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

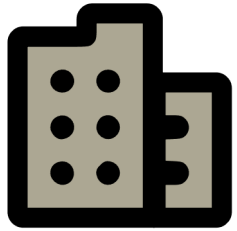
October, 4, 2022

IMPaCT

Spanish Precision Medicine Infrastructure associated with Science and Technology

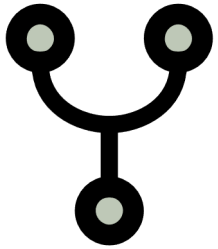
Proyect that aims to impulse precision medicine within the Spanish National Health System

15
Autonomous Communities



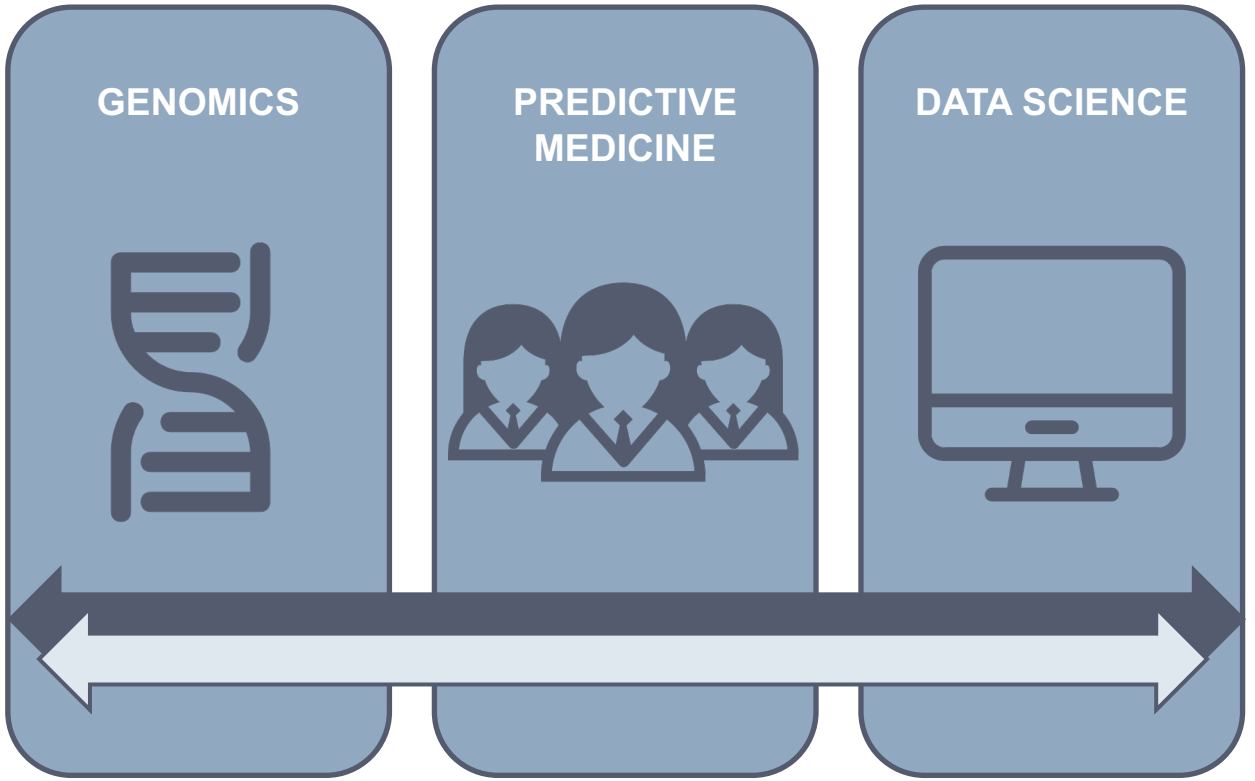
47
Participants

59
Research Groups



32
Associated Entities

IMPaCT



iMPaCT | Data

SOFTWARE

CLOUD

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?

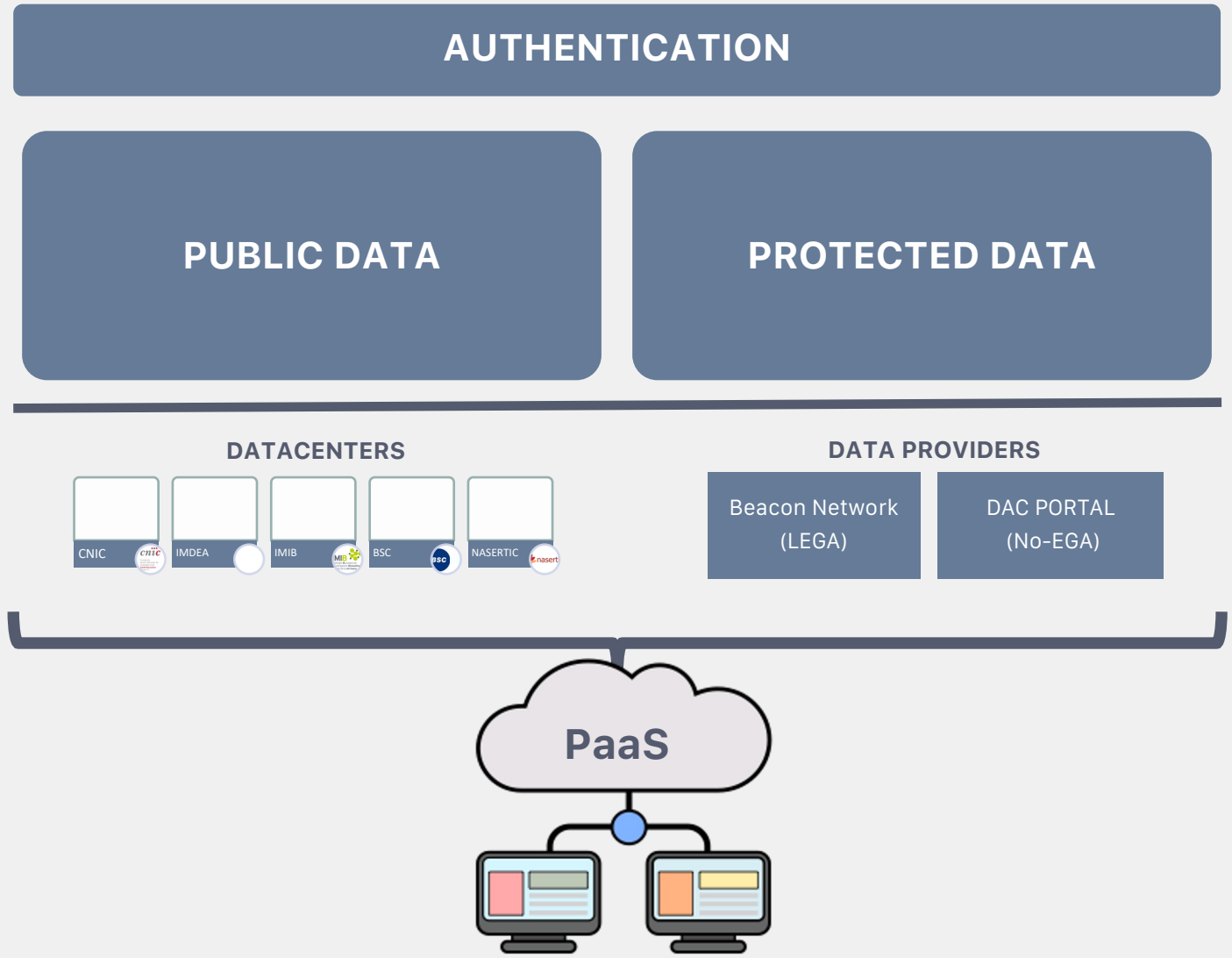
1. Barcelona Supercomputing Center - BSC. INB/ELIXIR-ES (Cataluña)
2. Centro Nacional de Investigaciones Cardiovasculares - CNIC (Madrid)
3. Instituto Murciano de Investigación Biosanitaria - IMIB (Murcia)
4. NASERTIC (Navarra)

HOW?

FEDERATED CLOUD

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

STRUCTURE AND CHARACTERISTICS



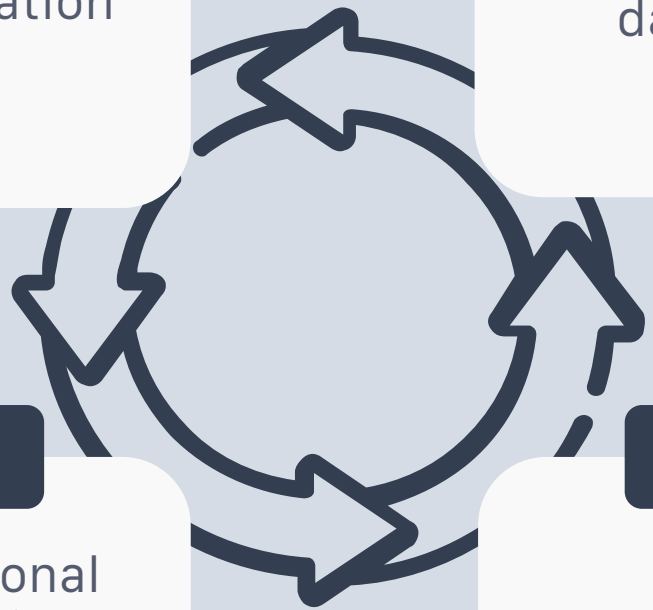
WHY GALAXY?

Rapidness

Easy initial implementation

Shared data space

Extensive library of public data thanks to CVMFS



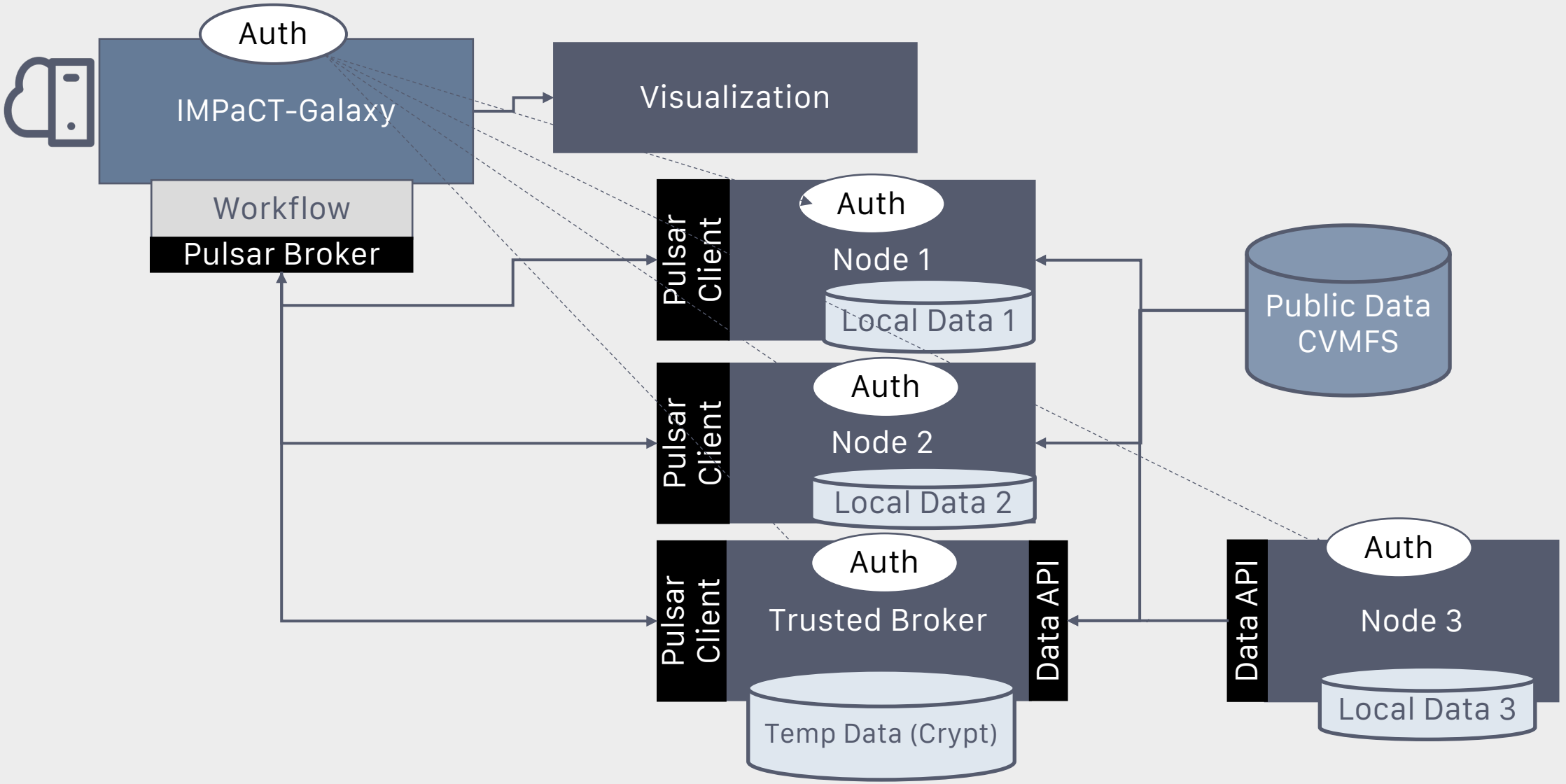
Uniformity

Allows having a functional analysis environment in a short time in different computational nodes

Reproducibility

Provides support for a broad set of bioinformatics analyses

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

iMPaCT - Data

Sign in to your account

Username or email

Password

Sign In

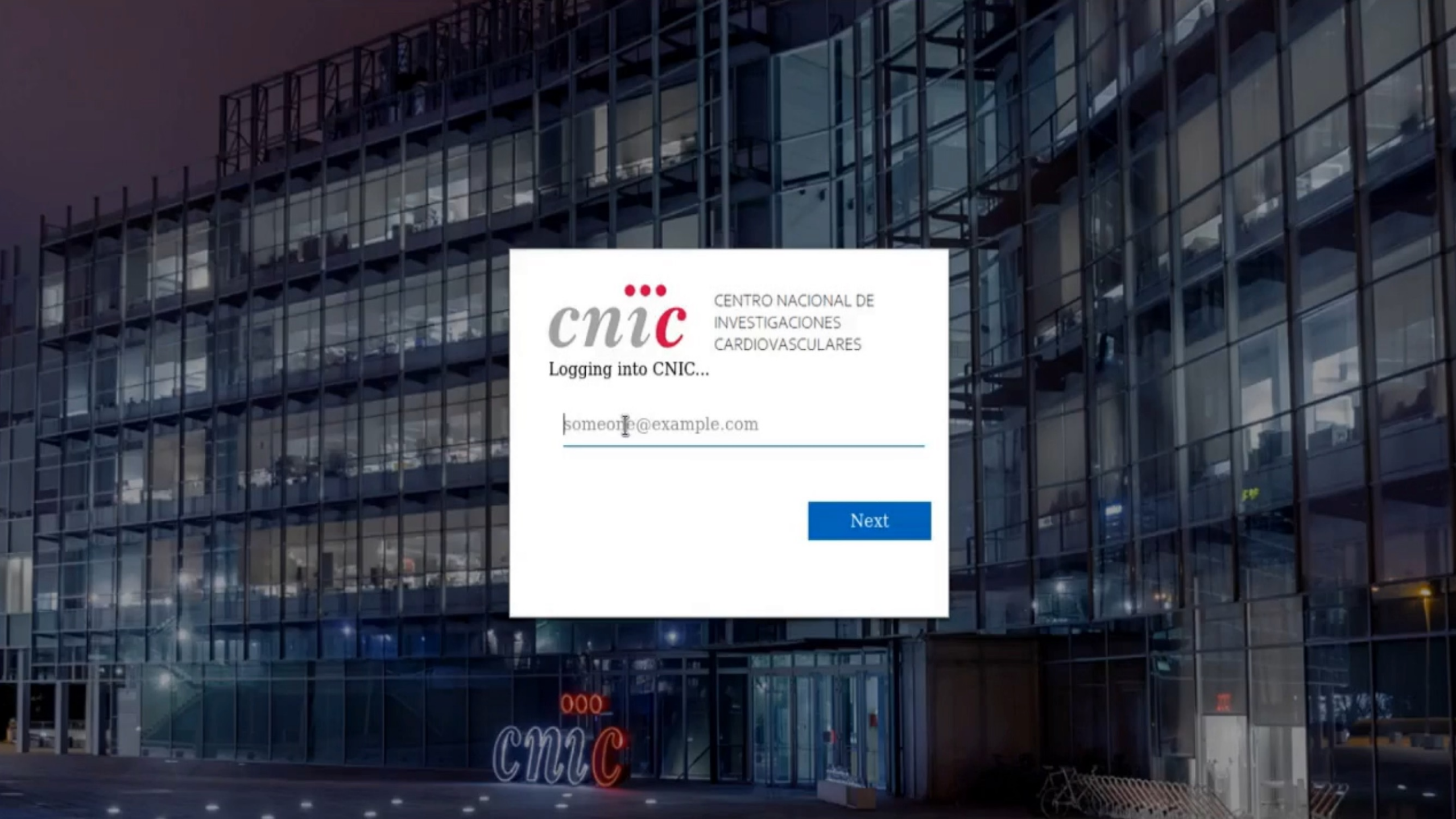
Or sign in with

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

nasertic

IMIB
Instituto Murciano de Investigación Biosanitaria

cnic


cnic
CENTRO NACIONAL DE
INVESTIGACIONES
CARDIOVASCULARES
Logging into CNIC...
someone@example.com
Next
cnic

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**

Galaxy PROJECT

PULSAR

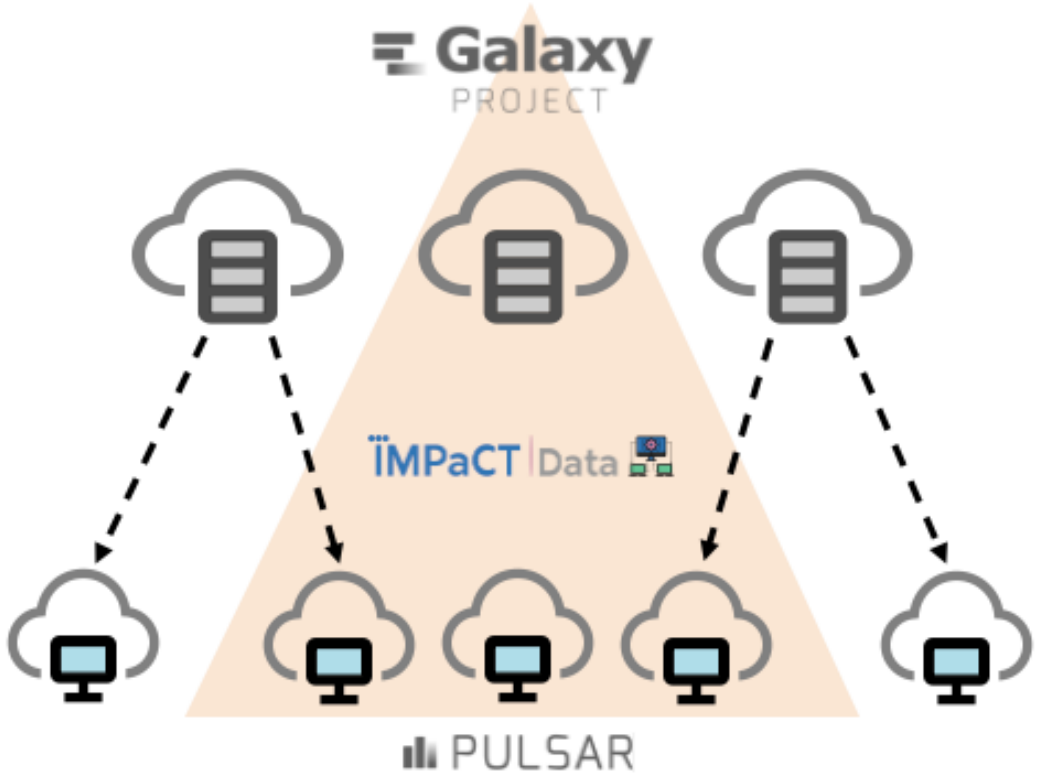
PROTECTED DATA

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

EUROPEAN GENOME-PHENOME ARCHIVE

CURRENT INFRASTRUCTURE

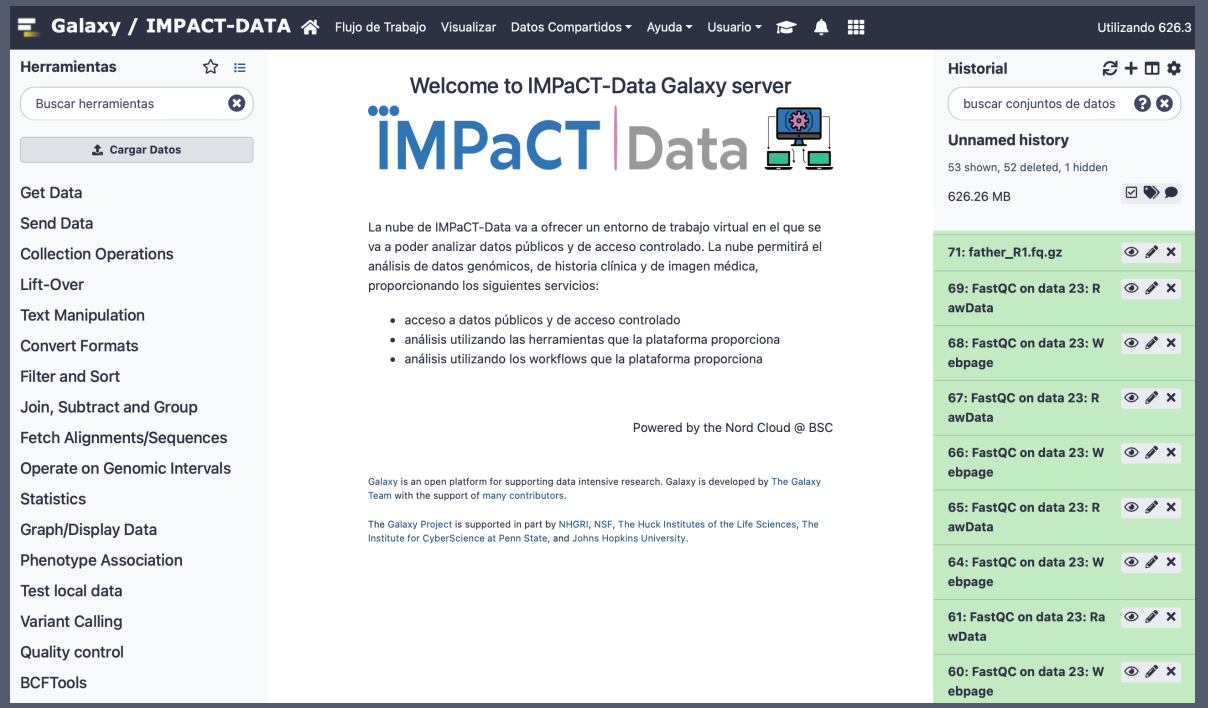
- Structure based on the connection of **one main node** (IMPaCT) and **three pulsar workers** (BSC, IMIB, CNIC)
- **Parallel jobs** proceeding from the execution of a **single tool** within the IMPaCT environment can be executed in **different cores.**
- Pulsar nodes can be connected to **several main**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

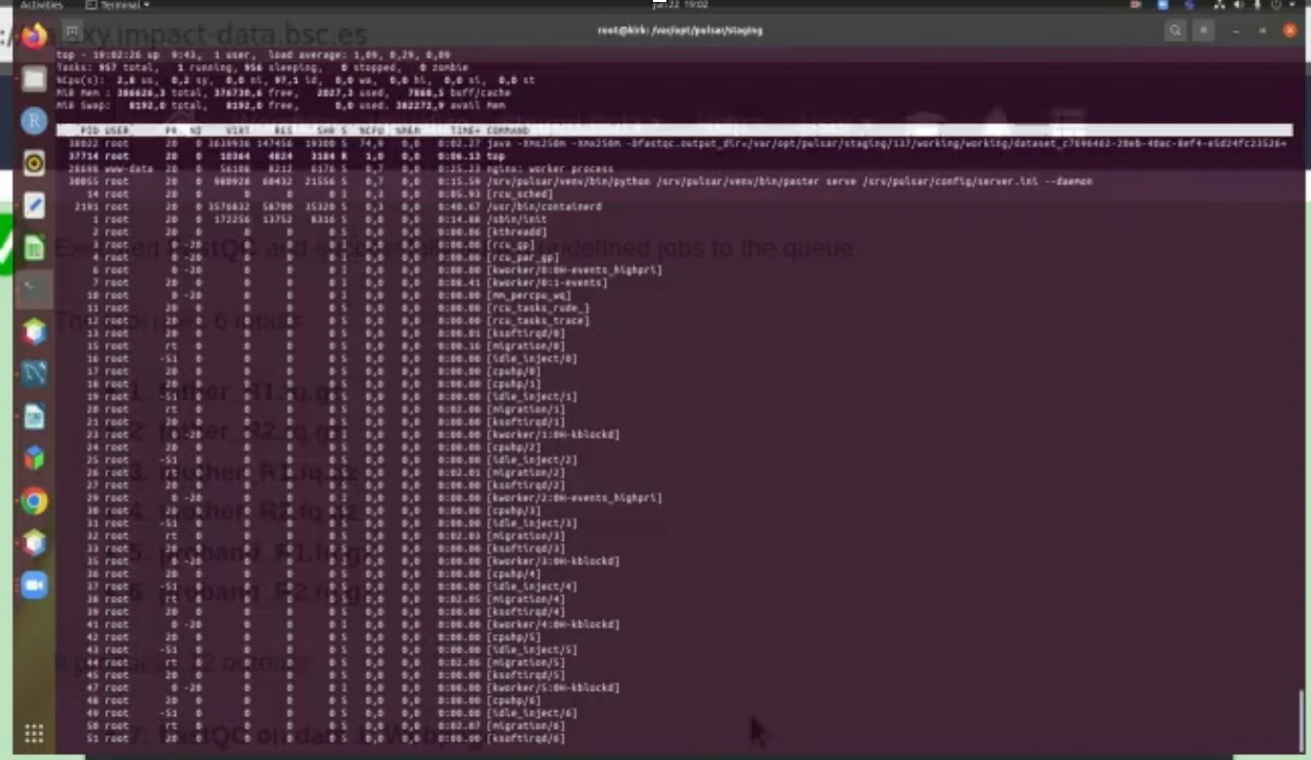


CURRENT GENERAL INFRAESTRUCTURE

- 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.

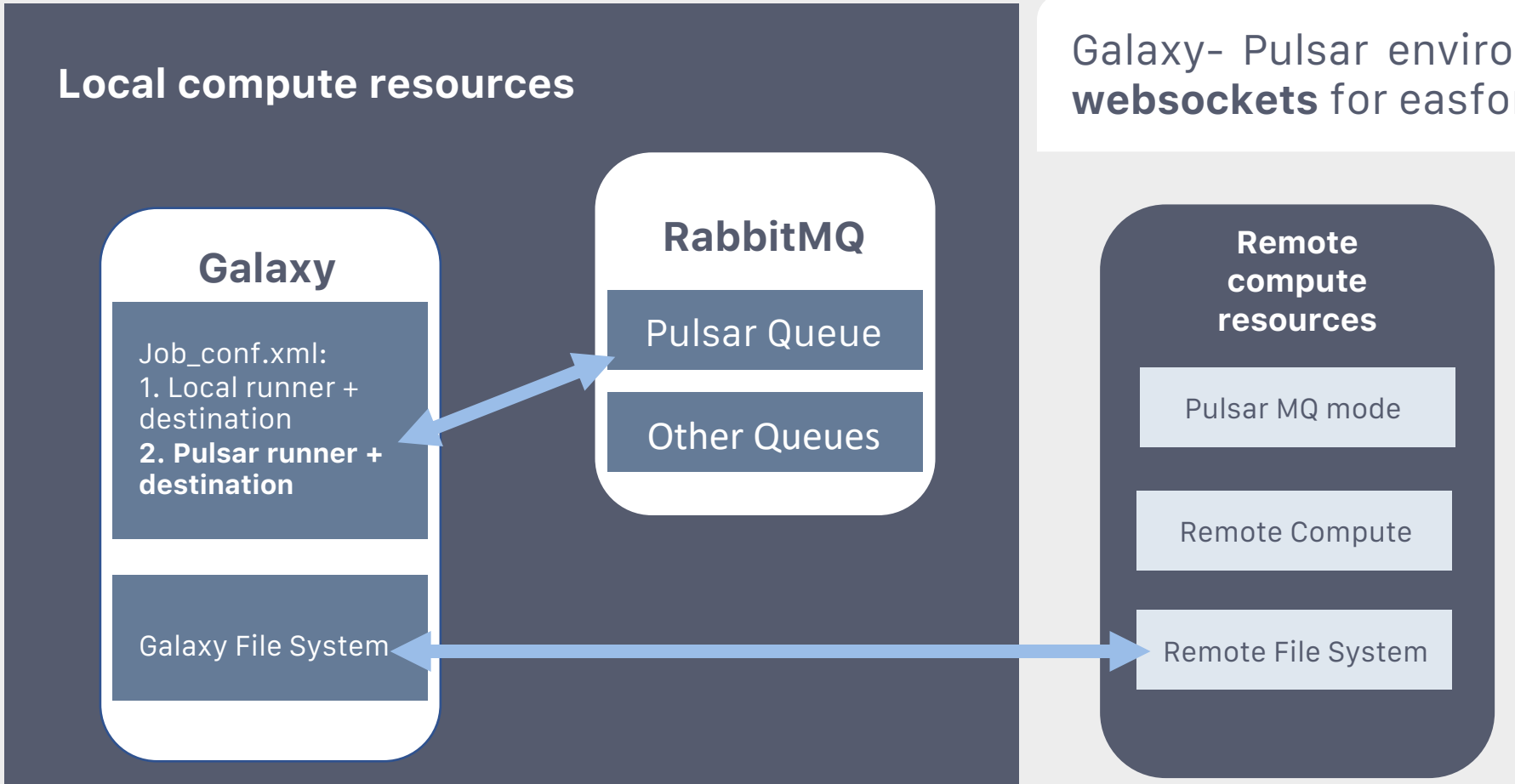




```
top - 17:02:22 up  6:53,  1 user,  load average: 0.49, 0.11, 0.04
Tasks: 226 total,   1 running, 225 sleeping,   0 stopped,   0 zombie
%Cpu(s):  6.2 us,  0.2 sy,  0.0 ni, 93.6 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
MiB Mem :  64318.4 total,  60959.3 free,    533.8 used,   2825.3 buff/cache
MiB Swap:      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	I	0.3	0.0	0:00.40	[kworker/9:1-events]
801	root	20	0	980984	61220	22280	S	0.3	0.1	0:11.89	/srv/pulsar/venv/bi+
1	root	20	0	168120	11440	8404	S	0.0	0.0	0:02.07	/sbin/init
2	root	20	0	0	0	0	S	0.0	0.0	0:00.00	[kthreadd]
3	root	0	-20	0	0	0	I	0.0	0.0	0:00.00	[rcu_gp]
4	root	0	-20	0	0	0	I	0.0	0.0	0:00.00	[rcu_par_gp]
6	root	0	-20	0	0	0	I	0.0	0.0	0:00.00	[kworker/0:0H-kbloc+
8	root	0	-20	0	0	0	I	0.0	0.0	0:00.00	[mm_percpu_wq]
9	root	20	0	0	0	0	S	0.0	0.0	0:00.00	[ksoftirqd/0]
10	root	20	0	0	0	0	I	0.0	0.0	0:00.28	[rcu_sched]
11	root	rt	0	0	0	0	S	0.0	0.0	0:00.08	[migration/0]
12	root	-51	0	0	0	0	S	0.0	0.0	0:00.00	[idle_inject/0]
13	root	20	0	0	0	0	I	0.0	0.0	0:00.44	[kworker/0:1-events]
14	root	20	0	0	0	0	S	0.0	0.0	0:00.00	[cpuhp/0]
15	root	20	0	0	0	0	S	0.0	0.0	0:00.00	[cpuhp/1]
16	root	-51	0	0	0	0	S	0.0	0.0	0:00.00	[idle_inject/1]

Next development implementation stages



Galaxy- Pulsar environment using **RabbitMQ** via **websockets** for easfor 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

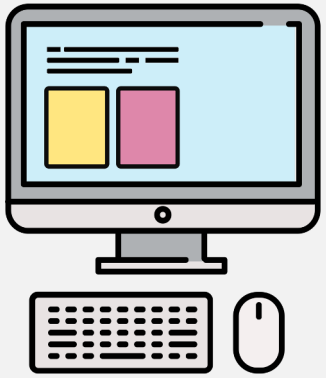
Authorization management system for access to both resources and data is delegated to the computing nodes or data providers, although the use of standards such as GA4GH Passport/Visa is proposed for the communication of credentials between nodes.

OTHER NOTES FOR THE FUTURE

1. Implementation of **data encryption protocols** for **protected** data handling.

2. **Standarization** 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

BSC
elixir SPAIN
INB Instituto Nacional de Bioinformática
GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN Instituto de Salud Carlos III
FEDER Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA "Una manera de hacer Europa"