UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

david

Master Internship Proposal

TOWARDS AN OPTIMIZED AND GENERIC STORAGE MODEL FOR ASTRONOMICAL DATA IN SPARK

Applications in universe science are among the most demanding of Big Data technology [6]. Indeed, recent new programs for sky and earth surveying (e.g LSST project) will produce peta bytes of data. Exploratory analysis of these data is crucial to enable scientists and practitioners to better understand their data and optimize various processes. This requires efficient database systems to manage and query these unprecedented amount of data [4].

Efficient query processing of astronomical data leads to optimize the data representation. Today, the most used formats in astronomy are FITS, HDF5, or simple csv, mainly for data exchange purpose. Besides, Parquet format[3], recommended by the Apache consortium, is becoming a de facto standard adopted by a large variety of Big Data tools, and NoSQL system [2]. However, there exists a gap between the astronomical standard formats and Parquet, as a matter of fact. More importantly, due to the amount of astronomical data, it adds a significant over-cost to the loading process in NoSQL systems like Spark [8], since the data should be converted from FITS to the format adopted in the target system.

The main objective of this internship is to fill this gap by proposing an optimized generic storage in Spark to represent at least FITS and HDF5 data formats into Spark `DataFrame`. A focus, in the proposed solution is to take the advantages of FITS/HDF5 data organization for optimizing current existing astronomical operators. The group ADAM hosting this internship has investigated this issue and proposed ASTROIDE [1], a framework for efficient query processing of astronomical data within a distributed architecture [5]. However, ASTROIDE does not support typical astronomical formats such as FITS and HDF5. The expected proposal may build on Spark-FITS [7], but must be implemented within ASTROIDE framework and be tested and compared with current implemented astronomical operators. The proposed design should be scalable and support incremental upload of large datasets, and optimize the related performance.

The internship will take place as follows:

- At first, the trainee will get acquainted with the team's knowledge about ASTROIDE and NoSQL technologies required by the project.
- Next, she/he will propose a baseline solution, not necessarily optimal from the querying point of view, but more optimal to load FITS and HDF5 into `DataFrame`.

- Finally, she/he will optimize further both the ingestion and the query performances and compare them to the baseline.

Required skills:
We seek highly motivated and ambitious candidates with a deep interest in working on big data technology, with strong object oriented programming skills. The candidate should be familiar with Unix scripting environment and tools like git, maven, ... This internship may open the way to a PhD thesis in collaboration between DAVID Lab at UVSQ/Paris-Saclay University and the CNES (Centre National d'Etudes Spatiales). A good background in data mining / machine learning is a plus for the purpose of the PhD thesis.

Duration: 5 to 6 months

Hosting lab: DAVID Laboratory (located in Versailles city - France), University of Versailles Saint-Quentin / University of Paris-Saclay. www.david.uvsq.fr

Advisors: Laurent Yeh and Karine Zeitouni
Contact: Laurent.Yeh@uvsq.fr, Karine.Zeitouni@uvsq.fr

References

[1] ASTROIDE - A distributed data server for big astronomical data. https://cnesuvsqastroide.github.io.

[2] NoSQL Databases. http://nosql-database.org/.

[3] Parquet. https://parquet.apache.org/.

[4] SciDB. https://www.paradigm4.com/try$_s$*cidb*/.

[5] Mariem Brahem, , Karine Zeitouni, and Laurent Yeh. Astroide: A unified astronomical big data processing engine over spark. *IEEE Transactions on Big Data*, 2018.

[6] Alyssa A Goodman and Curtis G Wong. Bringing the night sky closer: Discoveries in the data deluge. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, pages 39–44, 2009.

[7] Julien Peloton, Christian Arnault, and Stéphane Plaszczynski. Fits data source for apache spark. *Computing and Software for Big Science*, 2(1):7, 2018.

[8] Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.