

# IMPACT-Data Biomedical Cloud

An initial iteration for a federated virtual computing environment in the context of Precision Medicine in Spain using Galaxy

**María Chavero Díez**

Spanish National Bioinformatics Institute, ELIXIR-ES  
Barcelona Supercomputing Center

*October, 4, 2022*



# IMPACT

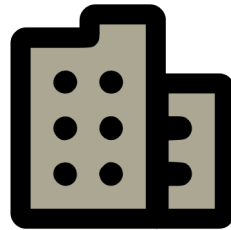
## Spanish Precision Medicine Infrastructure associated with Science and Technology

Project that aims to impulse precision medicine within the Spanish National Health System



**15**

**Autonomous  
Communities**



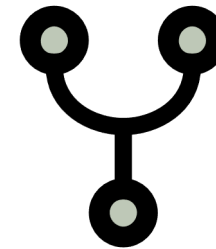
**47**

**Participants**



**59**

**Research Groups**

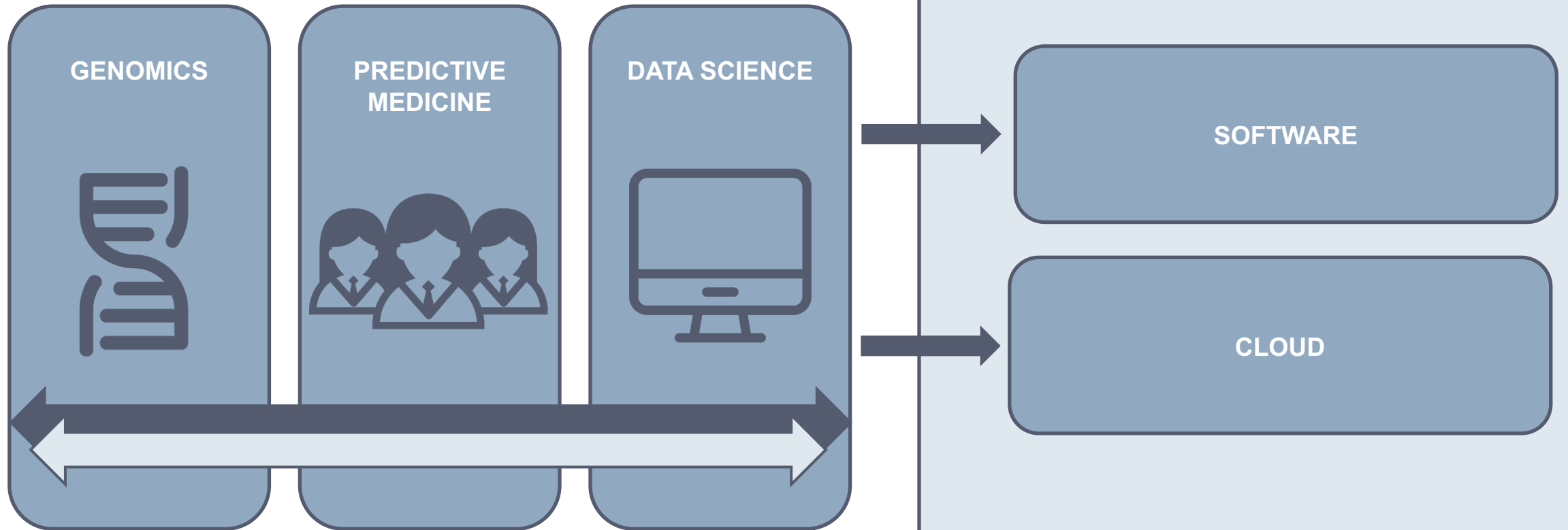


**32**

**Associated  
Entities**

# IMPACT

IMPACT | Data 

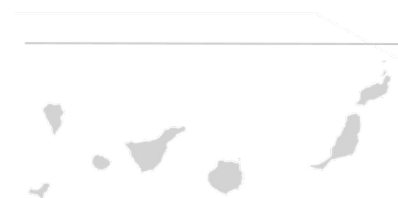


INSTITUTION	FUNCTION	INITIAL RESOURCES
BSC	Central Node + Computational Node	Cloud OpenStack. 5VMs 12VCPUs, 46GB RAM, 10TB Storage
CNIC	Computational Node + Data	Cloud OpenStack. 6VMs, 10VCPUs, 40GB RAM, 1TB storage.
IMIB	Computational Node + Data	1 VM, 16VCPUs, 64GB RAM, 1TB storage
NASERTIC	Computational Node + Data	Cluster. 48VCPUs, 256GB RAM, 5TB storage.

# WHO?



# HOW?

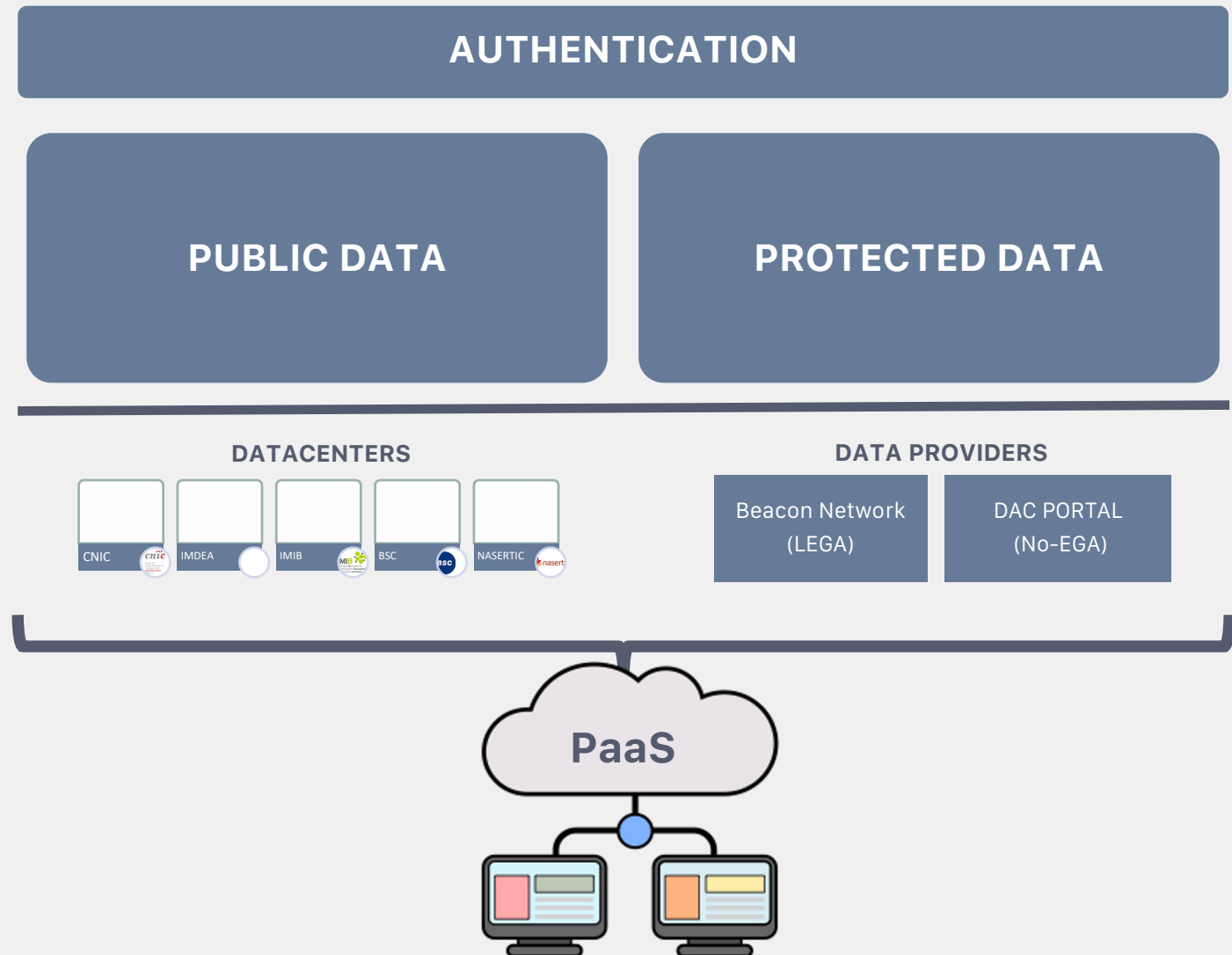


1. Barcelona Supercomputing Center - BSC. INB/ELIXIR-ES (Cataluña)
2. Centro Nacional de Investigaciones Cardiovasculares - CNIC (Madrid)
3. Instituto Murciano de Investigación Biosanitaria - IMIB (Murcia)
4. NASERTIC (Navarra)

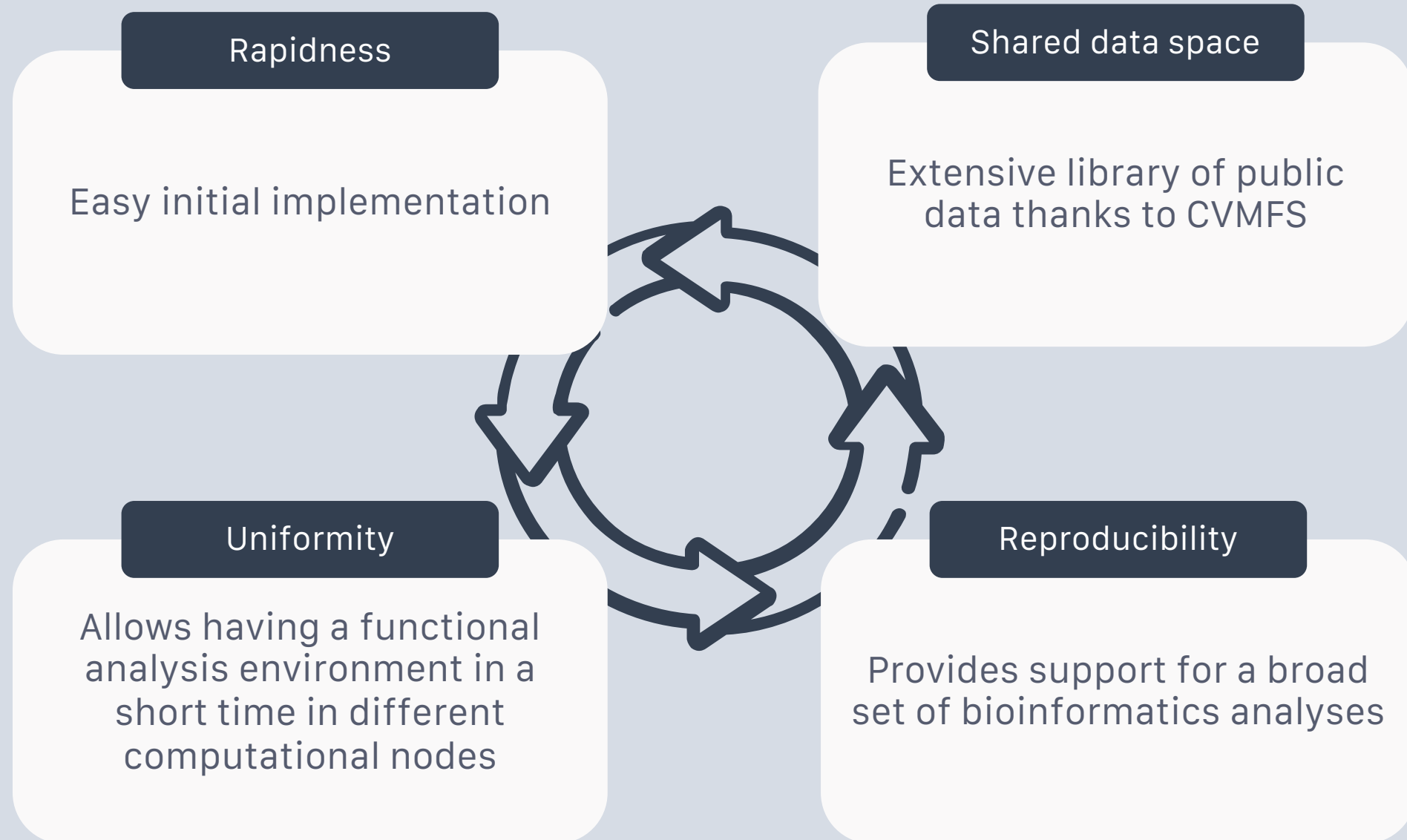
# FEDERATED CLOUD

- **Integrated authentication** system
- **Shared non-sensitive** data system
- **Sensitive data management** system
- **Distributed computing** environment
- **Development** environment

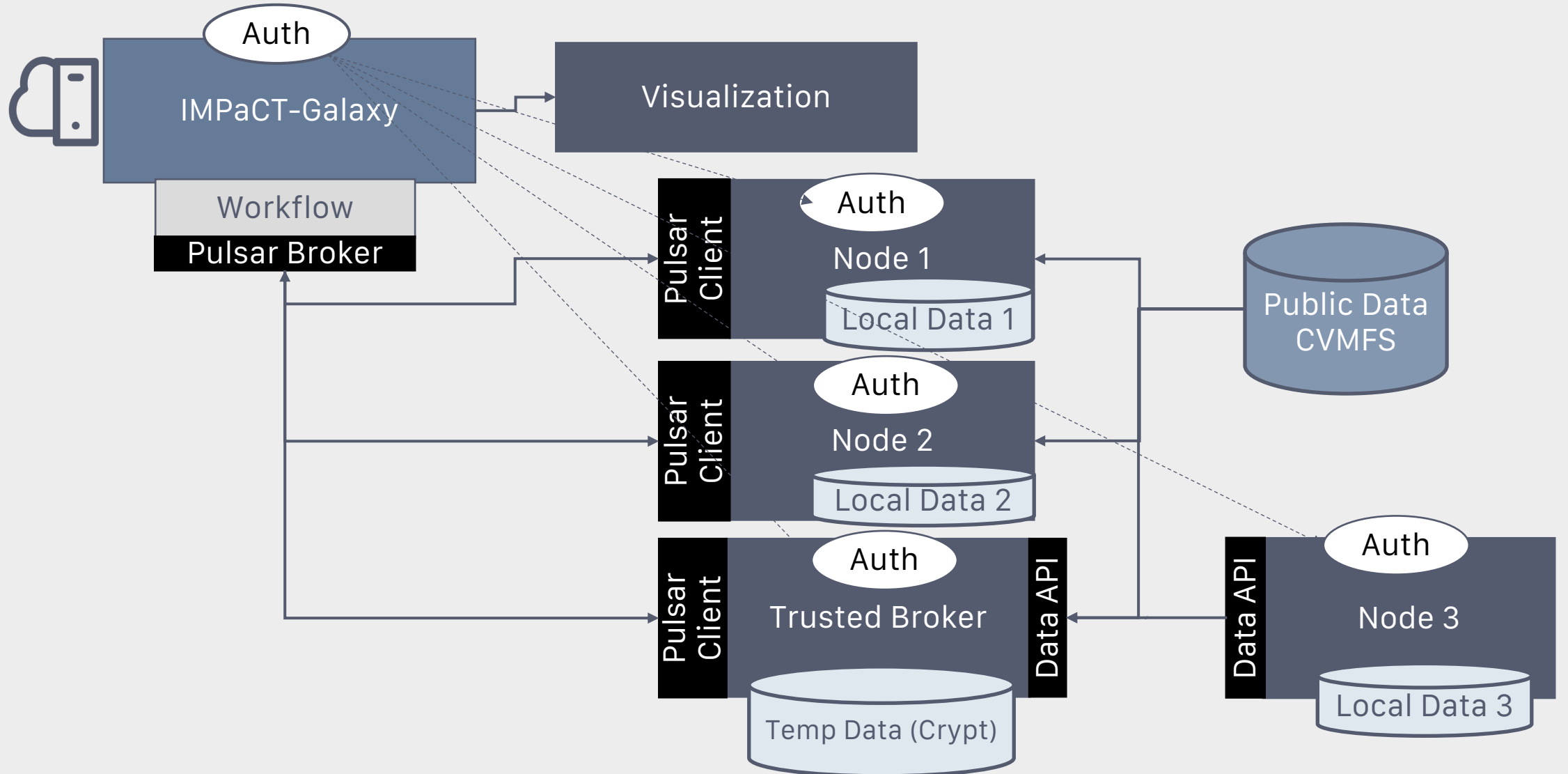
# STRUCTURE AND CHARACTERISTICS



# WHY GALAXY?



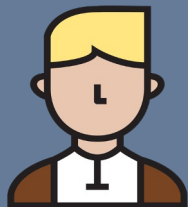
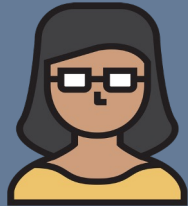
# DESIRED INFRAESTRUCTURE



# OUR DEVELOPMENT



# USERS



# AUTHENTICATION



- Single user account at the IMPaCT-Data level associated with **different identity providers**.
- Common **authentication** server to enable "**Single Sign-On**" (SSO) or common procedures.
- **Central** user administration system that would make it possible to implement a **shared authorization strategy** between the different centers.

Welcome to Galaxy, please log in

Public Name or Email Address

Password

Forgot password? [Click here to reset your password.](#)

Login



Don't have an account? [Register here.](#)

Welcome to IMPaCT-Data Galaxy server



La nube de IMPaCT-Data va a ofrecer un entorno de trabajo virtual en el que se va a poder analizar datos públicos y de acceso controlado. La nube permitirá el análisis de datos genómicos, de historia clínica y de imagen médica, proporcionando los siguientes servicios:

- acceso a datos públicos y de acceso controlado
- análisis utilizando las herramientas que la plataforma proporciona
- análisis utilizando los workflows que la plataforma proporciona

Powered by the Nord Cloud @ BSC

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

# IMPACT - Data

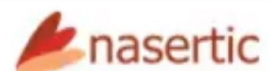
Sign in to your account

Username or email

Password

Sign In

Or sign in with





CENTRO NACIONAL DE  
INVESTIGACIONES  
CARDIOVASCULARES

Logging into CNIC...

someone@example.com

Next



# WHAT KIND OF DATA CAN WE WORK WITH INSIDE IMPaCT-DATA?

## PUBLIC DATA

ANY DATA THAT CAN BE ACCESSED BY ANYONE  
WITH A **VALID IMPaCT ACCOUNT**



## PROTECTED DATA

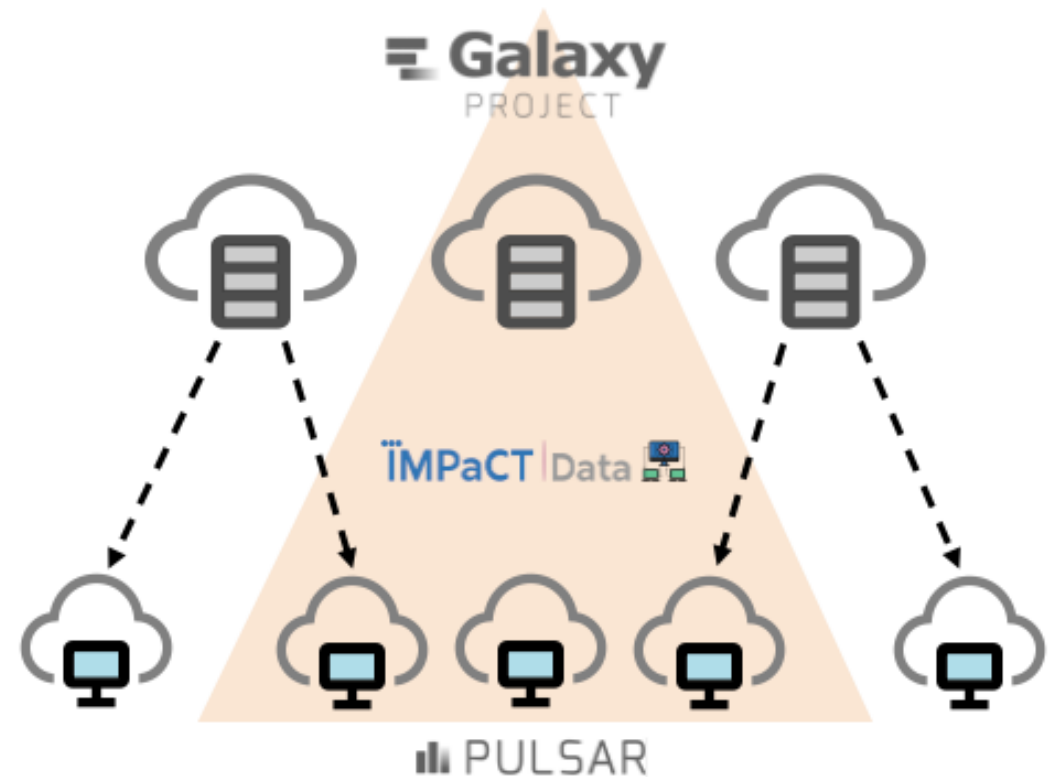
ONLY **APPROVED USERS** WILL BE ABLE TO  
INTERACT WITH **ACCESS CONTROLLED DATA**





# CURRENT INFRASTRUCTURE

- Structure based on the connection of **one main node** (IMPACT) and **three pulsar workers** (BSC, IMIB, CNIC)
- **Parallel jobs** proceeding from the execution of a **single tool** within the IMPACT environment can be executed in **different cores**.
- Pulsar nodes can be connected to **several main Galaxy servers**, even if they are outside of the IMPACT network and receive jobs from both of them **simultaneously**.



# CURRENT GENERAL INFRAESTRUCTURE

- Access **sample data** thanks to CVMFS, load **local** files or access information from a customizable **cloud space**.
- Users can also perform their analysis by using IMPaCT-Data **pre-installed** tools and workflows or **upload** their own.

The screenshot shows the IMPaCT-Data Galaxy server interface. The top navigation bar includes 'Galaxy / IMPACT-DATA', 'Flujo de Trabajo', 'Visualizar', 'Datos Compartidos', 'Ayuda', 'Usuario', and 'Utilizando 626.3'. The left sidebar lists various tools and operations: Herramientas, Get Data, Send Data, Collection Operations, Lift-Over, Text Manipulation, Convert Formats, Filter and Sort, Join, Subtract and Group, Fetch Alignments/Sequences, Operate on Genomic Intervals, Statistics, Graph/Display Data, Phenotype Association, Test local data, Variant Calling, Quality control, and BCFTools. The main content area displays a welcome message: 'Welcome to IMPaCT-Data Galaxy server' with the IMPaCT-Data logo. Below the logo, a paragraph describes the server's purpose: 'La nube de IMPaCT-Data va a ofrecer un entorno de trabajo virtual en el que se va a poder analizar datos públicos y de acceso controlado. La nube permitirá el análisis de datos genómicos, de historia clínica y de imagen médica, proporcionando los siguientes servicios:'. A bulleted list follows: 'acceso a datos públicos y de acceso controlado', 'análisis utilizando las herramientas que la plataforma proporciona', and 'análisis utilizando los workflows que la plataforma proporciona'. Below this, it says 'Powered by the Nord Cloud @ BSC'. At the bottom, there are two lines of text: 'Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.' and 'The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.' The right sidebar shows a 'Historial' section with a search bar and a list of history items, including '71: father\_R1.fq.gz', '69: FastQC on data 23: R awData', '68: FastQC on data 23: W ebpage', '67: FastQC on data 23: R awData', '66: FastQC on data 23: W ebpage', '65: FastQC on data 23: R awData', '64: FastQC on data 23: W ebpage', '61: FastQC on data 23: R awData', and '60: FastQC on data 23: W ebpage'.

# Galaxy / IMPACT-DATA

Tools

search tools

Upload Data

Filter and Sort

Join, Subtract and Group

Fetch Alignments/Sequences

Operate on Genomic Intervals

Statistics

Graph/Display Data

Phenotype Association

Test local data

Testing Tool

Variant Calling

Quality control

MultiQC aggregate results from bioinformatics analyses into a single report

FastQC Read Quality reports

BCFTTools

```
top - 17:02:28 up 0:43, 1 user, load average: 1.09, 0.79, 0.69
tasks: 937 total, 1 running, 906 sleeping, 0 stopped, 0 zombie
%cpu(s): 3.8 us, 0.2 sy, 0.0 ni, 97.2 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
MiB Mem : 288428.3 total, 139730.6 free, 2027.3 used, 3886.5 buff/cache
MiB Swap: 0.0 total, 0.0 free, 0.0 used, 282272.0 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR  S  %CPU  %MEM     TIME+ COMMAND
 2802 root        20   0 283936 147456 19320  S  1.4  5.1  0:00.13 java -Xms250m -Xmx250m -Dfactqc_output_dir=/var/job/pulsar/staging/11/working/working/dataset_c780440-20b-40ac-b0f4-6524f7c2152a-
37754 root        20   0 20364 4824 3284  R  1.0  0.0  0:00.13 top
28486 www-data  20   0 56108 8212 6176  S  0.7  0.0  0:25.23 nginx: worker process
30855 root        20   0 908928 60432 21556  S  0.7  0.0  0:15.58 /srv/pulsar/venv/bin/python /srv/pulsar/venv/bin/paster serve /srv/pulsar/config/server.ini --daemon
  34 root        20   0 0 0 0  S  0.1  0.0  0:00.93 [rcu_sched]
2181 root        20   0 1576832 58768 15328  S  0.1  0.0  0:40.67 /usr/bin/contalord
  1 root        20   0 172256 13752 8316  S  0.0  0.0  0:14.88 /sbin/init
  2 root        20   0 0 0 0  S  0.0  0.0  0:00.08 [kthreadd]
  3 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [rcu_gp]
  4 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [rcu_tasks_trace]
  6 root        0 -20 0 0 0  S  0.2  0.0  0:00.00 [kworker/0:0H-events_highpri]
  7 root        20   0 0 0 0  S  0.1  0.0  0:00.41 [kworker/0:1-events]
 10 root        0 -20 0 0 0  S  0.2  0.0  0:00.00 [mm_percpu_wq]
 11 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [rcu_tasks_rude_]
 12 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [rcu_tasks_trace]
 13 root        20   0 0 0 0  S  0.5  0.0  0:00.01 [ksoftirqd/0]
 15 root        rc  0 0 0 0 0  S  0.5  0.0  0:00.16 [migration/0]
 16 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/0]
 17 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [cpuhp/0]
 18 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/1]
 20 root        rc  0 0 0 0 0  S  0.5  0.0  0:02.00 [migration/1]
 21 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [ksoftirqd/1]
 23 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [kworker/1:0H-kblockd]
 24 root        20   0 0 0 0  S  0.0  0.0  0:00.00 [cpuhp/2]
 25 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/2]
 26 root        rc  0 0 0 0 0  S  0.5  0.0  0:02.01 [migration/2]
 27 root        20   0 0 0 0  S  0.0  0.0  0:00.00 [ksoftirqd/2]
 29 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [kworker/2:0H-events_highpri]
 30 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [cpuhp/3]
 31 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/3]
 32 root        rc  0 0 0 0 0  S  0.5  0.0  0:02.03 [migration/3]
 33 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [ksoftirqd/3]
 35 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [kworker/3:0H-kblockd]
 36 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [cpuhp/4]
 37 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/4]
 38 root        rc  0 0 0 0 0  S  0.5  0.0  0:02.05 [migration/4]
 39 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [ksoftirqd/4]
 41 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [kworker/4:0H-kblockd]
 42 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [cpuhp/5]
 43 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/5]
 44 root        rc  0 0 0 0 0  S  0.5  0.0  0:02.06 [migration/5]
 45 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [ksoftirqd/5]
 47 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [kworker/5:0H-kblockd]
 48 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [cpuhp/6]
 49 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/6]
 50 root        rc  0 0 0 0 0  S  0.5  0.0  0:02.07 [migration/6]
 51 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [ksoftirqd/6]
```

```
top - 17:02:22 up 6:53, 1 user, load average: 0.49, 0.11, 0.04
Tasks: 226 total, 1 running, 225 sleeping, 0 stopped, 0 zombie
%cpu(s): 6.2 us, 0.2 sy, 0.0 ni, 93.6 id, 0.0 wa, 0.0 hi, 0.0 si, 0.0 st
MiB Mem : 64318.4 total, 60959.3 free, 533.8 used, 2825.3 buff/cache
MiB Swap: 0.0 total, 0.0 free, 0.0 used, 63088.7 avail Mem

  PID USER      PR  NI    VIRT    RES    SHR  S  %CPU  %MEM     TIME+ COMMAND
 3393 root        20   0 3306976 252720 19364  S 102.3  0.4  0:05.98 java -Xms250m -Xmx2+
  181 root        20   0 0 0 0  S  0.3  0.0  0:00.40 [kworker/9:1-events]
  801 root        20   0 988984 61220 22280  S  0.3  0.1  0:11.89 /srv/pulsar/venv/bi+
  1 root        20   0 168120 11440 8484  S  0.0  0.0  0:02.07 /sbin/init
  16 2 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [kthreadd]
  3 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [rcu_gp]
  4 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [rcu_par_gp]
 18 6 root        20   0 -20 0 0  S  0.1  0.0  0:00.00 [kworker/0:0H-kbloc+
  8 root        0 -20 0 0 0  S  0.1  0.0  0:00.00 [mm_percpu_wq]
  9 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [ksoftirqd/0]
 10 root        20   0 0 0 0  S  0.1  0.0  0:00.28 [rcu_sched]
 11 root        rc  0 0 0 0 0  S  0.5  0.0  0:00.08 [migration/0]
 12 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/0]
 13 root        20   0 0 0 0  S  0.1  0.0  0:00.44 [kworker/0:1-events]
 14 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [cpuhp/0]
 15 root        20   0 0 0 0  S  0.5  0.0  0:00.00 [cpuhp/1]
 16 root        -51  0 0 0 0  S  0.5  0.0  0:00.00 [idle_inject/1]
```

Using 1.4 GB

## History

search datasets

## Unnamed history

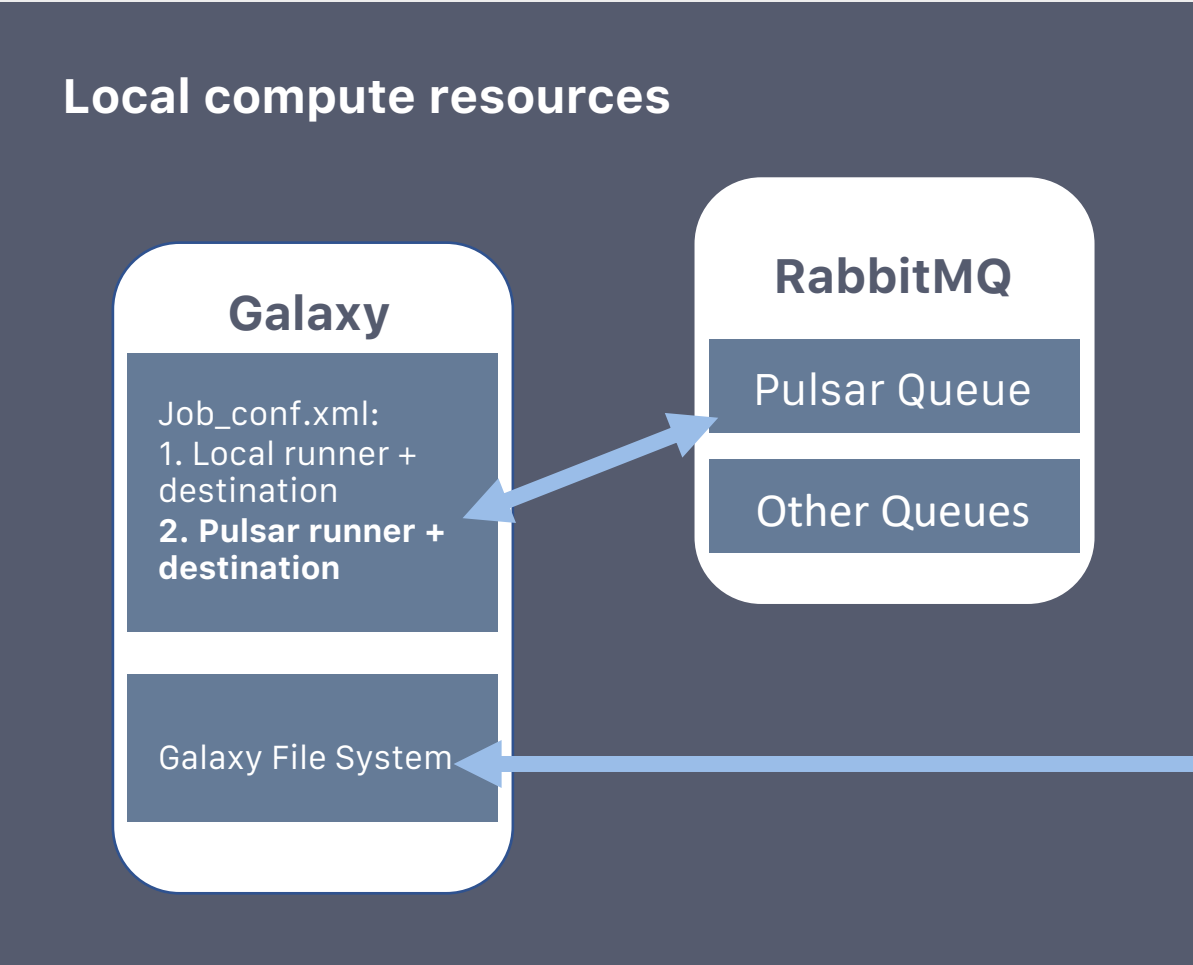
18 shown

867.83 MB

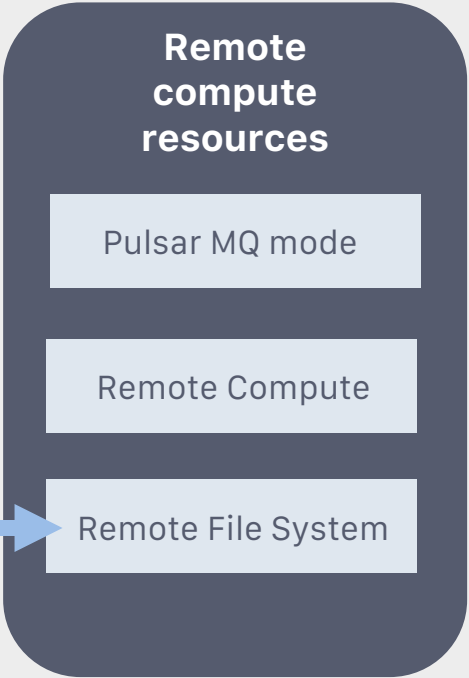
- 18: FastQC on data 6: RawData
- 17: FastQC on data 6: Webpage
- 16: FastQC on data 5: RawData
- 15: FastQC on data 5: Webpage
- 14: FastQC on data 4: RawData
- 13: FastQC on data 4: Webpage
- 12: FastQC on data 3: RawData
- 11: FastQC on data 3: Webpage



# Next development implementation stages



Galaxy- Pulsar environment using **RabbitMQ** via **websockets** for easier, faster integration.



# WHAT ABOUT WORKING WITH SENSITIVE DATA?

## **EGA-LIKE TECHNOLOGY (Federated LocalEGA concept)**

EGA provides a storage solution, a management structure composed of an "inbox" with the corresponding data ingestion protocol, an API providing metadata describing the hosted data, to be used for synchronization with the central data catalog, a access credential management module synchronized with the central access management module and finally a distribution module that, among other uses, will provide encrypted data to the computing nodes.

## **GA4GH AUTHORIZATION PASSPORTS**

Authorization management system for access to both resources and data is delegated to the computing nodes or data providers, although the use of standards such as GA4GH Passport/Visa is proposed for the communication of credentials between nodes.

# OTHER NOTES FOR THE FUTURE

1. Implementation of **data encryption protocols** for **protected** data handling.
2. **Standardization** of ansible recipes for **easier environment reproductibility**.
3. Inclusion of **new platforms** for data **analysis** and **processing**.



**bio.tools**  
**BioContainers**  
**WorkflowHub**  
**OpenEBench**  
**EGA**

# CONCLUSION

The IMPaCT-Data Biomedical Cloud is designed to enable the execution of **analytical workflows** assembled by any of the consortium's organizations, which **analyze** individual or combined clinical, genomic and medical imaging **datasets** for research purposes.

It represents an opportunity to disseminate software development **best practices** and promote the adoption of standards developed in ELIXIR, GA4GH and similarly **community-driven efforts**

# THANK YOU

twitter: @mchdiez  
linkedin: María Chavero Díez

