Open-source and cloud-native solutions for managing and analyzing heterogeneous and sensitive clinical Data

Daniele Spiga INFN-Perugia
spiga@pg.infn.it
On behalf of PLANET Team

This Talk

Introduction to the PLANET project
- The use case

A Cloud native platform for heterogeneous data
- Requirements and technical goal
- Architecture overview

Status and next step
- Enhancement toward EPIC-Cloud integration

Summary

A highly multidisciplinary Team, Many diverse competences are needed.

- Pasquale Lubrano
- Davide Salomoni
- Mirco Tracolli
- Diego Ciangottini
- Giuseppe Ambrosio
- Fabrizio Stracci
- Paolo Reboldi
- Cristina Duma
- Alessandro Costantini
- Sara Cutini
- Diego Ciangottini
- Barbara Martelli
- Giusy Sergi
- Jacopo Gasparetto
- Loriano Storchi
- Elisabetta Ronchieri

The PLANET Project in a nutshell

PLANET (**P**ollution **L**ake **AN**alysis for **E**ffective **T**herapy) a INFN-funded research initiative aiming at developing

- An observational (ecological) study to evaluate the association between air pollution and Covid19, taking care of a variety of components that are supposed to influence rates of SARS-COV-2 diffusion and infection
  - A synergy between INFN and epidemiological and medical knowledge University of Perugia

The study is based on the hypothesis that pollution may contribute to the spread and/or the severity of COVID-19, through 2 possible mechanisms:

- **Acute**: microparticles derived from fossil fuels might act as airborne carriers of virus;
- **Chronic**: exposure to microparticles and chemical pollutants might cause chronic lung injury, exacerbating the consequences of viral infection.

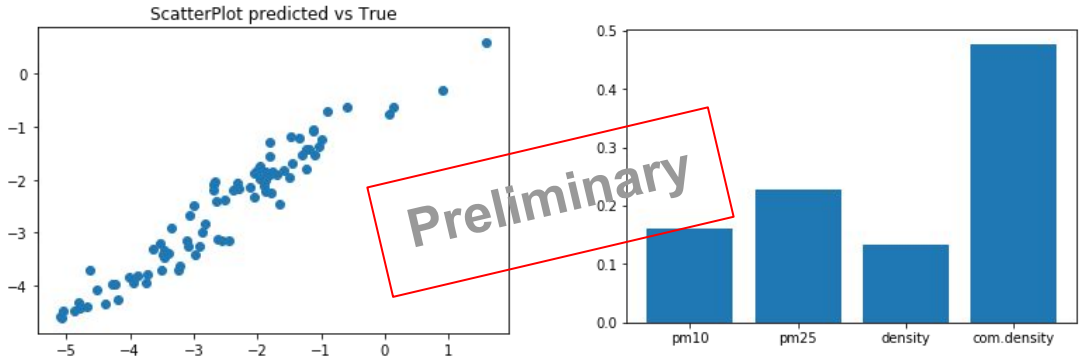
The Analysis Strategy

A key element of PLANET is to include a wide variety of components that are expected to influence rates of SARS-COV-2 diffusion and infection

- atmospheric data, population density, urban vs rural environment, mobility, socio-economic conditions…
- take into account also the fact that severity COVID-19 disease and deaths are also influenced by many variables (age, gender, comorbidities, frailty, etc..).

Current focus on Feature importance evaluation: assign a score to input features based on how useful they are at predicting a target variable (Covid19)

- Models such as: Random Forest; k-nearest neighbors;



INFN key expertise

INFN is a pioneer in the design and implementation of large-scale computing infrastructures and applications

- Primarily developed to meet the needs of the latest generations of high energy physics (HEP) experiments
- **now rapidly extending to other communities.** INFN@PLANET:
  - At this conference: ML_INFN; CYGNO; INFN-Cloud

- **Data Analysis**
  To develop models and to implement statistical data analysis
- **Data Curation and Data Management**
  Organization and integration of data collected from heterogeneous sources; Enabling FAIR (Findable, Accessible, Interoperable, Reusable) data repositories
- **Integrated and Certified Computing Infrastructure**
  A ISO 27001 / 27017 Certified Data-Lake to manage confidential data (ISS, Hospital, ASL)
  An-easy-to-use computing platform fully integrated with the **INFN-Cloud national infrastructure**

Heterogeneous DataSet

Raw data collected so far by the project:

- Different types
- Different schemas
  - Data originally acquired with a different purpose from that for which they are used here

What metter here is mainly the variety, partially also the volume (i.e. pollutants data)

- In principle PLANET could manage data in "silos" or "Data Warehouses", but it is considered strategic to be proactive, moving towards more advanced approaches: the Data Lake is one of these

Category	Source	Area	Period	Frequency	Format
Meteorological	Copernicus Atmosphere Monitoring Service (CAMS)	Italy	1 January 2017 - 31 December 2020	daily	Numerical Control (nc)
Atmosphere	Copernicus Atmosphere Monitoring Service (CAMS)	Italy	25 December 2017 - 31 July 2021	hourly	Numerical Control (nc)
Atmosphere	ARPA Umbria	Umbria, Italy	1 June 2017 - 31 December 2020	hourly	Numerical Control (nc)
Sociological	ISTAT	Italy	2020	yearly	Comma separated values (csv)
Demographic	ISTAT	italy	2020	yearly	Comma separated values (csv)
COVID-19	Istituto Superiore di Sanità **ISS**	Italy	2020 - 2022	daily	Comma separated values (csv)
COVID-19	ASL VT			daily	Comma separated values (csv)
COVID-19	Regione Marche	Marche Region	2020 - 2022	daily	Comma separated values (csv)
COVID-19	Regione Liguria	Liguria Region	2020 - 2021	weekly	Comma separated values (csv)

Main requirements and objectives

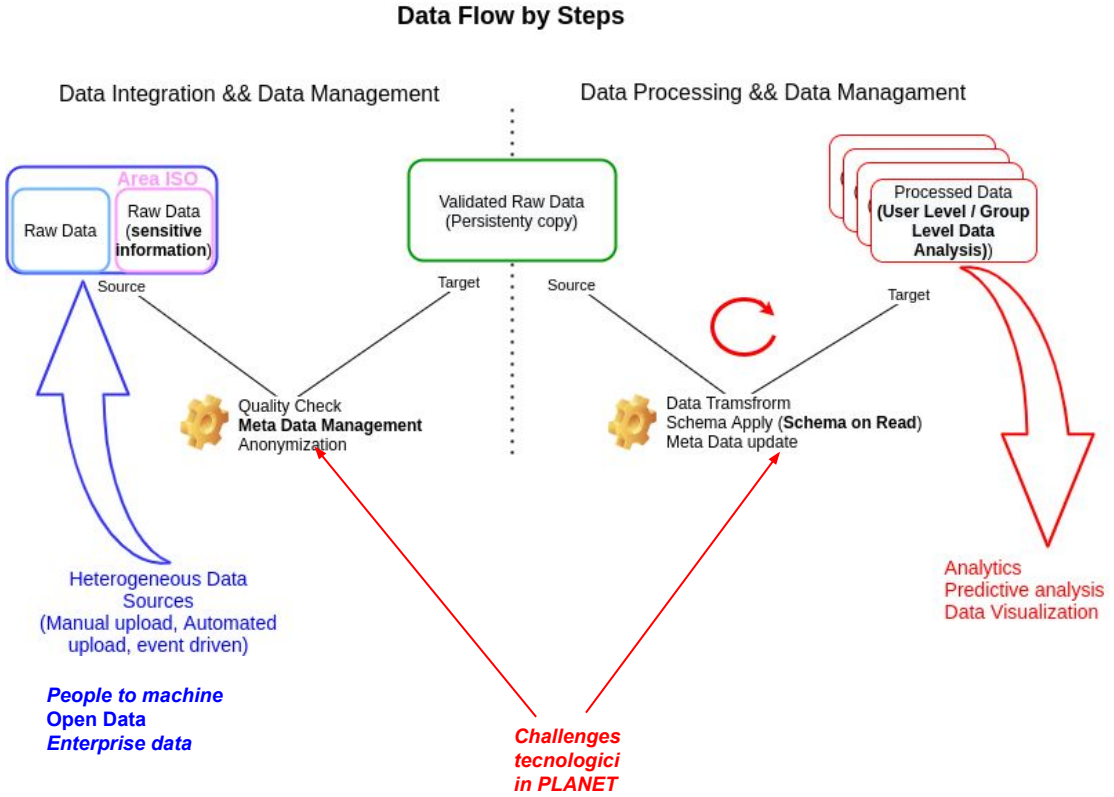
Use PLANET project as a case study to develop a generic, reusable and extendible platform in order to cope with:

- **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

Toward a Data Lake model

- **A data storage environment allow to keep data in their native format.**
  - They remain in this condition until it is necessary to define a structure
- **Enable automated data validation (and organization)**
- **Data scheme is defined at the time of analysis and not at the time of archiving**



A Storage Centric Solution

Minio Storage solution has been chosen as core component to build our the cloud native solution

- S3 compliance, Powerful WebUI
- Proven scalability
- Native integration with AWS STS credentials, external OIDC IdP's

Bucket notifications allow to send events to supported external services on certain object or bucket events

Metadata: writes and operates on metadata and data together to provide granularity at the level of individual objects

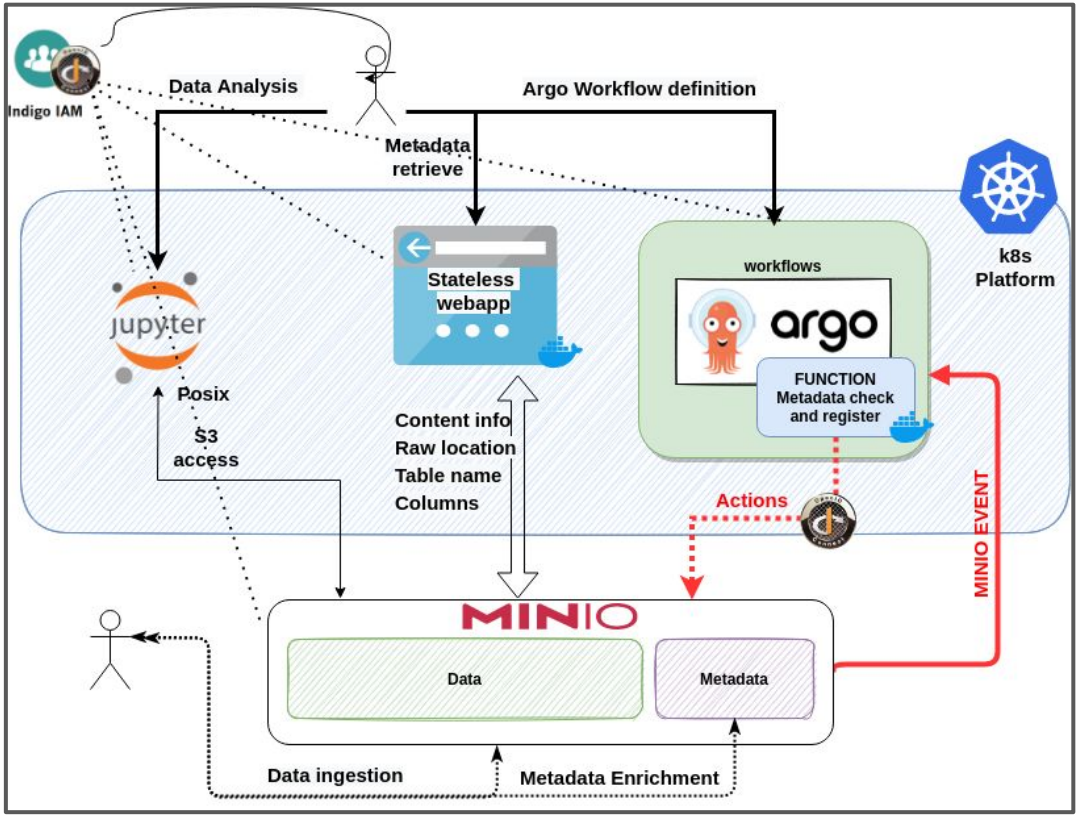
- Support for user defined metadata

Support for **customizable authorization policies** with OpenPolicyAgent

Architectural Schema and technologies

Connecting the dots..

- Metadata management and enrichment
- Stateless approach
- Automation (events and workflows), and self-healing
- Highly integration between data and compute
- Scope based AuthN/Z



Metadata enrichment

Uploader: a lightweight MinIO go client (CLI/GUI) utility to upload files and enrich related metadata

GUI uploader

MinIO stored metadata (example)

MinIO metadata

User defined metadata **enrichment**

```
{
  ETag:124f5ff208b5bcfd1c097014498a07a9
  Key:umbria-54050-7
  LastModified:2021-10-08 07:53:42 +0000 UTC
  Size:16668
  ContentType:json
  Expires:0001-01-01 00:00:00 +0000 UTC
  UserMetadata:map[
      Info_field:coronavirus
      Info_region:Umbria
      Info_state:Italy
      Len:15
      Result_fields:[note denominazione_regione residenti data codice_geo
                     lat_geo guariti guariti_clinici tasso_positivi_x1000
                     deceduti stato long_geo tamponi_eseguiti in_isolamento_domiciliare
                     isolamento_volontario tipo_geo tamponi_positivi casi_positivi
                     denominazione_geo sign_positivi_x1000 attualmente_positivi
                     nuovi_positivi di_cui_ricoverati_con_sintomi
                     codice_regione usciti_da_isolamento ricoverati_totale
                     di_cui_ricoverati_in_terapia_intensiva status]
      Result_key:results
      Type:json]
      UserTags:map[]
      UserTagCount:1
      Owner:{DisplayName: ID:}
      Grant:[]
      StorageClass:
      IsLatest:false
      IsDeleteMarker:false
      VersionID:
      ReplicationStatus:
      Expiration:0001-01-01 00:00:00 +0000 UTC
      ExpirationRuleID: Err:<nil>
}
```

Automated data validation through Argo Workflow

Every file uploaded to the PLANET DataLake is automatically validated

The EventSource: Each upload generate a Minio EVENT

```
spec:
  minio:
    example:
      endpoint: 'planet-store.cloud.cnaf.infn.it:9000'
      bucket:
        name: demo-raw
      accessKey:
        name: artifacts-minio
        key: accesskey
      secretKey:
        name: artifacts-minio
        key: secretkey
      events:
        - 's3:ObjectCreated:Put'
```

Argo Sensor detect the event and **trigger the validation**

```
triggers:
  - template:
      name: minio-workflow-trigger
      k8s:
        source:
          resource:
            apiVersion: argoproj.io/v1alpha1
            kind: Workflow
            metadata:
              generateName: artifact-workflow-2-
              namespace: argo-events
            spec:
              entrypoint: hook
              templates:
                - container:
                    args:
                      - THIS_WILL_BE_REPLACED
                    command:
                      - hook
                    env:
                      - name: ACCESSKEYID
                        value: admin-creds
                      - name: ENDPOINT
                        value: 'planet-store.cloud.cnaf.infn.it:9000'
                      - name: SECRETACCESSKEY
                        value: '223^sU#O!ksss'
                    image: 'dodasts/planet-demo-hook:v0'
                    imagePullPolicy: Always
                  name: hook
```

A customizable validation function is automatically executed:

If metadata OK then:

tell Minio to move data to the validated bucket

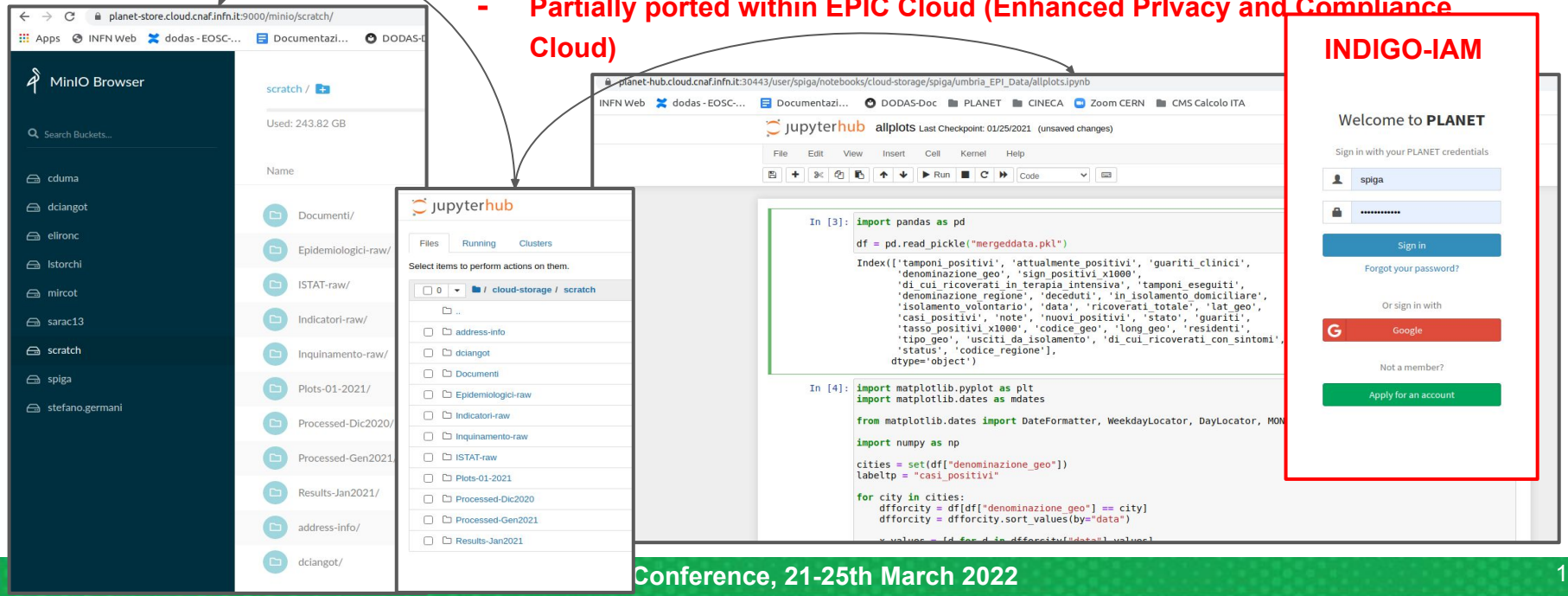
else

tell MinIO to move data to triage && notify

fi

Integrated Analysis Platform

- A JupyterHub integration with MinIO and INDIGO-IAM Authorization Server is ready: embedded posix access is provided
  - Partially ported within EPIC Cloud (Enhanced PrIvacy and Compliance Cloud)

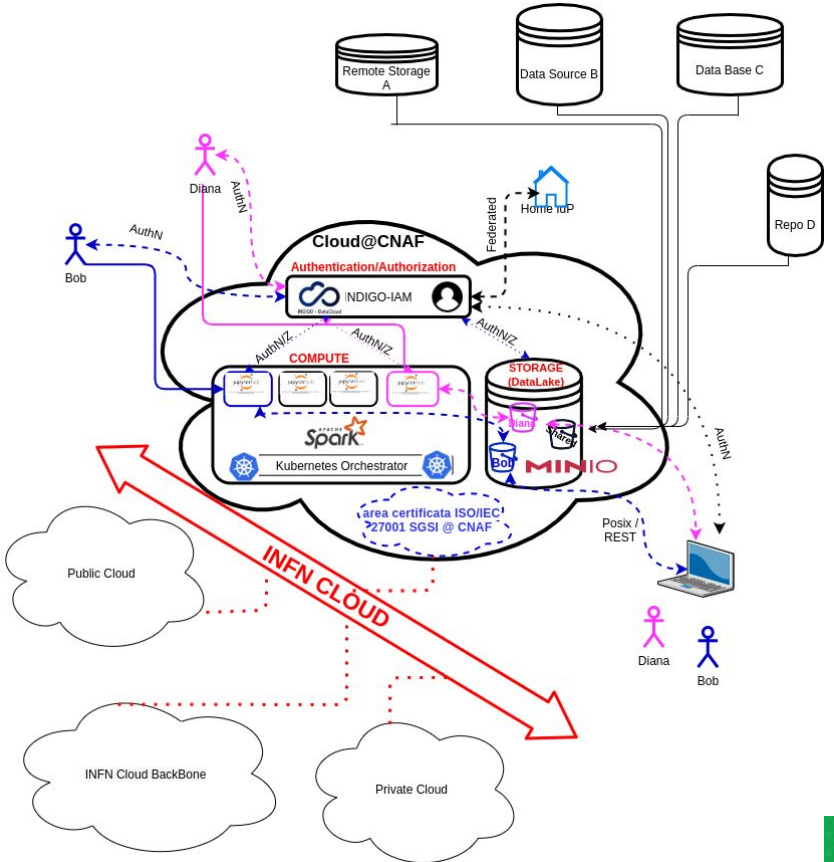


Future evolution: The vision

To integrate and make available an "open" and generic and reusable platform EPIC compliant with the INFN Cloud ecosystem [1]

- Integration of data from multiple information sources
- Processing (descriptive, predictive and real time analysis)
- Federation of computing resources through INFN-Cloud enabling technology

[1] see here



Summary and Conclusion

The PLANET project has been presented. A Multidisciplinary project aiming at analyzing heterogeneous and sensible data

- A generic cloud native platform for managing and enabling data analysis has been successful prototyped
- Highly centred on Metadata management and enrichment

The event based system for workflow automation represent a key element

- Further enhancement is expected with new used cases

There is an ongoing activity is moving the PLANET platform inside the EPIC-Cloud environment.

- To harden the current platform applying the technical measures required to improve the security of the whole system.