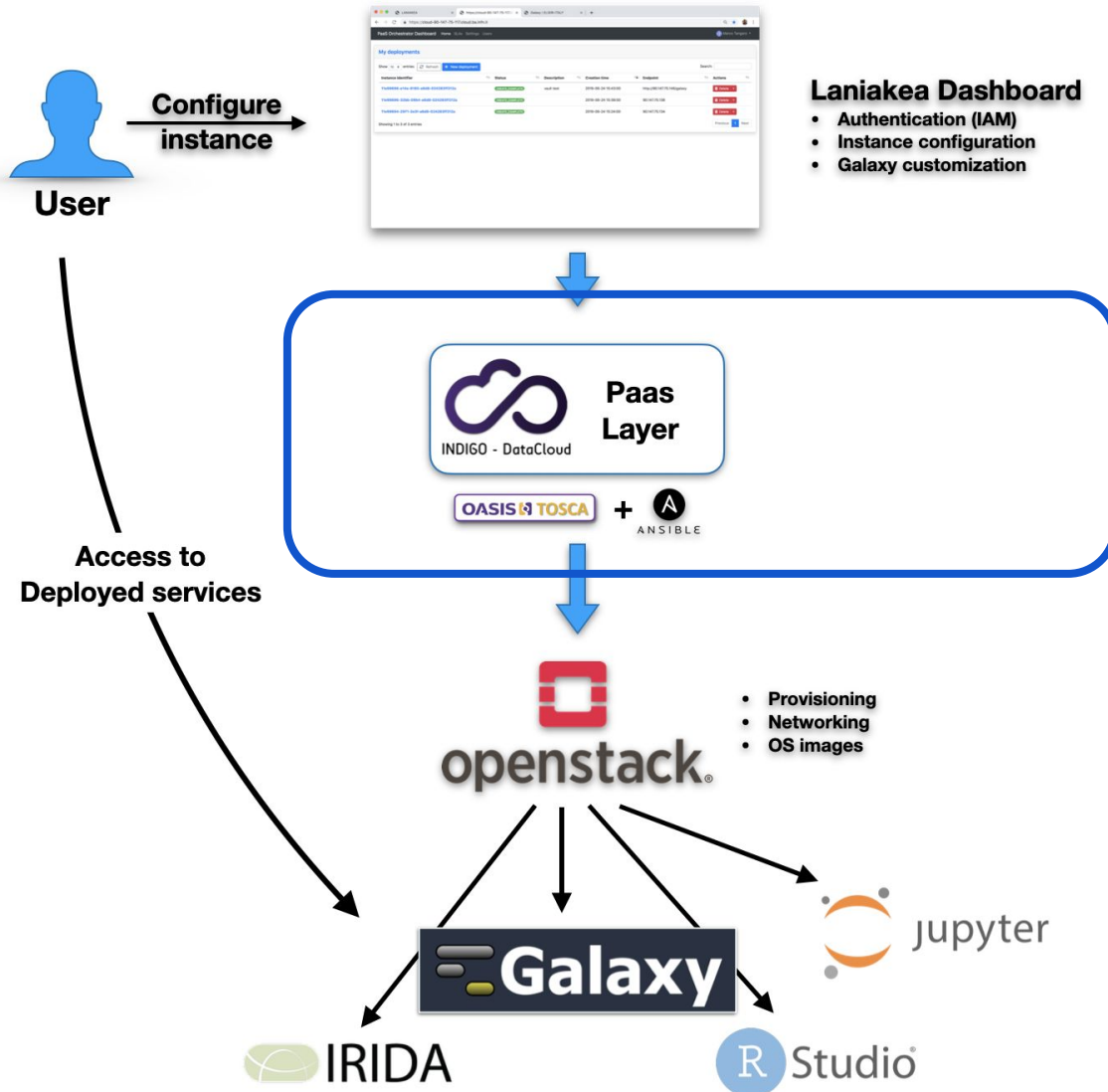
- **Dashboard** - User friendly access to configure and launch a Galaxy instance
- **INDIGO PaaS** - Galaxy automatic d
- **Cloud Providers** - ReCaS-Bari

Laniakea architecture



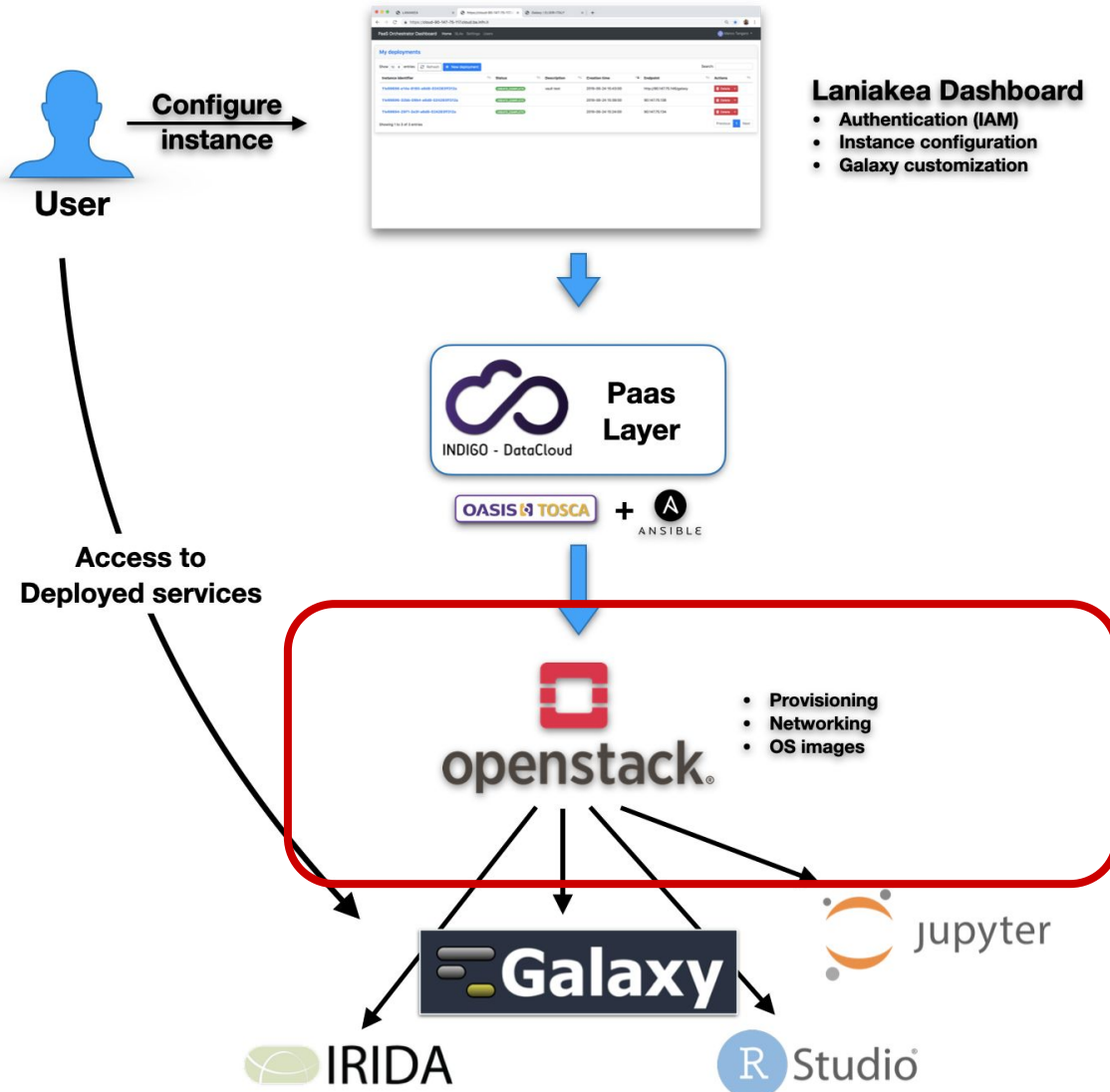
- **Dashboard** - User friendly access to configure and launch a Galaxy instance
- **INDIGO PaaS** - Galaxy automatic d
- **Cloud Providers** - ReCaS-Bari

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”.



OPENVPN

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

Laniakea encryption

Laniakea encryption

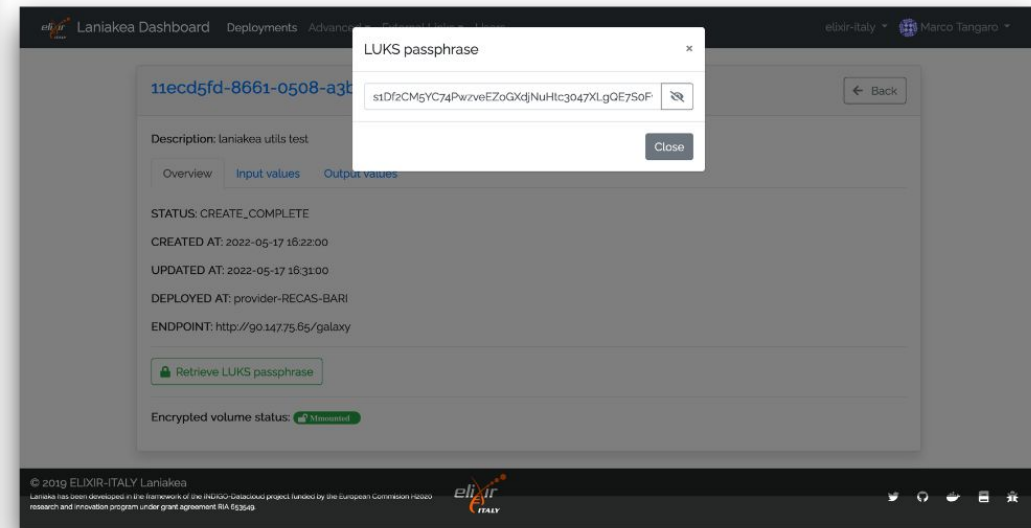
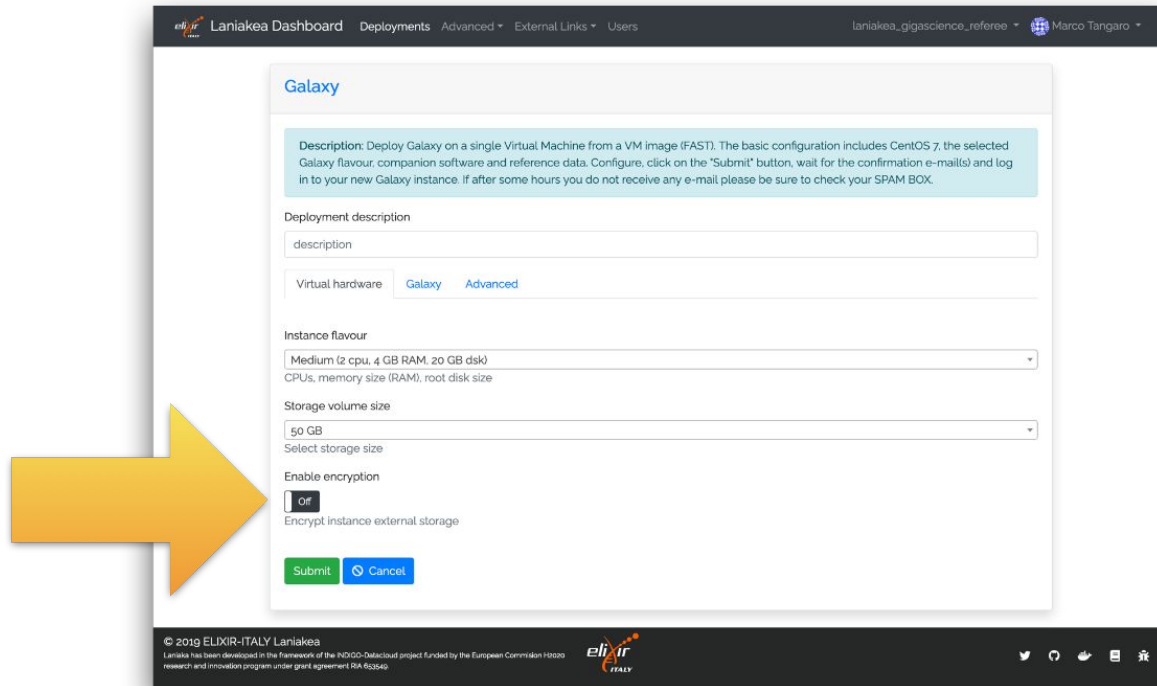


The user data privacy is granted through **LUKS** storage encryption as a service: the encryption procedure is automated in order to simplify the user experience, each user can encrypt storage on-demand, using a strong random alphanumerical passphrase.

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**.

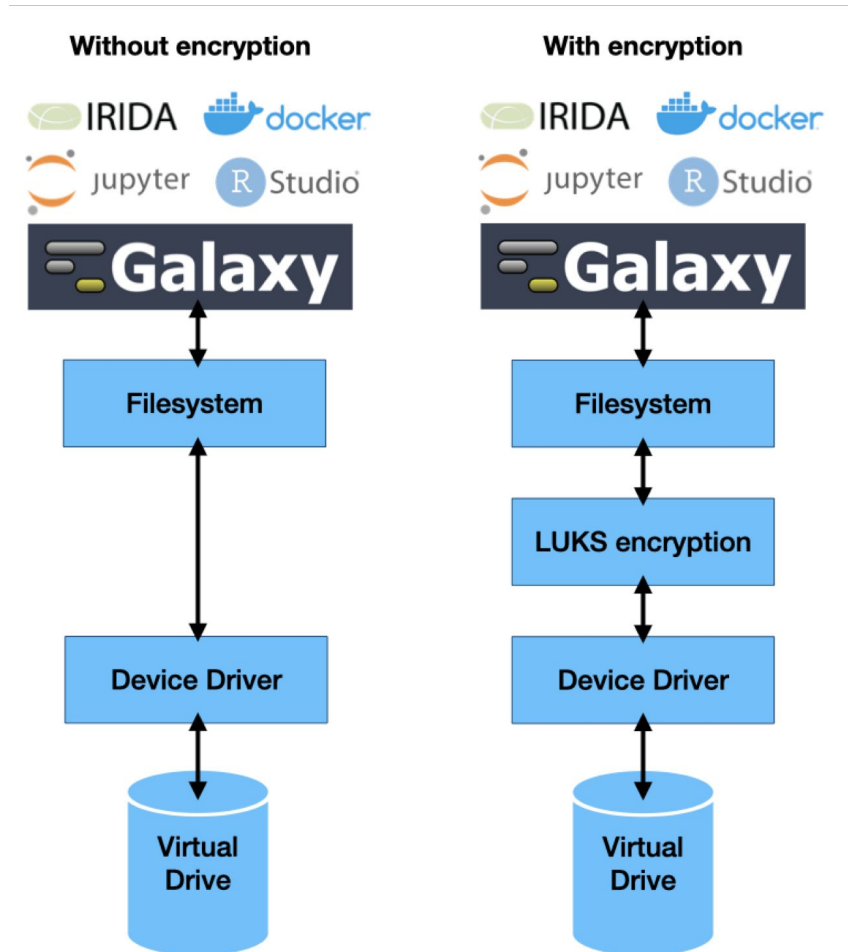
The encryption layer sits between the physical disk and the file system.

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:

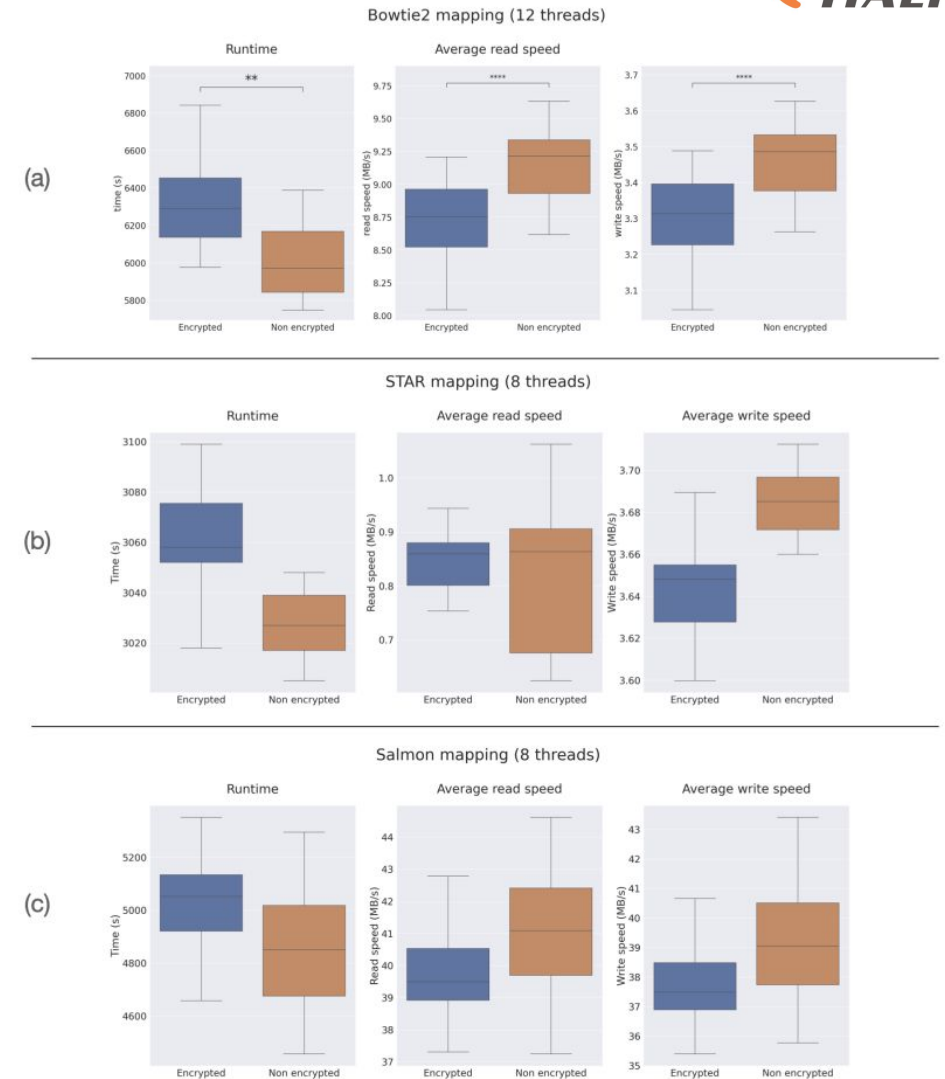
- aes-xts-plain64 encryption
- 256 bit key
- sha256 as hash algorithm used for key derivation.



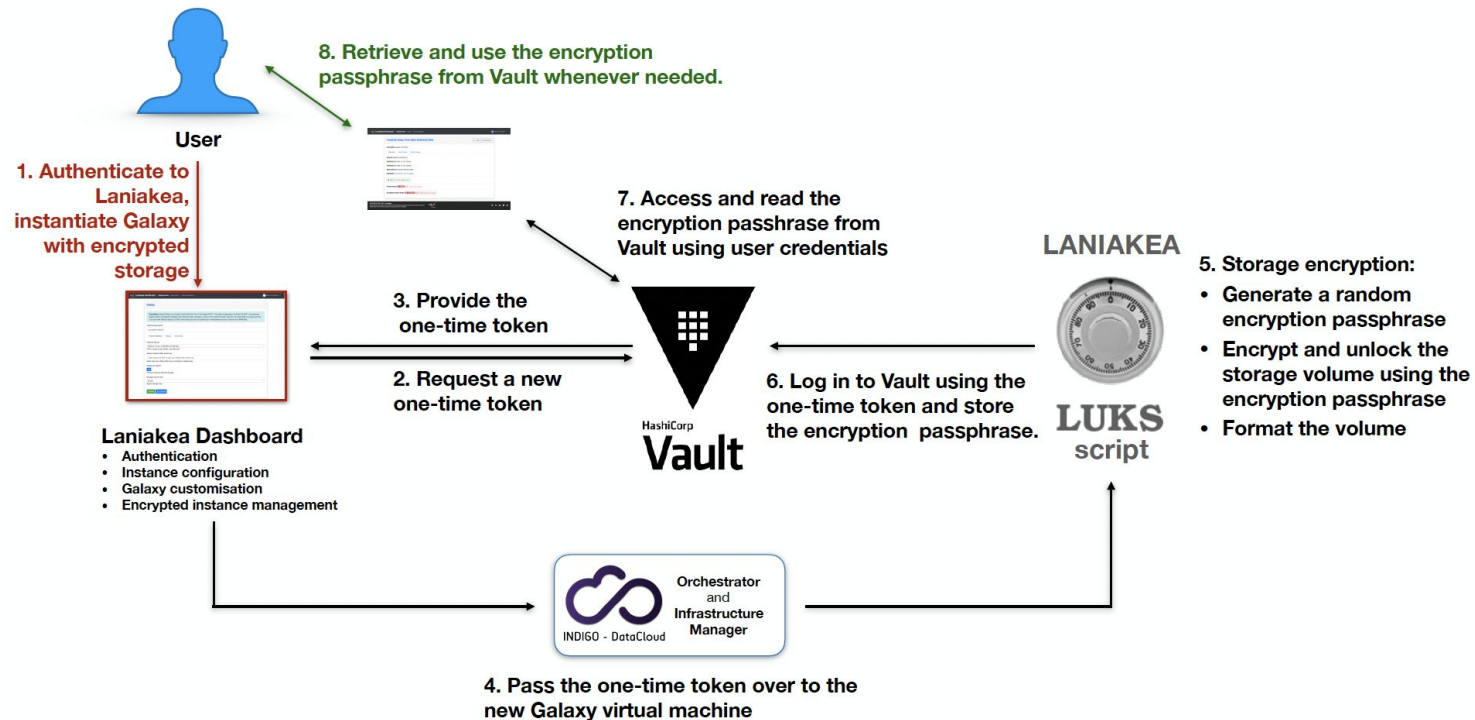
The underlying infrastructure

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

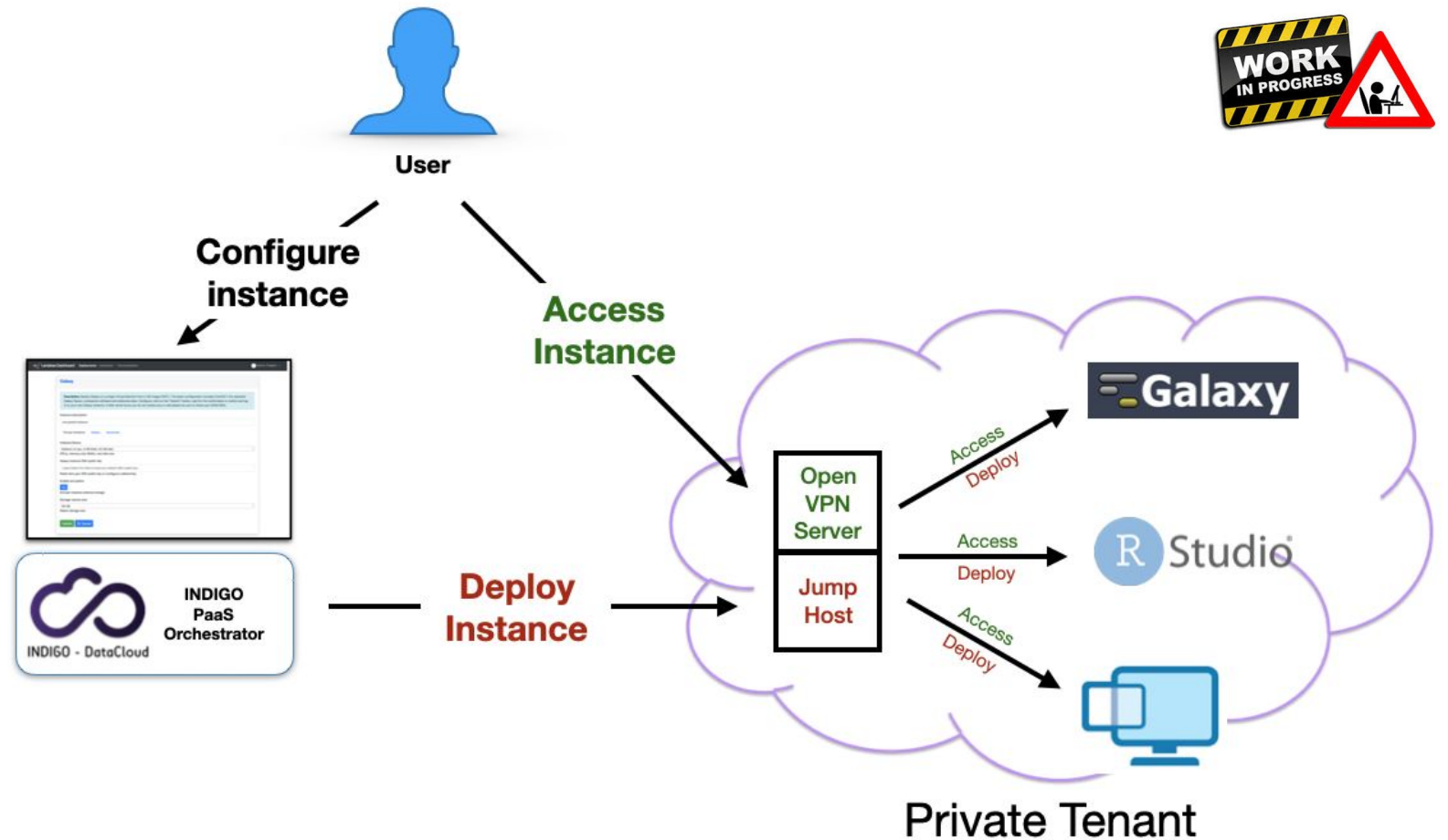
Deployments under VPN



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.



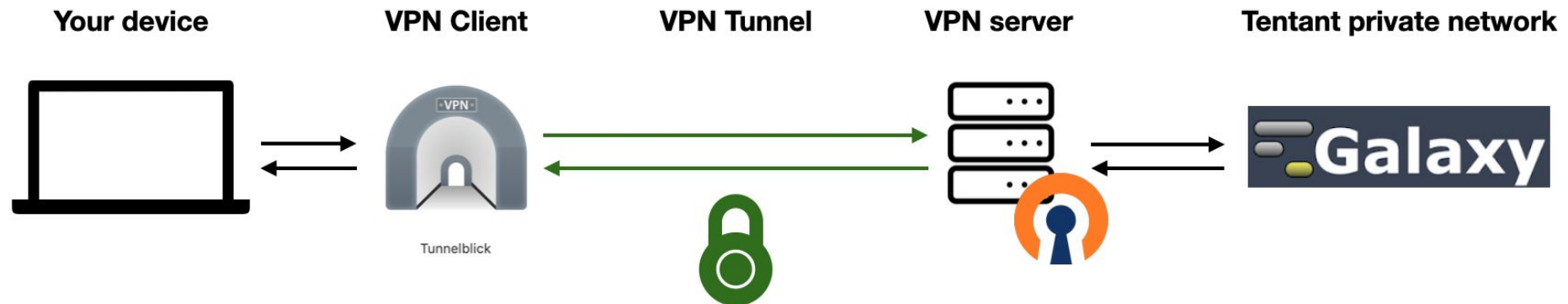
Deployments under VPN



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

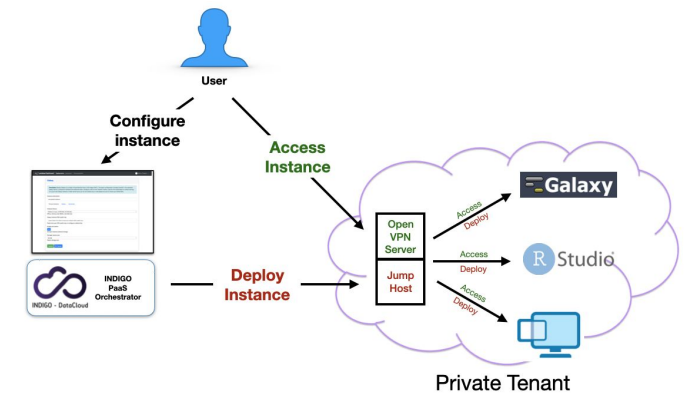
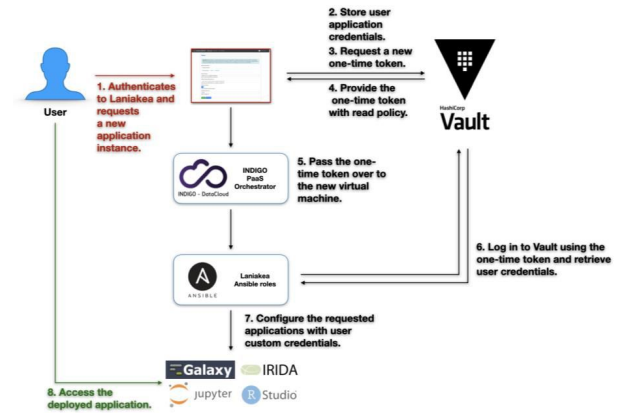


Conclusions

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

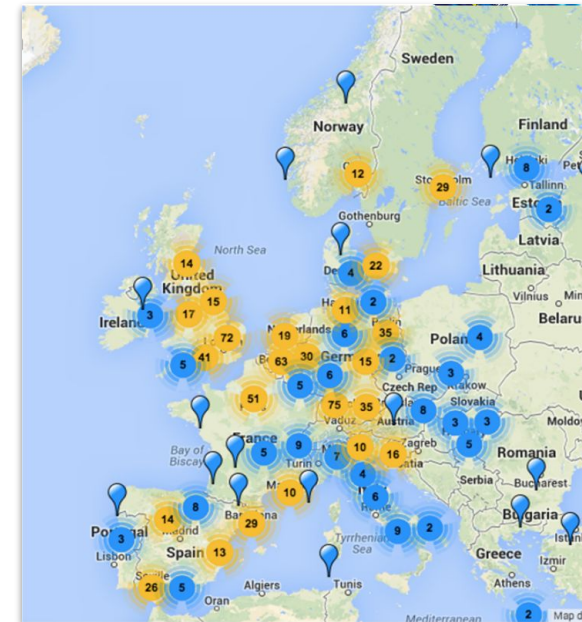
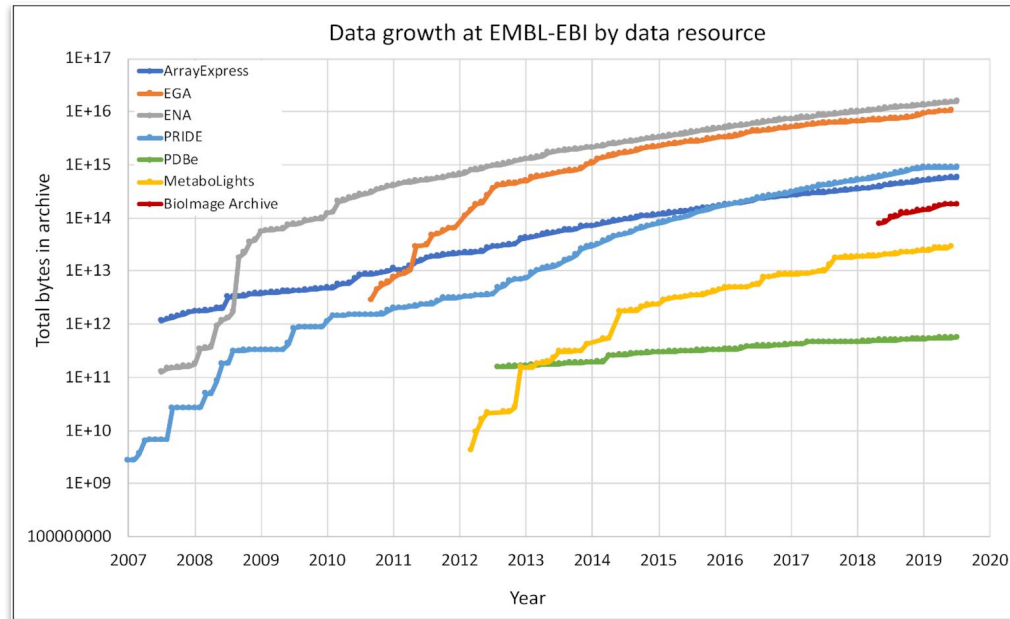
Nadina Foggetti (Legal expert) nadina.foggetti@ba.infn.it

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

4 Data packaging

Reading in count-data

```
MacBook-Pro-di-Marco:~$ macos perl -export/V2/N1-Galaxy/optimizer/optimizer_genetic.pl --file file --file file -lqt /export/galaxy/database/files/000/dataset_544.dot --smlnd /export/galaxy/database/files/000/dataset_543.dot --dataset_cordony ...
```

Table Of Contents

BWA example pipeline

Introduction

Genomics Pipeline: DNA-Seq Analysis

```
bash fastq \
collater \
exclude=@FASTQ,SECONDARY,SUPPLEMENTARY \
filter <input_bam \
gzi \
inputformat=bam \
level=1 \
outputdir=output_path \
outputperreadgroup \
outputperreadgroupffix=1_fix.gz \
outputperreadgroupffix=2_fix.gz \
outputperreadgroupffix=3_fix.gz \
outputperreadgroupffix=4_fix.gz \
outputperreadgroupffix=5_fix.gz \
tryopt \
```

Output artifacts:

- demux-details.gsv: [view](#) | [download](#)
- demux.gsv: [view](#) | [download](#)

After demultiplexing, it's useful to generate a summary of the demultiplexing results. This allows you to determine how many sequences were obtained per sample, and also to get a summary of the distribution of sequence qualities at each position in your sequence data.

```
q1ize demux summarize \
--data demux.gsv \
--visualization demux.gsv
```

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

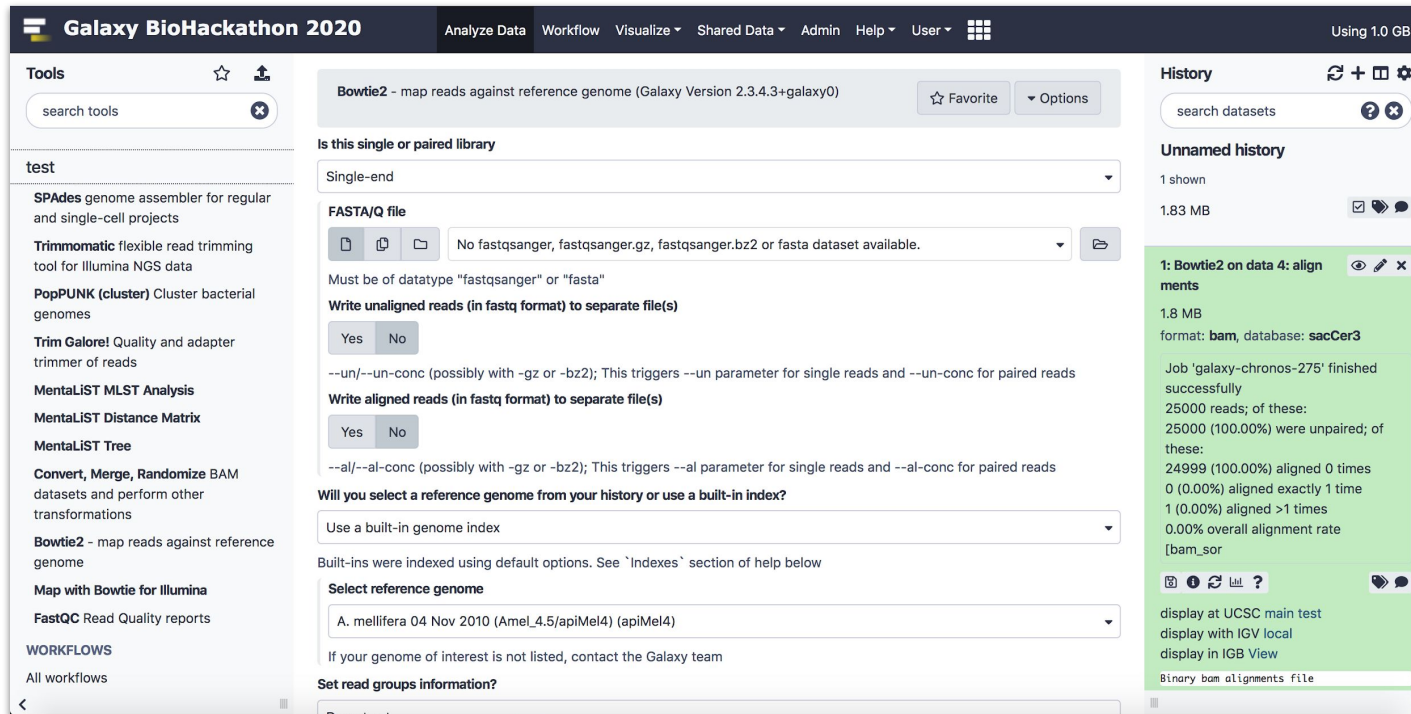
Workflows involve more than one tools!

Tools are usually manually installed.

GDPR

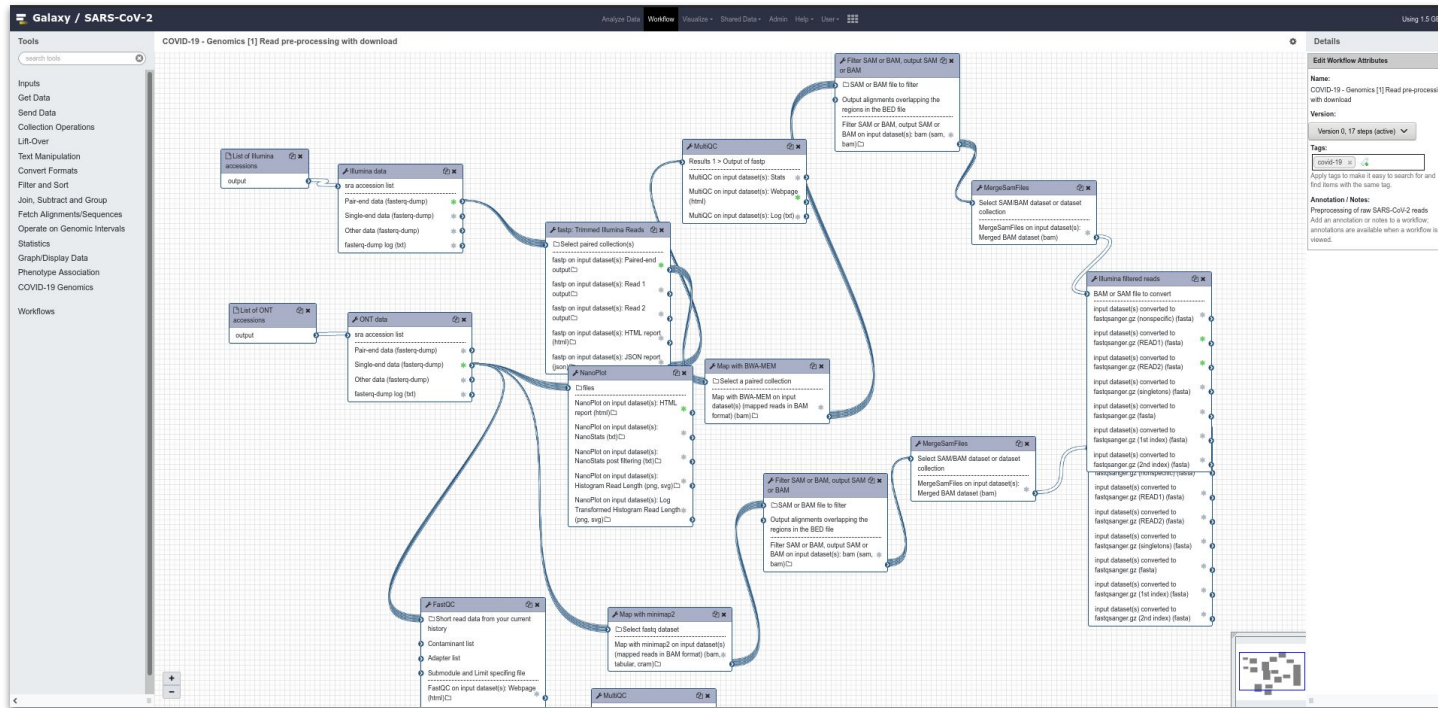
The GDPR explicitly recognizes the sensitive nature of the collected genetic data (Article 9), but at the same time permits sensitive genetic data processing for scientific research purposes (Article 89(1)), provided this is allowed by EU or Member States law framework and appropriate safeguards measures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





The screenshot shows the Galaxy BioHackathon 2020 interface. The main tool is Bowtie2, configured for a single-end library. The FASTA/Q file field is empty, with a message indicating no files are available. The reference genome is set to 'A. mellifera 04 Nov 2010 (AmeL4.5/apiMel4) (apiMel4)'. The job history panel on the right shows a completed job '1: Bowtie2 on data 4: alignments' with a size of 1.8 MB. The job details indicate that 25,000 reads were unpaired, and 24,999 reads (100.00%) were aligned 0 times, with 1 read (0.00%) aligned exactly 1 time and 1 read (0.00%) aligned more than 1 time. The overall alignment rate is 0.00%.

- 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 BioHackathon 2020 Analyze Data Workflow Visualize Shared Data Admin Help User Using 1.0 GB

Galaxy version 20.05

Server

Data Types

Data Tables

Display Applications

Jobs

Workflow Invocations

Local Data

User Management

Users

Groups

Roles

Forms

Tool Management

Install and Uninstall

Manage Metadata

Manage Whitelist

Manage Dependencies

Manage Dependencies (legacy)

View Lineage

View Migration Stages

View Error Logs

javascript:void(0)

Search All Installed Only

6035 repositories available at <https://toolshed.g2.bx.psu.edu/>

Name	Owner	Downloaded	Updated
bowtie2 Bowtie2: Fast and sensitive read alignment	devteam	>19k	today

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 supports gapped, local, and paired-end alignment modes. Bowtie 2 outputs alignments in SAM format, enabling interoperability with a large number of other tools.

Show additional details and dependencies.

Revision	Tools and Versions	Requires	Tests	
25	bowtie2 2.3.4.3+galaxy0	+18.01	✓	Uninstall
24	bowtie2 2.3.4.3	+18.01	✓	Install
23	bowtie2 2.3.4.2	+18.01	✓	Install
21	bowtie2 2.3.4.1	+17.01	✓	Install

Galaxy ToolShed

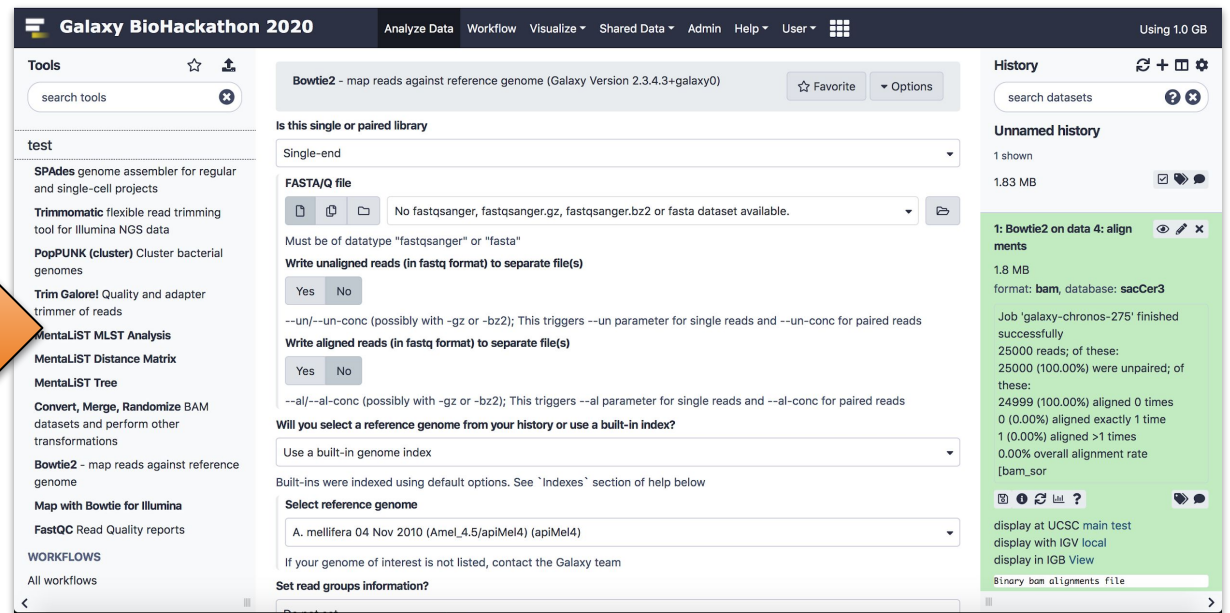
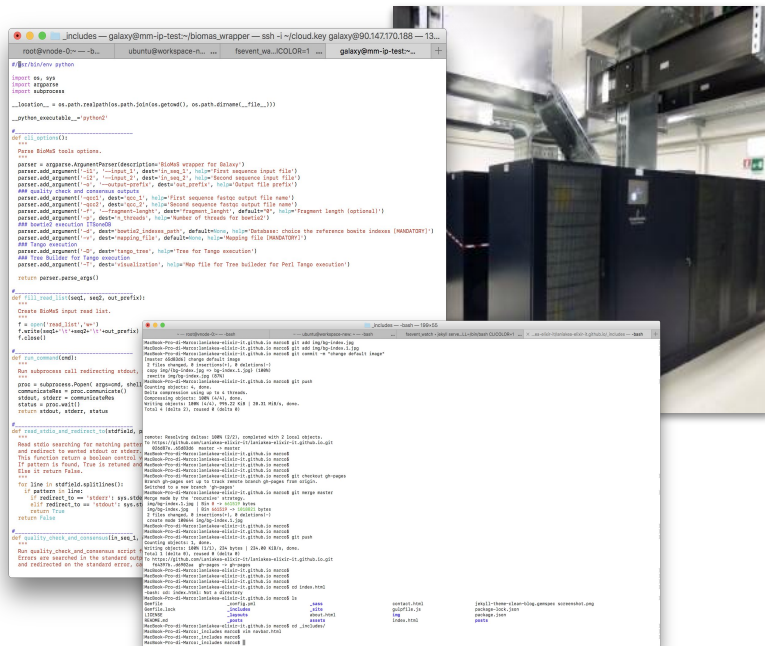
Serves as an "app store" to all Galaxies worldwide.

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.

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



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

Laniakea

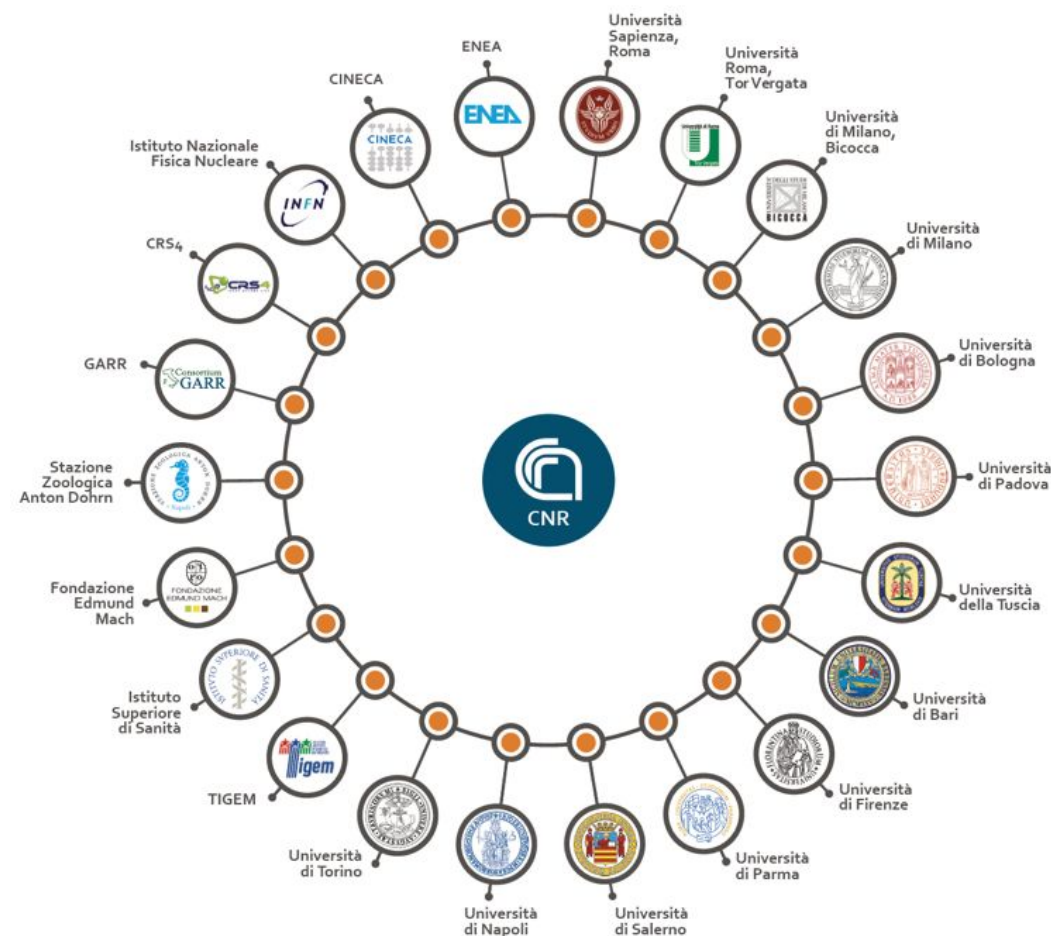


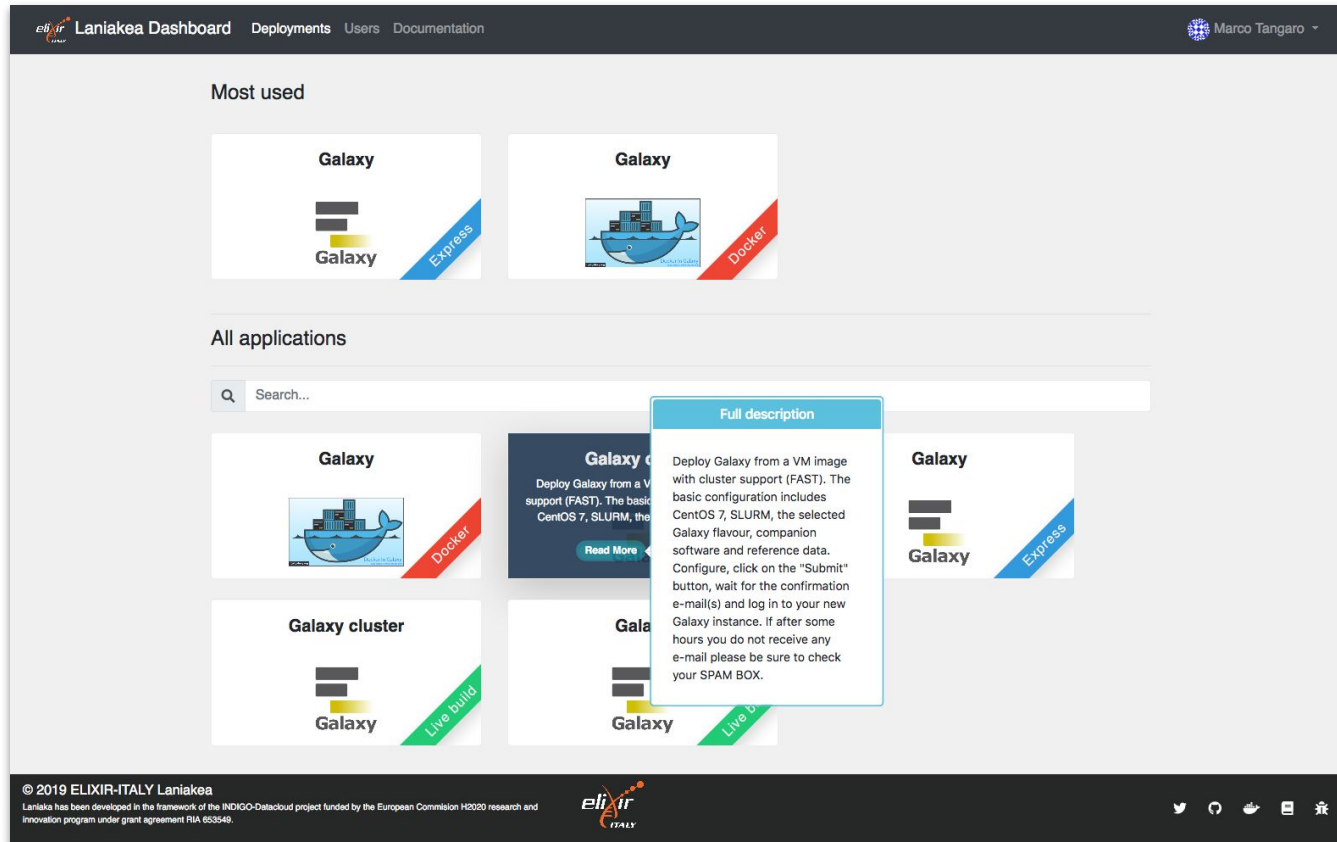
ELIXIR-Italy partners are actively involved in the service development and/or also contribute with cloud resources.

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

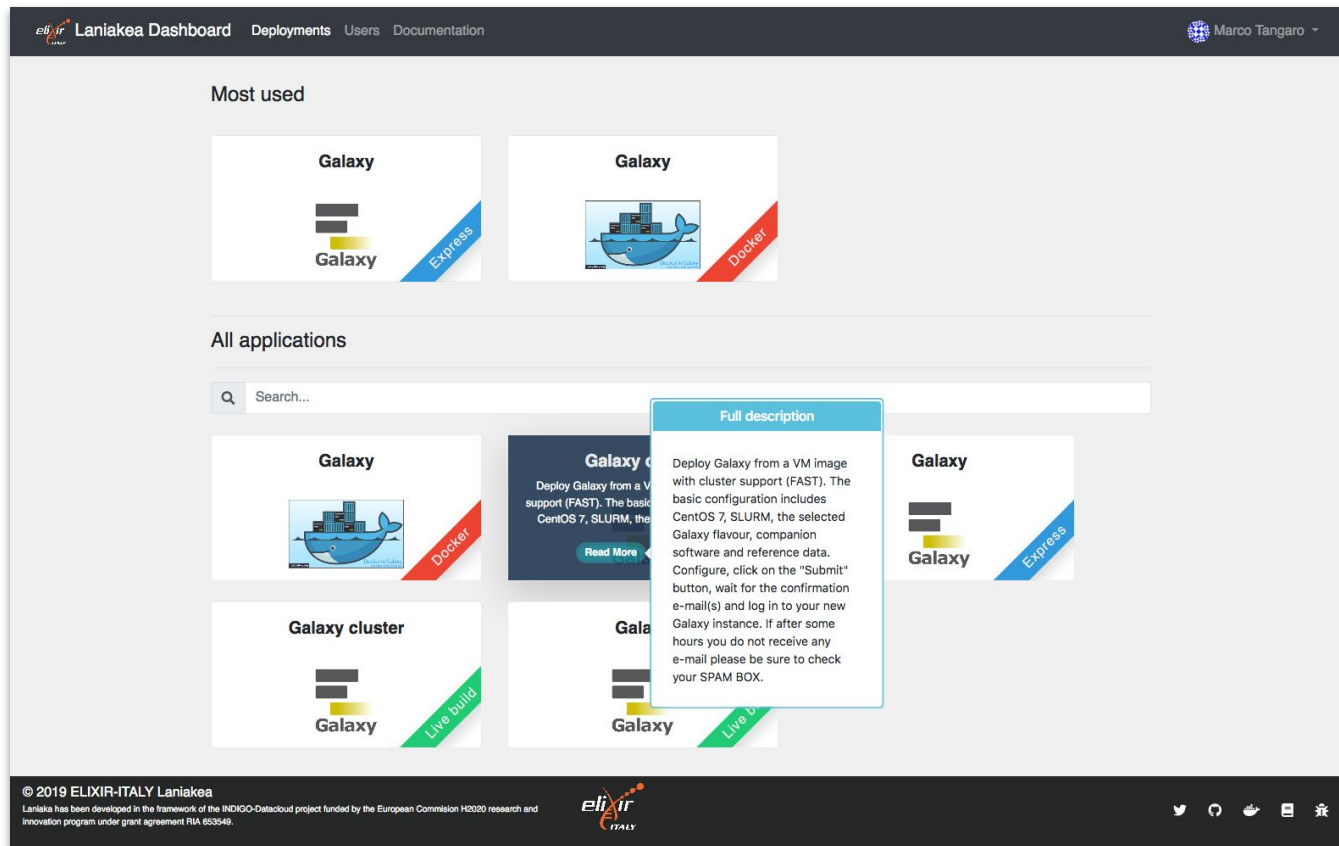


A screenshot of the Laniakea Dashboard home page. The page has a dark header with the "elixir Laniakea Dashboard" logo, navigation links for "Deployments", "Users", and "Documentation", and a user profile for "Marco Tangaro". The main content area is divided into two sections: "Most used" and "All applications". The "Most used" section contains two tiles for "Galaxy": one with a "Galaxy Express" badge and another with a "Galaxy Docker" badge. The "All applications" section has a search bar and a grid of application tiles. One tile for "Galaxy" is highlighted with a "Full description" tooltip that reads: "Deploy Galaxy from a VM image with cluster support (FAST). The basic configuration includes CentOS 7, SLURM, the selected Galaxy flavour, companion software and reference data. Configure, click on the 'Submit' button, wait for the confirmation e-mail(s) and log in to your new Galaxy instance. If after some hours you do not receive any e-mail please be sure to check your SPAM BOX." The footer contains copyright information for 2019 ELIXIR-ITALY Laniakea, a description of the project's funding, and social media icons.

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, ...



Different deployment strategies:

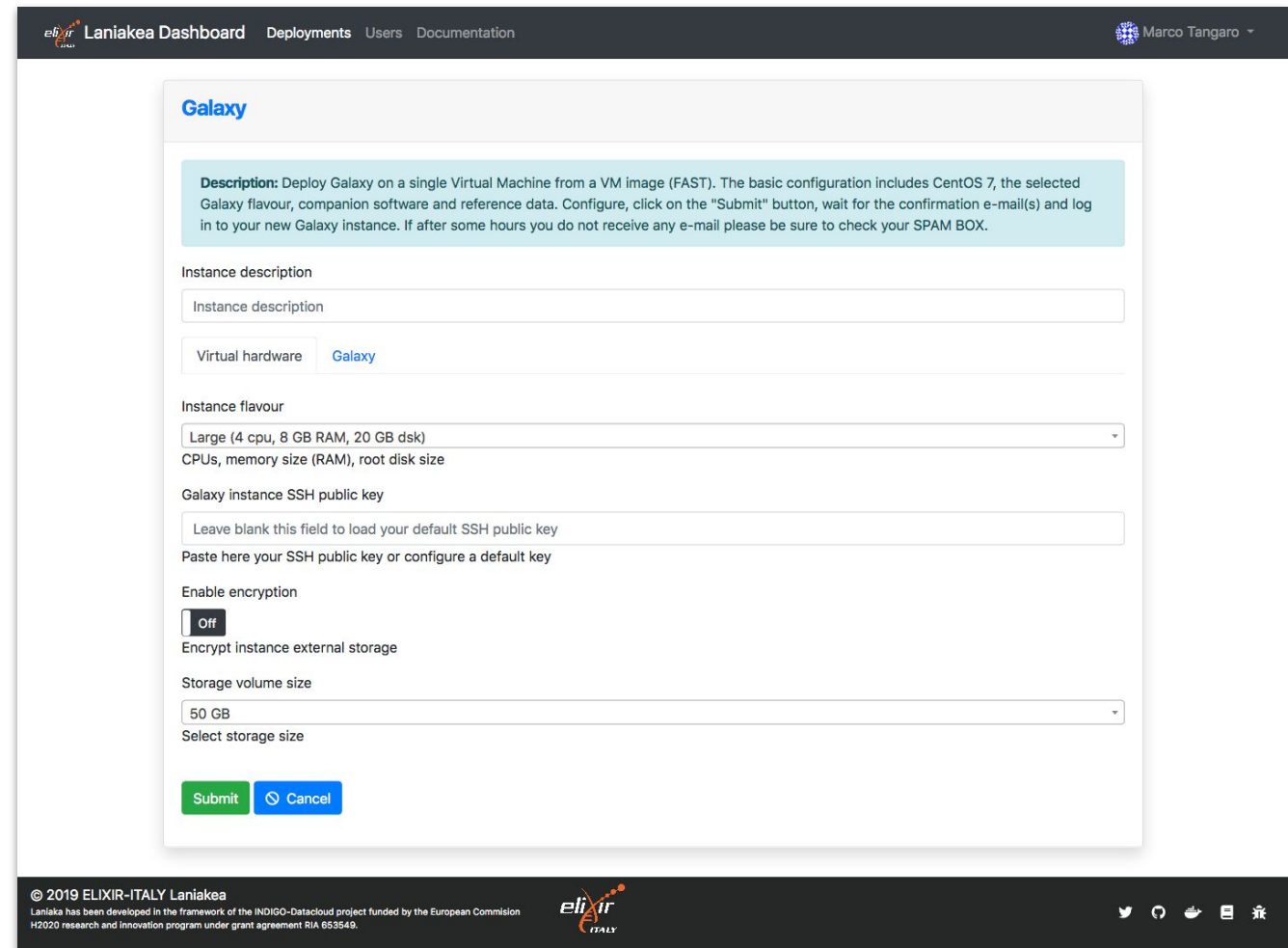
Live Build: build Galaxy from scratch -> always up-to-date (deployment time depending by the tools number).

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

Docker: fast deployment of new flavours.

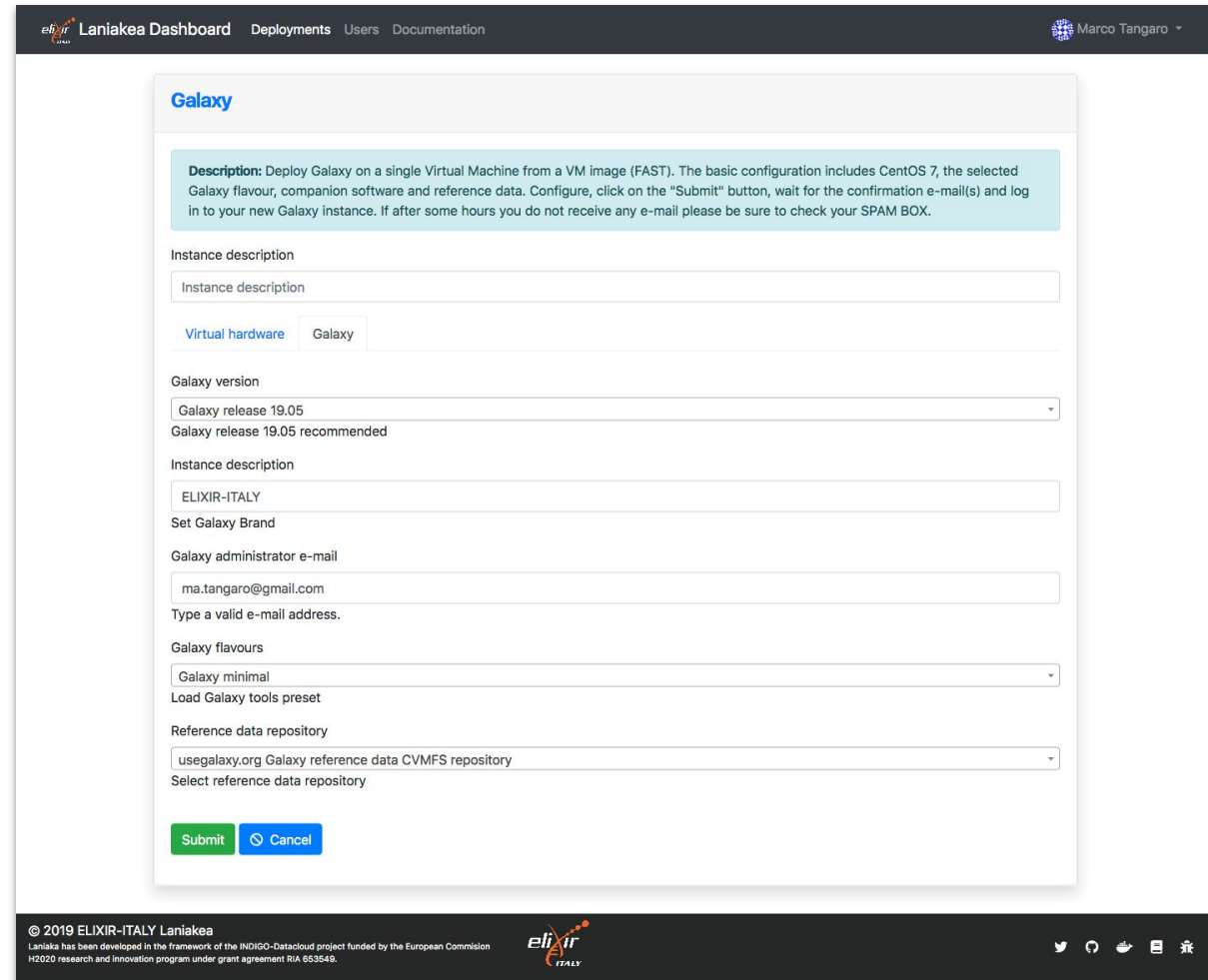
The web front-end provides different tabs to configure your Galaxy.

Virtual hardware: CPU, RAM and Storage

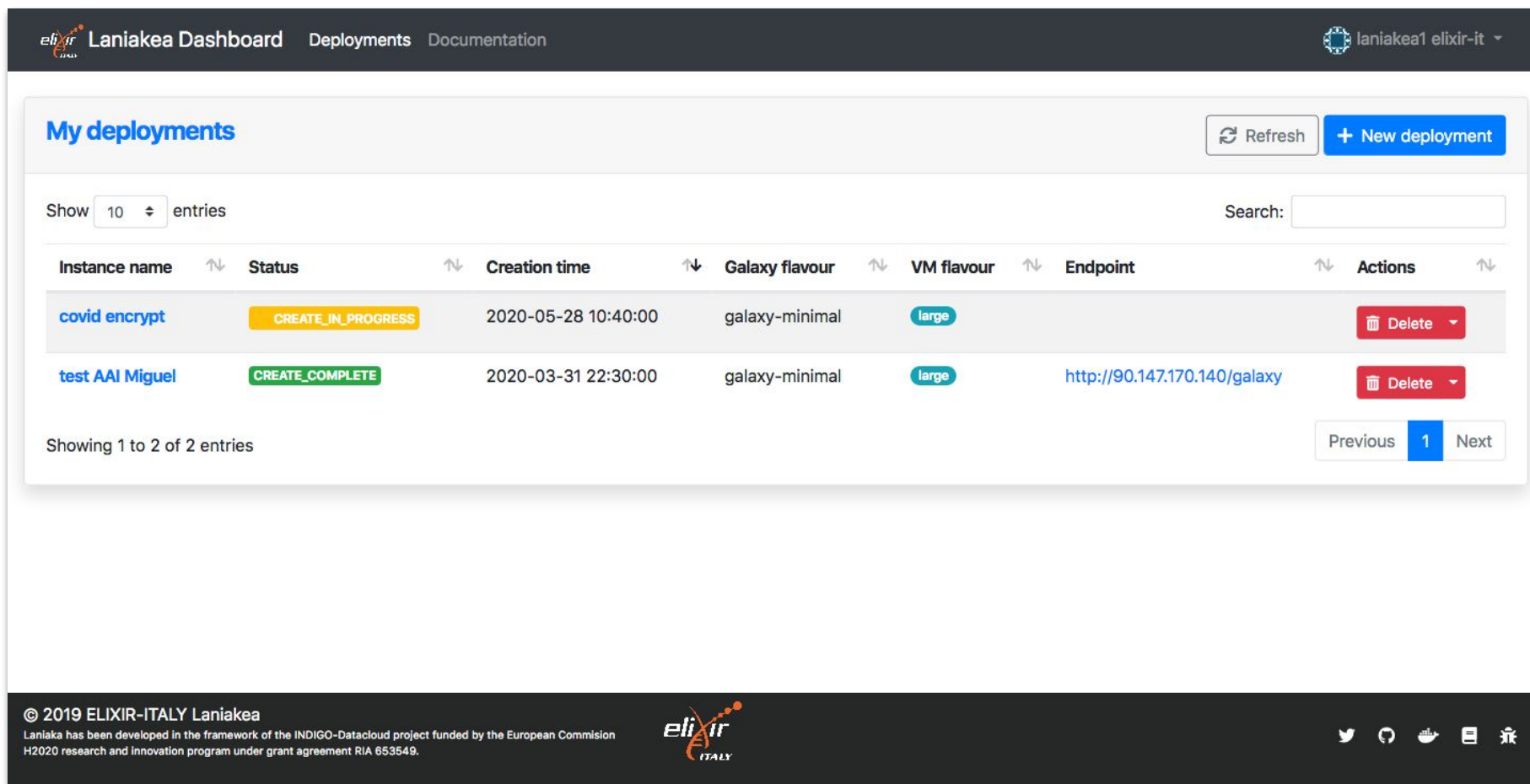
A screenshot of the Laniakea Galaxy configuration dashboard. The page has a dark header with navigation links: "Laniakea Dashboard", "Deployments", "Users", and "Documentation". A user profile "Marco Tangaro" is visible in the top right. The main content area is titled "Galaxy" and contains a description, an "Instance description" text field, tabs for "Virtual hardware" and "Galaxy", an "Instance flavour" dropdown menu set to "Large (4 cpu, 8 GB RAM, 20 GB dsk)", an "SSH public key" text field, an "Enable encryption" toggle set to "Off", and a "Storage volume size" dropdown menu set to "50 GB". At the bottom are "Submit" and "Cancel" buttons. The footer contains copyright information for ELIXIR-ITALY Laniakea and social media icons.

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

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



The screenshot shows the 'Galaxy' configuration page in the Laniakea Dashboard. The page has a dark header with navigation links: 'Laniakea Dashboard', 'Deployments', 'Users', and 'Documentation'. The user 'Marco Tangaro' is logged in. The main content area is titled 'Galaxy' and contains a description box with instructions on how to deploy Galaxy on a VM. Below this are several form fields: 'Instance description' (empty), 'Virtual hardware' (selected) and 'Galaxy' (unselected) tabs, 'Galaxy version' (dropdown menu showing 'Galaxy release 19.05' and 'Galaxy release 19.05 recommended'), 'Instance description' (text input with 'ELIXIR-ITALY'), 'Set Galaxy Brand' (empty), 'Galaxy administrator e-mail' (text input with 'ma.tangaro@gmail.com'), 'Galaxy flavours' (dropdown menu showing 'Galaxy minimal'), 'Load Galaxy tools preset' (empty), and 'Reference data repository' (dropdown menu showing 'usegalaxy.org Galaxy reference data CVMFS repository'). At the bottom are 'Submit' and 'Cancel' buttons. The footer contains copyright information for 2019 ELIXIR-ITALY Laniakea, a small ELIXIR ITALY logo, and social media icons.





My deployments Refresh + New deployment

Show 10 entries Search:

Instance name	Status	Creation time	Galaxy flavour	VM flavour	Endpoint	Actions
covid encrypt	CREATE_IN_PROGRESS	2020-05-28 10:40:00	galaxy-minimal	large		Delete
test AAI Miguel	CREATE_COMPLETE	2020-03-31 22:30:00	galaxy-minimal	large	http://90.147.170.140/galaxy	Delete

Showing 1 to 2 of 2 entries Previous 1 Next

© 2019 ELIXIR-ITALY Laniakea
Laniakea has been developed in the framework of the INDIGO-Datacloud project funded by the European Commission H2020 research and innovation program under grant agreement RIA 653549.

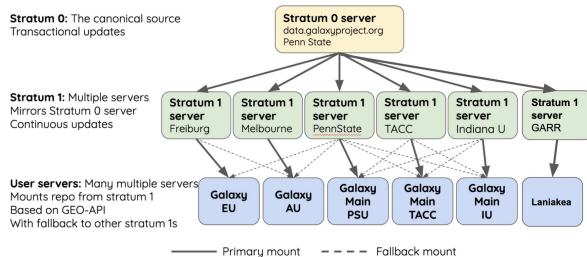
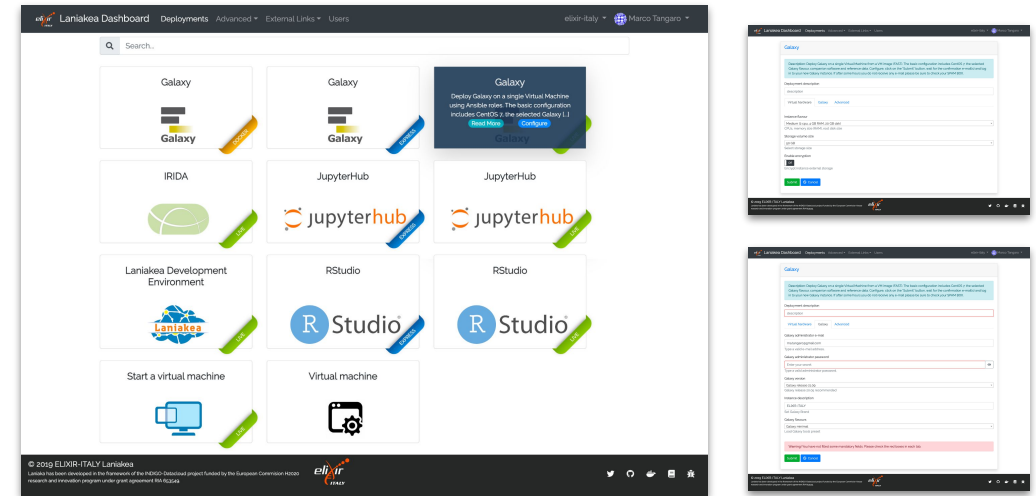


Laniakea main features

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.

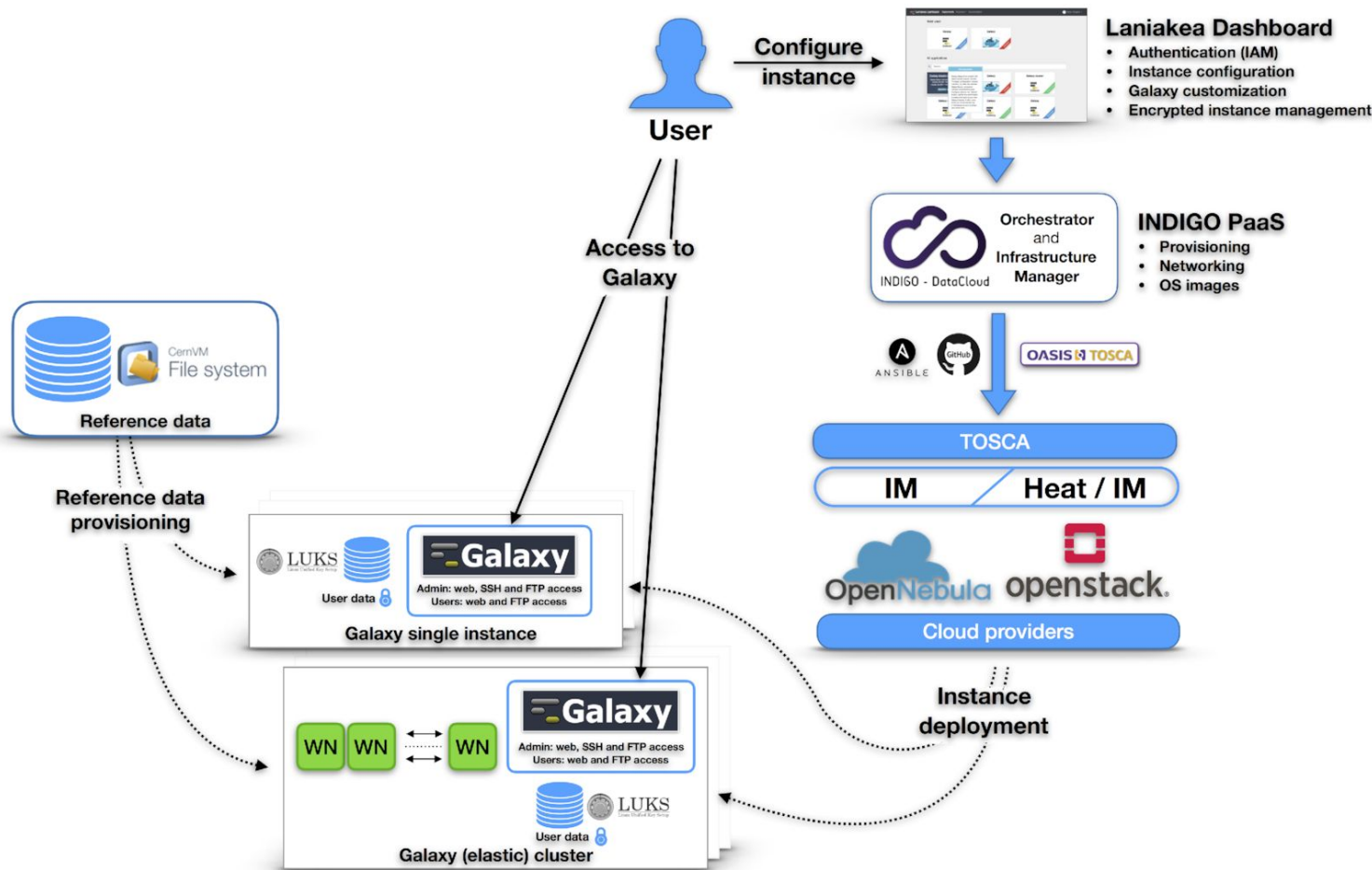


More Applications - No more limited to Galaxy. Jupyter Notebooks, RStudio and IRIDA available.

Environment with NextFlow, CWLtool and other development tools available.

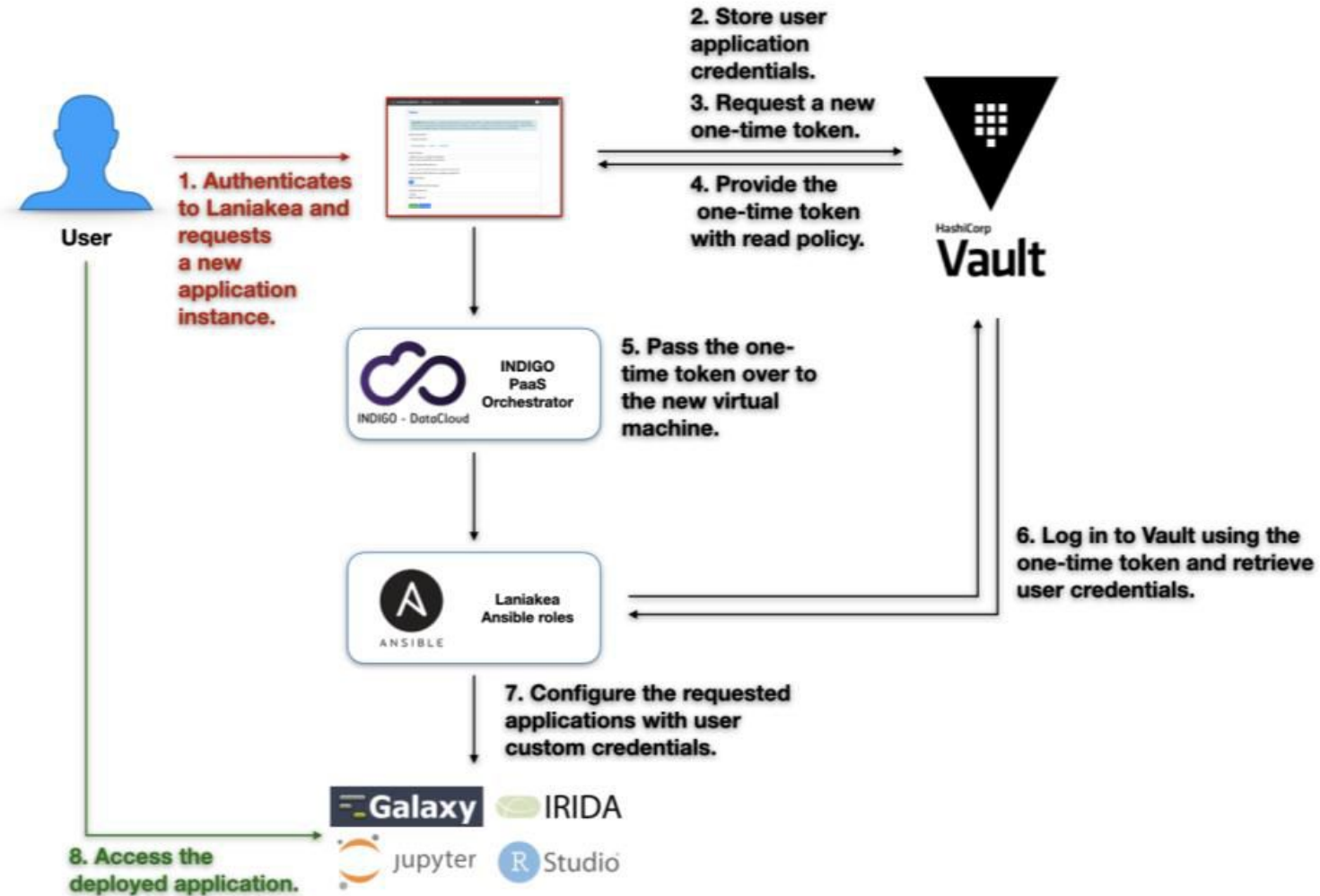


Laniakea architecture



- **Dashboard** - User friendly access to configuration and 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



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.

