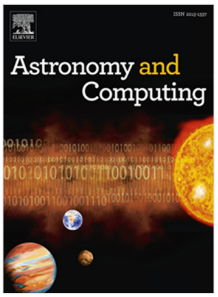
Accepted Manuscript



JOVIAL: Notebook-based astronomical data analysis in the cloud

M. Araya, M. Osorio, M. Díaz, C. Ponce, M. Villanueva, C. Valenzuela, M. Solar

PII: S2213-1337(18)30018-0
DOI: https://doi.org/10.1016/j.ascom.2018.09.001
Reference: ASCOM 246

To appear in: *Astronomy and Computing*

Received date : 26 February 2018
Accepted date : 4 September 2018

Please cite this article as: Araya M., Osorio M., Díaz M., Ponce C., Villanueva M., Valenzuela C., Solar M., JOVIAL: Notebook-based astronomical data analysis in the cloud. *Astronomy and Computing* (2018), https://doi.org/10.1016/j.ascom.2018.09.001

JOVIAL: Notebook-based Astronomical Data Analysis in the Cloud

Mauricio Araya[a,*], Maximiliano Osorio[a], Matías Díaz[a], Carlos Ponce[a], Martín Villanueva[a], Camilo Valenzuela[a], Mauricio Solar[a]

[a] *Universidad Técnica Federico Santa María, Avenida España 1680, Valparaíso, Chile*

Abstract

Performing astronomical data analysis using only personal computers is becoming impractical for the very large data sets produced nowadays. As analysis is not a task that can be automatized to its full extent, the idea of moving processing where the data is located means also moving the whole scientific process towards the archives and data centers. Using Jupyter Notebooks as a remote service is a recent trend in data analysis that aims to deal with this problem, but harnessing the infrastructure to serve the astronomer without increasing the complexity of the service is a challenge. In this paper we present the architecture and features of JOVIAL, a Cloud service where astronomers can safely use Jupyter notebooks over a personal space designed for high-performance processing under the high-availability principle. We show that features existing only in specific packages can be adapted to run in the notebooks, and that algorithms can be adapted to run across the data center without necessarily redesigning them.

Keywords: Jupyter Notebooks, Cloud Computing, Docker, Kubernetes, Dask, High-Performance Computing

1. Introduction

The fast-paced technology improvements on astronomical telescopes, instruments and archives, plus the accumulative nature of astronomical data, contribute to generate an overwhelming data space waiting to be analyzed and exploited. The data deluge problem in astronomy is not a menacing challenge in the horizon, but a very real challenge for current astronomical data analysis. This is the case for the ALMA Data, where some single data products can reach hundreds of GigaBytes, being expensive to transport, store, process, and load them into memory (Testi et al., 2010). The next generation of projects such as Large Synoptic Survey Telescope - LSST (Ivezic et al., 2008), Extreme Large Telescope - ELT (Gilmozzi and Spyromilio, 2007), and the Square Kilometer Array - SKA (Dewdney et al., 2009) will increase the data generation rate from two to three orders of magnitude, so scientists and engineers are preparing themselves to cope with this new reality.

The somewhat mundane problem of personal computers not being able to store and process the astronomical files produced nowadays, have been forcing a paradigm-shift in astronomical data analysis. It is not enough to develop new algorithms and allocate computing infrastructure; modern software and services need to be fostered to equip astronomers with the tools to deal with the data deluge problem.

It is clear that processing tasks need to be moved towards the large archives where the data resides, minimizing the data transfer to the end user and taking advantage of the high-performance computing infrastructure available at the data centers (Djorgovski et al., 2003; Moolekamp and Mamajek, 2015). However, the efficient use of the high-performance computing and storage infrastructure requires a good understanding of technical details that are not in the domain of expertise of astronomers. Moreover, new challenges arise when these services are deployed under the high-availability principle in order to compete with the convenience of local processing.

This paper presents our approach to provide a Cloud service for astronomical data analysis called JOVIAL, which provides Jupyter notebooks to astronomers in a transparent and high-available fashion. Section 2 discuss why Jupyter notebooks are a suitable tool for astronomical data analysis, while Section 3 presents the architecture that we used for offering them in the Cloud. Then, Sections 4 and 5 presents our current efforts to provide a consistent set of libraries for astronomical data analysis and how the computing tasks can be distributed along the available infrastructure. We conclude in Section 6 with our next steps towards improving the services provided by JOVIAL.

2. Notebook-based Astronomical Data Analysis

Modern data analysis strongly relies on computational methods, models and algorithms. These are usually grouped in software packages that provide from very general to very specific functionality that support the analysis process. Despite the large variety of them, one can identify the key elements that these set of tools must cover.

Interactivity. Most of the visualizations and plots that are made during the analysis process are not published nor shared, and only the key results are published. Any discovery process include some trial and error actions, fine tuning and data exploration tasks. Interactive tools help to speedup this discovery process, and that is the main reason why astronomers usually use specialized interactive desktop applications.

*Corresponding author e-mail: mauricio.araya@usm.cl