Contribution ID: **55** Type: **Oral Presentation**

Secure deployments of Galaxy Servers for analyzing personal and Health Data leveraging the Laniakea service (Remote presentation)

Tuesday, March 21, 2023 4:40 PM (20 minutes)

Data security issues and legal and ethical requirements on the storage, handling and analysing genetic and medical data are becoming increasingly stringent. Some regulatory obstacles may represent a gap to data sharing and the application of Open Science and Open Access principles. In this perspective, Task 6.6 of the EOSC-Pillar (https://www.eosc-pillar.eu/) project aimed to analyze the regulatory compliance of the integrated and interoperable PaaS level data analysis service (Laniakea) for ELIXIR and the Life Science community in general, the result of an interaction between Galaxy services and data repositories. Starting from the research activity carried out in the context of Task 6.6, the work aims to define the ethical and legal requirements that must be respected in order to guarantee an adequate balance between data and privacy protection and effective application of the FAIR, OS and OA principles.

From a technological point of view, we have implemented the necessary measures to improve the security of the entire service. In particular, the goal is to guarantee the creation of isolated and secure environments to carry out data analyses. To do this, we have focused on two critical aspects: data access and storing and network access control to the service.

User data isolation is accomplished by encrypting the entire storage volume associated with the virtual machine, using the Linux kernel encryption module. The level of disk encryption is completely transparent to software applications, in this case Galaxy. The procedure has been completely automated through the web Dashboard of the PaaS orchestration service (https://github.com/indigo-dc/orchestrator), taking advantage of Hashicorp Vault for storing user passphrases. After authenticating on the Dashboard, the user enables data encryption when setting up a new instance. The Dashboard contacts Hashicorp Vault to get a token that can only be used once. The token is passed to the encryption script on the virtual machine, a random passphrase is generated, the volume is encrypted, unlocked and formatted. Finally, the encryption script accesses Vault using the one-time token and stores the passphrase which will be accessible, at any time, only to the user via the Dashboard. This strategy makes it possible to create secure encryption keys and at the same time prevent user credentials or the encryption passphrase from being transmitted unencrypted to the virtual infrastructure, compromising its security.

The INDIGO Cloud orchestration system (PaaS layer) that allows the automatic deployment of the data analysis service has been extended to be able to manage the creation of virtual environments on a private network, taking advantage of the isolation of L2 virtual networks at the tenant level guaranteed by the cloud provider and automatically configuring appropriate security groups that monitor network traffic. In this way, access to the various analysis environments is blocked from the external network and also the traffic between virtual machines instantiated on the same private network, but belonging to different deployments, is filtered. The access to the service provided to users leverage VPN server at the tenant level. To improve the user experience, VPN authentication has been integrated with the authentication and authorization system, INDIGO IAM (https://github.com/indigo-iam/iam), used by the entire PaaS/IaaS stack and based on OpenID Connect. This way, users don't have to create additional accounts/credentials, but can use federated authentication. In particular, the solution implemented for the VPN is based on the OpenVPN open source software and on a PAM module developed ad-hoc to allow authentication via IAM.

The solutions described have been tested and validated on the ReCaS-Bari (https://www.recas-bari.it/index.php/en/) cloud and on the INFN-Cloud (https://www.cloud.infn.it/) multi-site distributed infrastructure.

Primary authors: DONVITO, Giacinto (INFN); Dr TANGARO, Marco Antonio (CNR and INFN, Italy); ANTONACCI, Marica (INFN); Dr FOGGETTI, Nadina (INFN, Italy)

Presenter: Dr TANGARO, Marco Antonio (CNR and INFN, Italy)

Session Classification: Health & Life Sciences (including Pandemic Preparedness Applications)

Track Classification: Track 2: Health & Life Sciences (including Pandemic Preparedness Applications)