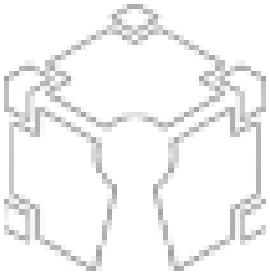

Contribution ID: **7** Type: **Poster**

Full Stack Data Science: Using Python to download, clean, analyze and visualize Gaia data

Python is not only one of the most important programming languages in the software industry[1][2], but also shows the most solid growth among the widely used ones thanks to its popularity in the Data Science field[3][4]. Python gained lots of traction in the scientific computing field in the 90s and the 2000s thanks to its simplicity and its power as a "glue language", allowing scientists to unify their tasks and use only one tool[3]. However, with the explosion of the Big Data movement in the 2010s and the evolution of web browsers and personal computers, it was no longer possible to use one single language for Data Science: efficient data access has to be done through databases using SQL, all the Hadoop stack is written in Java or Scala, while modern visualization is usually developed in JavaScript. Even if some current solutions, such as PySpark, enable the use of Python for Big Data, the deployment is often complex, the workflow fragile, and a gap between Small and Big Data still exists.

We argue that the aforementioned complexity is an obstacle for scientists without formal training in Software Engineering and also for the general public, particularly in Astronomy. In recent times though, the Python ecosystem has evolved to catch up with the Big Data world, and new tools have appeared to process, analyze and visualize big amounts of data (where we will freely define "big" as "not fitting in RAM", i.e. in the order of tens of gigabytes or more). On the one hand, libraries like numba allow Python to scale up by compiling a subset of the language to assembly code in a Just-in-Time fashion (hence improving the performance of single node programs), and on the other hand projects like Dask are bringing distributed and out-of-core capabilities to the usual "small data" tools of the Python in Data Science stack (numpy, pandas, scikit-learn), allowing easy reuse of existing codebases and easy scaling to tens to thousands of nodes with minimal changes in the deployment.

In this talk we will showcase the power of these modern Python packages to perform Full Stack Data Science on the Gaia Data Release 2 (DR2), a recently released astrometry dataset with observations of 1.3 billion stars by the Gaia mission. We will demonstrate how to use Pyia and Astroquery to easily download Gaia data[4][5], how to process big amounts of data on a modest laptop using Dask, and how to scale that to a distributed cloud computing cluster to increase performance with minimal cost and minimal code changes. We will also feature special Python libraries for Big Data visualization that avoid common visualization pitfalls such as Datashader, Holoviews and Bokeh. To finish the presentation, we will justify how these tools simplify the access to Astronomical data in general, discuss the limitations of the approach and talk about future developments.

References

[1]: TIOBE index https://www.tiobe.com/tiobe-index/

[2]: GitHut https://madnight.github.io/githut/#/pull_requests/2017/4

[3]: The Incredible Growth of Python https://stackoverflow.blog/2017/09/06/incredible-growth-python/

[4]: KDNuggets Poll https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

[3]: Beazley, David M. "Scientific computing with Python." In Astronomical Data Analysis Software and Systems IX, vol. 216, p. 49. 2000.

[4]: Adrian Price-Whelan. "Adrn/pyia: V0.3". Zenodo, June 5, 2018. doi:10.5281/zenodo.1275923.

[5]: Price-Whelan, A. M., B. M. Sipőcz, H. M. Günther, P. L. Lim, S. M. Crawford, S. Conseil, D. L. Shupe et al. "The Astropy Project: Building an inclusive, open-science project and status of the v2. 0 software." arXiv preprint arXiv:1801.02634 (2018).

Primary authors: Mr CANO RODRÍGUEZ, Juan Luis; Mr IZQUIERDO AMO, Roberto

Co-author: Ms BADENAS AGUSTÍ, Mariona

Presenters: Mr CANO RODRÍGUEZ, Juan Luis; Mr IZQUIERDO AMO, Roberto

Session Classification: Posters and Demos

Track Classification: Space Science Data and Software