Big Data Analytics to Enable Integrated Research of Biodiversity and Climate Datasets in the Amazon Basin

Pedro Luiz Pizzigatti Corrêa[1], Giri Prakash[2], Mike Frame[3], Bhargavi Krishna[2], Luciana Rizzo[4], Ricardo Oliveira[5], Wesley Barbosa[5], André Batista[5], Paulo Artaxo[6], Solange Alves-de-Souza[5], and Katia Ferraz[5]

[1]University of Tennessee
[2]Oak Ridge National Laboratory
[3]USGS Headquarters
[4]Universidade Federal de São Paulo
[5]Universidade de São Paulo
[6]USP University of Sao Paulo

November 24, 2022

Posted on 24 Nov 2022 — CC-BY 4.0 — https://doi.org/10.1002/essoar.10500800.1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Abstract

With the mass adoption of data analysis in several scientific fields such as climatology, medicine, astronomy and astrophysics, the availability of an appropriate analytics infrastructure has become a necessity increasingly recognized by the scientific community. However, appropriate tools and applications are required to process the large volume of data collected and generated by researchers. One of the biggest challenges lies in the fact that these tools need to be gathered to be applied in specific domains. The area of bioclimatic data is a scientific field that still has much to improve in this matter. It is a field of study that lacks great efforts in the direction to provide methodologies and tools to facilitate the understanding of the complex phenomena involved in the influence that environmental variables have on biodiversity on the planet. Thus, the purpose of this work is to propose a big data analytics architecture that presents an ecosystem that systematizes and facilitates the task of the scientists to deal with the complexity in the bioclimatic data analysis, providing tools for storage, management, analysis using machine learning algorithms and data mining, and visualization tools. The methodological approach of this work was to make a thorough bibliographical study to verify the most used tools and the suitability of each one to the purpose of the work. In addition, the literature provided indications of software ecosystem implementations methodologies that served as a guide in the architecture design. Within the architecture, we attempted to gather a set of bioclimatic data based on a subset of data obtained from the Atmospheric Radiation Measurement (ARM) data repository for climatic data, and the Brazilian Biodiversity Portal for biodiversity data. As a result, we were able to gather a series of tools to access data such as Cassandra, distribution of processing such as Spark, programming interface represented by Jupyter Notebook, system modules for data format conversion, machine learning algorithms libraries and software for data visualization. This research discuss the importance of a domain purpose design of a data analysis architecture for bioclimatic data. We concluded that this type of ecosystem is imperative to facilitate the research process and increase the quality of the results.

AGU Fall Meeting 2018

Big Data Analytics to Enable Integrated Research of Biodiversity and Climate Datasets in the Amazon Basin

Pedro Pizzigatti-Corrêa - University of São Paulo
Giri Prakash - Oak Ridge National Laboratory
Mike Frame - United States Geological Survey
Bhargavi Krishna - Oak Ridge National Laboratory
Luciana Rizzo - Federal University of Sao Paulo
Andre Batista - University of Sao Paulo
Wesley Barbosa - University of Sao Paulo
Ricardo Oliveira - University of Sao Paulo
Katia Ferraz - University of Sao Paulo
Paulo Artaxo - University of Sao Paulo
Solange Souza - University of Sao Paulo

INTRODUCTION

With the mass adoption of data analysis in several scientific fields such as climatology, medicine, astronomy and astrophysics, the availability of an appropriate analytics infrastructure has become a necessity increasingly recognized by the scientific community. However, appropriate tools and applications are required to process the large volume of data collected and generated by researchers. One of the biggest challenges lies in the fact that these tools need to be gathered to be applied in specific domains. The area of bioclimatic data is a scientific field that still has much to improve in this matter. It is a field of study that lacks great efforts in the direction to provide methodologies and tools to facilitate the understanding of the complex phenomena involved in the influence that environmental variables have on biodiversity on the planet. Thus, the purpose of this work is to propose a big data analytics architecture that presents an ecosystem that systematizes and facilitates the task of the scientists to deal with the complexity in the bioclimatic data analysis, providing tools for storage, management, analysis using machine learning algorithms and data mining, and visualization tools. Within the architecture, we attempted to gather a set of bioclimatic data based on a subset of data obtained from the Atmospheric Radiation Measurement (ARM) data repository for climatic data, and the Brazilian Biodiversity Portal for biodiversity data. As a result, we were able to gather a series of tools to access data such as Cassandra, distribution of processing such as Spark, programming interface represented by Jupyter Notebook, system modules for data format conversion, machine learning algorithms libraries and software for data visualization.

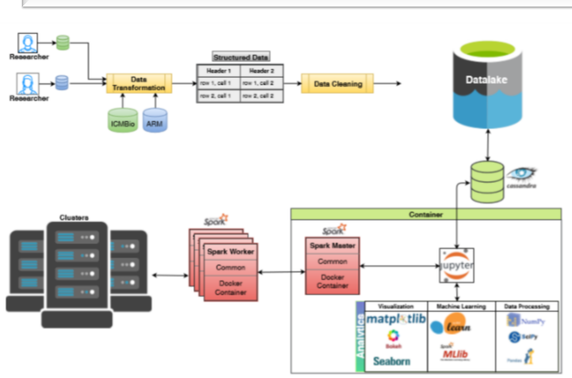

Figure 1 – Second version of the Architecture - BioClimate Big Data Analytics.
Source: Authors

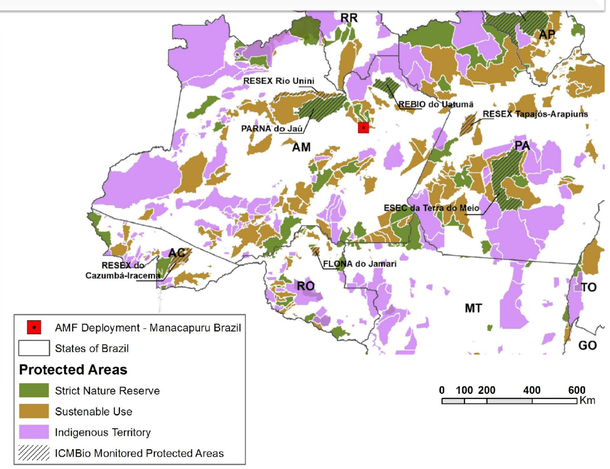

Figure 2 – Protected Areas and ARM deployment at regional scale.
Source: Authors

RESULTS

The first experiments used the dataset of Brazilian Biodiversity Portal for biodiversity data about occurrences of butterflies, plants, birds, and mammals for regional-scale located around the GOAmazon experimental sites to develop a biodiversity-climate hybrid data visualization and analysis tools.

Class	Most Influential Parameters
Birds	Humidity, NO_2, NO
Amphibians	Temperature, Humidity, NO_2
Insects	Humidity, Temperature, NO_2
Mamals	Temperature, Humidity, N_2
Fishes	Temperature, Precipitation, NO_2
Reptiles	Temperature, Humidity, NO_2

Table 1 – Preliminary results with the most influential variables for each class of families.
Source: Authors

NEXT STEPS

Run the same experiment in the second version of Bioclimate analysis Architecture;

Develop an analysis pipeline to automate the first experiment and calculate performance metrics;

Integrate Datasets from T3 Tower in the Biodiversity Data from this region;

Development of new analysis to the correlation effects of pollution in species distribution.

REFERENCES

[1] ICMBio – Biodiversity Monitoring. Methodological procedures applied to monitoring - *Monitoramento da Biodiversidade Roteiro metodológico para aplicação de monitoramento*. Authors: Rodrigo de Almeida Nobre, Marcelo Rodrigues Kinouchi, Pedro de Araujo.

[2] CORREA, P.L.P. *et al. Information system architecture for the integrated management of the biodiversity monitoring data in protected areas* ., 2015 ICMBio proceedings.

[3] DASGUPTA, A.; POCO, J.; WEI, Y.; COOK, R.; BERTINI, E.; SILVA, C. T. Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison. IEEE Transactions on Visualization and Computer Graphics, v. 21, n. 9, p. 996-1014, 2015. ISSN 1077-2626.

ACKNOWLEDGEMENT

This research is supported by grant #2016/04982-0, São Paulo Research Foundation (FAPESP) - Brazil - Research Program on eScience.