A DATA LAKE PROTOTYPE SYSTEM FOR MULTIWAVELENGTH (MWA) AND MULTIMESSENGER ASTRONOMY (MMA)

Matheus Bento[1], Reinaldo Roberto Rosa[1,2] and Ulisses Barres de Almeida[3]
[1]*Applied Computing Graduate Program (CAP), INPE-MCTI*
Av. dos Astronautas, 1758, S. J. dos Campos, SP, Brazil, CEP 12227-010
[2]*Lab for Computing and Applied Mathematics, COPDT-INPE-MCTI*
Av. dos Astronautas, 1758, S. J. dos Campos, SP, Brazil, CEP 12227-010
[3]*Centro Brasileiro de Pesquisas Físicas (CBPF)*
R. Dr. Xavier Sigaud, 150 – Urca, Rio de Janeiro, RJ, Brazil, CEP 22290-180

ABSTRACT

At the beginning of the 21st century, large telescopes are coming into operation, making astronomy a Big Data science like few others. Recently, given this scenario, applied computing to astronomy and space sciences has pointed out major challenges involving data science and data engineering. Faced with such a challenge, this article presents a first prototype of a data lake as a strategic solution for heterogeneous data astronomy such as multi-wavelength astronomy (MWA), also including multi-messenger astronomy (MMA). The proposed prototype presents a three-layer architecture, composed of storage, processing and application layers. The storage layer incorporates S3 for raw data storage, MariaDB for simple data storage (crud) and TimescaleDB for complex data storage. The processing layer includes Laravel for CRUD operations and Python for handling large volumes of data. The application layer uses ReactJS with the Cubes.dev library for analytical data visualization. The solution proposes to deliver, when necessary, agility and concatenation between heterogeneous data (light curve, spectra and images) coming from different silos (instruments and telescopes). Thus, the solution, within the MWA and MMA paradigms, incorporates data cube technology and can be customized for applications other than astronomy, such as space weather monitoring and debriefing in the context of space medicine.

KEYWORDS

Data Lake, Multi-Wavelength Astronomy, Heterogeneous Data Science, Big Data, Agile Analytics, Data Cubes

1. INTRODUCTION

Modern astronomy faces a growing challenge to deal with the large quantity and diversity of data generated by observatories around the world (Ahmed et al., 2017). The exploration of the cosmos requires the analysis and integration of information from different kinds of sources, such as terrestrial telescopes, satellites, and space detectors. However, these data are often stored in a fragmented and isolated way in systems known as “silos” of row data which in practice can be strategically analyzed in a concatenated and agile way within the concept of a data lake (Rosa, R. R. 2020). This solution, compared to the traditional data warehouse approach, has been explored in several areas in which agile analysis of raw data, based on machine learning, is required (Che and Duan, 2020). Given the typical data warehouse limitations, there is a gap for a solution that allows efficient integration and analysis of multiple sources of astrophysical data, overcoming the barriers imposed by data silos. It is in this context that the Data Lake solution is useful also in space physics and astronomy. The Data Lake solution provides a centralized and flexible environment for the storage and processing of raw heterogeneous data. It allows the gathering of dispersed information in different silos, promoting the concatenation and integration of different kinds of data into a unified structure. Furthermore, the Data Lake can handle the heterogeneity of data formats, helping to analyze and obtain insights from large volumes of information.

Recently, the strategies in the analysis of large amounts of data for MMA have been established. Following these guidelines, we seek to develop a Data Lake Prototype that allows the agile and related concatenation of data from different observatories, respecting the heterogeneity of the data and handling the challenges imposed by the various silos of information present in the context of MMA and MWA, but that can be applied (and customized) to other areas that involve the same challenges that involve agile Big Data analysis. In this context, this paper presents a solution based on a Data Lake, with a focus on the application of MMA and MWA, which deals with the analysis of various heterogeneous astronomical data sources (silos). Furthermore, it stands out for the management of large volumes of data, seeking to simplify the management, security, sharing, and agile and concatenated big data analysis incorporating the data cube concept. Next, the details of the methodology used in the implementation of Data Lake will be presented in-depth, covering the layers of storage, processing, and application (Chen et al, 2012; Allen et al., 2019).

2. THE PROTOTYPE

The methodology adopted in this study involves the implementation of a data lake prototype (DLP) to be validated and assist the visual analysis of raw data in several projects that have the participation of Brazilian groups (e.g. LSST, SKA, CTA, among others) (see, Rosa 2018).

The DLP structure is composed of three main layers: storage, processing, and application as we can see in Figure 1. In the storage layer, we use different technologies to help us handle different types of data. Amazon S3 (Simple Storage Service) is employed as the primary storage service, enabling scalability and secure storage of astronomical data collected from concatenated observatories and other instruments. MariaDB is used as a relational database to store simple data such as Organization, Security data, etc., while TimescaleDB is used to store complex data in time series format, such as long-term astronomical event data (Zhang et al. 2018; Dean and Ghemawat. 2008).

In the processing layer, we adopted a hybrid approach to handle the manipulation and analysis of simple data/astronomical data. We use the Laravel framework, a PHP web development framework, to implement create, read, update, and delete (CRUD) operations on stored data, database as MariaDB. In addition, we use the Python programming language, along with libraries such as NumPy and Pandas, to perform statistical analysis, data transformations, and processing of large volumes of astronomical data, Laravel is able to orchestrate the connection with the databases and call Python API's to process data (Schreiber et al. 2018; Vouk 2008; Barchi et al., 2020; Jin & Finkel, 2018).

In the application layer, we are using the ReactJS framework to develop an interactive and friendly user interface. Additionally, we've incorporated the Cubes.dev library to make it easy to visualize astronomical data through dashboard analytics as charts, tables, and other visual analytics representations. This layer allows users to interact with the data stored in the data lake, performing custom queries, exploring relationships between data, gaining valuable insights, and also generating astrophysical reports that offer the user management the queries and data input following a given structure (Sidhu and Sehgal, 2015; Giommi et al., 2020).

This DLP is designed to allow a quick concatenation of information among counterpart observatories and heterogeneous astronomical data sources. The data lake and data cube concept offers flexibility and scalability to deal with the multitude of existing data silos in astrophysics, enabling integrated and multidimensional analysis, being possible to create different cubes to handle different data (Cuzzocrea, 2010; Ivezic et al., 2020; Buchschacher et al., 2019; Barres De Almeida et al., 2020).

Furthermore, as part of the methodology adopted, it will be necessary to establish a process for integrating and transforming data from different sources. This will involve using tools and techniques to deal with the heterogeneity of data formats, such as using Python libraries such as Pandas to handle structured data, Pickle Format, CSV, and other common formats. It will also be necessary to explore different technologies of each used storage system, such as Amazon S3, MariaDB, and TimescaleDB, to efficiently extract, transform, and load data. This approach will allow the creation of a flexible and adaptable environment, capable of handling the large volume of data characteristic of astronomy from multiple sources (Kelleher & Tierney, 2018; Allen et al., 2019; Carvalho et al., 2010; Bartos et al., 2017; Rosa, R. R. 2020).

The effectiveness of the proposed methodology is demonstrated through the implementation of a prototype, the DL-MMA (Data Lake for Multi-Messenger Astronomy), which is in testing and being

validated in the context of the mockup data products which will come from the LSST ecosystem. However, we emphasize that the data lake with the data cube approach presented in this study can be applied in other areas of interest that face similar challenges related to the agile and concatenated analysis of data from different silos (Rosa, R. R. 2020). An analytical dashboard, developed in dash.py (there is also a version in react), is incorporated into the datalake front-end to quantify the percentage of data used in a given application (Figure 2).

A typical minimal data cube input organized within the data lake is shown in Table 1. The prototype is called ADALA (Analytical Data Lake for Big Data Astronomy).

2.1 Figures and Tables

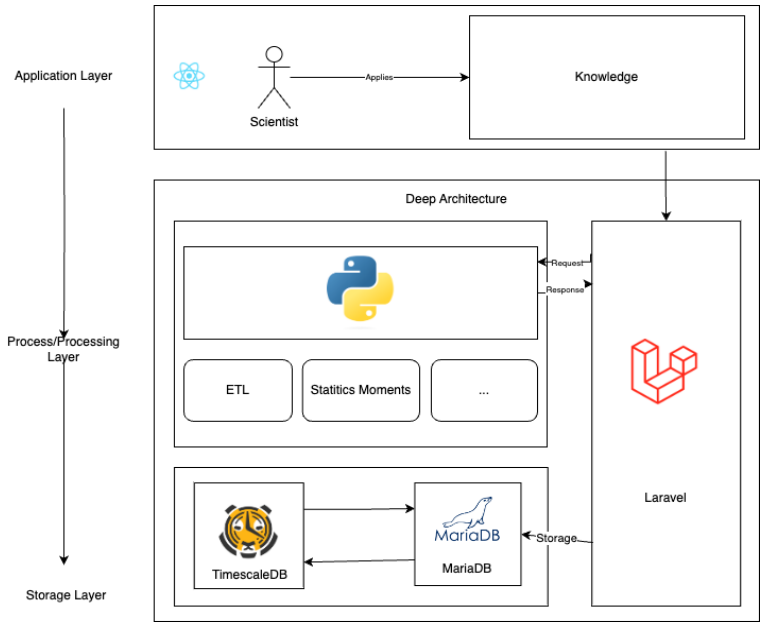

Figure 1. The DLP (also called ADALA in its preliminary version using Python)

Table 1. A hypothetical table based on five instruments that observe the same astronomical event within the MWA/MMA paradigm. The instruments are distributed as instances and the use of their respective dataframes incorporated into the Data Cube work as attributes. A Big Data usage analysis of this primary framework fed by the raw data from each silo is the input to the ADALA dashboard shown in Figure 2

INSTRUMENT	DF1(%) 5.1GB	DF2(%) 6.5GB	DF3(%) 31GB	DF4(%) 112GB
#001	98	95	89	71
#002	90	87	78	56
#003	67	46	45	43
#004	67	40	40	35
#005	59	42	35	30
#006	55	42	35	28
#007	45	38	27	17
#008	40	28	16	08

2.1.1 Results

The results obtained at this time are based on the implementation of a data lake for an active and flexible data system, following the MMA paradigm. The developed solution will offer access to a user-friendly interface, facilitating administration, security, sharing, and agile and concatenated analysis of different types of data.

Considering the deficiency of solutions for the concatenation of information between observatories and the heterogeneity of different kinds of data, the data lake makes it possible to integrate different sources of

data in a single environment, overcoming a fragmented approach and data "silos". Through the layer of processing and application, the concatenation of information was implemented by means of ETL techniques (Extract, Transform, and Load) and analytical analysis (Allen et al., 2019; Rosa, R. R. 2020).

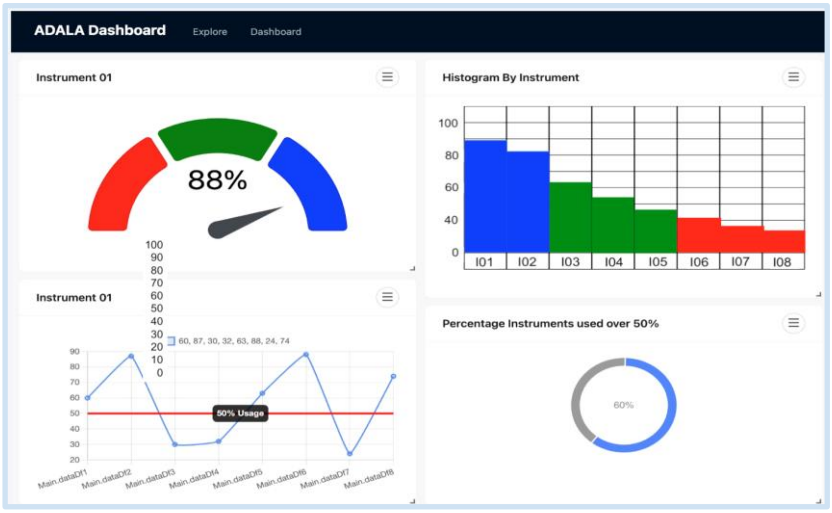

Figure 2. The 2nd order analytical dashboard, implemented in react on the front-end, applied to a hypothetical dataframe shown in Table 1. The info-graphs on the left measures, for each instrument (instances), the amount of data used (upper dash) in a given analysis of 1st order (ex. Figure 3) over time (bottom dash). The infographics on the right inform the percentage of data used by each instrument (upper dash) and the percentage of those whose data used were greater than 50% (lower dash)

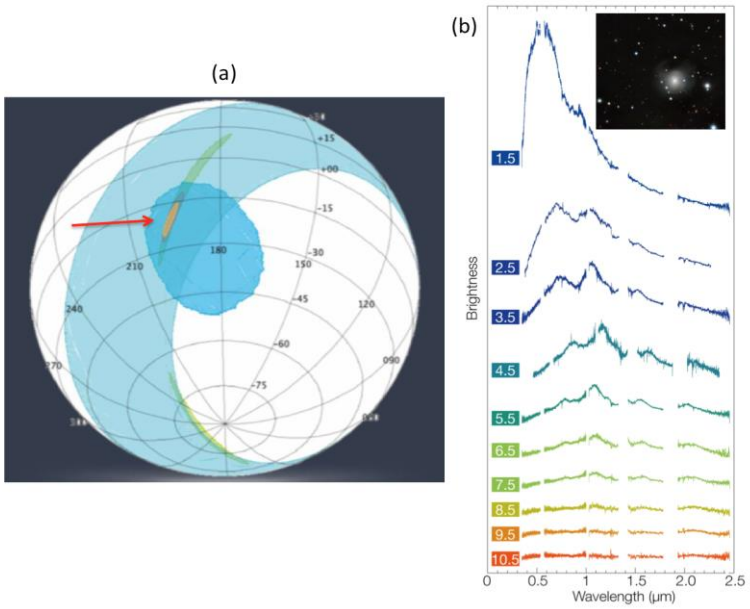

Figure 3. (a) Overlapping sky maps from the final localization of GW170817 via GWTC-1 (red arrow); (b) X-shooter composed spectra of kilonova NGC 4993 with the respective GROND image embedded. From ESCAPE Virtual Observatory. In the context of the developed prototype, the incorporation of the data cube concept allows the aggregation and multidimensional analysis of heterogeneous data. This approach facilitates the exploration and visualization of the two data more efficiently, allowing a deeper understanding of the astrophysical area

As a preliminary result of this application, MMA data were selected within the silo concept for future validation of the prototype, scheduled for the second half of 2023. The validation has been considering the following steps:

1. Loading sky maps with DSS2 in HEALPix FITS format;
2. Location of event GW170817 observed by LIGO and VIRGO;
3. Input of correlated data in other sources/frequencies (via ALADIN Desktop): PGC, GWGC, HyperLEDA, and 2MASS;
4. Generation of the Data Cube by Category (eg Sources and/or Telescopes);
5. Rendering of Location Maps by category (eg Telescopes) (Figure 3.1) and MWA (Figure 3.2).
6. Generation of the integrated MMA dashboard for the GW170817 event.

3. CONCLUDING REMARKS

The DLP presented in this paper is a comprehensive strategy for the development of a Data Lake capable of combining data from multiple sources in an agile way, overcoming the challenges imposed by the existence of different data silos. The implementation of the Data Cube concept embedded in a Data Lake, called DL-MMA (Rosa, 2018), shows promise for dealing with the analytical concatenation of redundant and valuable information in a continuous data cycle. At the conclusion of this test project, we recognize that the continuous evolution of two computational resources, such as programming languages, messaging services, and data banks for time series and other types of data, is essential to maintain the flexibility and portability of the developed solution. We believe that the IT concept based on the Data Cube immersed in the Data Lake has the potential to benefit the scientific community, allowing a more agile and concatenated approach to the analysis of information from different data silos (Beyer & Laney, 2012).

It is worth noting that the concept of a Data Lake with a Data Cube studied and developed in this work is not only limited to the DL-MMA prototype. What makes the concept specific are the attributes that make up the Dice Cube, allowing its application in other astronomical contexts, such as the demands of Virtual Observatories (VOs), as well as in other areas that require agility and analytical concatenation from silos of heterogeneous dice (Hajjat et al. 2011, Allen et al., 2019 and Rosa, R. R. 2020).

As a perspective for future work, it is intended to explore the application of the developed methodology in other astronomical Data Cubes and in areas of interest that involve agility and analytical concatenation. A promising example is the monitoring of Space Weather, which requires analysis of different types of solar, geomagnetic, and ionospheric data. Such data can be integrated with other data systems that involve visual analysis of raw data as is the case of NASA's Artemis mission, the initial step which will establish a sustainable exploration of the Moon to prepare for missions to Mars. In this context, digital platforms that can evaluate astronaut training remotely can be based on DLP like the one we present here. Therefore, the solution presented here has direct applications in other data paradigms, for example in environmental (Velho et al., 2001, Bolzan et al. 2005), and aerospace sciences (Lara et al. 2007, Mattedi et al., 2004), which also demand agility in analytics.

ACKNOWLEDGEMENT

The authors thank CNPq and CAPES for partial financial support. MB thanks for the data available from ALADIN and ESCAPE. RRR thanks FAPESP under Process No 2021/155114-8.

REFERENCES

Ahmed, E., Yaqoob, I., Hashem, I. A. T., Khan, I., Ahmed, A. I. A., Imran, M., & Vasilakos, A. V. (2017). The role of big data analytics in Internet of Things. Computers & Networks, 129(2), 459-471.

Allen, M. G., Dowler, P., Evans, J. D., Cui, C., & Jenness, T. (2019). The International Virtual Observatory Alliance. Proceedings of Astronomical Data Analysis Software and Systems XXVIII. Retrieved from arxiv.org/abs/1903.06636.

Barres de Almeida, U., Krone-Martins, Diaz, M. P., Do Nascimento, J. D., Leo, W. V., Rosa, R. R., & Saito, R. K. (2020). Information technology & astronomical data in Brazil: Perspectives and proposals. Boletim da Sociedade Astronômica Brasileira, 32(1), 142-146.

Barchi, P. H., Carvalho, R. R., Rosa, R. R., Sautter, R. A., & Soares-Santos, M. (2020). Machine and Deep Learning applied to galaxy morphology: A comparative study. Astronomy and Computing, 30, 100334.

Bartos, I., Kowalski, M., Multimessenger Asytonomy. (2017). IOP Publishing.

Beyer, M. A., & Laney, D. (2012). The importance of 'big data': A definition. Gartner, 6(13), 1-3.

Bolzan, M.A., Sahai, P.R., Fagundes, P.R., Rosa, R.R., Ramos, F.M. & Abalde, J.R. (2005). Intermittency analysis of geomagnetic storm time-series observed in Brazil. Journal of Atmospheric and Solar-Terrestrial Physics 67 (14), 1365-1372.

Buchschacher, N., Alesina, F., & Burnier, J. (2019). No-SQL Databases: An Efficient Way to Store and Query Heterogeneous Astronomical Data. In DACE 523, Astronomical Data Analysis Software and Systems XXVIII (pp. 405). ASP.

Carvalho, R. R. et al. (2010). The Brazilian Virtual Observatory - A New Paradigm for Astronomy. Journal of Computational Intelligence in Sciences, 1(3), 187-206.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165-1188.

Cuzzocirea, A. (2010). OLAP Data Cube Compression Techniques: A Ten-Year-Long History. Lecture Notes in Computer Science, Springer.

Giommi, P., Arrigo, G., Barres de Almeida, U., De Angelis, M., Del Rio, J. V., Di Ciaccio, S., Di Pippo, S., Iacovoni, S., & Pollock, A. (2020). The Open Universe Initiative. In Space Capacity Building in the XXI Century (pp. 377-386). Springer.

Hajjat, M., Liu, X., Nguyen, T. N., Wang, C., & Zhang, Z. (2011). Cloudward bound: Planning for beneficial migration of enterprise applications to the cloud. IEEE Network, 25(4), 28-35.

Ivezic, Z., Connolly, A. J., Vanderplas, J. T., & Gray, A. (2020). Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton University Press.

Jin, Z., & Finkel, H. (2018). Power and performance tradeoff of a floating-point intensive kernel on OpenCL FPGA platform. IEEE Symposium on Parallel and Distributed Processing (pp. 716-720).

Kelleher, J. D., & Tierney, B. (2018). Data Science. MIT.

Lara, A., Borgazzi, A., Mendes Jr., O., Rosa, R.R. & Domingues, M.O. (2008). Solar Physics 248, 155-166.

Mattedi, A., Ramos, F.M., Rosa, R.R. & Mantegna, R.N. (2004), Value-at-risk and Tsallis statistics: risk analysis of the aerospace sector. Physica A: Statistical Mechanics and its Applications 344 (3-4), 554-561.

Rosa, R. R. (2020). Data Science Strategies for Multimessenger Astronomy. Anais da Academia Brasileira de Ciências, 93. DOI: 10.1590/0001-3765202020200861

Schreiber, R., Carpentier, M., Tchounikine, A., & Bresson, G. (2018). Towards a big data architecture for learning analytics. IEEE Transactions on Learning Technologies, 11(2), 198-208.

Sidhu, P., & Sehgal, R. (2015). A survey on data warehousing and big data analytics. International Journal of Computer Applications, 125(2), 28-33.

Velho, H.F.C., Rosa, R.R., Ramos, F.M., Pielke, R.A., Degrazia, G.A. & Neto, C.R. (2001). Multifractal model for eddy diffusivity and counter-gradient term in atmospheric turbulence. Physica A: Statistical Mechanics and its Applications 295 (1-2), 219-223.

Vouk, M. A. (2008). Cloud computing: Issues, research, and implementations. Journal of Computing and Information Technology, 16(4), 235-246.

Zhang, M., Zhang, Y., & Zhang, D. (2018). Big data storage systems: A survey. IEEE Access, 6, 18329-18352.

Zhang, Z., & Zhao, W. (2015). Astronomy in the Big Data Era. Data Science Journal, doi:10.5334/dsj-2015-011.