When astronomy meets big data: the clustering of matter in the Universe

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.

This internship focuses more specifically on the distributed computing framework Apache Spark [3]. 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. In addition Spark optimizes data transfer, memory usage and communications between processes based on a graph analysis (DAG) of the tasks to perform, largely relying on the functional programming.

While there are many packages based on Apache Spark to manipulate spatial 2D datasets (e.g. Geospark, Geomesa, Magellan, GeoTrellis), there are very few initiatives to process and analyse 3D data sets which were hitherto too costly to be processed efficiently. Hence, we extended the functionalities of the Apache Spark SQL module to ease the manipulation of 3D data sets and perform efficient queries: partitioning, data sets join and cross-match, nearest neighbors search, spatial queries, and more. Our developments are packaged in spark3D [4].

In this internship, we propose to tackle the challenge of the clustering of very large spatial 3D data sets. The search for clusters of matter (over-densities) is a typical task in galaxy surveys, but the size of current and future datasets makes traditional tools completely ineffective. With this work, you will walk away with an understanding of modern challenges in astronomy and contribute to it, appreciate some beautiful night skies, and use big data to help pushing further the frontiers of Science!

2 Proposed tasks

We propose first to study the popular DBSCAN clustering algorithm [5]. Our group developed a prototype version for Apache Spark, and the student will make the integration within spark3D. A special attention will be put in performing reproducible benchmarks on existing datasets. Second, the student will focus on methods based on graphs and graph-parallel computation to describe the interaction between astronomical objects. This work will be mainly exploratory. Ultimately, all the developments of the student will be integrated in the Apache Spark-based framework for processing large-scale spatial 3D data, spark3D.

3 Requirements

- At least one of the following programming language: Scala or Python.

- Knowledge of parallel and distributed computing.
- Being aware of the big data challenges and issues.
- Working with or willing to learn Apache Spark.
- Working with or willing to learn graphs.

[1] https://github.com/astrolabsoftware
[2] https://en.wikipedia.org/wiki/Vera_C._Rubin_Observatory
[3] http://spark.apache.org/
[4] https://github.com/astrolabsoftware/spark3d
[5] https://en.wikipedia.org/wiki/DBSCAN