Title: DESIGN AND IMPLEMENTATION OF AN AUTOMATED CI/CD PIPELINE FOR GENOMICS DATA

Supervisor: Stefano Cozzini/ Alberto Cazzaniga/ Danilo Licastro

Affiliation: Area Science Park

Requirements: proficiency in python, sound knowledge of Ci/CD procedures and Kubernetes. Knowledge of Kafka/Apache Airflow and genomic pipelines a plus.

Duration: 5 months

Other info: in collaboration with LAGE

Project AIM

Introduction, brief description of the context

At the Laboratory for Genomics and Epigenomics (LAGE) there is the need to automatize as much as possible the genomic computational pipelines that elaborate the row data created by the NGS sequencer. The aim of this project is to develop an automated service that takes care to execute such pipelines and it's easily scalable on the ORFEO datacenter kubernetes resources.

Description of the activities that the student will have to carry out

- literature analysis and context study;
- define the requirements of the CI/CD systems either functional (what the system shall do) or non-functional requirements (system property, constraints, technologies to use). produce a System requirements document.
- Design the architecture of the system with a blueprint and also use Data Flow Diagrams (DFD) to describe the flow of the data through the various processes.
- Testing the architecture: Use testing in order to validate the requirements, find defects. Both unit and integration tests should be defined. The latter will be implemented thanks to the CI/CD technologies used.
- Deploy the application on the ORFEO Kubernetes clusters
- Perform some scalability tests and associated benchmark