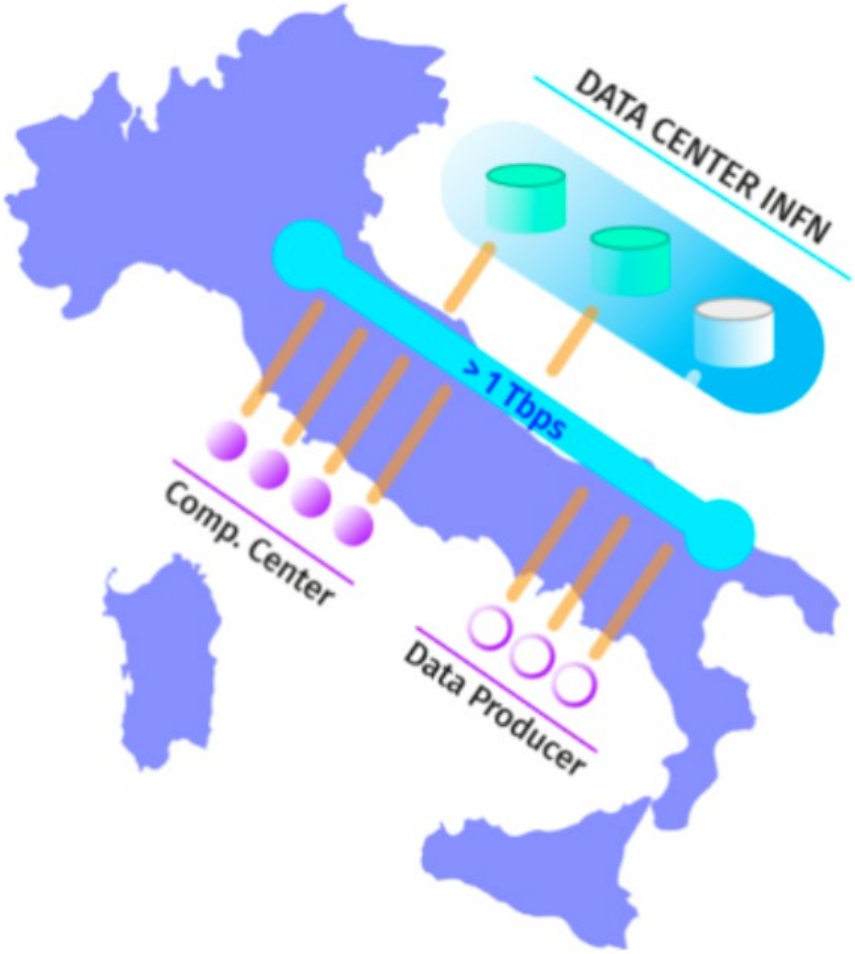
Stato della piattaforma EPIC nella prospettiva del Tecnopolo

Barbara Martelli
(per il team DataCloud)

INFN4LS
14/07/2023

INFN Cloud

- A **multi-site, federated Cloud** infrastructure integrating HPC and HTC resources
- In **production** since March 2021.
- The **seed** of a National Datalake for research and beyond, building on existing, renewed or new e-Infrastructures.
- Architectural foundations:
  - No vendor lock-in (open-source, vendor-neutral)
  - Federation of existing resources (computing and data management)
  - Dynamic orchestration of resources via INDIGO PaaS Orchestrator
  - Consistent AuthN/AuthZ at all cloud levels via OpenID-Connect/OAuth2
- Includes a secure, GDPR-compliant region dedicated to life-science use cases: EPIC (**E**nhanced **P**r**I**vacy and **C**ompliance) Cloud



https://www.cloud.infn.it/

Enhanced **Pr**I**v**acy and **C**ompliance **C**loud is an ISO certified cloud platform

A region of INFN Cloud with a certified Information Security Management System

EPIC Cloud offers a secure **Community Cloud**

Biomedical and genomic researchers

Industrial researchers

Site locations: CNAF Bologna (active now), Bari and Catania will be added in october enhancing the high availability and disaster recovery capabilities

Resource available today: 1PB storage, 2k CPU, 16 TB RAM, 4 GPU

Ongoing expansion with 3.5M euro form NRRP resources and 4M euro of other projects

Why EPIC Cloud

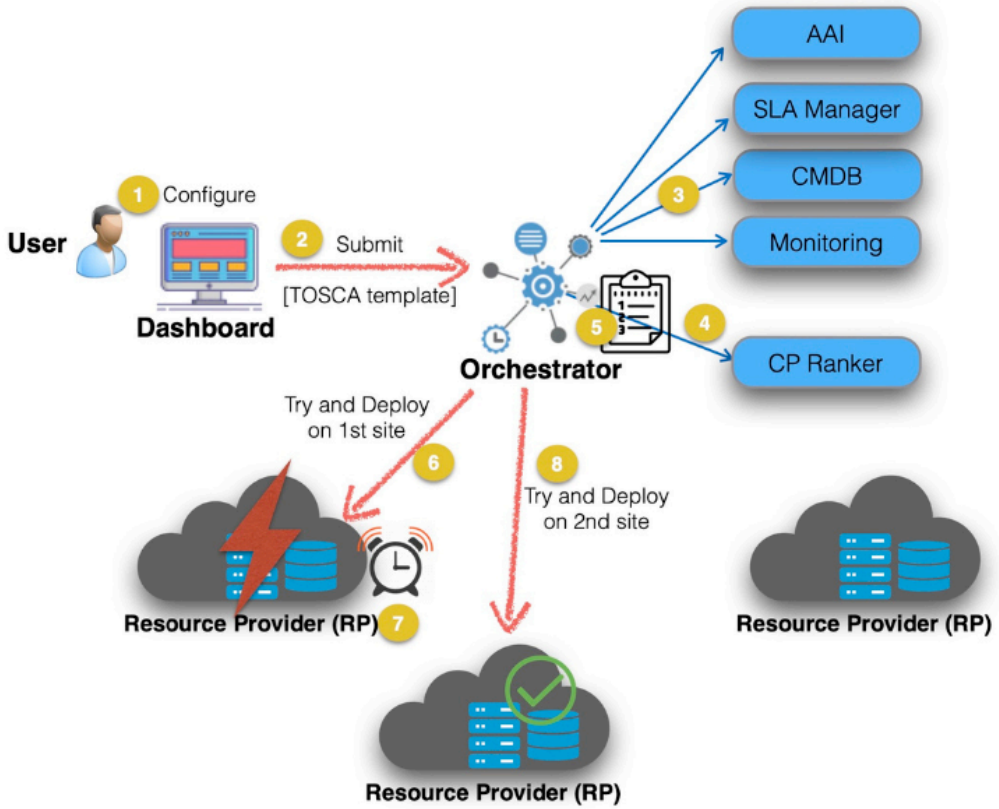
- The GDPR states that Clinical and medical data (for instance, genomic) is personal data; i.e., it fits in the Art.9 special categories of personal data.
  - Genomic data is mostly impossible to be anonymized → GDPR shall always be applied
  - ISO/IEC 27001 is the main certification mechanism compliant with GDPR requirements (Art. 43, 58, 63)
- In order to comply with the requirements of health research projects INFN is involved in, we created at CNAF **a region of the INFN Cloud infrastructure**, applied specific organizational and technical security measures, and certified it ISO/IEC 27001, 27017, 27018.
  - This is **EPIC Cloud**: a *reference Cloud implementation for the treatment of sensitive data at INFN*

From the Data Controller side, the fact that EPIC Cloud is ISO-certified is a way to demonstrate that processing is performed in accordance with the GDPR

THE FEDERATION MIDDLEWARE

The INDIGO PaaS Orchestrator enables the federation of distributed and heterogeneous compute environments: clouds, docker orchestration platforms, HPC systems.

- **Smart scheduling** → Automatic selection of the best provider
  - based on compute/storage requirements vs provider capabilities including the following criteria:
    - Resource quotas (SLA)
    - Monitoring data
    - Support for specialized hardware (GPU, Infiniband)
    - Data location
- Support for **hybrid deployments** and network orchestration
- **Client interfaces** for advanced users (REST APIs, CLI, python bindings) and end-users (web dashboard - no skills required)



THE SERVICE IMPLEMENTATION STRATEGY

The employed strategy is based on the **Infrastructure as Code paradigm**.
Users describe "**What**" is needed rather than "**How**" a specific service or functionality should be implemented.
The adopted technologies enable a Lego-like approach: **services can be composed and modules reused** to create the desired infrastructure.

TOSCA is used to model the topology of the whole application stack

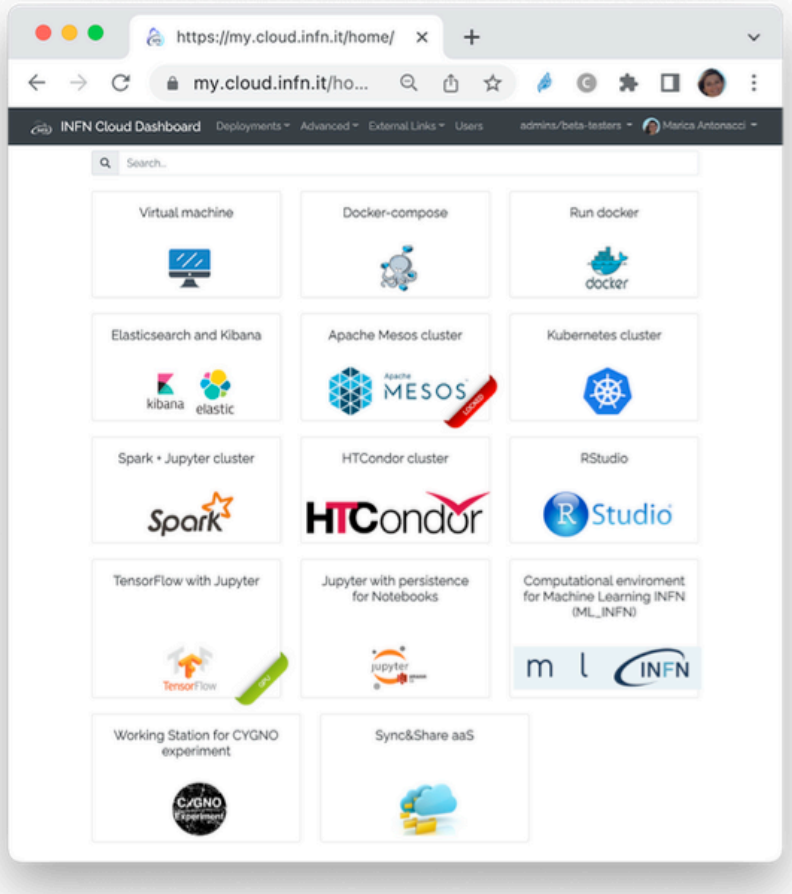
Ansible is used to automate the configuration of the virtual environments

Docker is used to encapsulate the high-level application software and runtime

THE PAAS DASHBOARD

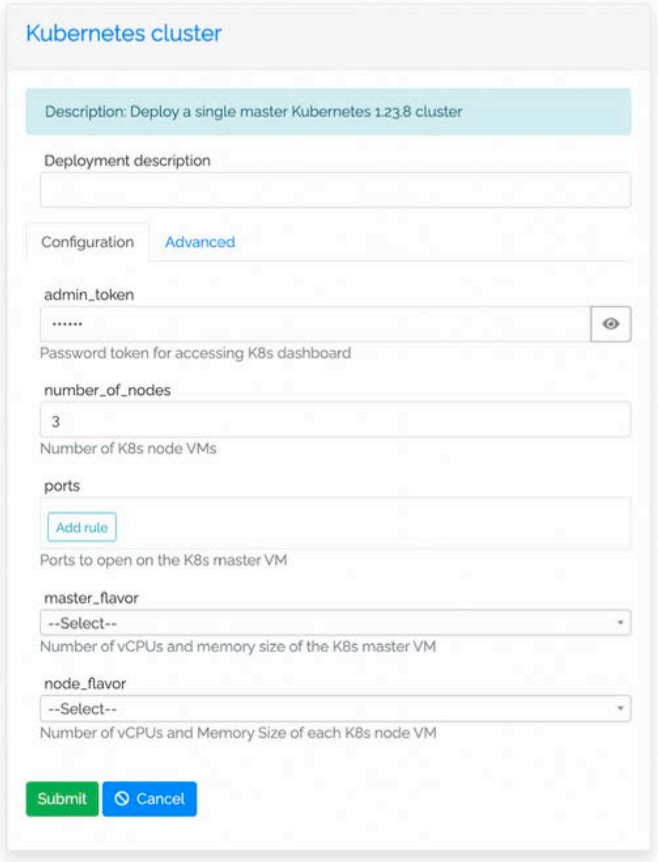
The INDIGO PaaS Dashboard is a web-based user interface that enables users to manage and monitor their deployments without requiring any TOSCA knowledge.
The dashboard hides all technical details and provides an intuitive interface for managing service deployments.

- OpenID-Connect Authentication
- Multi-tenancy
- Secrets management (via Vault integration)
- Dynamic view of service catalog (depending on the user group membership)



Self-provisioning

REQUEST SERVICES WITH JUST A FEW CLICKS

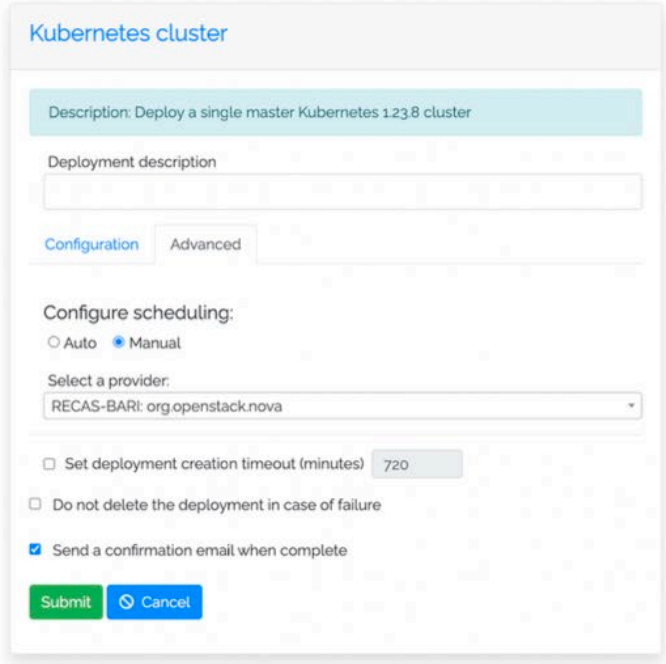


Customize your deployment

through the deployment input parameters

Choose the Scheduling strategy

- automatic: let the Orchestrator select the best provider
- manual: choose the provider from the drop down menu automatically created by the Dashboard with the list of providers returned by the SLA Manager service



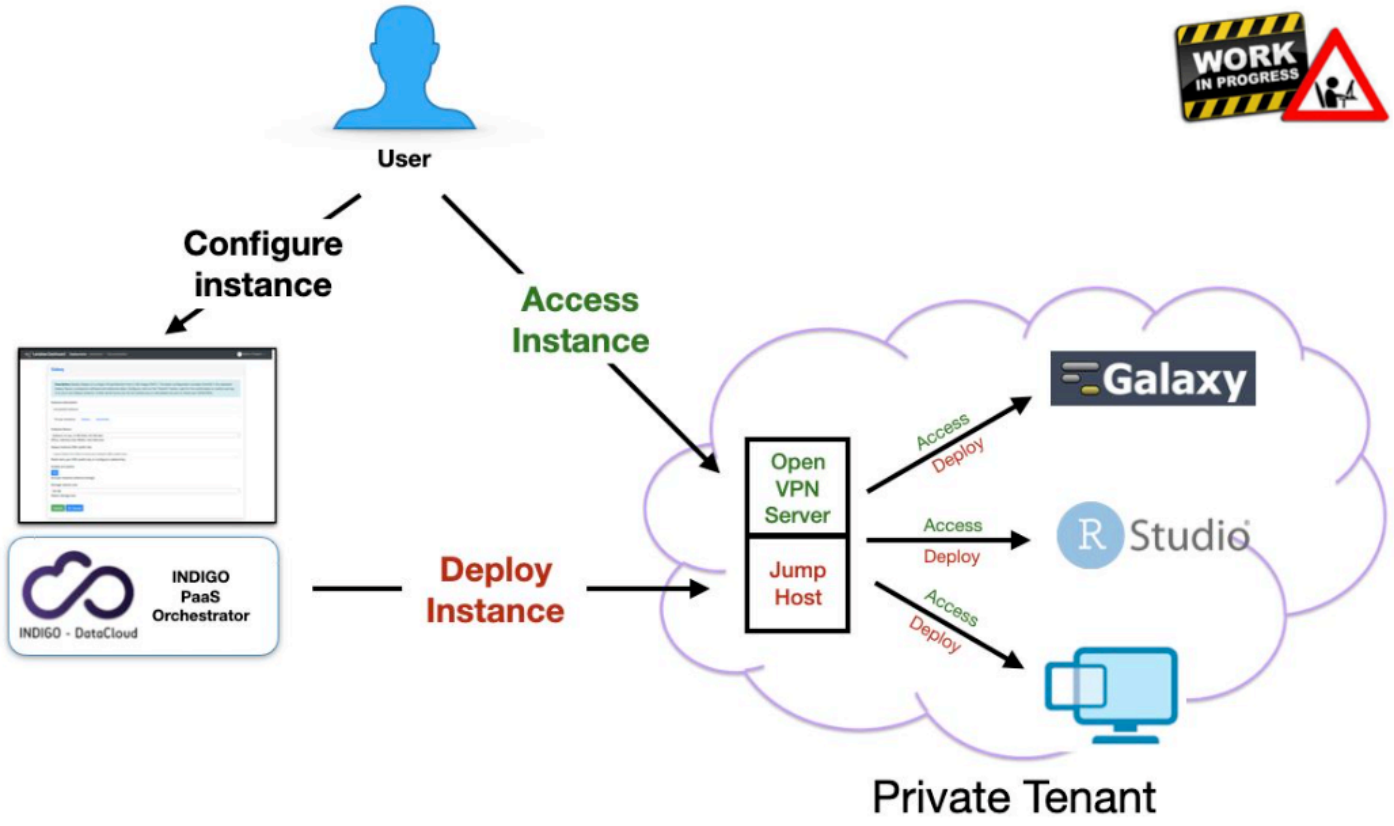
Security: multiple isolation levels

Deployments under VPN

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

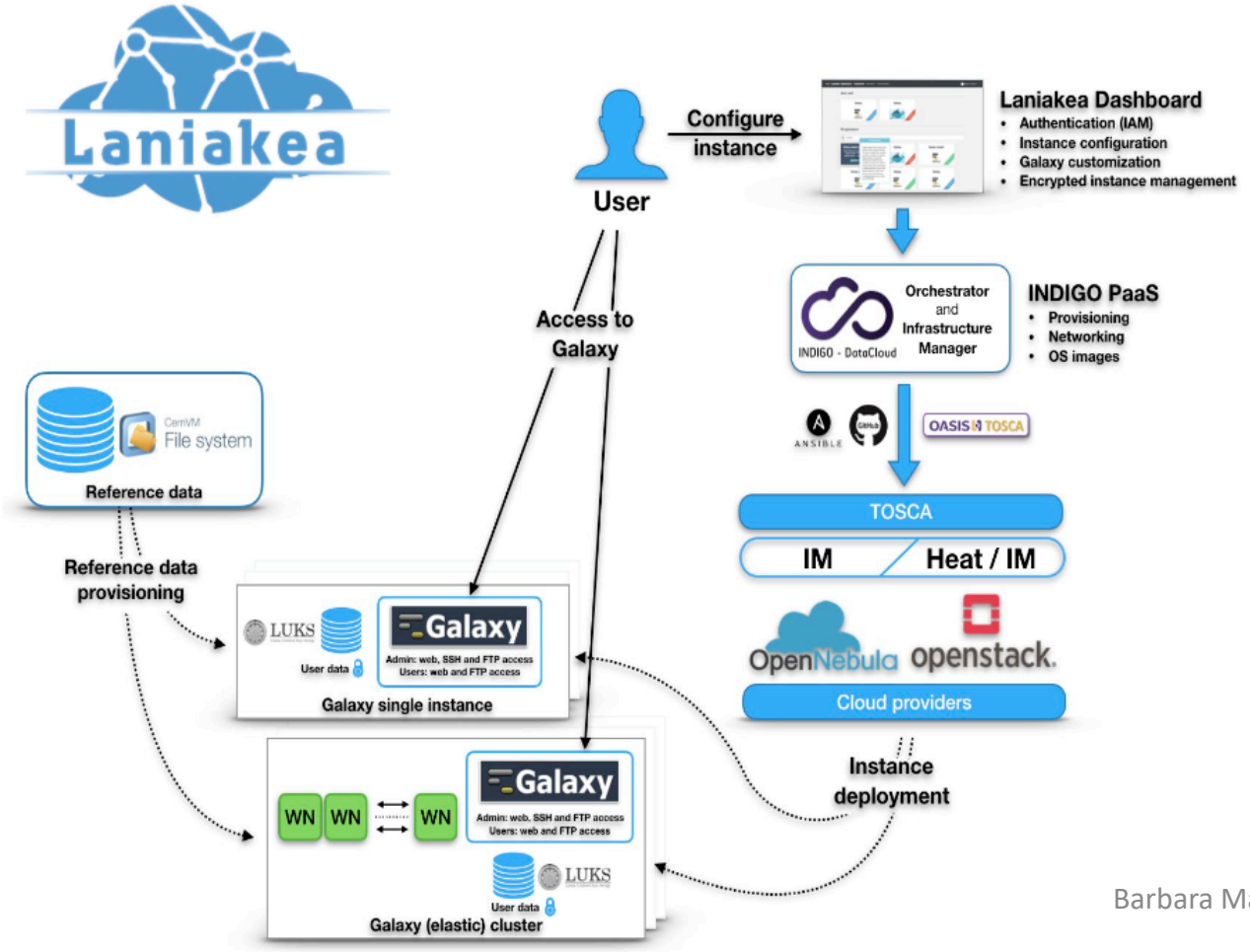
Isolation is reached using Tenant and security groups properties, granting the access only through VPN authentication.

User authentication to the VPN using the same Laniakea credentials.



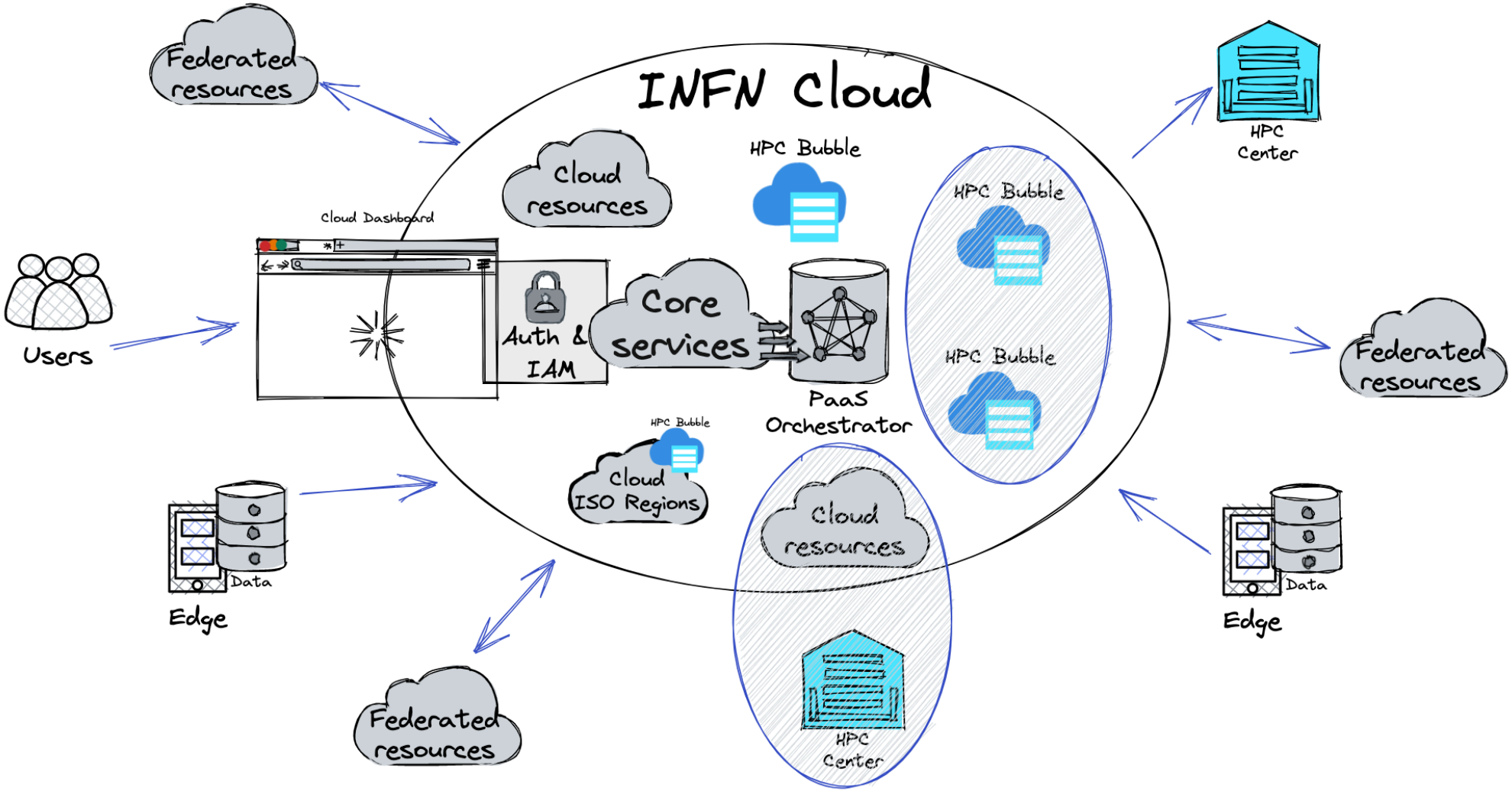
Integration of bioinformatics tools

Laniakea architecture



- **Dashboard -** User friendly access to configuration and and launch of a Galaxy instance.
- **IAM -** Authentication and Authorization system.
- **INDIGO PaaS -** Galaxy automatic deployment.
- **Cloud Providers -** (INFN) ReCaS-Bari and others.
- **Persistent storage** - With/without encryption.
- **Reference data availability** - With CERN-VM FileSystem.
- **CLUES -** Elasticity manager.

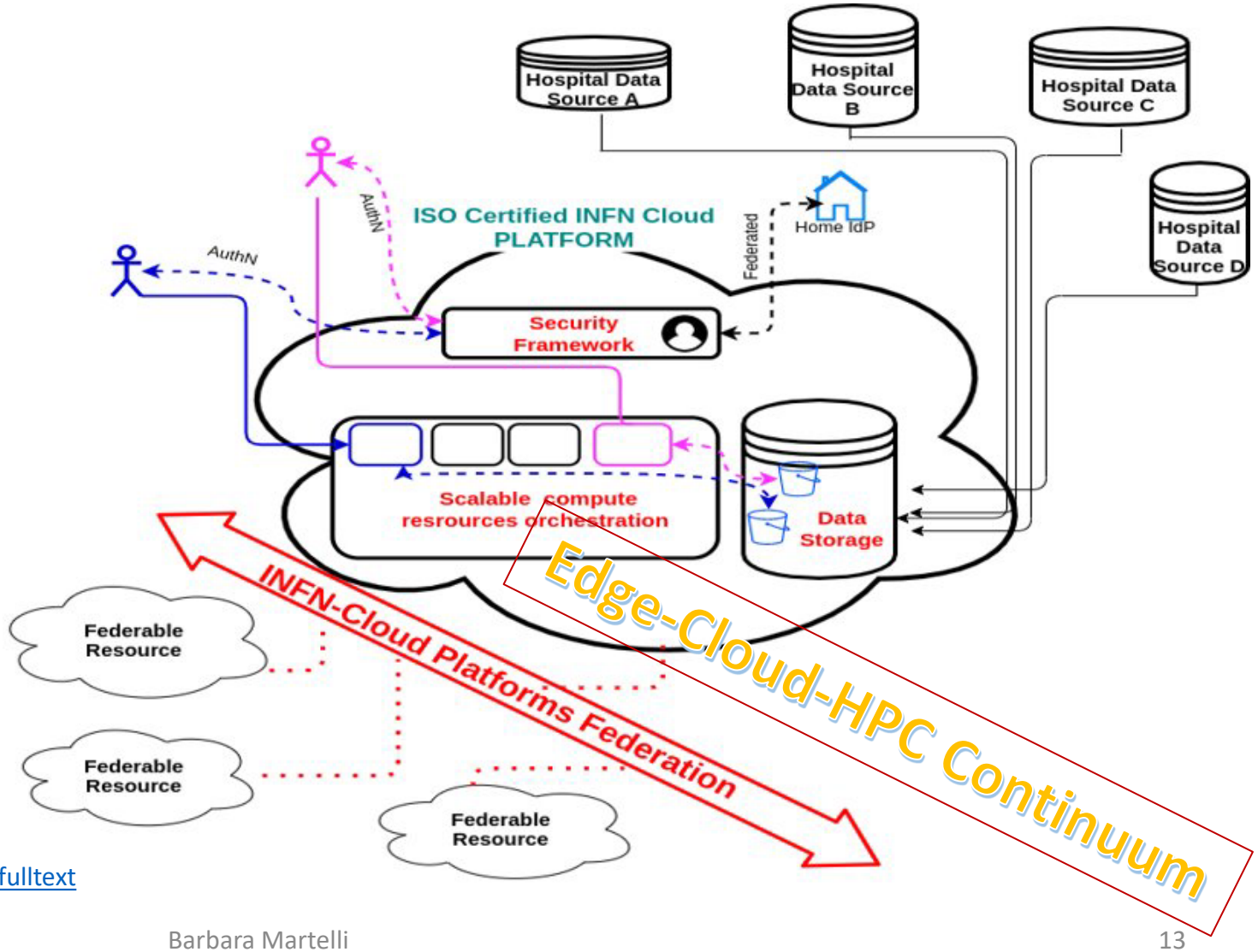
The *continuum* from Edge, to Cloud, to HPC



The goal: a federated datalake

Multiple ways to ingest and process data are possible. For example, to handle sensitive data (e.g., in the nation-wide Health Big Data project), we are working on supporting these options:

1. **Central harvesting** of data generated remotely
2. **Edge-level anonymization**, followed by central ingestion and analysis of data
3. **Edge-level feature extraction**, followed by central ingestion and analysis of features
4. **Federated learning** based on edge-level training, followed by publishing of the trained methods and by inference performed either centrally or at other edge locations.



https://www.physicamedica.com/article/S1120-1797(21)00320-3/fulltext

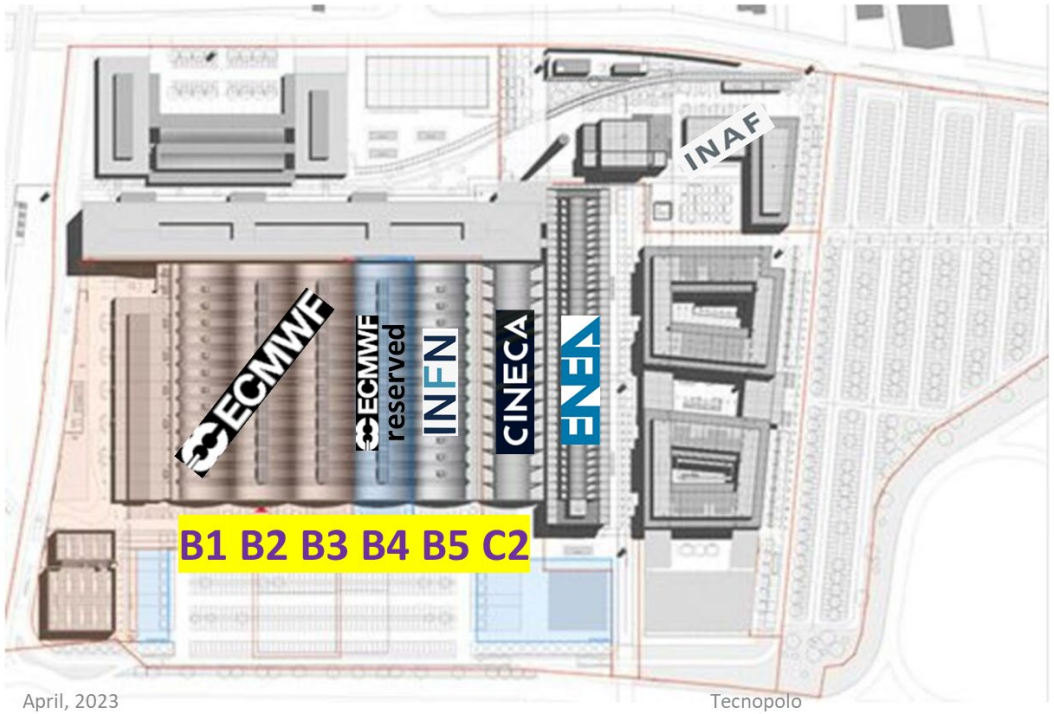
Estensione di EPIC a Bari e Catania

- Obiettivo: certificazione INFN entro ottobre ‘23
- Siti coinvolti:
  - Amministrazione Centrale
  - Bari
  - Catania
  - CNAF
- ISO 27001 27017 27018
- Impostazione del sistema in modo da rendere semplice l’aggiunta di ISO 9001 (obiettivo 2024)

Coprogettazione, sviluppo e manutenzione di soluzioni software di DataCloud per il settore della ricerca.

Erogazione di servizi di DataCloud IaaS, SaaS e PaaS in community deployment model.

Trasferimento del CNAF al Tecnopolo



April, 2023

Tecnopolo

July 2023 Room Ready
- Electrical and cooling system available
- Racks installed

Oct 2023 Network Ready
- Cabling
- Active core systems
- Management network
- DCI with legacy site

Gen 2023 Resource Relocation starts
- New Storage Installed
- Legacy CPU and Storage migration starts

Mar 2024 Migration Finalized
- Tape Library migrated
- Site commissioned

Credits: D.Cesini and Tier1 team

Trasferimento di EPIC al tecnopolo

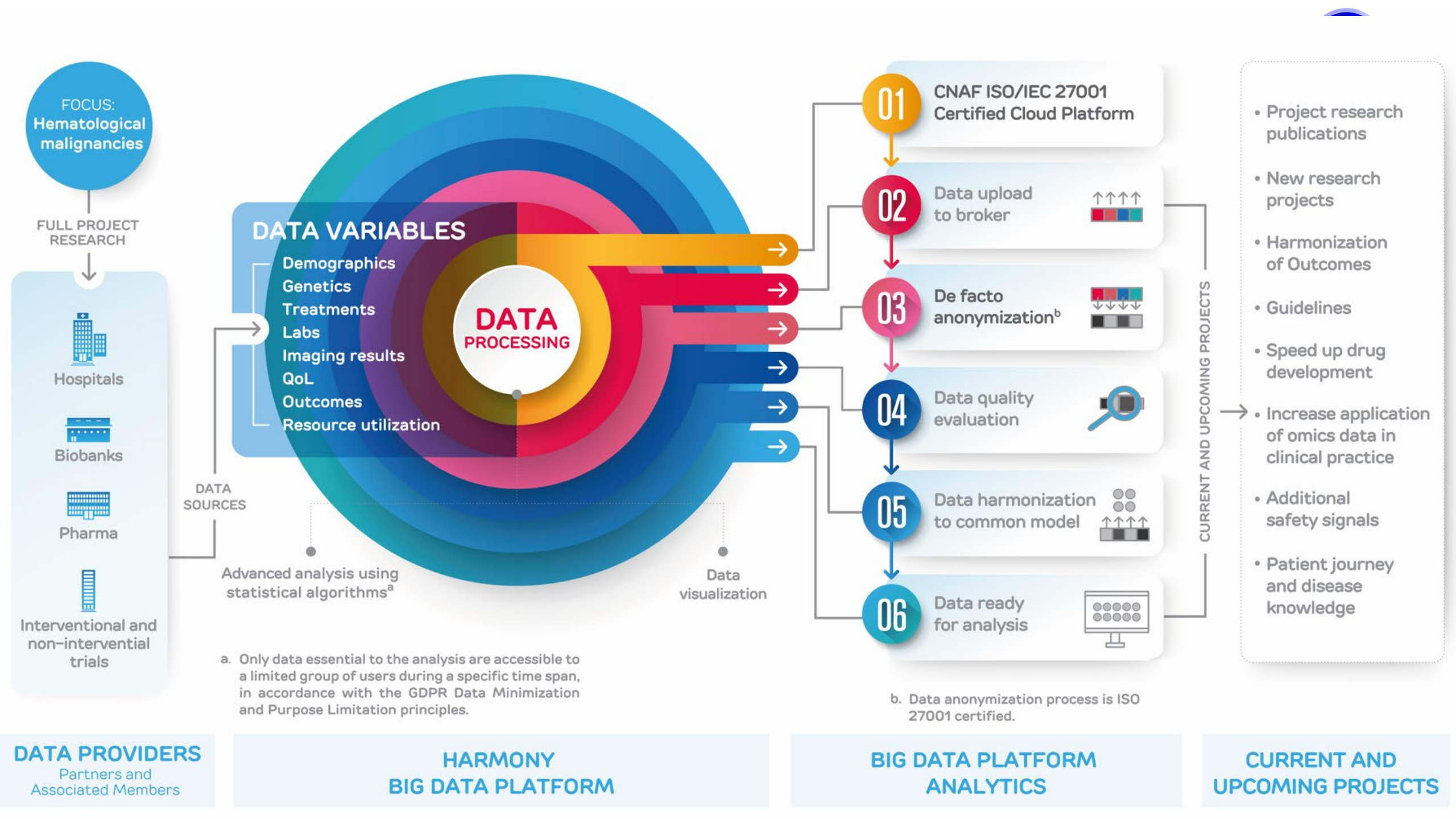
- EPIC ultima infrastruttura da trasferire (presumibilmente non prima dell'estate '24)
- In arrivo a settembre hardware (SSD storage) su gara già aggiudicata per 500k euro -> sarà installato presso l'attuale Tier1
- Disponibili 465k + 330k euro, in discussione se avviare le gare a settembre '23, se si avviano, da discutere dove installare l'HW
- Hardware da gara TeRABIT previsto per inizio '24, presumibilmente installato al tecnopolo
- Hardware DARE (3.5 Meuro) da definire

Progetti attualmente ospitati su EPIC

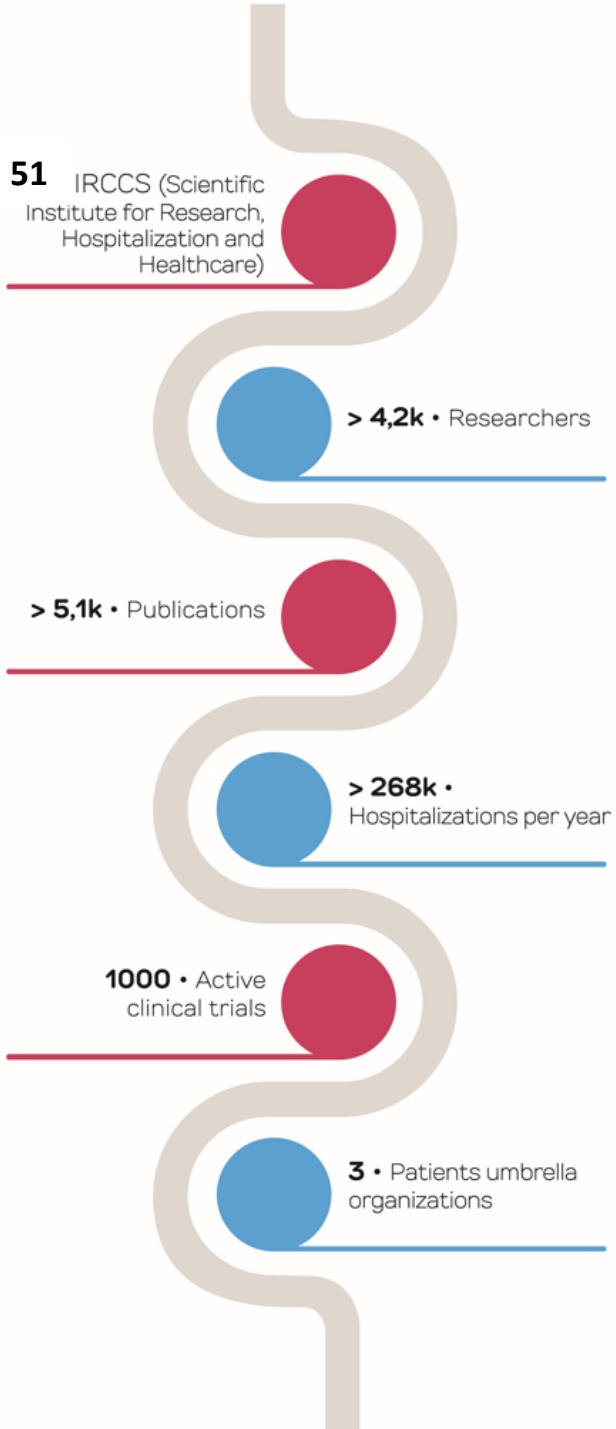
Harmony Alliance

- HARMONY (Healthcare Alliance for Resourceful Medicine Offensive against Neoplasms in hematologY)
  - Harmony is a public-private partnership involving more than 100 organizations from 18 European countries, such as hospitals, universities, research institutes, medical associations, patient organizations, pharmaceutical companies and IT companies
  - Harmony is aimed at exploiting Big Data to develop more personalized treatments for blood cancer patients
  - Most of data analyzed in Harmony are genomic data
  - The project big data platform is located at INFN-CNAF

100k patients' datasets


FOCUS:
Hematological malignancies
FULL PROJECT RESEARCH
Hospitals
Biobanks
Pharma
Interventional and non-interventional trials
DATA SOURCES
DATA VARIABLES
Demographics
Genetics
Treatments
Labs
Imaging results
QoL
Outcomes
Resource utilization
DATA PROCESSING
Advanced analysis using statistical algorithms[a]
Data visualization
01 CNAF ISO/IEC 27001 Certified Cloud Platform
02 Data upload to broker
03 De facto anonymization[b]
04 Data quality evaluation
05 Data harmonization to common model
06 Data ready for analysis
CURRENT AND UPCOMING PROJECTS
• Project research publications
• New research projects
• Harmonization of Outcomes
• Guidelines
• Speed up drug development
• Increase application of omics data in clinical practice
• Additional safety signals
• Patient journey and disease knowledge
a. Only data essential to the analysis are accessible to a limited group of users during a specific time span, in accordance with the GDPR Data Minimization and Purpose Limitation principles.
b. Data anonymization process is ISO 27001 certified.
DATA PROVIDERS
Partners and Associated Members
HARMONY BIG DATA PLATFORM
BIG DATA PLATFORM ANALYTICS
CURRENT AND UPCOMING PROJECTS

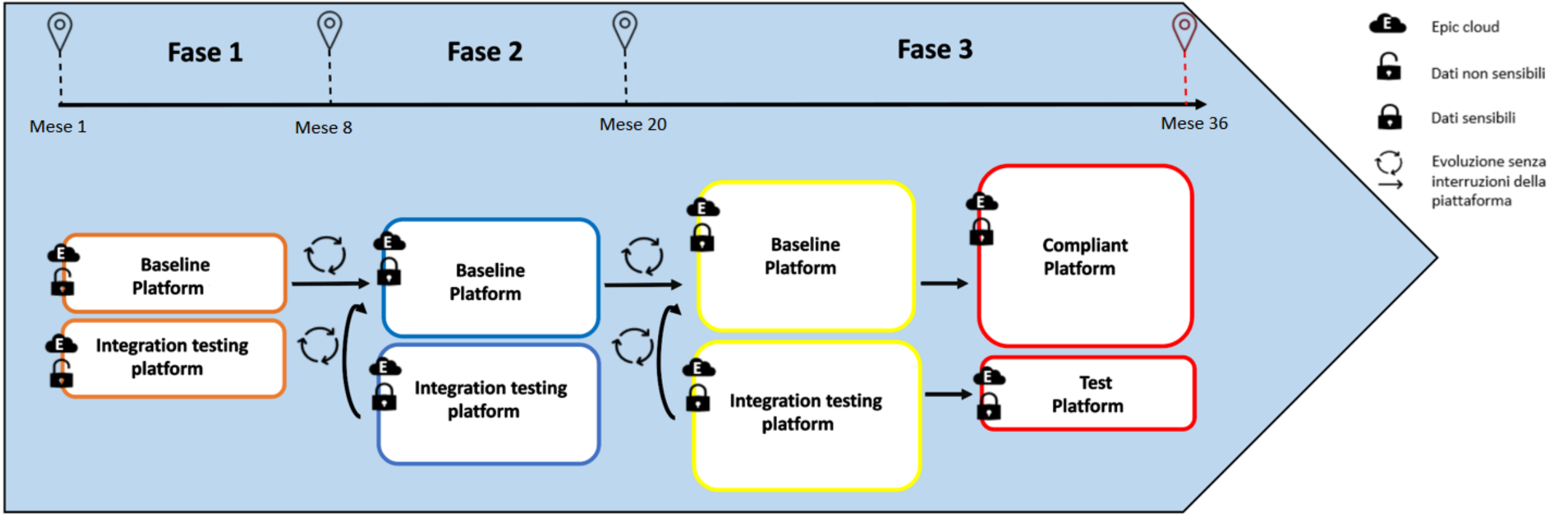
Health Big Data



- A 10-year project founded by the Italian Health Ministry
- Goal: develop a federated cloud platform enabling the sharing of patients' data at national level (EHR, omics, clinical, epidemiology data)
- Almost all Italian Scientific Institutes for Research, Hospitalization and Healthcare involved
- The project platform is HBD-DataCloud, a community cloud federation *based on INFN Cloud technologies*

Sant'Orsola

- Co-sviluppo della piattaforma di genomica computazionale dell'IRCCS AOU Sant'Orsola (settembre '22 – agosto '25)



Altri progetti attivi

- PLANET (CSN5, in dismissione)
- Covid-stat (per la parte di analisi dati)

Alcune richieste di collaborazione ricevute

Collaborazione con San Raffaele

- In ambito HBD
  - Cbioportal https://www.cbioportal.org/
    - Non risultano installazioni in Italia, Reply ne installerà una su EPIC
  - Molecular Tumor Board
    - Reply sviluppa l'appicativo di proprietà del progetto HBD, quindi anche nostra
  - Il lavoro verrà svolto da Reply, noi dobbiamo monitorare e verificare che risponda ai nostri requisiti.
- In discussione accordo di ricerca INFN + San Raffaele + Elixir
  - Gestione di una piattaforma di medie dimensioni (1PB storage, 17 nodi dedicati, alcune GPU
  - Efficientamento e standardizzazione di pipeline analisi in snakemake
  - Utili ambienti tipo HPC Bubbles -> sinergia con TeRABIT

Accordo di collaborazione con IFOM

- Galaxy, Nextflow, slurm, jupyter, container-> sinergia con HBD, Sant'Orsola, ICSC Spoke8, DARE
- Studiare o sviluppare tool per risolvere problematiche relative alla riproducibilità e tracciabilità delle analisi
- Studiare tool per la gestione del ciclo di vita del dato
- Browser Genomici (es. https://github.com/igvteam/igv.js/ che presenta il dato appena generato da pipeline -> sinergia con Sant'Orsola e ICSC Spoke8
- Algoritmi per la segmentazione es. https://github.com/MouseLand/cellpose -> sinergia con attività di segmentazione di immagini radiomiche (contornare gli organi per esempio) (Retico)?

Attività previste nell'imminente futuro

- Accordi richiesti:
  - Istituto Europeo di Oncologia (rete ACC, genomica)
  - Gaslini di Genova (rete IDEA, malattie pediatriche rare)
  - Bellaria (rete RIN, radiomica)
- Progetti PNRR in corso:
  - ICSC Spoke 8: identificati diversi use case che necessitano di INFN cloud sicura (in particolare Istituto Ortopedico Rizzoli, INFN Ferrara e Università di Catania)
  - DARE: identificati 40 piloti, alcuni di essi necessiteranno dell'infrastruttura cloud sicura

Macro attività previste

- Creare figure esperte di life science da vari punti di vista
  - Applicazioni come l'ecosistema Elixir (Galaxy, Laniakea) OpenCGA, FEGA, sistemi HPC come Clara Parabrics di NVIDIA o simili, LIMS, RedCap, con l'obiettivo di renderle interoperabili e scalabili su infrastrutture cloud
  - Sistemi di autenticazione GDPR compliant (es. IAM con MFA e feature di auditability)
  - Data management GDPR compliant (es. FTS, RUCIO e integrazione di RUCIO con cataloghi esterni)
  - Gestione dei dati personali (conoscenza interdisciplinare tecnico-legale in ottica R&D)