WORKSHOP GARR 2021

NET MAKERS

Data storage, protection and management for the Life Science community

M. A. Tangaro, G.Donvito, F. Zambelli

CNR-IBIOM and INFN-Bari

IBIOM

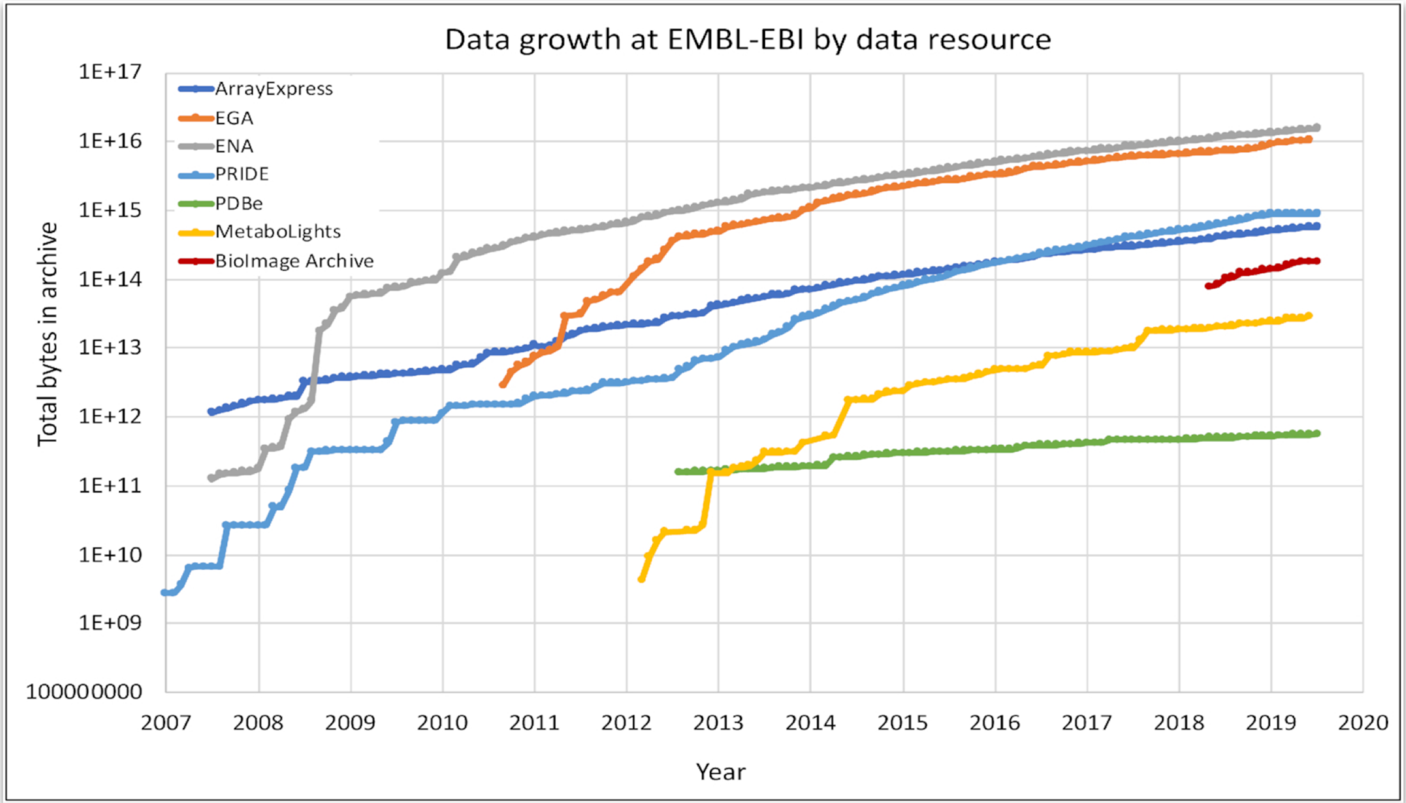
Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari

INFN

Data in Life Science

Data volume is doubling every six months.

Not only in quantity but also in variety.

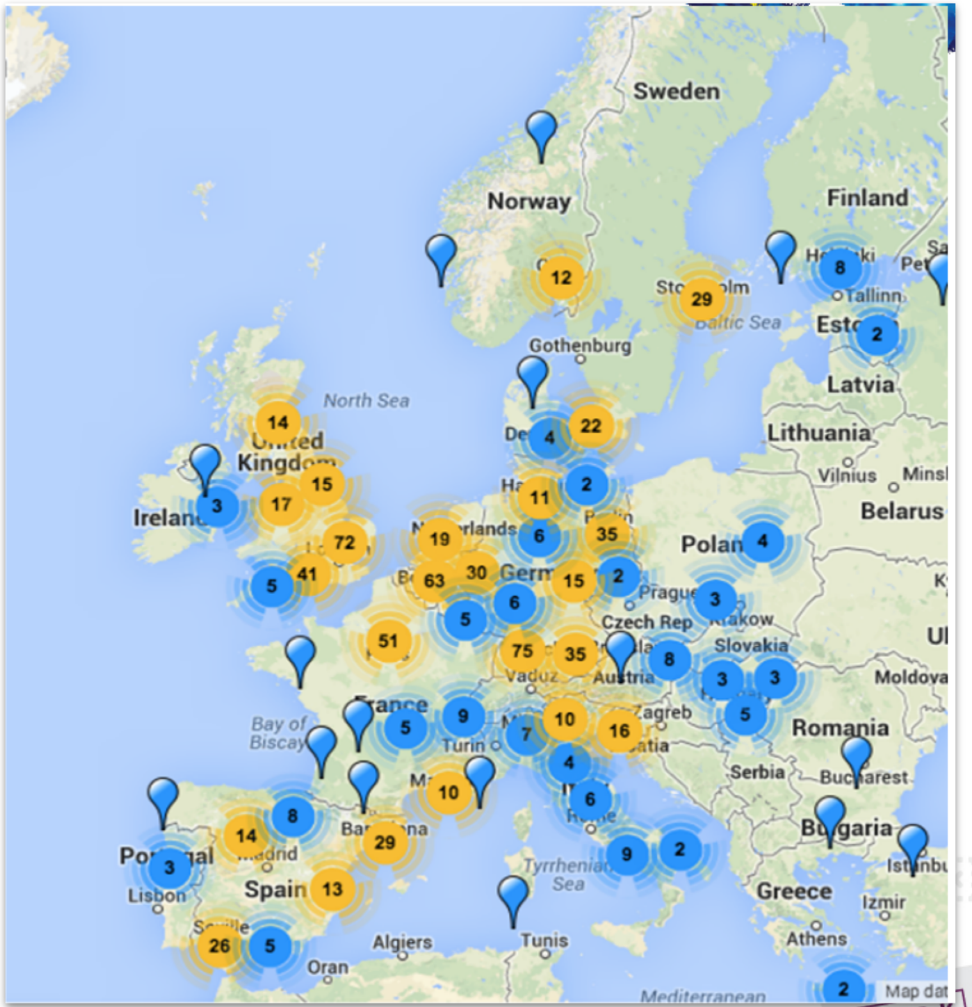


Data growth at EMBL-EBI Source: Charles E. Cook et al. Nucl. Acids Res. 2020; Volume 48, Issue D1, Pages D17-D23

Data in Life Science

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

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

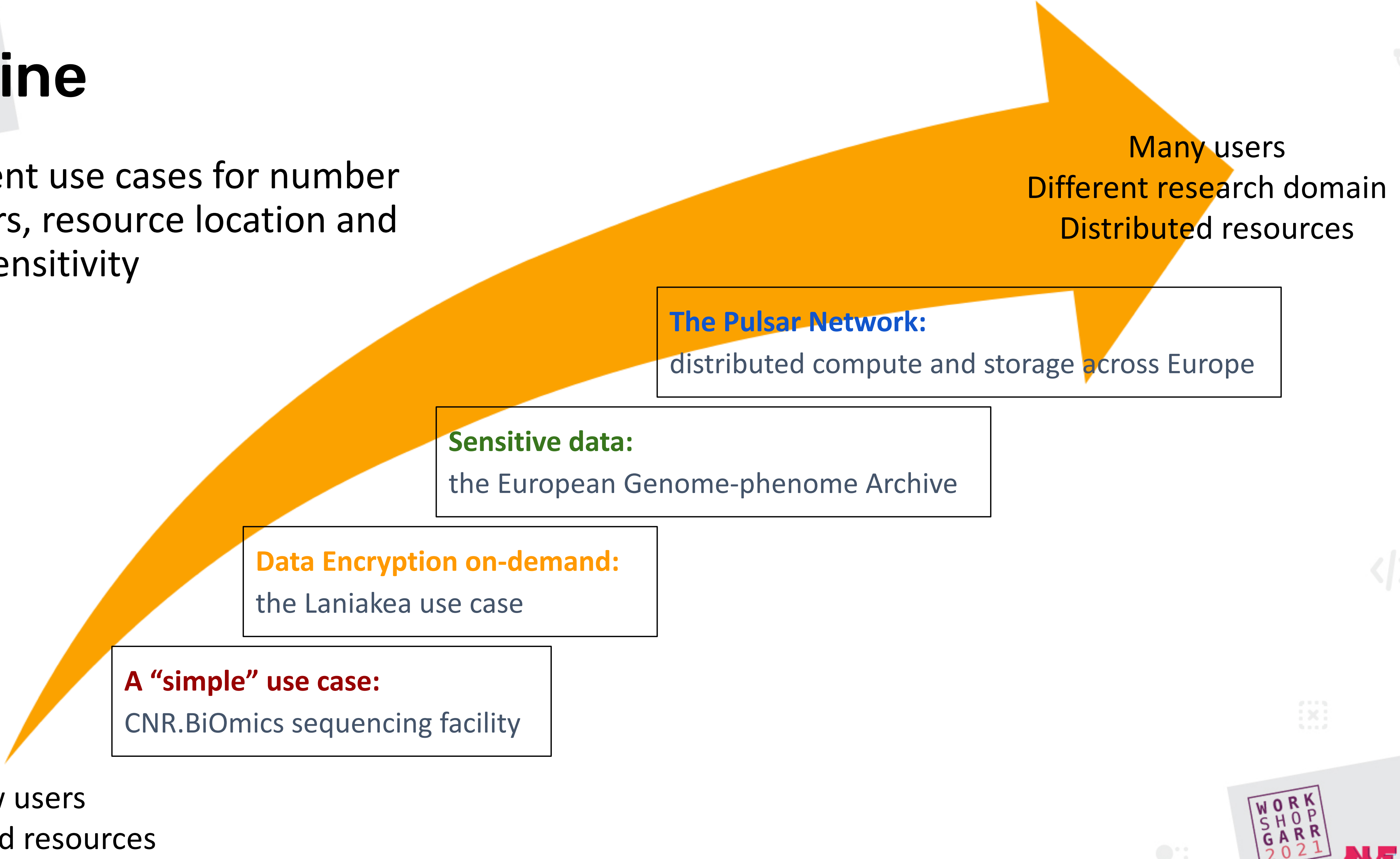


Data in Life Science

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)) without explicit consent, provided this is allowed by EU or Member States law framework and appropriate safeguards measures are in place.

Outline

Different use cases for number of users, resource location and data sensitivity



Many users
Different research domain
Distributed resources

The Pulsar Network:
distributed compute and storage across Europe

Sensitive data:
the European Genome-phenome Archive

Data Encryption on-demand:
the Laniakea use case

A "simple" use case:
CNR.BiOmics sequencing facility

Few users
dedicated resources

CNR.BiOmics

The CNR.BiOmics project ("National Research Center in Bioinformatics for Omics Sciences", PIR01_00017 14.5 M€ and CIR01_00017 2M€), currently on-going, which aims to strengthen the Italian node of the European Research Infrastructure ELIXIR in the southern regions.

10X Genomics

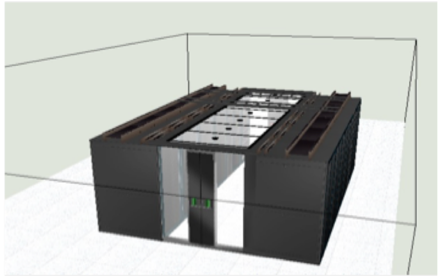
Illumina NovaSeq 6000

Oxford Nanopore GridION

PacBio Sequel

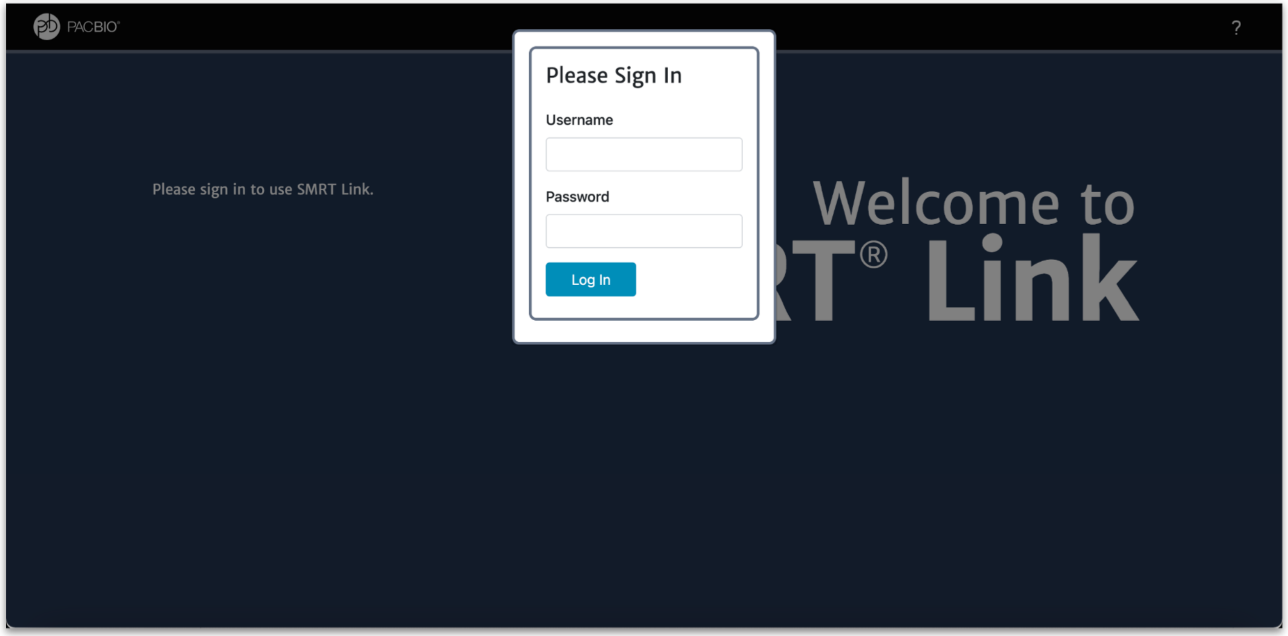
Orbitrap Fusion

BioNano Genomics



ICT Facility
10K core – 15 Pb

CNR.BiOmis



PacBio Sequel2:

- 2-3 TB output per sample
- 4 sample per run
- 1 run per week.

The sequencing facility is housed at IBIOM-CNR (Bari)

The analysis facility is housed at ReCaS-Bari datacenter.

Data are moved from the PacBio to the analysis server, where PacBio SMRT-Link have been configured for analysis.

70 TB Lustre storage are currently supporting the config.

Laniakea

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

LANIAKEA is a cloud based Galaxy instance provider.

Developed by ELIXIR-ITALY

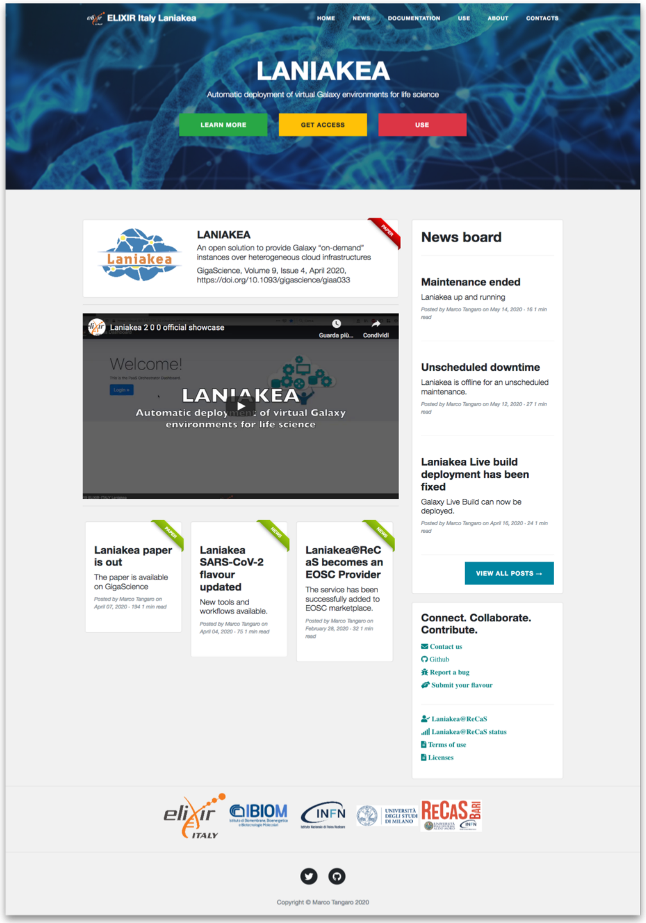
By hiding the technical complexity behind a user-friendly web front-end, Laniakea allows its users to configure and deploy virtual Galaxy instances with a handful of clicks.

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

No need for maintenance of the hardware and software infrastructure.

Laniakea paper

Laniakea@ReCaS paper



LANIAKEA

Automatic deployment of virtual Galaxy environments for life science

LANIAKEA

An open solution to provide Galaxy "on-demand" instances over heterogeneous cloud infrastructures

GigaScience, Volume 9, Issue 4, April 2020, https://doi.org/10.1093/gigascience/giaa033

LANIAKEA

Automatic deployment of virtual Galaxy environments for life science

News board

Maintenance ended

Laniakea up and running

Posted by Marco Tangaro on May 14, 2020 · 16 1 min read

Unscheduled downtime

Laniakea is offline for an unscheduled maintenance.

Posted by Marco Tangaro on May 12, 2020 · 27 1 min read

Laniakea Live build deployment has been fixed

Galaxy Live Build can now be deployed.

Posted by Marco Tangaro on April 16, 2020 · 24 1 min read

Laniakea paper is out

The paper is available on GigaScience

Posted by Marco Tangaro on April 07, 2020 · 194 1 min read

Laniakea SARS-CoV-2 flavour updated

New tools and workflows available.

Posted by Marco Tangaro on April 04, 2020 · 75 1 min read

Laniakea@ReCaS becomes an EOSC Provider

The service has been successfully added to EOSC marketplace.

Posted by Marco Tangaro on February 28, 2020 · 32 1 min read

Connect. Collaborate. Contribute.

- Contact us
- Github
- Report a bug
- Submit your flavour

- Laniakea@ReCaS
- Laniakea@ReCaS status
- Terms of use
- Licenses

Copyright © Marco Tangaro 2020

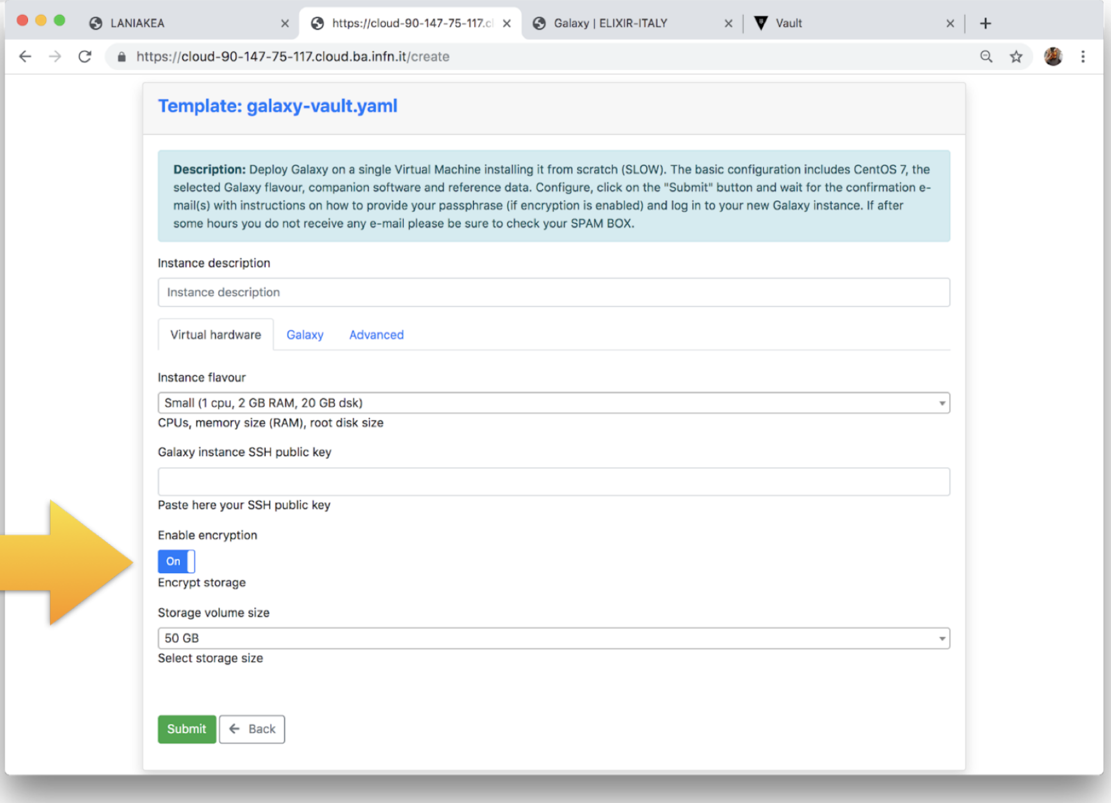
Laniakea – Storage encryption on-demand

The user data privacy is granted through LUKS storage encryption as a service: data are isolated from any other instance on the same platform and from the cloud service administrators.

The encryption procedure has been completely automated in order to simplify the user experience: the user can encrypt storage on-demand, using a strong random alphanumerical passphrase.

This has been achieved integrating 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.

On-demand encryption: user experience



https://cloud-90-147-75-117.cloud.ba.infn.it/create

Template: galaxy-vault.yaml

Description: Deploy Galaxy on a single Virtual Machine installing it from scratch (SLOW). The basic configuration includes CentOS 7, the selected Galaxy flavour, companion software and reference data. Configure, click on the "Submit" button and wait for the confirmation e-mail(s) with instructions on how to provide your passphrase (if encryption is enabled) 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.

Instance description

Instance description

Virtual hardware | Galaxy | Advanced

Instance flavour

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

CPUs, memory size (RAM), root disk size

Galaxy instance SSH public key

Paste here your SSH public key

Enable encryption

On

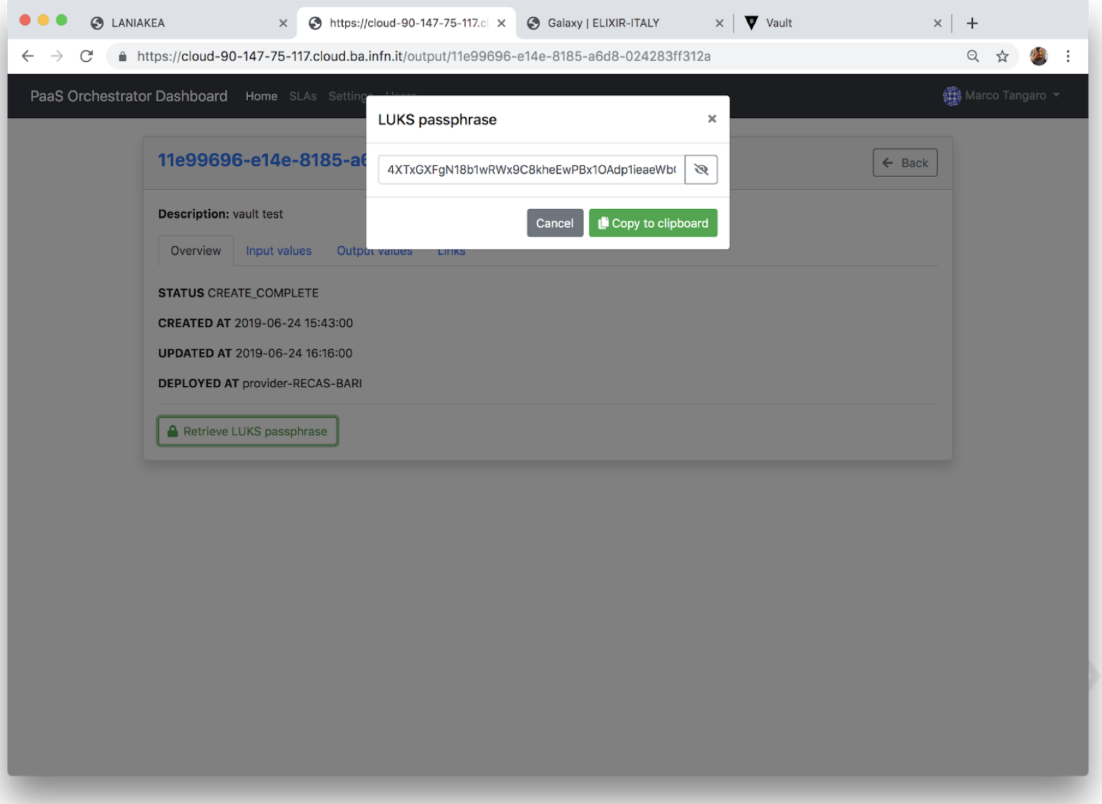
Encrypt storage

Storage volume size

50 GB

Select storage size

Submit | Back



https://cloud-90-147-75-117.cloud.ba.infn.it/output/11e99696-e14e-8185-a6d8-024283ff312a

LUKS passphrase

4XTxGXFgN18b1wRWx9C8kheEwPBx1OAdp1ieaeWb(

Cancel | Copy to clipboard

11e99696-e14e-8185-a6

Description: vault test

Overview | Input values | Output values | Links

STATUS CREATE_COMPLETE

CREATED AT 2019-06-24 15:43:00

UPDATED AT 2019-06-24 16:16:00

DEPLOYED AT provider-RECAS-BARI

Retrieve LUKS passphrase

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.

On-demand encryption: the infrastructure

A bash script (now pypi package) is used to encrypt the storage using a random passphrase and then store it on Hashicorp Vault.

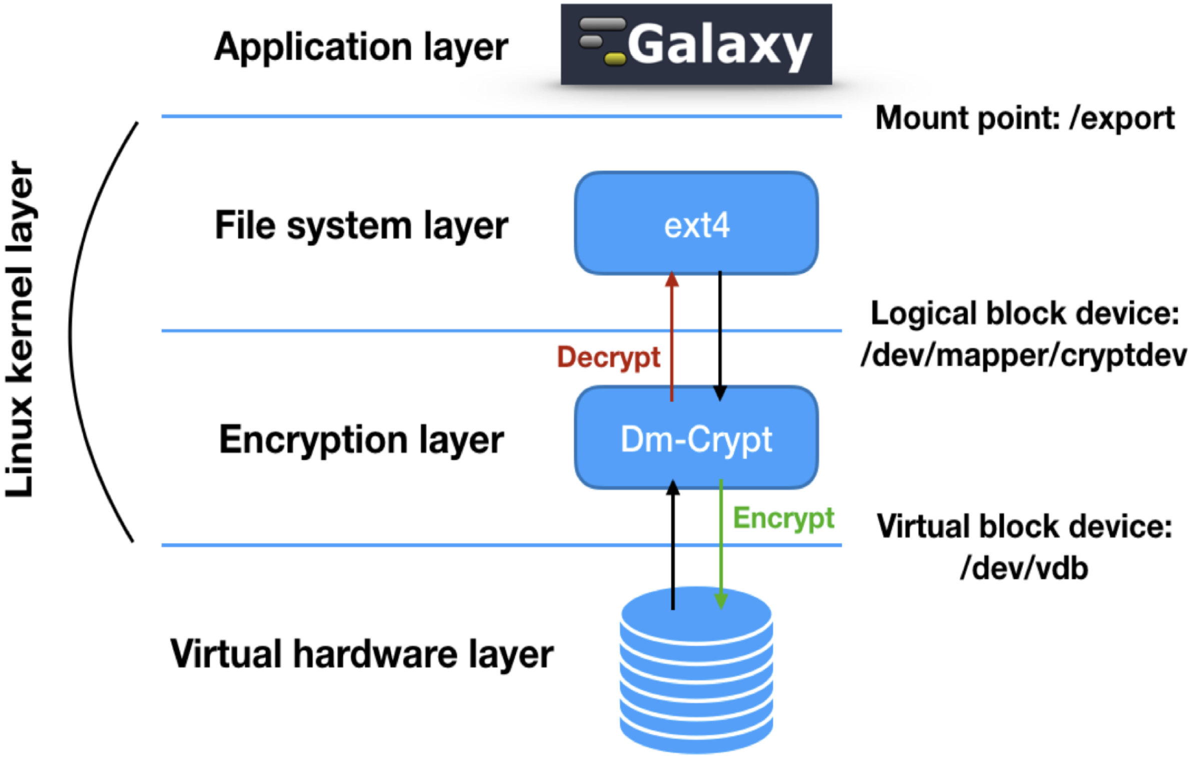
The encryption layer sits between the physical disk and the file system.

Galaxy 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:

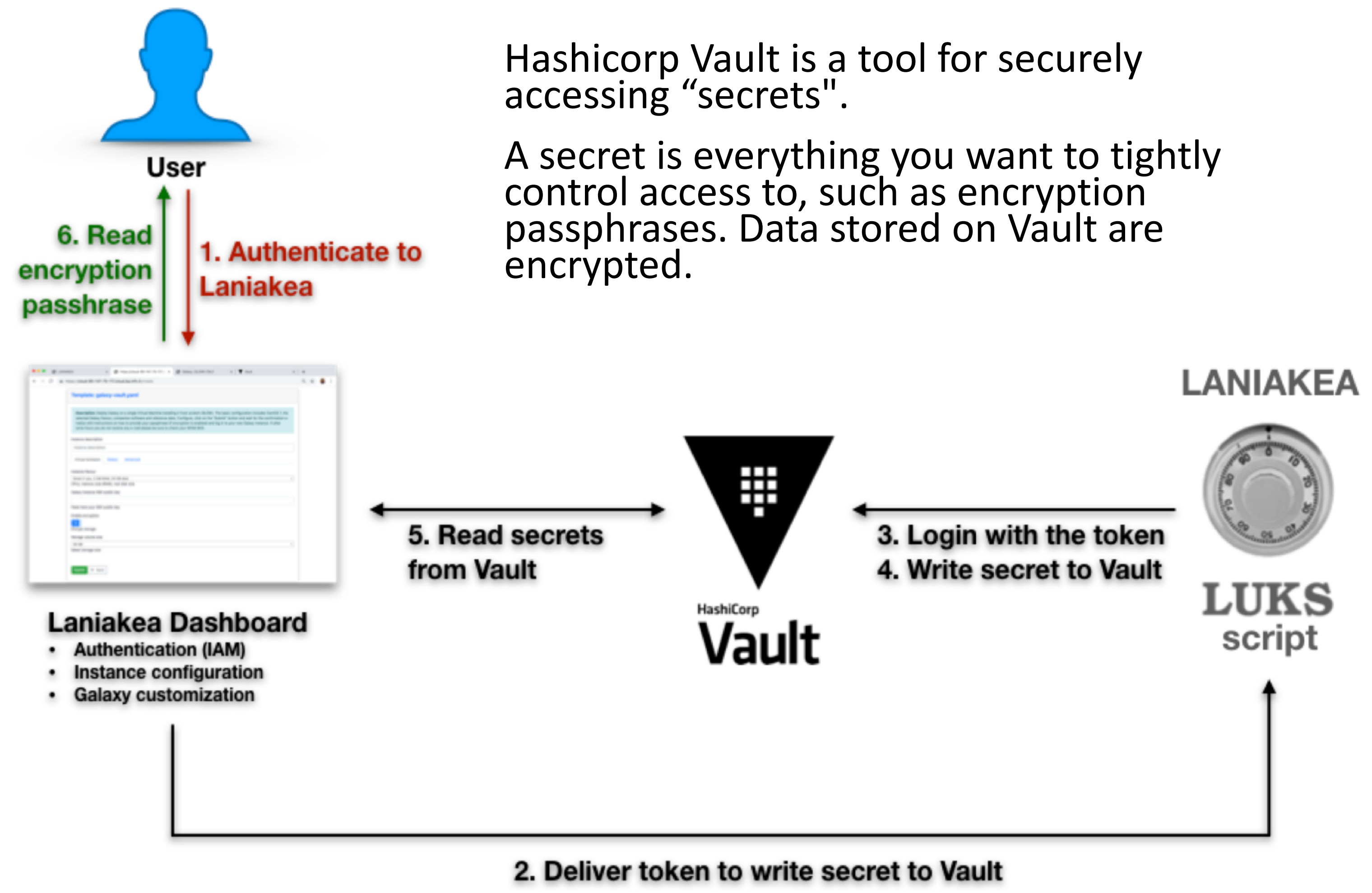
- aes-xts-plain64 encryption
- 256 bit key
- sha256 as hash algorithm used for key derivation.



On-demand encryption: the infrastructure

Hashicorp Vault is a tool for securely accessing “secrets".

A secret is everything you want to tightly control access to, such as encryption passphrases. Data stored on Vault are encrypted.



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 LUKS script.
4. The passphrase is sent to Vault by Laniakea LUKS script
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.

The European Genome-phenome Archive

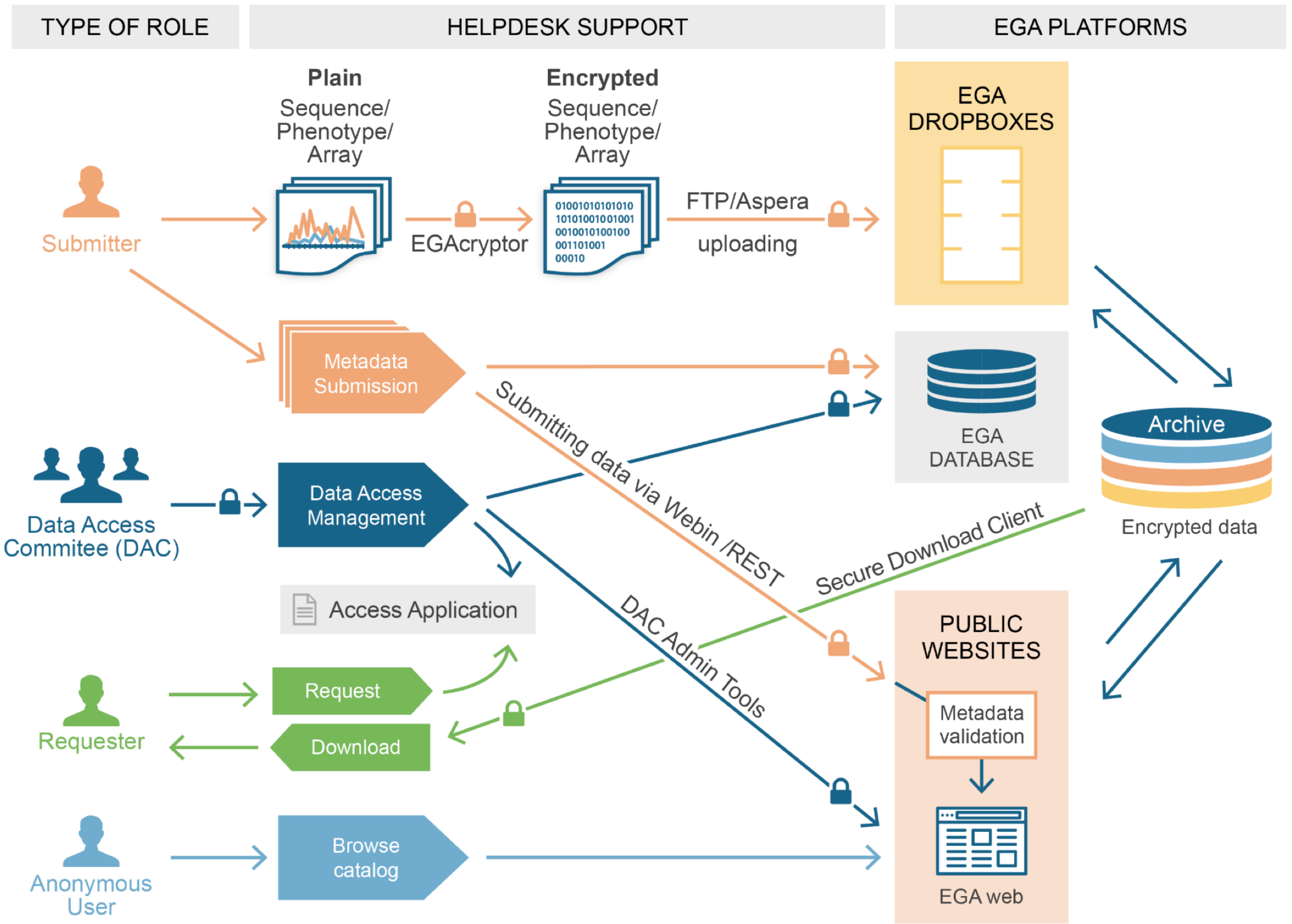
EUROPEAN GENOME-PHENOME ARCHIVE

The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects.

Data at EGA was collected from individuals whose consent agreements authorise data release only for specific research use to bona fide researchers.

Strict protocols govern how information is managed, stored and distributed by the EGA project.

The European Genome-phenome Archive



Studies and datasets can be browsed by anonymous users.

Data access committee is responsible for approving access to single of multiple datasets.

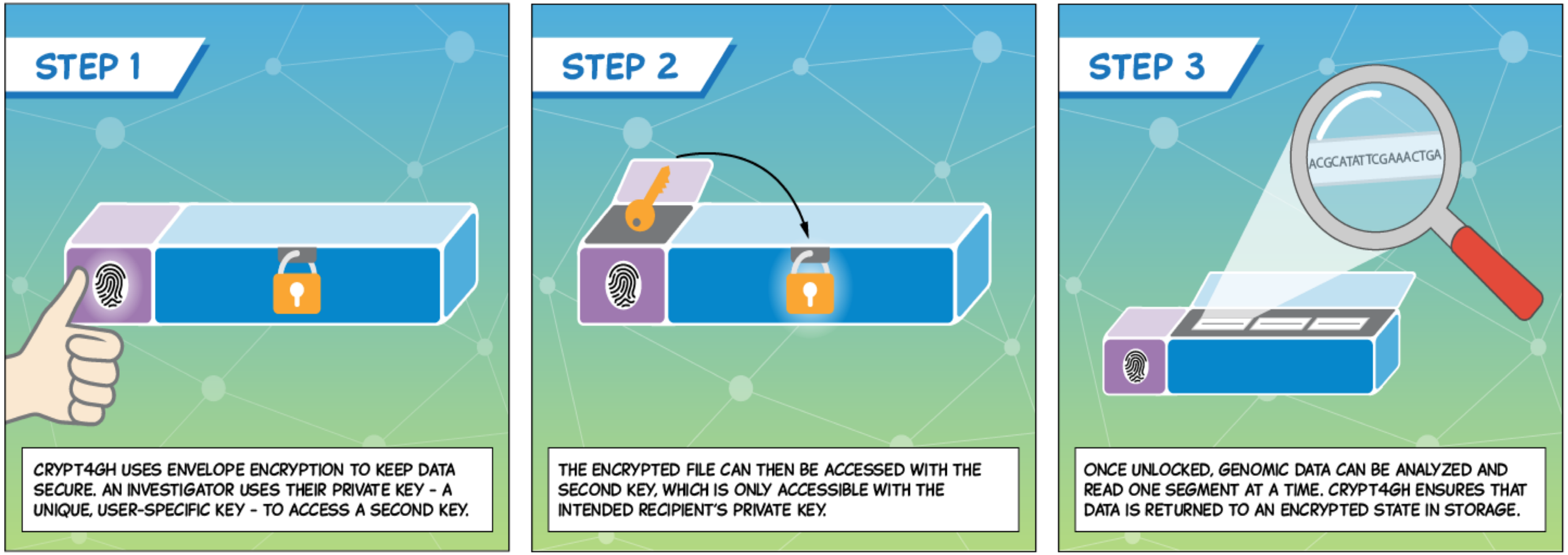
Data are encrypted. Trusted users exploit a user-specific key to decrypt data.

Local EGA

It aims at solving the issue where sensitive data cannot move across borders (cf to GDPR), while public metadata can. Files will be stored encrypted in the Local EGAs located in different countries, while public metadata stays at Central EGA.

1. Submitters upload encrypted files into a Local EGA inbox, located in the relevant country.
2. Encrypted files are moved from to long-term storage, and information are saved in Local EGA database.
3. In the process, each ingested file obtain an Accession ID, which identifies it uniquely across the EGA.
4. The distribution system allows requesters to access securely the encrypted files in the long-term storage, using the accession id, if permissions are granted by a Data Access Commitee (DAC).

The European Genome-phenome Archive



CRYPT4GH USES ENVELOPE ENCRYPTION TO KEEP DATA SECURE. AN INVESTIGATOR USES THEIR PRIVATE KEY - A UNIQUE, USER-SPECIFIC KEY - TO ACCESS A SECOND KEY.

THE ENCRYPTED FILE CAN THEN BE ACCESSED WITH THE SECOND KEY, WHICH IS ONLY ACCESSIBLE WITH THE INTENDED RECIPIENT'S PRIVATE KEY.

ONCE UNLOCKED, GENOMIC DATA CAN BE ANALYZED AND READ ONE SEGMENT AT A TIME. CRYPT4GH ENSURES THAT DATA IS RETURNED TO AN ENCRYPTED STATE IN STORAGE.

The GA4GH Crypt4GH format:

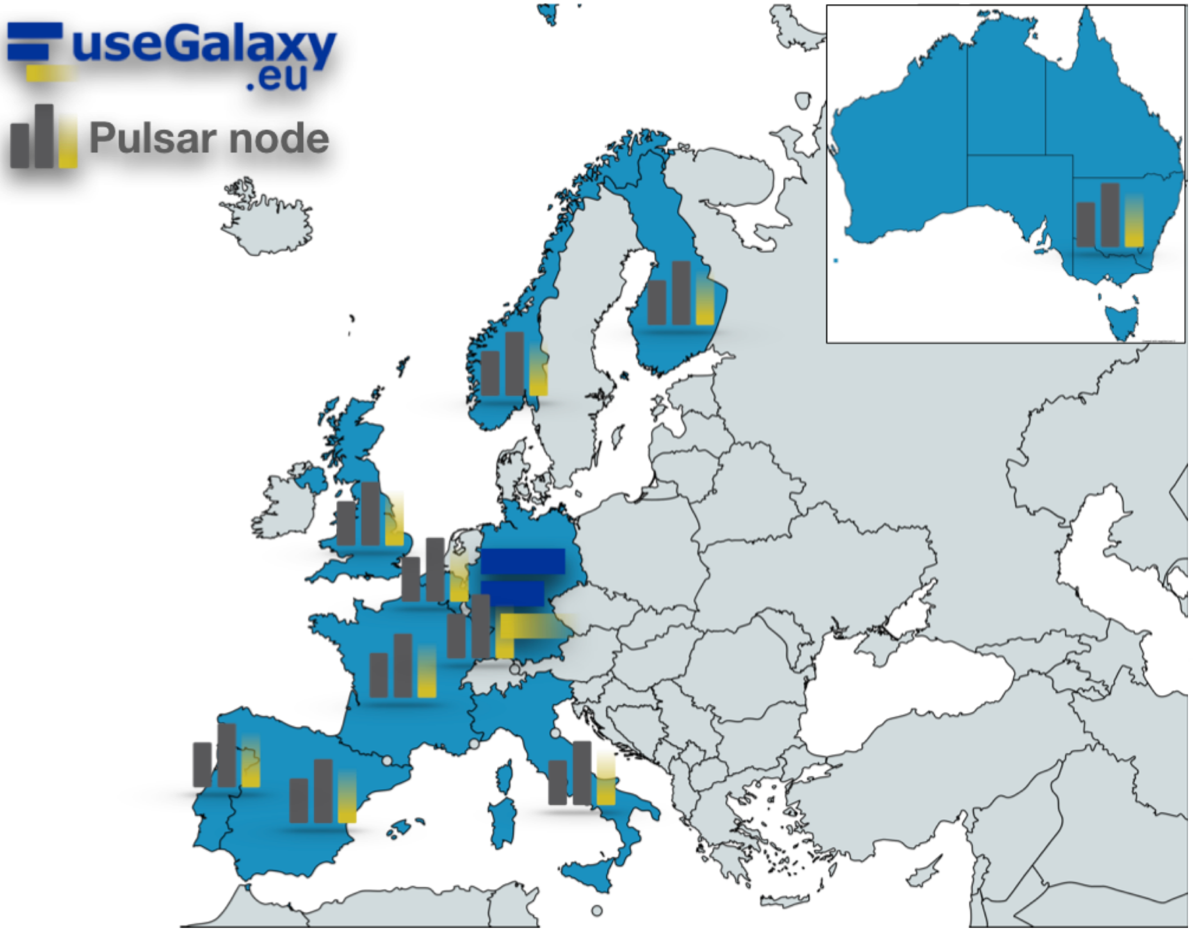
- No re-encryption upon ingestion (only decryption).
- Minimal re-encryption for data distribution.
- Shipping only selected segments for data distribution.

The European Pulsar Network

The most innovative computing centers across Europe are currently interested to share their remote computation power to support the UseGalaxy.eu load.

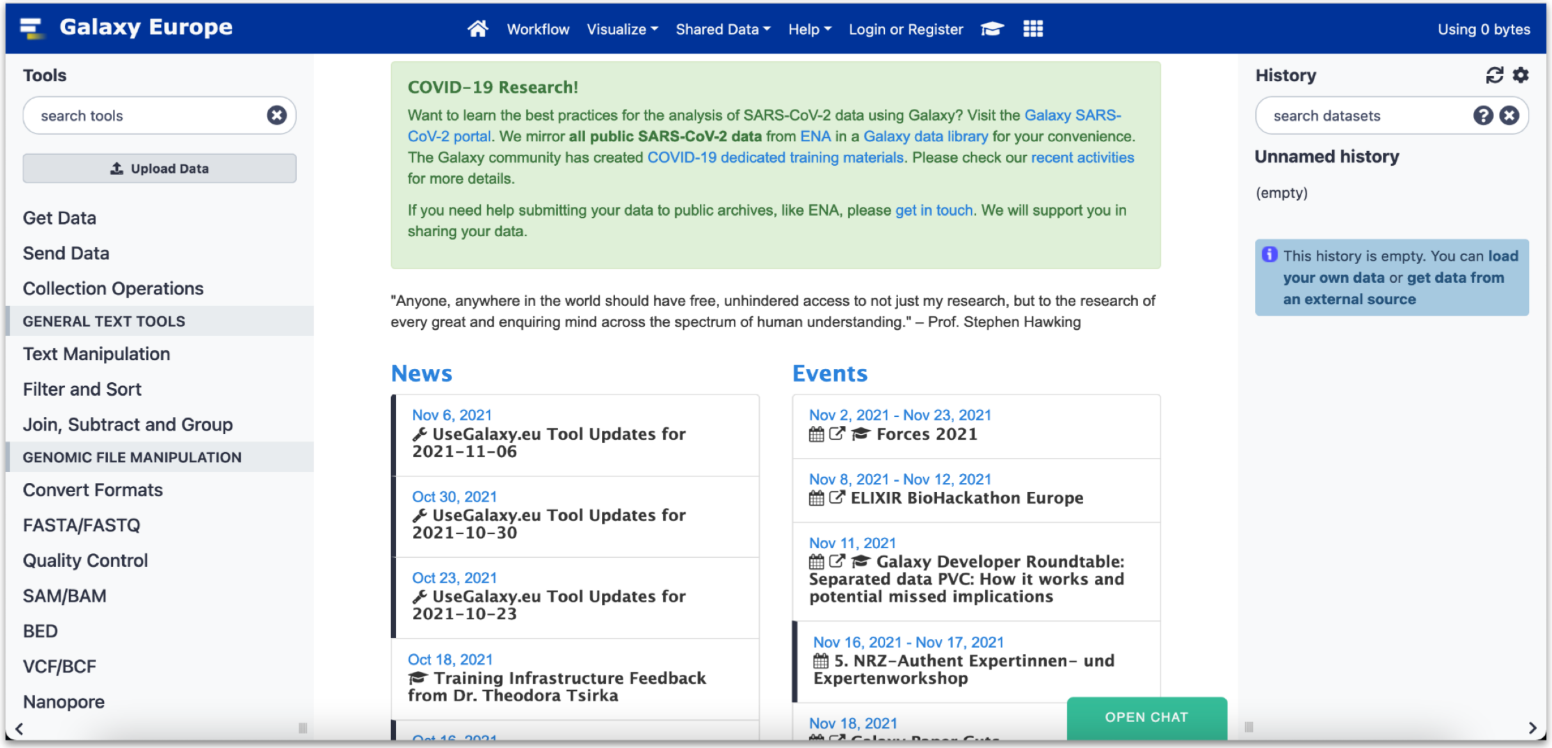
- DE, de.NBI cloud
- **IT, ReCaS-Bari and GARR Cloud**
- BE, Vlaams Supercomputer Centrum (VSC)
- PT, Tecnico ULisboa
- ES, Barcelona Supercomp. Center (INB-BSC)
- NO, University of Bergen
- CZ, CESNET
- FI, CSC
- UK, Diamond Light Source
- FR, GenOuest

https://pulsar-network.readthedocs.io/



The European Pulsar Network

The Galaxy front-end



This history is empty. You can load your own data or get data from an external source

COVID-19 Research!

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal. We mirror **all public SARS-CoV-2 data** from ENA in a Galaxy data library for your convenience. The Galaxy community has created COVID-19 dedicated training materials. Please check our recent activities for more details.

If you need help submitting your data to public archives, like ENA, please get in touch. We will support you in sharing your data.

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

News

Nov 6, 2021
UseGalaxy.eu Tool Updates for 2021-11-06

Oct 30, 2021
UseGalaxy.eu Tool Updates for 2021-10-30

Oct 23, 2021
UseGalaxy.eu Tool Updates for 2021-10-23

Oct 18, 2021
Training Infrastructure Feedback from Dr. Theodora Tsirka

Events

Nov 2, 2021 - Nov 23, 2021
Forces 2021

Nov 8, 2021 - Nov 12, 2021
ELIXIR BioHackathon Europe

Nov 11, 2021
Galaxy Developer Roundtable: Separated data PVC: How it works and potential missed implications

Nov 16, 2021 - Nov 17, 2021
5. NRZ-Authent Expertinnen- und Expertenworkshop

Nov 18, 2021

Galaxy is an **open**, web-based platform for **accessible**, **reproducible**, and **transparent** computational biomedical research.

The European Pulsar Network

Galaxy PROJECT

Pulsar allows a Galaxy server to automatically interact with remote systems, ensuring job and provenance information are correctly exchanged.

- Compute: submits and manages jobs.
- Data: stages user data.

Data needs to be moved from Galaxy to the remote compute nodes, impacting negatively on the job’s execution time in case of large amounts of data.

Provides reference data and Galaxy dependencies (read only).

WORKSHOP GARR 2021 NET MAKERS

The Cern-VM FileSystem

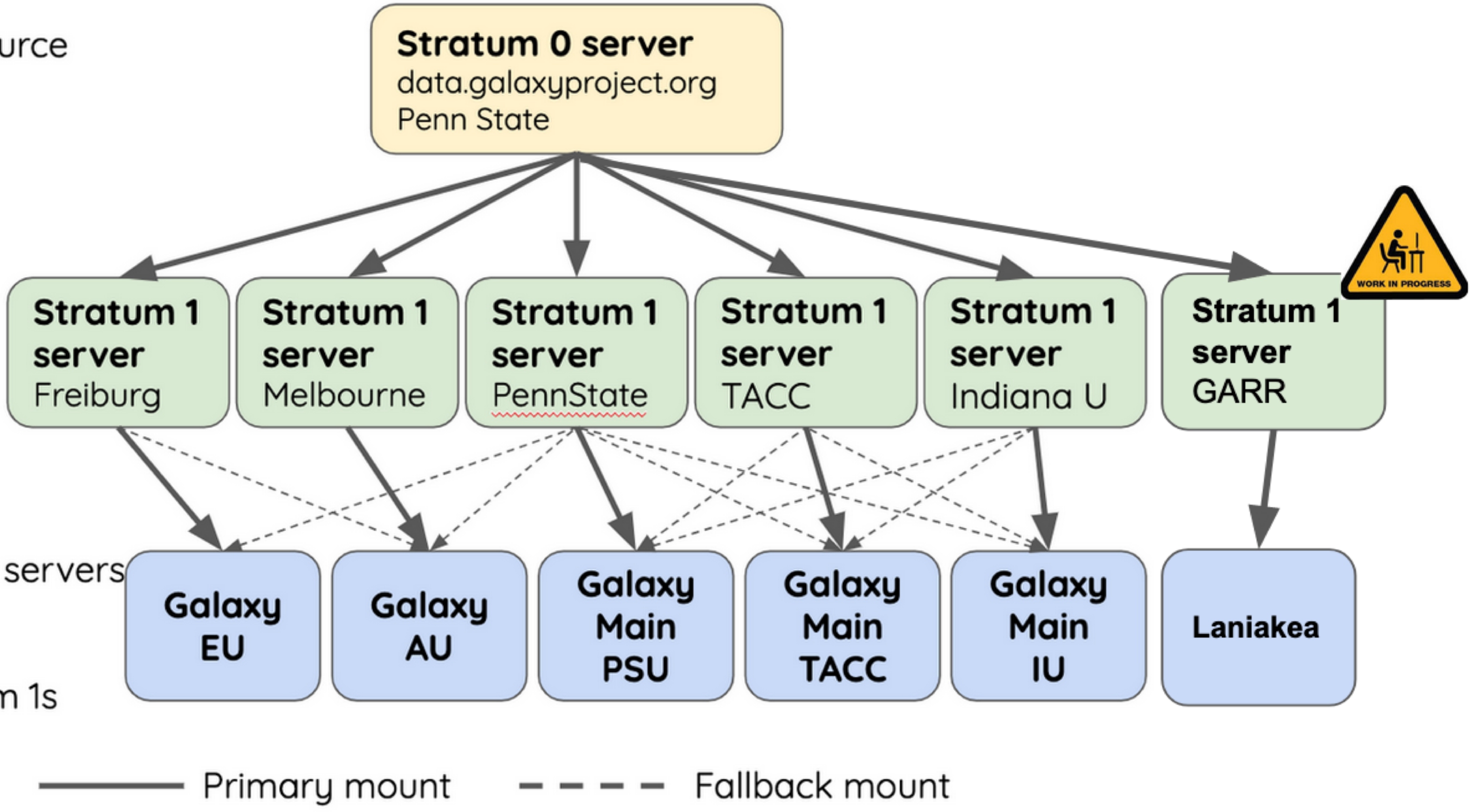
- Needed a method of sharing reference data, e.g. genomic sequences, across countries efficiently
- **CVMFS** is an efficient method for read only data sharing between systems
  - Originally designed for distributed software installation at CERN
  - Turns out it's really useful for read only data sets as well
  - HTTP-based, firewall friendly
- All nodes of Galaxy Main get their reference genomes and indices from CVMFS
  - Shared via mirroring and caching across the country
- It's also really useful to share data **globally**
  - The **usegalaxy.*** and **Laniakea** initiatives has taken full advantage of this.

The Cern-VM FileSystem

Stratum 0: The canonical source
Transactional updates

Stratum 1: Multiple servers
Mirrors Stratum 0 server
Continuous updates

User servers: Many multiple servers
Mounts repo from stratum 1
Based on GEO-API
With fallback to other stratum 1s



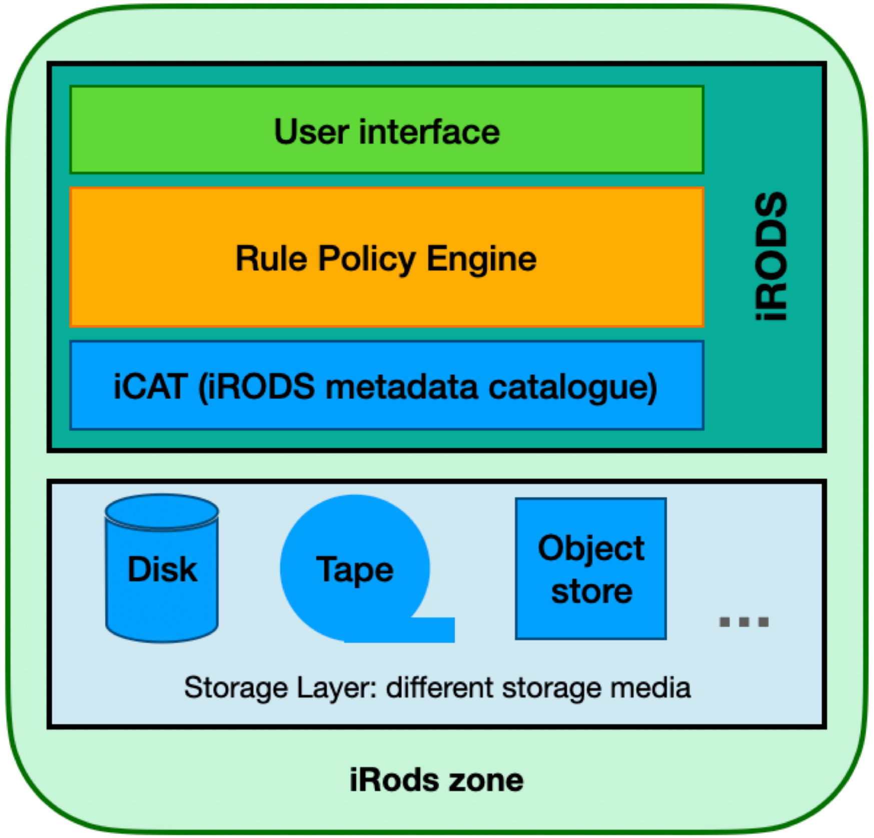
The Pulsar Network: optimizing resource usage

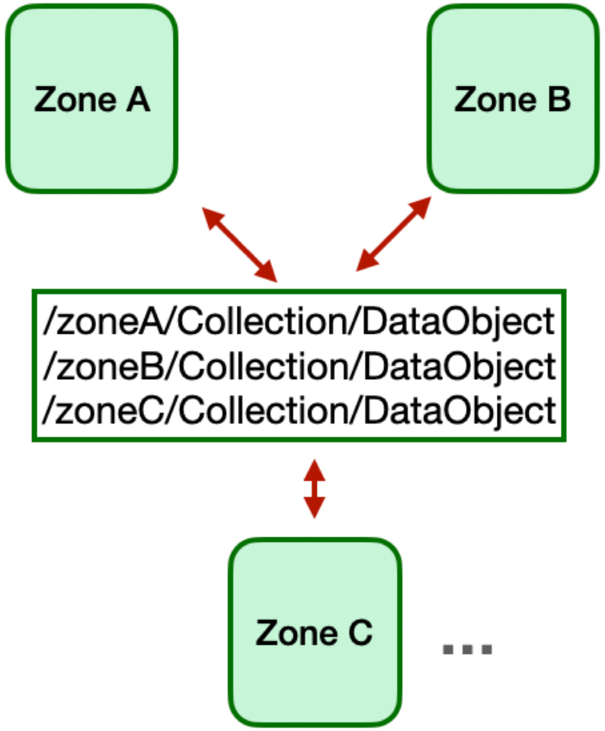
⚠

This work is still in early stage development. There is an ELIXIR Implementation plan and other initiatives at EOSC level working on this.

We are still evaluating technologies to support the Pulsar Network and also funding opportunities.

The Pulsar Network: optimizing resource usage




Zone A
Zone B
/zoneA/Collection/DataObject
/zoneB/Collection/DataObject
/zoneC/Collection/DataObject
Zone C
...

- Independent Administrative domains
- Users can access shared data and metadata in other zones
- No passwords needed, authenticated to local zone

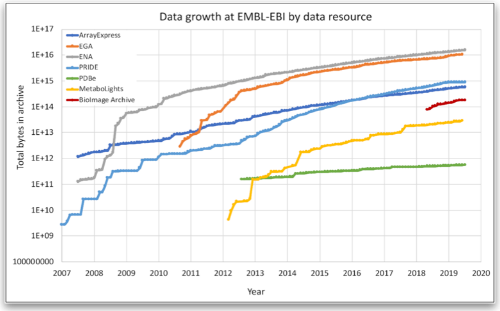
Distributed file systems like iRODS (or similar solution) are keeping track of objects (files, metadata, etc.) and their locations.

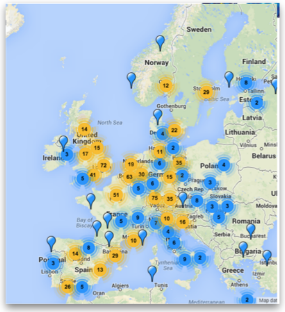
The Pulsar Network: optimizing resource usage

- The job-staging is dependent on the specific tool and analysis and can become the dominating part of the job runtime.
- A network-wide caching layer can accelerate job-staging times and thus job runtimes.
- Objects can be located in different physical locations and can be replicated into other locations before a job starts.
- Adding iRODS or similar solution support would allow replicating data wherever it's needed on the filesystem level without changing logic at the application level. However, this would require that all partners support a single distributed file system, which in turn would make it harder to contribute to the network.
- In case files are already available in a Pulsar destination, the staging step will take this into account.

Conclusions

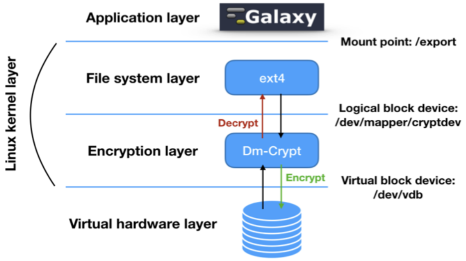
Data rapid growth, distributed across Europe and often under GDPR.





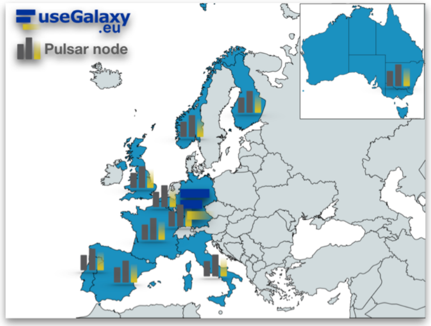
On-demand solution for rapidly store and analyse data, also in case of sensitive data.

Laniakea



Distributed Storage (and compute) solution across Europe.

EUROPEAN GENOME-PHENOME ARCHIVE



Thank you for your attention!

Contacts

- ELIXIR: elixir-europe.org
- ELIXIR Italy: elixir-Italy.og
- Head of Node: g.pesole@ibiom.cnr.it
- Technical Coordinator: federico.zambelli@unimi.it
- Compute platform : giacinto.donvito@ba.infn.it

Backup

ELIXIR Italy compute platform upgrade

- A multi-institution, distributed and federated compute and storage infrastructure
- Providing most advanced technologies in the fields of Cloud and HPC
  - Able to support the most modern Artificial Intelligence and Big Data analysis solutions
- Flexible and expandable in the future aiming to support future bioinformatics use cases
- The overall infrastructure will leverage about:
  - 27’000 Cpu/cores
  - About 20Pyte of disk storage
  - 20 NVIDIA V100 GPU
- Distributed over 4 different sites in Italy