Workshop sul Calcolo nell'INFN: Paestum, 23 - 27 maggio 2022

Ambienti virtuali di analisi dati in cloud creati on demand compliant con i requisiti tecnici e legali per l'analisi di dati genetici e sanitari

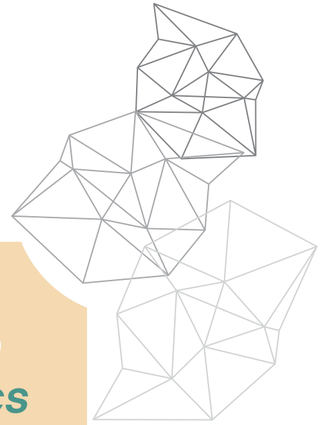
M. Antonacci (*), G. Donvito (*), N. Foggetti (*), M. Tangaro (**)

(*) Istituto Nazionale di Fisica Nuclare - Sezione di Bari
(**) CNR-IBIOM

EOSC-Pillar

Coordination and Harmonisation of National &Thematic Initiatives to support EOSC

Il progetto è finalizzato a coordinare e armonizzare le diverse iniziative nazionali relative alla EOSC in vari paesi come Austria, Belgio, Francia, Germania e Italia.



Use Case 6 - Exploring reference data through existing computing services for the bioinformatics community

Main challenges:

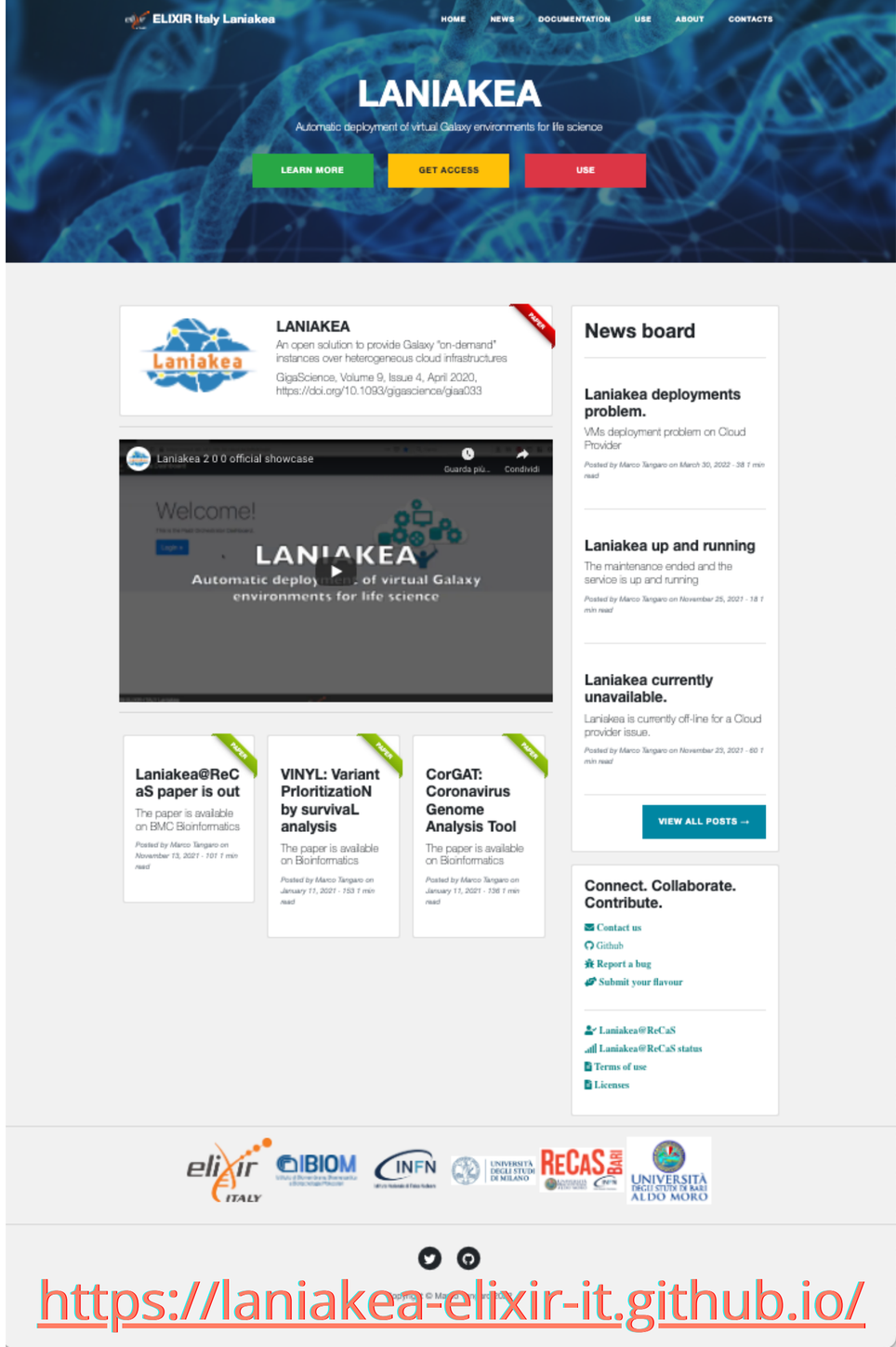
- **ensure the reproducibility and coherency of the data analysis (performed on different platforms)**
- **conform to data protection regulations concerning health personal data**

Laniakea

Galaxy as a Service

Galaxy è un workflow manager adottato in molti ambienti di ricerca nel campo delle scienze biologiche al fine di facilitare l'interazione con gli strumenti bio-informatici e la gestione di grandi quantità di dati biologici.

Laniakea è un framework software per la creazione di istanze Galaxy on-demand su infrastrutture cloud federate.



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

Analizzare la conformità normativa del servizio Laniakea, per le comunità ELIXIR e Life Science, in caso di dati clinici e sensibili

Legal framework

Secure infrastructure

Data Encryption

Aspetti legali ed etici

Legal Framework for the use and re-use of health data for scientific purposes.

N. Foggetti, G. Donvito, M.A. Tangaro

10.5281/zenodo.6334878

SCENARIO:
GARANTIRE CHE I REQUISITI DI SICUREZZA DEI DATI SANITARI SIANO SODDISFATTI DURANTE TUTTO IL PROCESSO

STUDIO DELLO SCENARIO E DEI GAP GIURIDICI ED ETICI (ES. PSEUDONIMIZZAZIONE)

Lo scopo dello studio è stato quello di gettare le basi per la stesura di blueprint/linee guida per ricercatori e istituzioni europee, che possano assisterli nella pubblicazione, condivisione e integrazione dei dati della ricerca.

DEFINIZIONE DELLA CHECKLIST PER LO SCENARIO

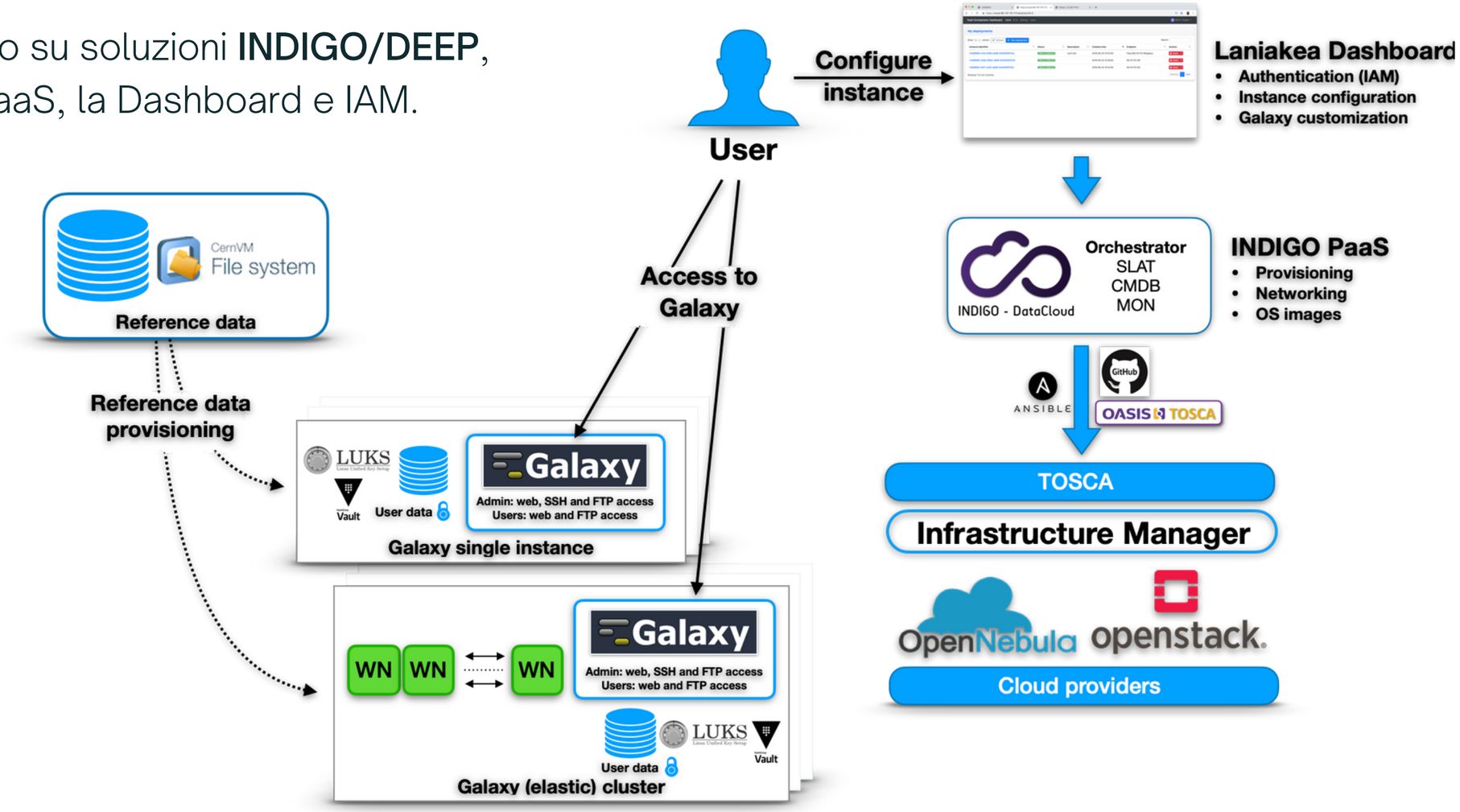
Le linee guida sono state tradotte in checklist pratiche che aiutino a garantire la conformità alla normativa sulla protezione dei dati in materia di dati personali sanitari.

APPLICAZIONE DEI PRINCIPI OS E OA, PRINCIPI FAIR AL TRATTAMENTO DEI DATI SANITARI ALL'INTERNO DELLO SCENARIO

E' fondamentale definire i requisiti etici e giuridici che devono essere rispettati al fine di garantire un adeguato bilanciamento tra tutela dei dati e della vita privata e l'effettiva applicazione dei principi FAIR, OS e OA.

Laniakea: Architettura

Il sistema è basato su soluzioni **INDIGO/DEEP**, in particolare la PaaS, la Dashboard e IAM.

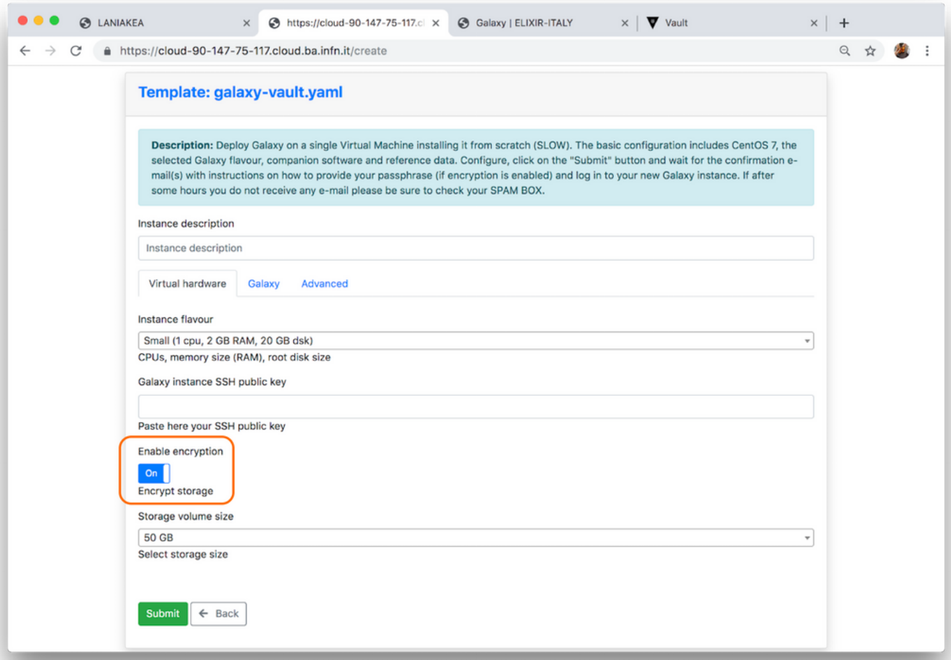


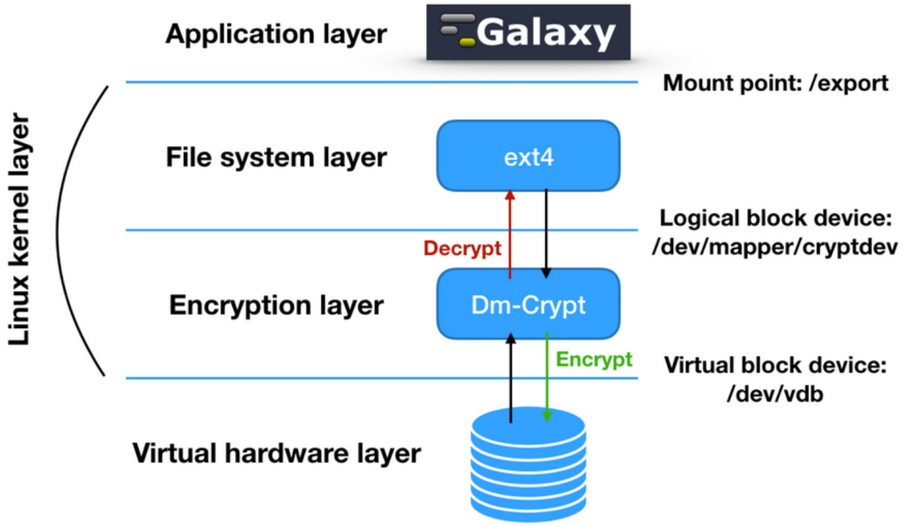
Storage Encryption

Protezione dei dati

"as a Service"

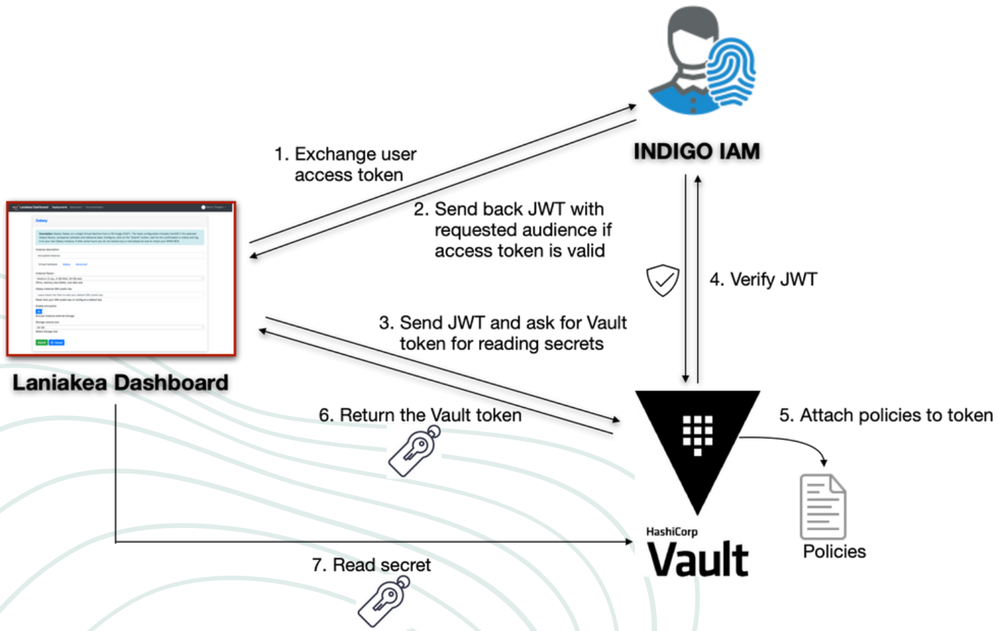
- Laniakea implementa una **procedura di cifratura del disco** su cui vengono memorizzati i dati gestiti da Galaxy **completamente automatizzata**
- Dal punto di vista dell'utente, la cifratura può essere abilitata tramite un semplice interruttore (toggle switch) disponibile nel form di configurazione del servizio
- La cifratura del disco viene effettuata usando lo standard **LUKS (Linux Unified Key Setup)** che utilizza il modulo *dm-crypt* del kernel (parte del framework device mapper) ed è quindi trasparente per l'applicazione (Galaxy in questo caso).
- Il modo in cui funziona LUKS è che una master key viene generata automaticamente per la crittografia e ci sono 8 slot per chiavi (passphrase) che proteggono la chiave principale. La gestione della passphrase è dunque un punto critico.

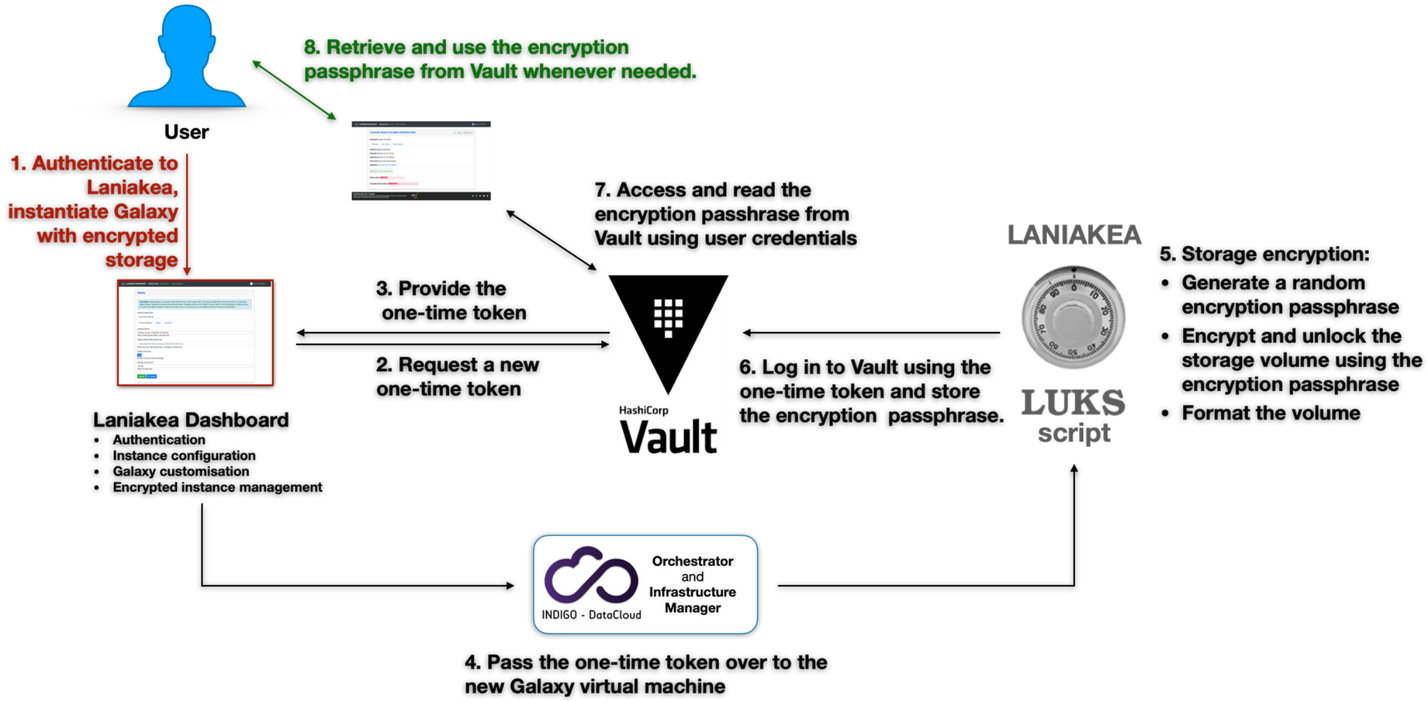




La gestione della passphrase

Per proteggere la passphrase abbiamo integrato Vault (il Secrets Manager della Hashicorp) nella nostra piattaforma.





AUTENTICAZIONE BASATA SU IAM E ISOLAMENTO DEGLI UTENTI

Su Vault abbiamo abilitato l'autenticazione tramite JWT integrando IAM. Ogni utente su Vault è autorizzato a leggere e scrivere secrets solo in path specifici, che dipendono dall'identità (token) dell'utente stesso.

POLICY DI AUTORIZZAZIONE

Vault mette a disposizione un meccanismo flessibile di autorizzazione basato su policy che consentono di stabilire cosa un utente, in possesso di un token valido, può fare e a quali path (secrets) può accedere.

La gestione della passphrase

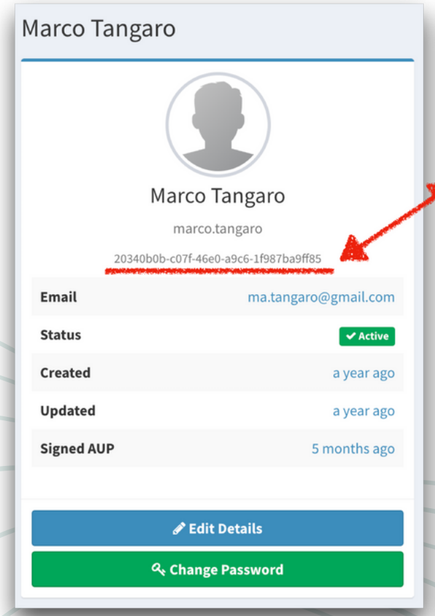
La passphrase viene generata automaticamente e salvata nel Vault in un path specifico per utente e per deployment.

GENERAZIONE DELLA PASSPHRASE

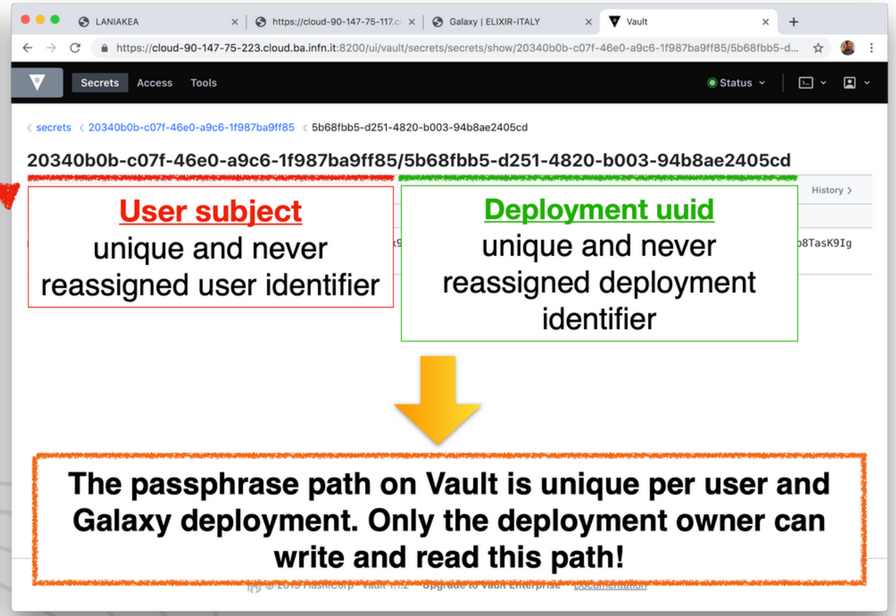
Durante il deployment, sulla VM uno script genera la password e la salva sul Vault usando un token di breve durata utilizzabile solo una volta (che viene passato dalla PaaS durante la contestualizzazione).

RECUPERO DELLA PASSPHRASE

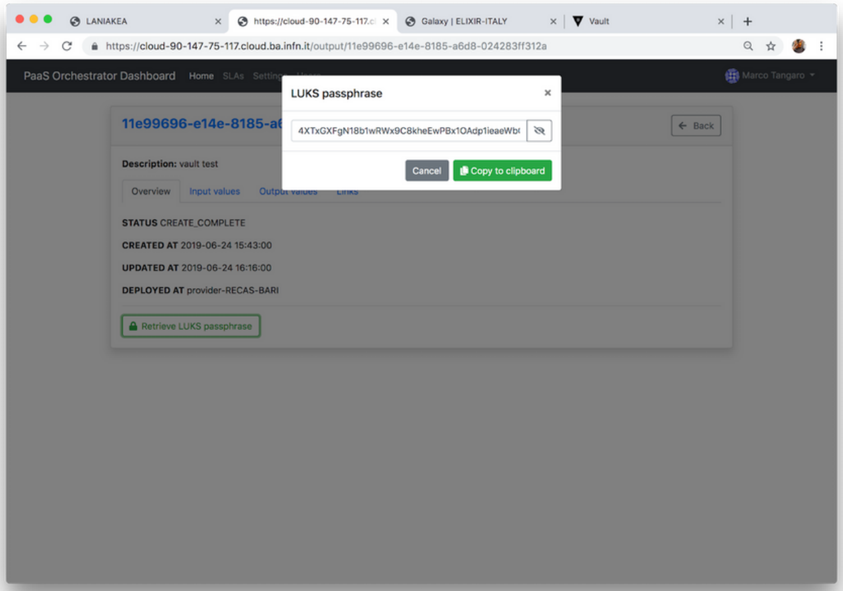
In ogni momento l'utente può recuperare attraverso la dashboard la passphrase generata per il proprio deployment.



User identity in IAM



Passphrase path on Vault



Ambienti virtuali isolati

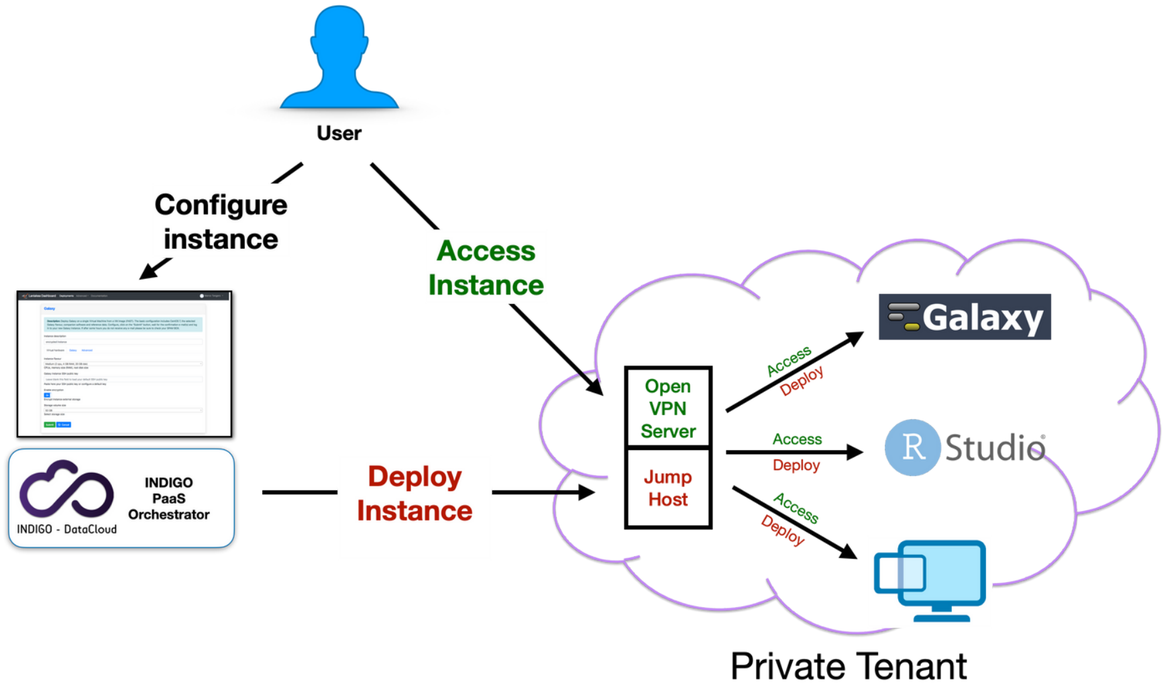
Deployment su rete privata

Nuova funzionalità della PaaS

- La PaaS di INDIGO/DEEP è stata estesa per consentire il deployment di VM su rete privata --> **sviluppo fatto in INFN Cloud**

- Il setup della piattaforma si complica leggermente perchè è necessario avere un "**jump host**" attraverso cui la PaaS può raggiungere le VM via ssh per eseguire la contestualizzazione

- Questo setup è completamente **trasparente per l'utente finale** in quanto la PaaS è in grado di gestire i jump host in maniera autonoma.



Deployment su rete privata

dietro le quinte

- L'Orchestrator legge il template TOSCA che descrive il deployment e riconosce il tipo di rete richiesto
- Recupera dal CMDB l'IP del proxy host e lo username da utilizzare e inietta queste informazioni nel template prima di passarlo all'Infrastructure Manager
- L'Infrastructure Manager usa il proxy host per raggiungere le VM su rete privata e configurarle tramite ansible

New derived tosca type

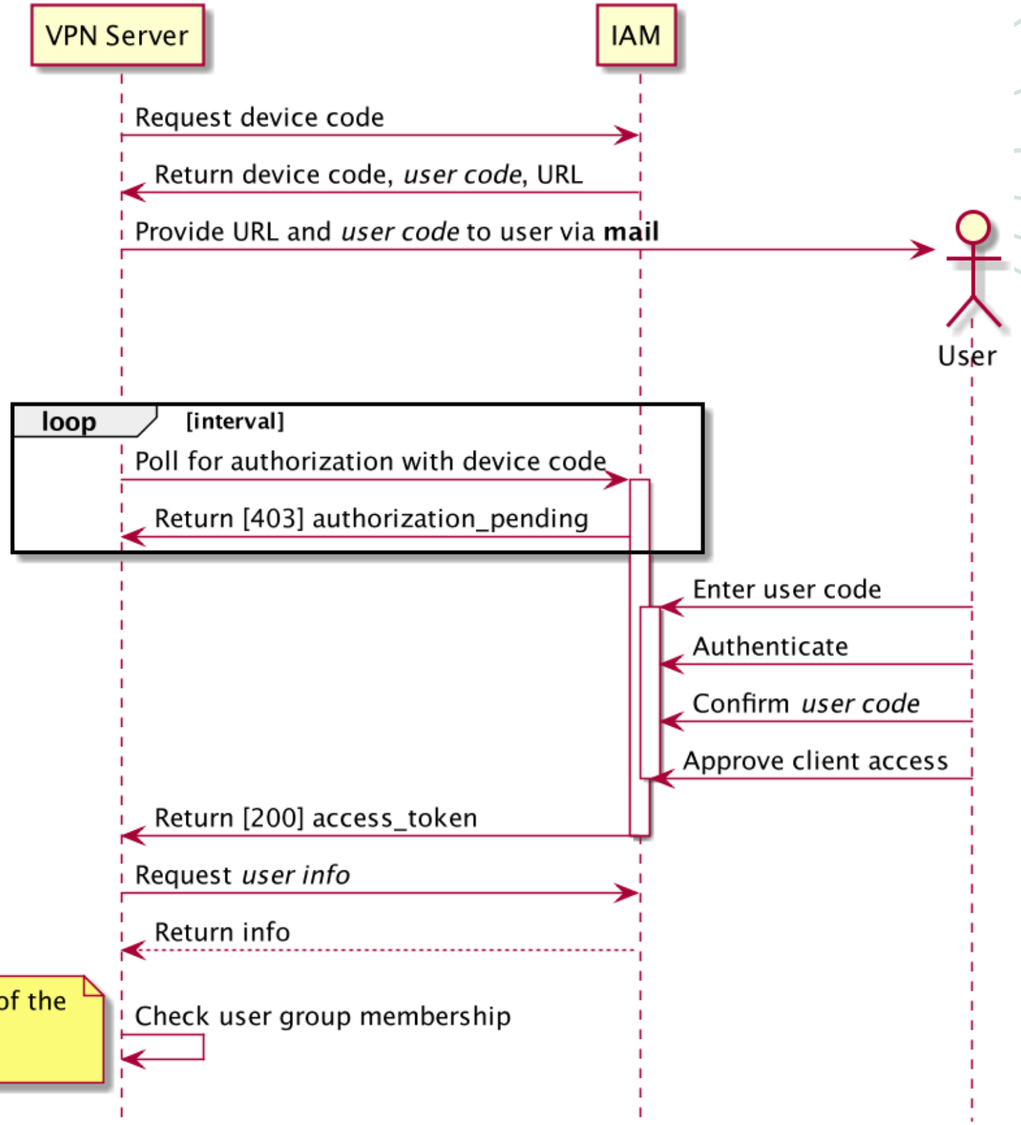
```
tosca.nodes.indigo.network.Network:
  derived_from: tosca.nodes.network.Network
  properties:
    proxy_host:
      type: string
      required: false
    proxy_credential:
      type: tosca.datatypes.Credential
      required: false
```

Extended CMDB *tenant* record


```
{
  "_id": "5b888af8-cb6e-4ff7-a5b0-3551203f9f55",
  "_rev": "6-ba6de81e66b97a5b05226497dcc42d09",
  "type": "tenant",
  "data": {
    "service": "RECAS-BARI_d83f79ed-497c-4921-91ee-45c03c94f892",
    "tenant_id": "c21b67024c724f559dc53ad530337d69",
    "tenant_name": "Laniakea",
    "iam_organisation": "laniakea",
    "private_network_proxy_host": "212.189.205.95",
    "private_network_proxy_user": "im"
  }
}
```

L'accesso ai servizi su rete privata

- Gli utenti possono accedere alla propria istanza **tramite VPN**
- Per l'autenticazione è stato sviluppato un **modulo PAM** che consente di usare IAM sfruttando il flusso di autorizzazione del "device code"
- L'implementazione di riferimento è basata su ***OpenVPN***



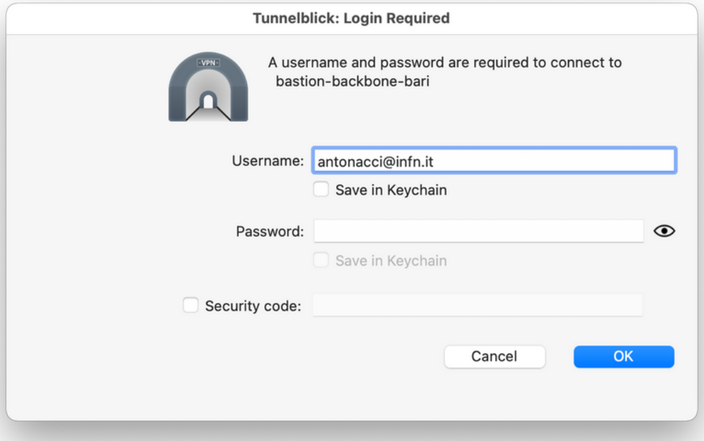
Integrazione con OpenVPN

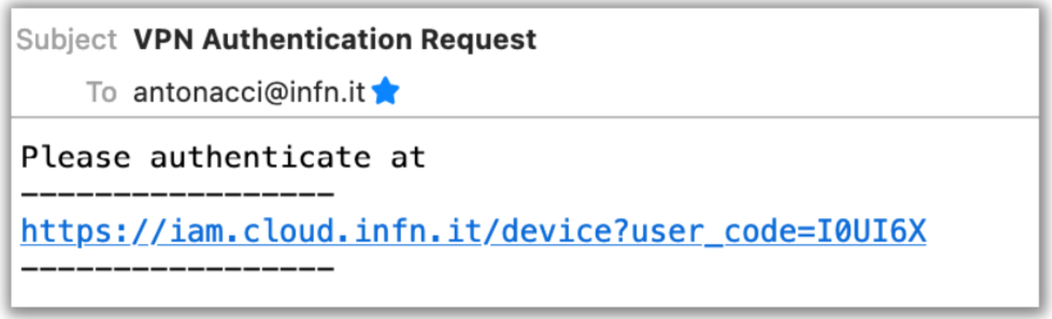
qualche dettaglio

- E' necessario utilizzare per il server una versione di OpenVPN >= 2.5 per abilitare la **deferred authentication**
- Abbiamo verificato la compatibilità con client di versione precedente (2.4.7)
- I test sono stati effettuati usando come **client Tunnelblick** e **OpenVPN**
- L'utente deve fornire un indirizzo email valido come username

server relevant params

```
plugin /usr/lib/x86_64-linux-
gnu/openvpn/plugins/openvpn-plugin-auth-
pam.so openvpn
duplicate-cn
setenv deferred_auth_pam 1
reneg-sec 0
hand-window 300
username-as-common-name
```





Conclusioni

Gli aspetti relativi alla sicurezza dei dati e i requisiti legali ed etici sulla conservazione e la gestione dei dati genetici e sanitari stanno diventando sempre più stringenti.

Il Task 6.6 del progetto EOSC-Pillar mirava ad analizzare la compliance normativa del servizio Laniakea di analisi dati di livello PaaS integrato e interoperabile per ELIXIR e la comunità di Life Science, frutto di un’interazione tra i servizi Galaxy e i repository di dati.

Dal punto di vista tecnologico, abbiamo implementato le misure necessarie per migliorare la sicurezza dell'intero servizio. In particolare, l'obiettivo è garantire la creazione di ambienti isolati e sicuri per svolgere le analisi dati. Per far questo, ci siamo concentrati su due aspetti critici: la gestione dei dati e il controllo dell'accesso al servizio.

Le attività presentate in questo talk sono state svolte in **forte singergia** con altri progetti e sfruttando anche sviluppi fatti in **INFN Cloud.**

Referenze

- Sito Laniakea: **https://laniakea-elixir-it.github.io/**
- Documentazione Laniakea: https://laniakea.readthedocs.io/en/latest/index.html
- Laniakea: an open solution to provide Galaxy “on-demand” instances over heterogeneous cloud infrastructures. https://doi.org/10.1093/gigascience/giaa033
- Laniakea@ReCaS: exploring the potential of customisable Galaxy on-demand instances as a cloud-based service. https://doi.org/10.1186/s12859-021-04401-3
- Legal Framework for the use and re-use of health data for scientific purposes. https://doi.org/10.5281/zenodo.6334878

Grazie
per l'attenzione

Contatti:
marica.antonacci@ba.infn.it
giacinto.donvito@ba.infn.it
nadina.foggetti@ba.infn.it
ma.tangaro@ibiom.cnr.it