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

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

AUTHENTICATION

PUBLIC DATA

PROTECTED DATA

DATA PROVIDERS

Beacon Network (LEGA)
DAC PORTAL (No-EGA)

DATA CENTERS

PaaS
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 INFRASTRUCTURE**

- IMPaCT-Galaxy
- Workflow
- Pulsar Broker
- Visualization
- Pulsar Client
- Node 1
  - Local Data 1
- Pulsar Client
- Node 2
  - Local Data 2
- Pulsar Client
  - Trusted Broker
- Temp Data (Crypt)
- Data API
- Public Data CVMFS
- Auth
- Node 3
  - Local Data 3
OUR DEVELOPMENT
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.
<table>
<thead>
<tr>
<th>PUBLIC DATA</th>
<th>PROTECTED DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANY DATA THAT CAN BE ACCESSED BY ANYONE WITH A <strong>VALID IMPaCT ACCOUNT</strong></td>
<td>ONLY <strong>APPROVED USERS</strong> WILL BE ABLE TO INTERACT WITH <strong>ACCESS CONTROLLED</strong> DATA</td>
</tr>
</tbody>
</table>

**PUBLIC DATA**
- Galaxy Project
- Pulsar

**PROTECTED 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.
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