# Disposable Apache Spark on Kubernetes clusters from Jupyter notebook service for distributed astronomy analysis

Dr Julien Peloton, IJCLab
peloton@lal.in2p3.fr

Internship Summer 2020 - Master level

## 1 Description

Our research group is investigating how to leverage the big data ecosystem tools to analyse current and future data sets in astronomy [1]. Among the future large experiments, the Large Synoptic Survey Telescope (LSST, [2]) will start soon collecting terabytes of data per observation night, and the efficient processing and analysis of both real-time and historical data remains a major challenge.

At University Paris Saclay, a cloud has been deployed by the LABEX P2IO, targeting scientific applications. This cloud (VirtualData, 3500 cores) is the main asset of the mésocentre, created in 2017. Open to all university users, including several astrophysics and cosmology use cases, it has been the basis for deploying advanced data processing services like Apache Spark [3] and JupyterHub. Apache Spark is an open-source framework for data analysis, based on the so-called MapReduce cluster computing paradigm, popularized by the Hadoop framework using implicit data parallelism and fault tolerance.

This project will develop the necessary integrations to use Apache Spark on Kubernetes clusters from Jupyter notebook service based on The Service for Web based ANalysis (SWAN, [3]). SWAN offers scalable interactive data analysis and visualizations using Jupyter notebooks, with Apache Spark computations being offloaded to compute clusters - on-premise YARN clusters and more recently to cloud-native Kubernetes clusters.

## 2 Proposed tasks

We propose the following tasks:

- the creation of Kubernetes clusters on VirtualData Openstack magnum interface from Jupyter notebooks
- Initialisation of the Apache Spark services (e.g. shuffle service, history server) on the Kubernetes cluster
- Intializing VirtualData services (e.g CVMFS) on Kubernetes cluster
- Development of web interface Jupyter plugin to attach the Kubernetes cluster to the Jupyter notebook (SWAN) based on user service account

## 3 Requirements

- Knowledge of Python.
- Knowledge of parallel and distributed computing, and cloud architecture.

- Being aware of the big data challenges and issues.
- Working with or willing to learn Apache Spark.

[1] https://github.com/astrolabsoftware
[2] https://en.wikipedia.org/wiki/Vera_C._Rubin_Observatory
[3] https://swan.web.cern.ch/