INFN CLOUD

Secure Platforms for INFN research projects' sensitive data

Giusy Sergi (INFN-CNAF),
Cristina Duma (INFN-CNAF),
Barbara Martelli (INFN-CNAF)

Big Data and healthcare

- Can we process this data?
  - Be aware of regulation requirements (GDPR, Italian law, Italian Data Protection Authority, local rules and regulations)
- Which type of service could we offer?
  - IaaS, PaaS, SaaS
  - Each solution foresees the application of different levels/types of responsabilities

Cloud provider point of view

Which type of service could we offer?

Responsibility model	IaaS (Infrastructure as a Service)	PaaS (Platform as a Service)	SaaS (Software as a Service)
GRC (Security Governance Risk & Compliance)	🟩	🟩	🟩
Data Security	🟩	🟩	🟩
Application Security	🟩	🟩	🟥
Platform Security	🟩	🟥	🟥
Infrastructure Security	🟥	🟥	🟥
Physical Security	🟥	🟥	🟥

Customer responsibility

Provider responsibility

In our cloud solutions (based on EPIC Cloud) the boundaries of responsibilities are clearly defined through contracts and agreements signed with each customer, project or collaboration

Different shared responsibility models

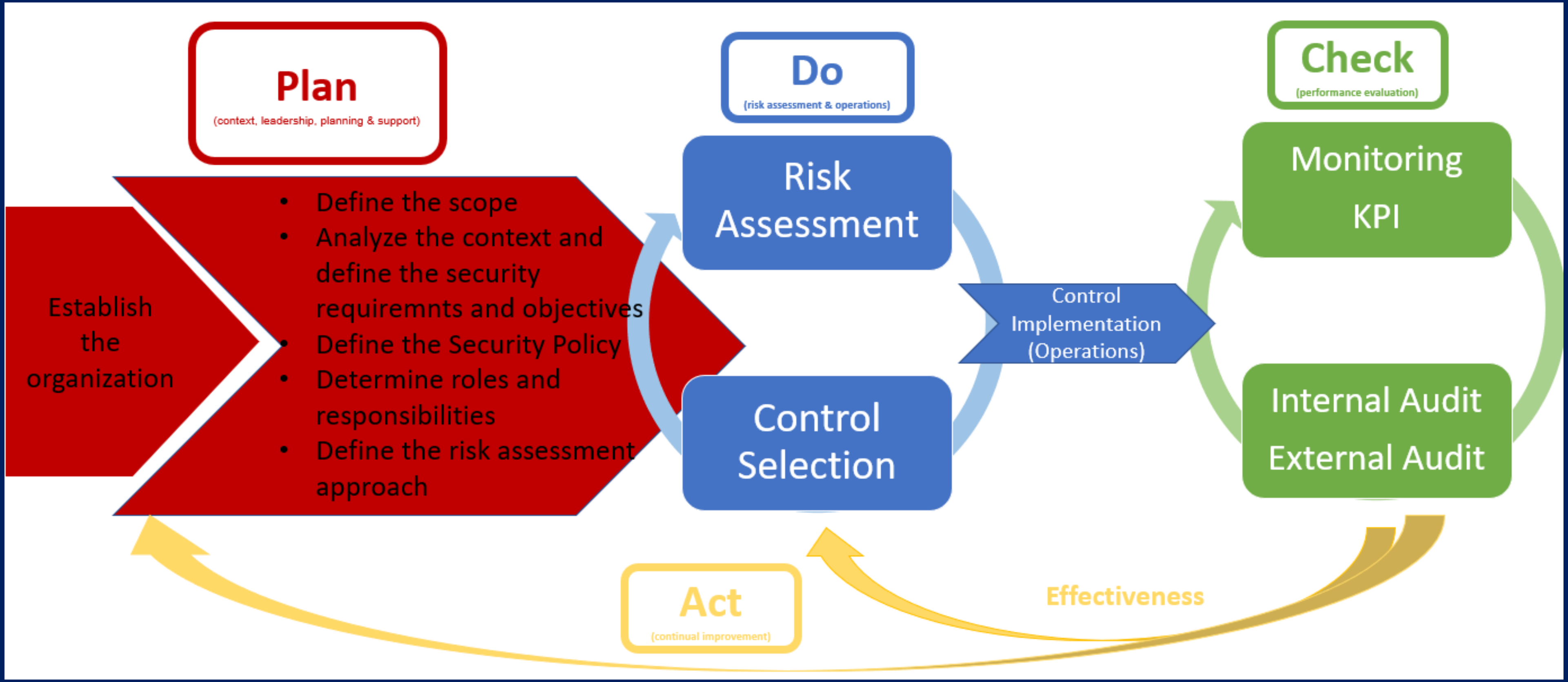
EPIC Cloud

Enhanced Privacy and Compliance Cloud – The INFN Cloud partition located at CNAF for personal and confidential data processing

- Motivation: GDPR states that Clinical and medical data (in particular the genomic ones) are personal data (fit in the Art.9 special categories of personal data).
  - Genomic data are 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 defined a region of INFN Cloud, applied specific organizational and technical security measures and certified it ISO/IEC 27001 27017 27018
  - This is the EPIC Cloud: a reference Cloud implementation for the treatment of sensitive data at INFN

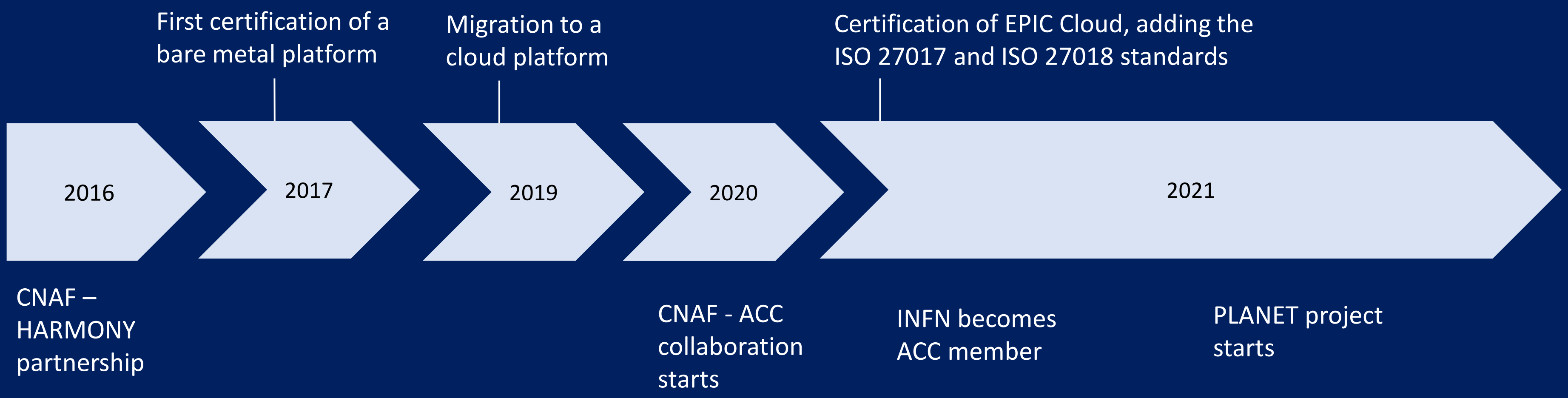
From the Data Controller side, the fact that EPIC Cloud is ISO certified is a way to demonstrate that processing is performed in accordance with GDPR

Not only technical measures, but an Information Security Management System in place


Plan
(context, leadership, planning & support)
Establish the organization
Define the scope
Analyze the context and define the security requiremnts and objectives
Define the Security Policy
Determine roles and responsibilities
Define the risk assessment approach
Do
(risk assessment & operations)
Risk Assessment
Control Selection
Control Implementation (Operations)
Check
(performance evaluation)
Monitoring
KPI
Internal Audit
External Audit
Act
(continual improvement)
Effectiveness

PLAN: Establish scope, objectives, resources, goals and plan how to undertake a risk assessment in order to deliver the desired results.
DO: Actions that need to be undertaken, i.e., risk analysis, risk treatment (implementing the necessary information security controls).
CHECK: Ensure the needs of interested parties have been met through metrics and measurements, audit and management review.
ACT: Look at results of measurements and improve where necessary, AND SO THE PROCESS STARTS AGAIN.

Certification Process



2016

CNAF – HARMONY partnership

2017

First certification of a bare metal platform

2019

Migration to a cloud platform

2020

CNAF - ACC collaboration starts

2021

Certification of EPIC Cloud, adding the ISO 27017 and ISO 27018 standards

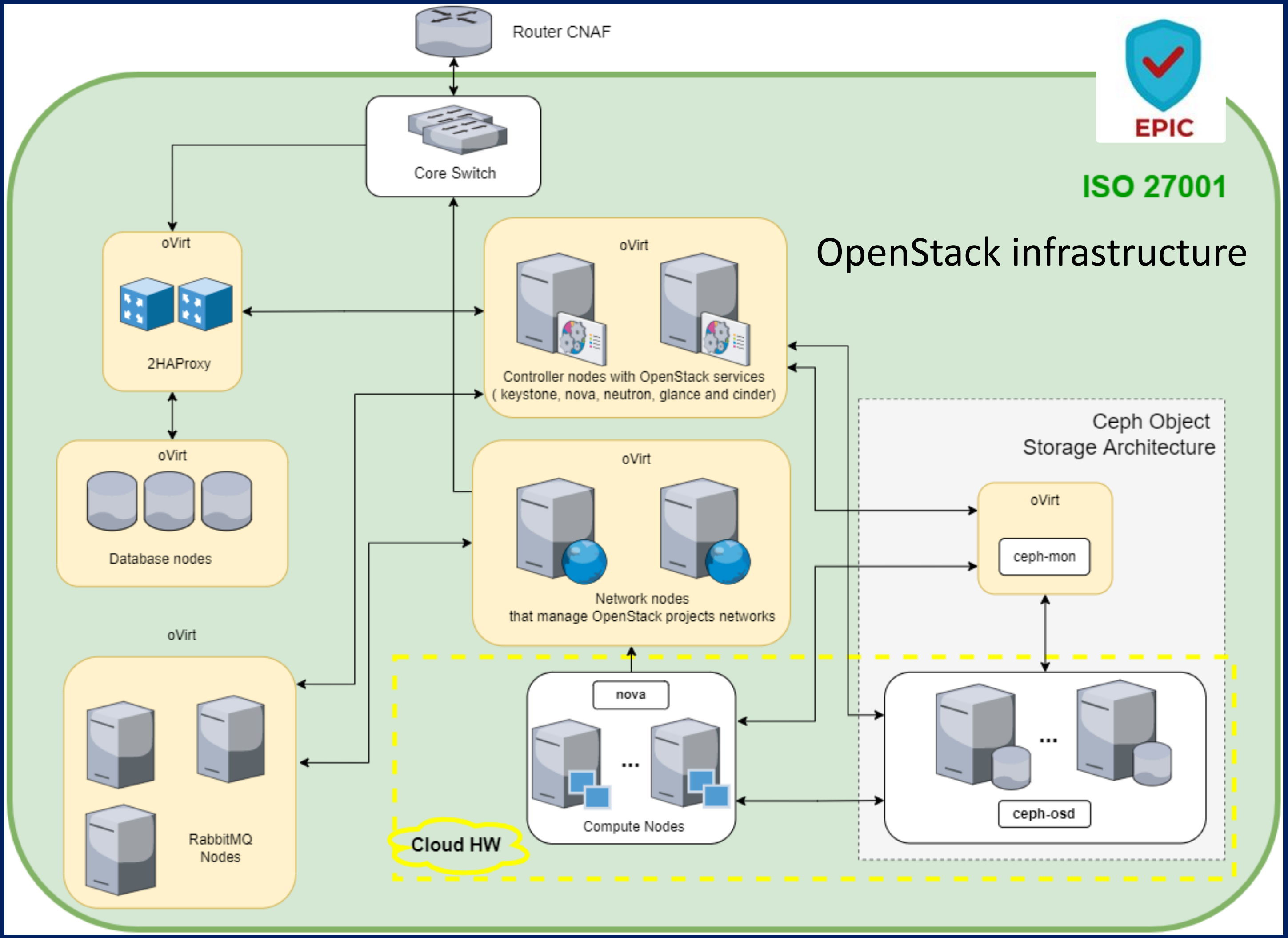
INFN becomes ACC member

PLANET project starts

- In 2017 we started with a bare-metal infrastructure that was ISO-27001 certified
- In 2019 we moved to a Cloud-based infrastructure, and EPIC Cloud was born
- In 2021 we added the cloud certifications ISO-27017 and ISO-27018

The EPIC Cloud infrastructure

- All services in HA (Availability is one aspect of security)
- It currently hosts 3 Projects
- ~7TB RAM, ~600TiB Disk raw
- Tenant/domain segregation
- Physical security (defined perimeters, controlled access to racks, TVCC)
- Network isolation from Tier-1 resources, Next-Generation Firewall (NGFW) in place



Evolution of services and related projects

- HARMONY and HARMONY Plus -> IaaS
- ACC -> mixed IaaS – SaaS model
- PLANET -> IaaS, evolving into SaaS

HARMONY / HARMONY PLUS

Healthcare Alliance for Resourceful Medicines Offensive against Neoplasms in Hematology

Two IMI-2 European Projects. HARMONY PLUS, build up on the success of HARMONY, involves 39 partners and 8 Associated Partners from 10 countries

Use of big data analytics to accelerate blood cancer research

Budget of 42.3M€ for HARMONY and 11.8M€ for HARMONY PLUS

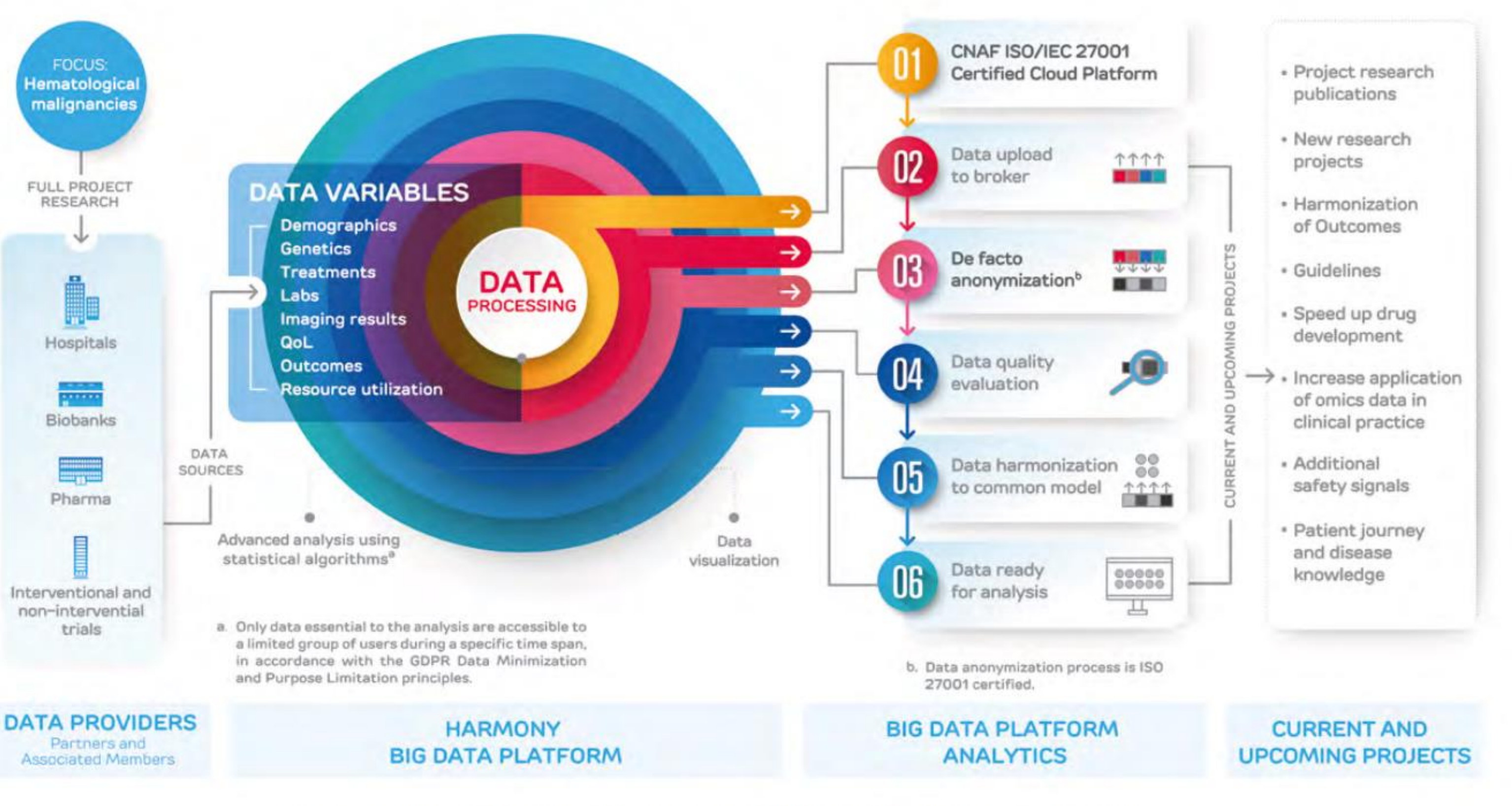
Over 100.000 patient datasets

Harmonization of datasets

Open and Standard interfaces

Our role: development of an ISO 27001-certified cloud platform to manage medical and genomic datasets in compliance with the GDPR Regulation

https://www.harmony-alliance.eu/



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

Figure 1 – Image based on an original idea by HARMONY (https://www.harmony-alliance.eu/).

HARMONY / HARMONY PLUS

Healthcare Alliance for Resourceful Medicines Offensive against Neoplasms in Hematology

- De facto anonymized data
  - GDPR does not apply, but the de-facto anonymization procedure assumes that the technology providers are ISO 27001 certified
- AgID measures (Standard and some of the Advanced)
- IaaS Service Model
  - HARMONY is the Data Controller
  - We are responsible for the infrastructure

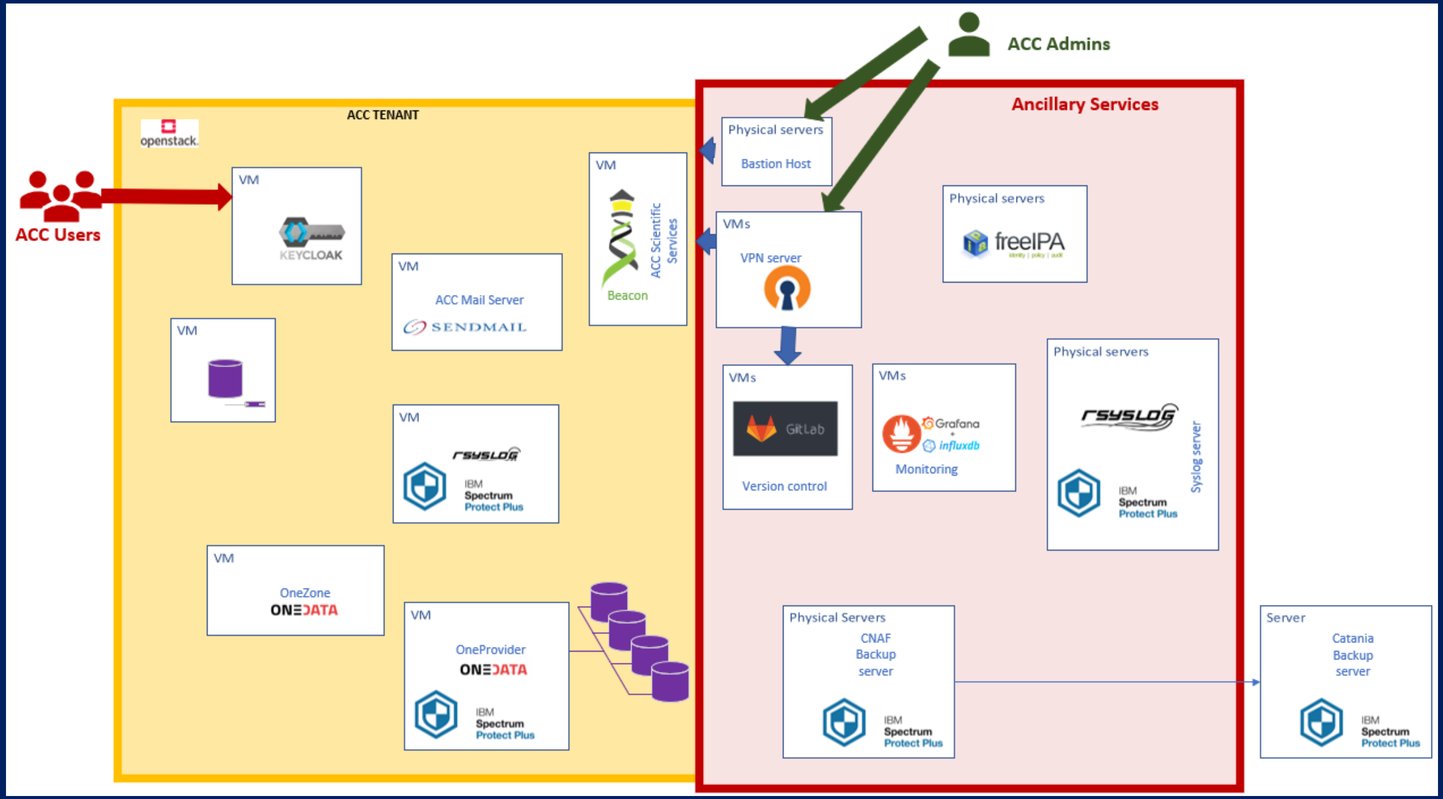
https://www.harmony-alliance.eu/

ACC

Alleanza Contro il Cancro

The National Oncology Network founded in 2002 by the Ministry of Health, joined by 51 IRCCS, ISS, AIFA, INFN and Politecnico di Milano and several patients' associations to perform translational research in the field of cancer research

- Genomic pseudonymized data
- GDPR applies
- AgID measures apply
- Italian Data Protection Authority rules apply



https://www.alleanzacontroilcancro.it/en/

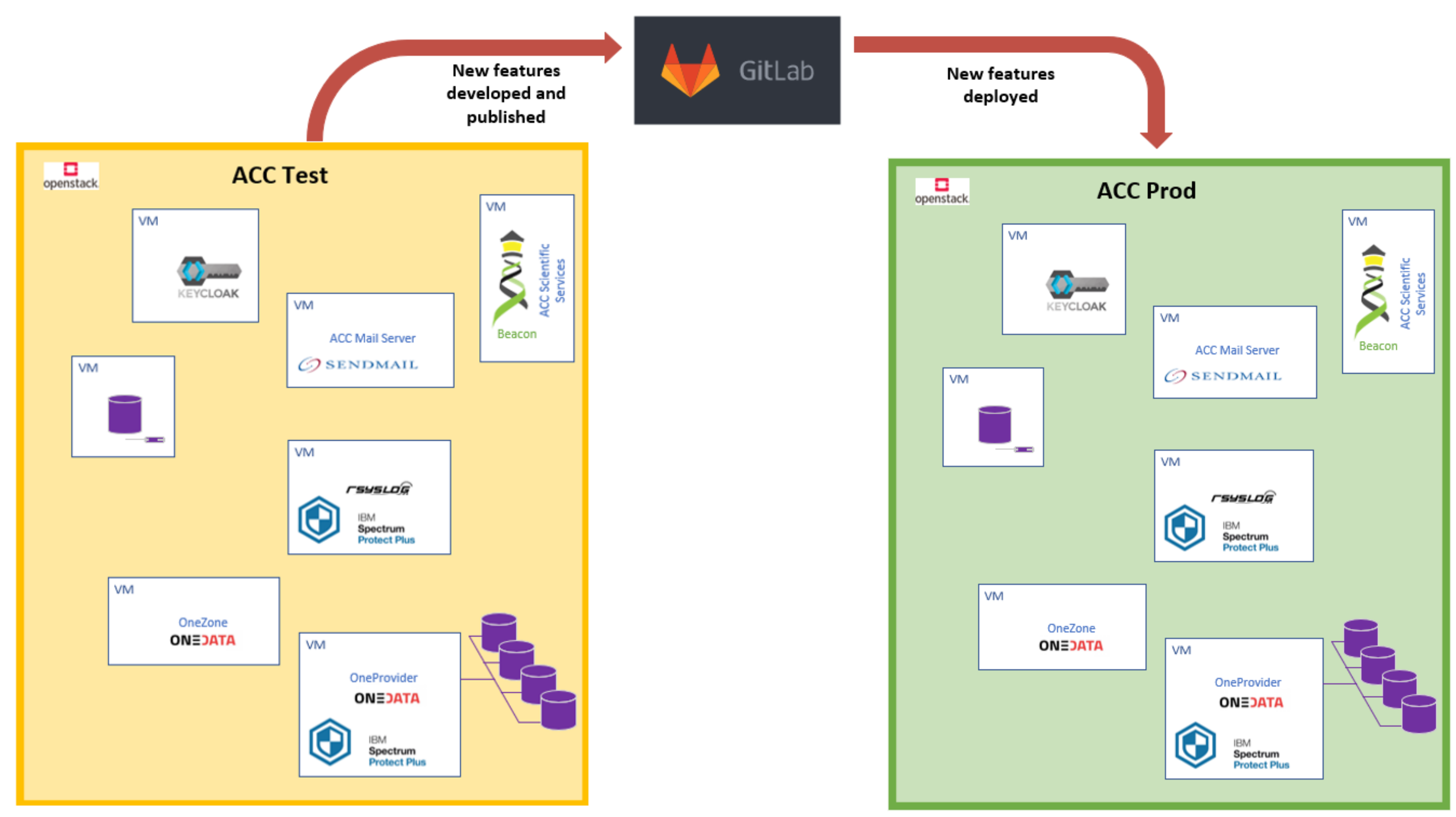
ACC - Services

Alleanza Contro il Cancro

Two separate OpenStack projects:

- **ACC-Test**, where services have been configured and tested and every change in configurations has been validated.
- **ACC**, where service configurations have been replicated and sensitive data is analyzed. Only ACC people have access to this project.

Each VM is hardened according to ISO 27001 Openscap profile



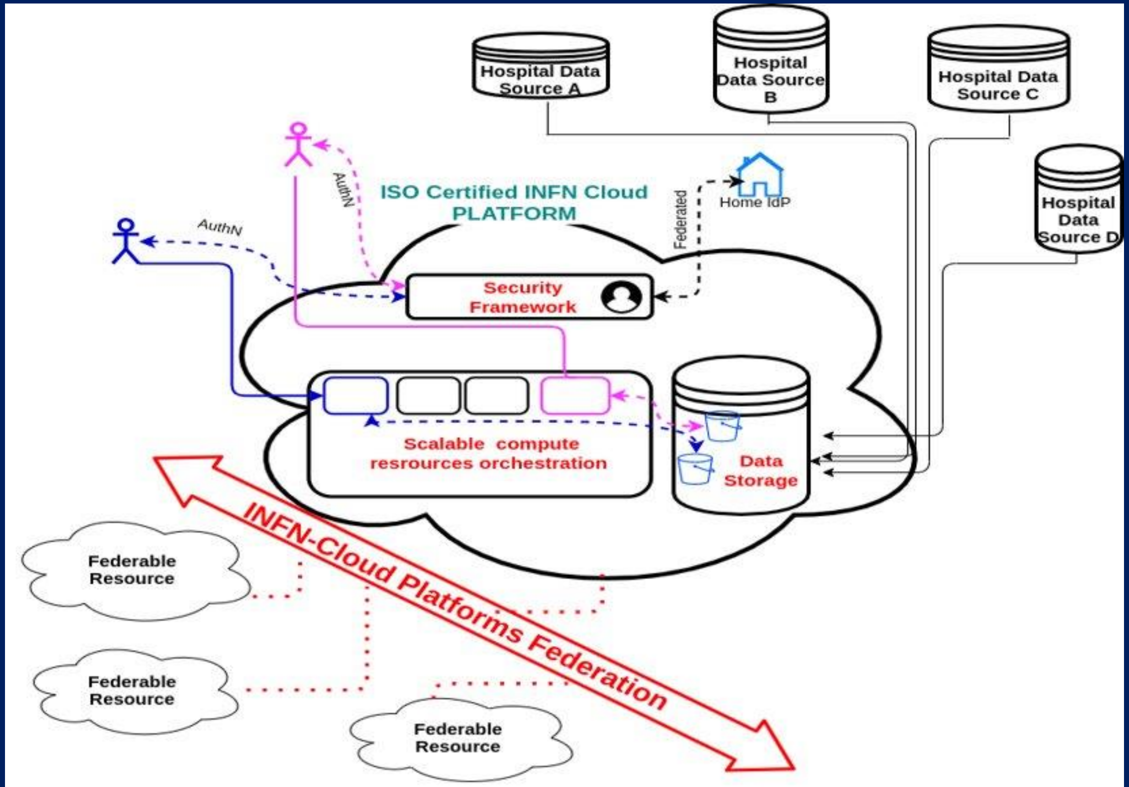
https://www.alleanzacontroilcancro.it/en/

PLANET

Pollution Lake ANalysis for Effective Therapy

An INFN/CSN5-funded research initiative aiming to implement an observational study to assess a possible statistical association between environmental pollution and Covid-19 infection, symptoms and course

- Data:
  - Pseudonymized clinical data (Covid-19 and Electronic Health Records from serveral hospitals)
  - Atmospheric data, population density, urban vs rural environment, mobility, socio-economic conditions
- Regulated by GDPR, Italian Data Protection Authority and ISS-INFN Convention
- First prototype developed on Cloud@CNAF



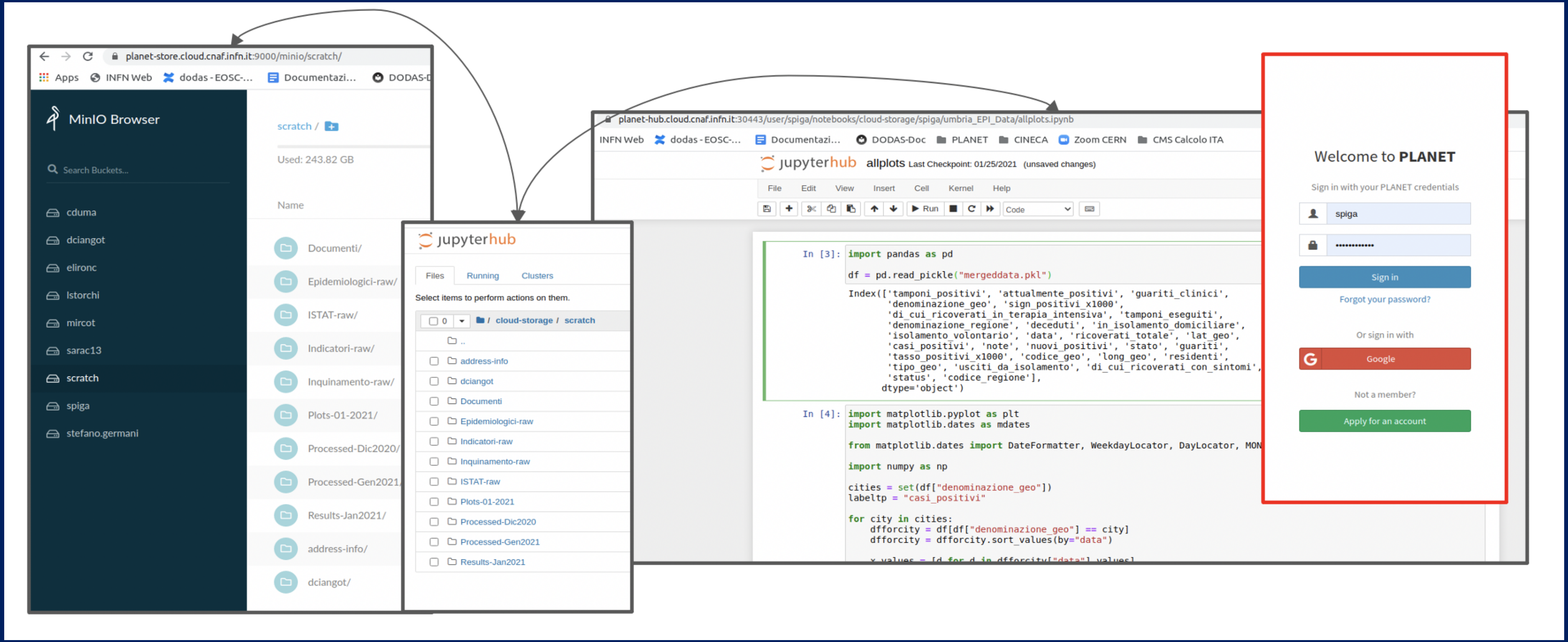
Our aim is to make this use case the first EPIC SaaS certified service (ambition/need to expand the scope of the ISO certificate to include SaaS services)

PLANET - Services

Pollution Lake ANalysis for Effective Therapy

First prototype on Cloud@CNAF:

- A JupyterHub integration with MinIO and INDIGO-IAM Authorization Server is ready



planet-store.cloud.cnaf.infn.it:9000/minio/scratch/

MinIO Browser

Search Buckets...

cduma
dciangot
elironc
lstorchi
mircot
sarac13
scratch
spiga
stefano.germani

scratch /

Used: 243.82 GB

Name

Documenti/
Epidemiologici-raw/
ISTAT-raw/
Indicatori-raw/
Inquinamento-raw/
Plots-01-2021/
Processed-Dic2020/
Processed-Gen2021/
Results-Jan2021/
address-info/
dciangot/

jupyterhub

Select items to perform actions on them.

/ cloud-storage / scratch

..
address-info
dciangot
Documenti
Epidemiologici-raw
Indicatori-raw
Inquinamento-raw
ISTAT-raw
Plots-01-2021
Processed-Dic2020
Processed-Gen2021
Results-Jan2021

planet-hub.cloud.cnaf.infn.it:30443/user/spiga/notebooks/cloud-storage/spiga/umbria_EPI_Data/allplots.ipynb

jupyterhub allplots Last Checkpoint: 01/25/2021 (unsaved changes)

```
In [3]: import pandas as pd

        df = pd.read_pickle("mergeddata.pkl")

        Index(['tamponi_positivi', 'attualmente_positivi', 'guariti_clinici',
               'denominazione_geo', 'sign_positivi_x1000',
               'di_cui_ricoverati_in_terapia_intensiva', 'tamponi_eseguiti',
               'denominazione_regione', 'deceduti', 'in_isolamento_domiciliare',
               'isolamento_volontario', 'data', 'ricoverati_totale', 'lat_geo',
               'casi_positivi', 'note', 'nuovi_positivi', 'stato', 'guariti',
               'tasso_positivi_x1000', 'codice_geo', 'long_geo', 'residenti',
               'tipo_geo', 'usciti_da_isolamento', 'di_cui_ricoverati_con_sintomi',
               'status', 'codice_regione'],
              dtype='object')

In [4]: import matplotlib.pyplot as plt
        import matplotlib.dates as mdates

        from matplotlib.dates import DateFormatter, WeekdayLocator, DayLocator, MON

        import numpy as np

        cities = set(df["denominazione_geo"])
        labeltp = "casi_positivi"

        for city in cities:
            dfforcity = df[df["denominazione_geo"] == city]
            dfforcity = dfforcity.sort_values(by="data")
```

Welcome to PLANET

Sign in with your PLANET credentials

spiga

••••••••••••

Sign in

Forgot your password?

Or sign in with

Google

Not a member?

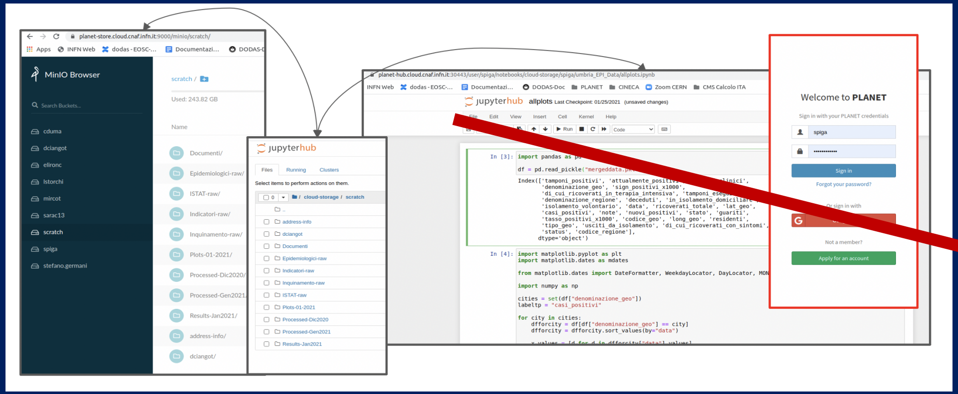
Apply for an account

PLANET - Services

Pollution Lake ANalysis for Effective Therapy

First prototype on Cloud@CNAF:

- A JupyterHub integration with MinIO and INDIGO-IAM Authorization Server is ready, embedded posix access is provided



planet-store.cloud.cnaf.infn.it:9000/minio/scratch/

planet-hub.cloud.cnaf.infn.it:30443/user/spiga/notebooks/cloud-storage/spiga/umbria_EPI_Data/allplots.ipynb

```
In [3]: import pandas as pd

        df = pd.read_pickle("mergeddata.pk

        Index(['tamponi_positivi', 'attualmente_positivi', 'clinici',
               'denominazione_geo', 'sign_positivi_x1000',
               'di_cui_ricoverati_in_terapia_intensiva', 'tamponi_esegu
               'denominazione_regione', 'deceduti', 'in_isolamento_domiciliare',
               'isolamento_volontario', 'data', 'ricoverati_totale', 'lat_geo',
               'casi_positivi', 'note', 'nuovi_positivi', 'stato', 'guariti',
               'tasso_positivi_x1000', 'codice_geo', 'long_geo', 'residenti',
               'tipo_geo', 'usciti_da_isolamento', 'di_cui_ricoverati_con_sintomi',
               'status', 'codice_regione'],
              dtype='object')

In [4]: import matplotlib.pyplot as plt
        import matplotlib.dates as mdates

        from matplotlib.dates import DateFormatter, WeekdayLocator, DayLocator, MON

        import numpy as np

        cities = set(df["denominazione_geo"])
        labeltp = "casi_positivi"

        for city in cities:
            dfforcity = df[df["denominazione_geo"] == city]
            dfforcity = dfforcity.sort_values(by="data")

            x_values = [d for d in dfforcity["data"].values]
```

Welcome to **PLANET**

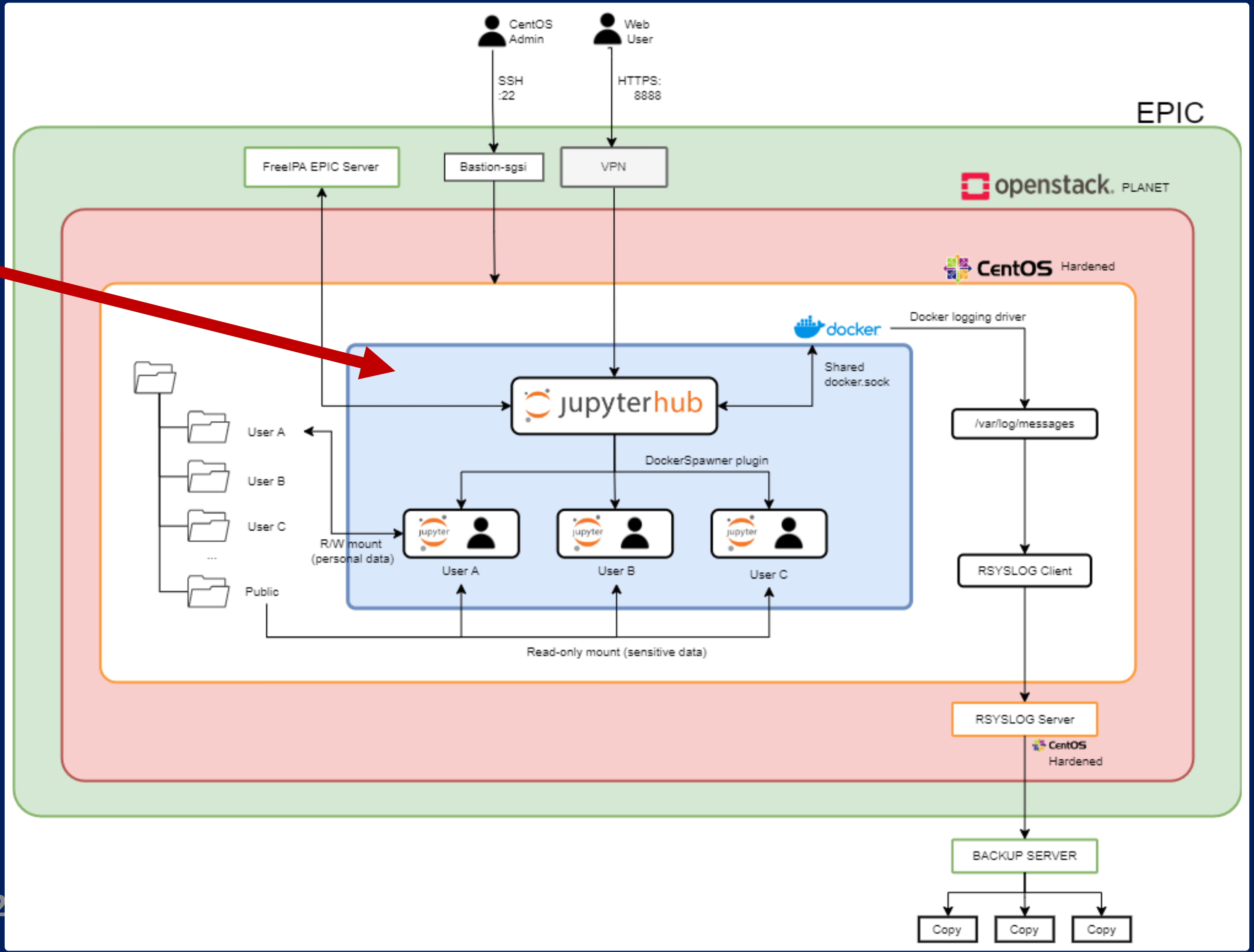
Sign in with your PLANET credentials

spiga

•••••••••••••

Partially ported within EPIC Cloud

- *Working on Minio porting and second prototype*

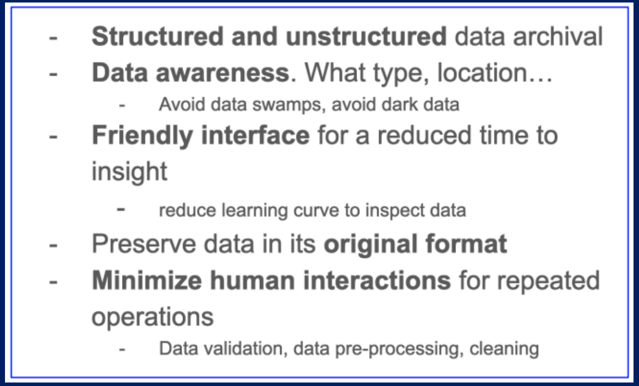


PLANET - Services

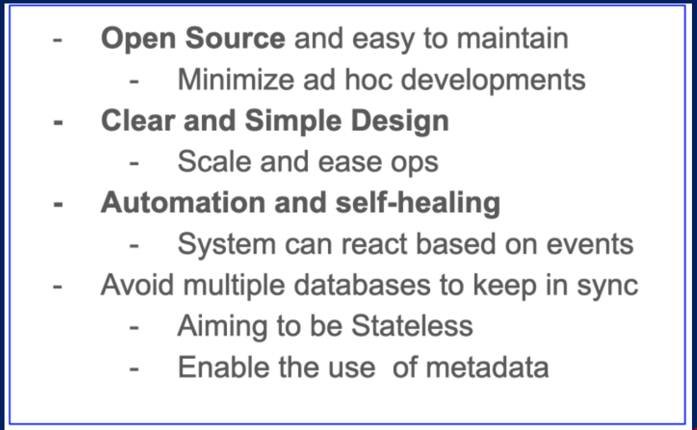
Pollution Lake ANalysis for Effective Therapy

Second prototype responding to final objectives of the project

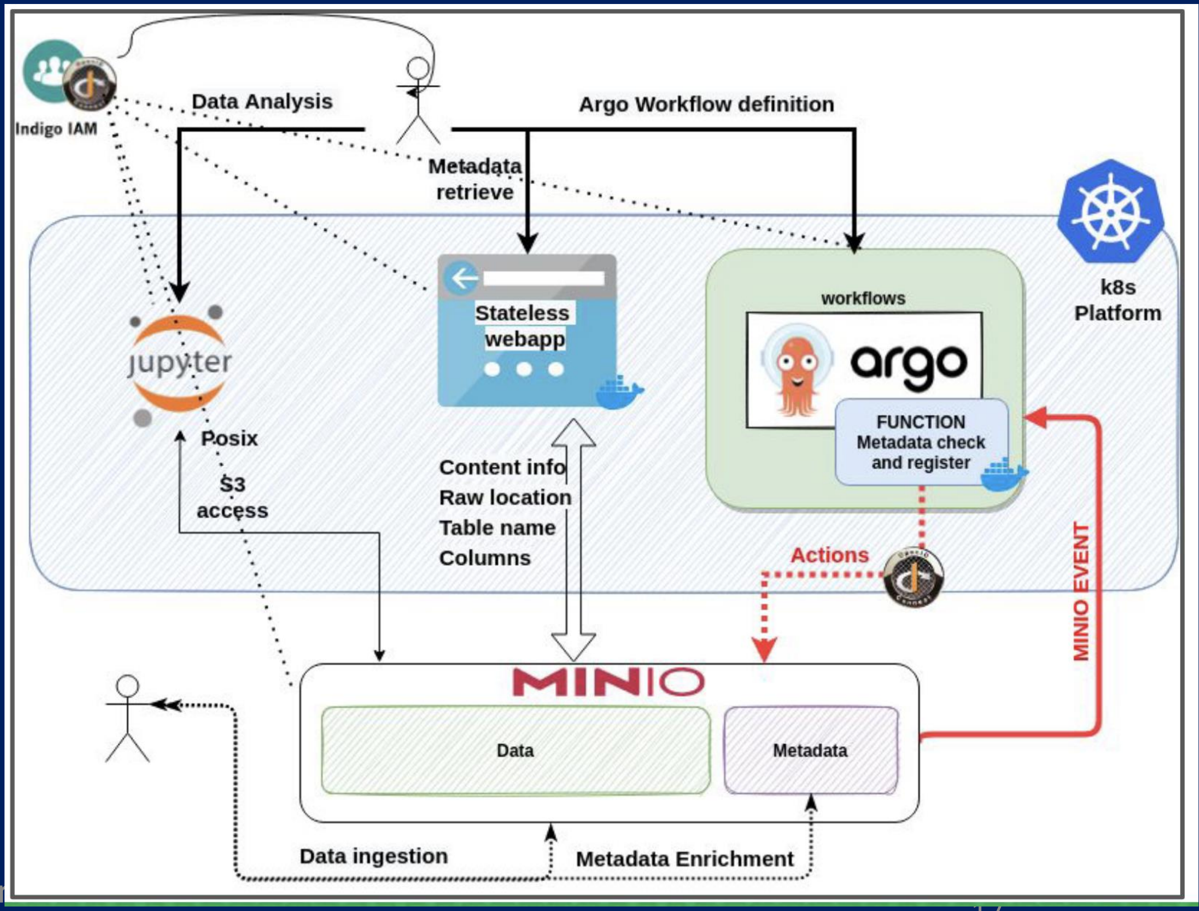
- Use PLANET project as a case study to develop a generic, reusable and extendible platform in order to cope with:



- **Structured and unstructured** data archival
- **Data awareness**. What type, location…
    - Avoid data swamps, avoid dark data
- **Friendly interface** for a reduced time to insight
    - reduce learning curve to inspect data
- Preserve data in its **original format**
- **Minimize human interactions** for repeated operations
    - Data validation, data pre-processing, cleaning



- **Open Source** and easy to maintain
    - Minimize ad hoc developments
- **Clear and Simple Design**
    - Scale and ease ops
- **Automation and self-healing**
    - System can react based on events
- Avoid multiple databases to keep in sync
    - Aiming to be Stateless
    - Enable the use of metadata



Slide courtesy to Daniele Spiga, PI PLANET

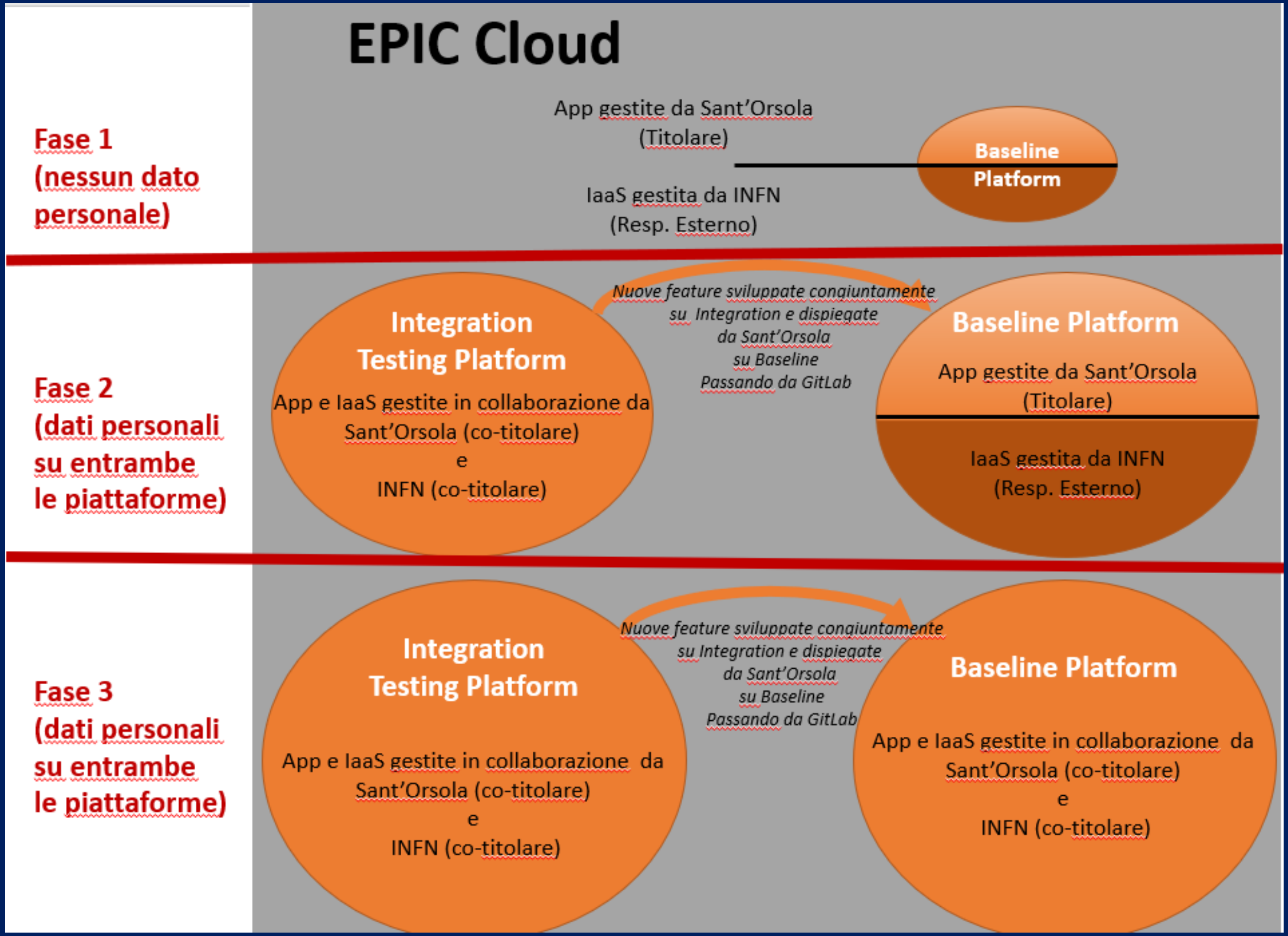
Health Big Data (HBD)

- Health Big Data is a 10-years project funded by the Italian Ministry of Health aiming at the creation of a federated and integrated big data platform for the health research at national level
  - 4 research netwoks: ACC, RIN, Cardio, IDEA
  - Research objectives: preventing diseases, personalizing treatments, improving the quality of life of patients
  - Budget: 55M€
- INFN is in the managing bodies of HBD. Its tasks include the definition of an integrated national platform and contributions to several Work Packages.
- The HBD architecture will provide solutions for several scenarios:
  1. Central harvesting of data collected remotely
  2. Edge anonymization, followed by central ingestion and analysis of data
  3. Edge feature extraction, followed by central ingestion and analysis of features
  4. Federated learning based on edge-based training, followed by publishing of the trained methods and by inference performed either centrally or at other edge locations

https://www.alleanzacontroilcancro.it/en/progetti/health-big-data/ https://doi.org/10.1016/j.ejmp.2021.10.005

INFN-IRCCS Sant’Orsola Collaboration

- Joint research agreement with the following objectives:
  - secure applications for genomic data
  - GPU -based solutions for genomic analysis methods
  - federated and integrated cloud platforms for homics data
  - adaptation of genomic pipelines to cloud and datalake architectures based on microservices
  - Integration of homics data and other clinical data like Electronic Medical Records (EMR)



EPIC Cloud

Fase 1
(nessun dato
personale)

App gestite da Sant’Orsola
(Titolare)

IaaS gestita da INFN
(Resp. Esterno)

Baseline
Platform

Fase 2
(dati personali
su entrambe
le piattaforme)

Integration
Testing Platform

App e IaaS gestite in collaborazione da
Sant’Orsola (co-titolare)
e
INFN (co-titolare)

Nuove feature sviluppate congiuntamente
su Integration e dispiegate
da Sant’Orsola
su Baseline
Passando da GitLab

Baseline Platform

App gestite da Sant’Orsola
(Titolare)

IaaS gestita da INFN
(Resp. Esterno)

Fase 3
(dati personali
su entrambe
le piattaforme)

Integration
Testing Platform

App e IaaS gestite in collaborazione da
Sant’Orsola (co-titolare)
e
INFN (co-titolare)

Nuove feature sviluppate congiuntamente
su Integration e dispiegate
da Sant’Orsola
su Baseline
Passando da GitLab

Baseline Platform

App e IaaS gestite in collaborazione da
Sant’Orsola (co-titolare)
e
INFN (co-titolare)

Conclusions

- The request for trusted Cloud solutions in health-related research fields is growing very fast, and INFN is at the forefront in many of these initiatives
- In order to fulfill the requirements expressed by the health research community we have built up a secure cloud region (EPIC Cloud) inside the INFN Cloud infrastructure and certified it ISO/IEC 27001 27017 27018
- EPIC Cloud is a reference architecture that has attracted national and International interest
- We started offering a IaaS platform, but we aim to evolve the Information Security Management System to include all the service models, including PaaS and SaaS
- Thanks to increasingly stronger collaborations with the health research community, we are evolving our service portfolio to include a growing number of applications and services, driven by use-cases needs
- We intend to extend this architecture to other INFN Cloud sites (Bari and possibly Catania)
- PNRR initiatives (not mentioned here) heavily rely on the availability of these INFN-backed solutions

Click to add text

Many thanks to all contributors

Alessandro Costantini, Andrea Chierici, Arianna Carbone, Cristina Vistoli, Daniele Cesini, Daniele Spiga, Davide Salomoni, Diego Ciangottini, Diego Michelotto, Elena Corni, Enrico Fattibene, Jacopo Gasparetto, Lorenzo Chiarelli, Luca dell'Agnello, Luigi Scarponi, Stefano Dal Pra, Stefano Longo, Stefano Zani, Patrizia Belluomo, Vincenzo Ciaschini