A Standardized GIS Web Service Oriented Shared-nothing Architecture - Taking Species Distribution Modeling as an Example

X Y Liu¹, J S D Yu¹*, D D Wei¹, R J Tong², D Y Li¹ and S Luo¹

¹The Academy of Digital China(Fujian), Fuzhou University, Fuzhou, China
²Department of Transportation, Fujian University of Technology, Fuzhou, China

*Corresponding author, e-mail: yyx350@126.com

Abstract. The rapid development of big data analytics technologies and tools, particularly the related distributed computing technologies, which divide large problems into smaller ones, provides important technical support for developing big earth data research. To realize the effective utilization of big data computing resources, it is important to overcome the potential incompatibility and inconsistency among the distributed computing nodes. Using standardized GIS web service interfaces is a promising approach to link existing heterogeneous GIS platforms through standardized protocols and establish a universal, platform-independent computing layer to improve the efficiency of computing resources for big earth data analysis. This research analyzed the requirements of distributed computing, and further proposed a standardized GIS web service-oriented shared-nothing architecture by comparing various mainstream distributed computing architectures and related web service specifications in the industry. For example, the prediction of Lophelia pertusa coral distribution by random forest is the application of the method discussed in this paper.

Keywords- Big Data; parallel computing; GIS Web services

1. Introduction
The rapid development of emerging big data analytics technologies and tools, especially the related distributed computing technologies, which divide large problems into smaller ones, provides important technical support for the development of big earth data research. Big data analytics always refers to processing datasets of a size that cannot be done on a single computing unit and for traditional data processing and algorithms that are inadequate.

Distributed computing always refers to a technology that carries out many calculations or processes simultaneously by multiple computing devices. There are several different types of distributed computing architectures, mainly including shared-memory architecture, shared-disk architecture, shared-nothing architecture, etc. In a shared-memory architecture, memory may be simultaneously accessed by multiple programs without redundant copies. It may also improve performance for passing data between programs. However, when several processors try to access the same memory location simultaneously, the memory bus is impeded due to memory contention. In a shared-disk architecture, each node has its own private memory, and the disks are accessible from all of the computing nodes. In a shared-nothing architecture, all nodes have sole access to distinct memory or storage. Hence, single points of failure and resource contention among nodes will be avoided. As nothing is shared in each node under a shared-nothing architecture, computing nodes of different communities can have their own processing framework without knowing each other. Hence, consistent functionality of interface compatibility for communication cannot always be guaranteed.

Normally, distributed computing technologies are not particularly designed for use over a wide-area network, such as the global internet, while web services provide the potential for creating and deploying loosely coupled applications
globally. However, web services are different from traditional distributed computing architectures. Different services tend to be heterogeneous when the applied interface specifications are neither compatible nor consistent.

Therefore, to realize the effective coordination of big earth data computing resources, it is important to overcome the potential incompatibility and inconsistency among nodes. Standardized interfaces are a promising approach to link existing heterogeneous GIS platforms through standardized protocols and establish a universal, platform-independent computing layer to improve the efficiency of big earth data analytics resources. Additionally, a series of related standardized works that have been delivered by international standardization organizations, including ISO, IETF, W3C, etc., where ISO TC211, in collaboration with Open Geospatial Consortium (OGC), are currently the main driving forces in open standards for interoperable GIS web services [1-4].

Regarding standardized GIS web service-oriented distribution architecture, Owonibi proposed a related D-WCPS framework to support several WCPS servers to share a service request for parallel computing [5]. The approach may still be enhanced by introducing ad hoc computing algorithms under unified standardized web interfaces. Yu proposed a standard-based collaborative analysis method based on Web Coverage Processing Service (WCPS) and R to effectively improve the interoperability and processing of spatial data infrastructure [6]. However, the method does not cover the distribution issues. Efforts have been investigated on the improvement of GIS parallel computing [7][8] without concerning the standardization issue. Some studies have realized the parallel management and visualization of multi-source heterogeneous spatial data based on standardization, but have not further completed the analysis and calculation [9][10]. This research proposes a standardized GIS web service-oriented shared-nothing architecture to address the issue. A study case based on species distribution modeling is examined to explore the application realization method.

The rest of the paper is organized as follows: Section 2 presents the study case. Section 3 presents the configuration of the proposed architecture. Section 4 describes the experiment and includes the discussion. Section 5 analyzes acceleration performance. Section 6 concludes the paper.

2. Case study

Marine species distribution models are often applied to predict the potential distributions of species across geographic space or time using environmental variables. The environmental variables may include temperature, precipitation or water depth. When modeling a large area with dozens of high-resolution environmental variables, a large amount of computing resources is needed to process the big data and build the model. When modeling with terabyte or petabyte data, a series of circumstances occur, such as having to take several hours or even more time for data retrieval, having a large amount of model calculation, having insufficient memory, and having to wait a long calculation time for prediction models. Species distribution prediction can be divided into a model training process and a model prediction process. Thus, parallel computing can be used in the above two processes to improve the computational efficiency, but this study only focuses on the model prediction process for parallel computing acceleration processing.

In this study, modeling the potential distribution of Lophelia pertusa coral in the North Atlantic [11] was taken as an example to develop the application framework. Biomod2 is an R language package for species distribution modeling [12]. Based on the environmental variables (EVas) and training samples (TSse), a random forest (RF) model was built through calculation and further evaluated based on testing samples. The RF model, together with EVas, was used to complete the projection for the prediction of species distribution. The serial calculation data flow chart is shown in Figure 1.

Species distribution prediction algorithms involve two processes, including RF model establishment and projection. With reference to the "divide-and-conquer" concept of big data processing [13]. The approach is to build an RF model
for EVas and TSse, split EVas, project each node in parallel based on the RF model and EVas partition data, and finally combine the partitioned results. The parallel computing data flow diagram is shown in Figure 2.

![Figure 2. Parallel computing data flow](image)

3. Configuration of the sharded-nothing architectural

Based on a Rasdaman and R parallel approach, the proposed method is improved to support distributed computing of the RF model by using submarine habitat variables and *Lophelia pertusa* coral distributed/non-distributed data.

Rasdaman [14] is a reference implementation of OGC WCS 2.0. In this study, the RF model was trained at the master node, which was delivered to the slave nodes based on the SCP protocol. Then, the shell script of distributed computing was activated through the SSH protocol. Slave nodes retrieved the partitioned EVas based on OGC WCS 2.0. The parallel projection was completed by R scripts, and then the partitioned results were sent back to the master node based on the SCP protocol. Finally, the complete result was combined by GDAL. The architectural is shown in Figure 3.

![Figure 3. Distributed parallel computing service framework](image)

4. Experiment

This experiment completes the *Lophelia Pertusa* coral distribution prediction based on submarine habitat variables [15], such as water depth, chlorophyll, alkalinity, etc. The computing node configuration is listed in Table 1.

| CPU                  | RAM | Operating System | Mechanical hard disk |
|----------------------|-----|-------------------|----------------------|
| Intel(R) Core (TM) i7-9700 | 16G | Ubuntu 18.04 LTS | 1T                   |

Based on the same submarine habitat variables, different sizes of data (as shown in Table 2) were used to assess the performance of a single computing node in the prototyped parallel architecture. The performance is illustrated in Figure 4.
Table 2. Data size table

| Experiment NO. | Data Size(M) |
|---------------|-------------|
| 1             | 9216        |
| 2             | 4608        |
| 3             | 2355.2      |
| 4             | 1126.4      |
| 5             | 563.3       |
| 6             | 281.1       |
| 7             | 141.7       |
| 8             | 69.7        |
| 9             | 35.4        |
| 10            | 18.3        |
| 11            | 9.1         |

Figure 4. Node computing time

5. Analysis

To explore the theoretical speedup of this study, modeling the speedup function of a single computing node is required. The node computing time in Figure 4 was used to calculate the theoretical speedup ratio of a single computing node under the proposed parallel architecture. The curve is consistent with the exponential semivariogram curve; refer to Figure 5. Its mathematical expression is as follows:

\[
y(x) = \begin{cases} 
0, & x = 0 \\
C_0 + C(1 - e^{-\frac{x}{a}}), & x > 0 
\end{cases}
\]

where \( C_0 \) denotes nugget value; \( C + C_0 \) represents sill.
The calculated fitting function is as follows:

\[
y(x) = \begin{cases} 
0, & x = 0 \\
1.326 + 95.66 \times (1 - e^{-\frac{x}{77.94}}), & x > 0 
\end{cases}
\]  

(2)

The fitting curve is depicted in Figure 6, and the characteristic parameters are listed in Table 3. R-square is 0.9941. The result reveals that the mathematical function is well fitted with the experiment data. When more than 234 nodes are theoretically adopted for parallel computing, the speedup effect will mostly remain constant when approaching the maximum value of 96.986.

**Table 3. Characteristic parameters**

| Nugget | Sill   | Range | R-square |
|--------|--------|-------|----------|
| 1.326  | 96.986 | 234   | 0.9941   |

6. Conclusion

This study delved into the random forest prediction of the distribution of *Lophelia Pertusa* coral in the North Atlantic Ocean. A standardized GIS web service framework was proposed based on the Rasdaman, shared-nothing architecture,
and the distributed computing method. The study involved the decomposition of the prediction model considering operator dependence, splitting the heterogeneous data of multiple sources, realizing interoperation and parallel processing, and exploring the theoretical speedup model of a single computing node under the proposed parallel architecture. In this framework, incompatibility and inconsistency among computing nodes are obviated, the problem of insufficient memory is addressed, and the prediction time of the *Lophelia Pertusa* coral distribution is reduced. Therefore, it can potentially provide a faster solution for cold-water coral prediction on a global scale among heterogeneous GIS platforms. However, this research does not split both RF model establishment and projection processes. The next step will focus on both process distribution and optimization to achieve better performance by considering the accuracy tolerance. This study is part of developing a standardized GIS web service-oriented distributed computing architecture for species distribution modeling. In this way, an improved species distribution modeling performance under evolving implementations is expected.

References

[1] International Organization for Standardization 2005 ISO 19123:2005 Geographic information -- Schema for coverage geometry and functions[CrossRef]
[2] Open Geospatial Consortium 2012 OGC 09-110r4 OGC® WCS 2.0 Interface Standard- Core: Corrigendum[CrossRef]
[3] Open Geospatial Consortium 2012 OGC 09-146r2 OGC® Implementation Schema for Coverages[CrossRef]
[4] Open Geospatial Consortium 2021 OGC 08-068r3 Web Coverage Processing Service (WCPS) Language Interface Standard[CrossRef]
[5] Michael O Resource-Aware Decomposition of Geoprocessing Services Based on Declarative Request Languages(Jena: Friedrich Schiller University Jena)[CrossRef]
[6] Yu J S D, Li J X, Peter B, Tong R J and Chen Y J 2018-9-22 A Standard-Based Collaborative Analytics Service of Remote Sensing Imagery 2018 Inter. Workshop on Big Geo. Data and Data Sci. (BGDDS) (Wuhan) pp 1-4[CrossRef]
[7] Yu J, Wu J X and Mohamed S 2015-11-3 GeoSpark: a cluster computing framework for processing large-scale spatial data Proc. of the 23rd ACM SIGSPATIAL Int. Conf. on Advances in Geo. Info. Sys. (Washington) pp 1-4[CrossRef]
[8] Yue P, Wu Z Y and ShangGuan B Y 2018 Design and Implementation of a Distributed Geospatial Data Storage Structure Based on Spark Geomatics and Info. Sci. of Wuhan University 43 pp 2295-2302[CrossRef]
[9] Liu Z W, Li S Y, Yu H J and Hao Z W 2016 Key Technology Research of massive multi-source heterogeneous spatial data visualization and management system based on 3D digital earth 2016 Int. Conf. on Info. Sys. and Artificial Intelligence (Hong Kong) pp 241-246[CrossRef]
[10] Huang Z C, Zhong A R and Li G Q 2016-6-24 One-stop Service for Multi-Source Heterogeneous Remote Sensing Data 7th Int. Conf. on Mechatronics, Control and Materials (ICMCM 2016) pp 417-423[CrossRef]
[11] Tong R J, Autun p, Janine G, Vikram U, Yu J S D and Zhang C C 2016-3-15 Quantifying relationships between abundances of cold-water coral *Lophelia pertusa* and terrain features: A case study on the Norwegian margin Continental Shelf Research 116 pp 13-26[CrossRef]
[12] Wilfried T, Damien G, Maya G, Robin E and Frank B 2021 Package ‘biomod2’--Ensemble Platform for Species Distribution Modeling[CrossRef]
[13] Hong S 2016 Big data analytics: six techniques Geo-spatial Info. Sci. 19 pp 119-128[CrossRef]
[14] Baumann P, Misev D, Merticariu V, Huu B P, Bell B 2018-11-6 rasdaman: Spatio-Temporal Datacubes on Steroids Proc. of the 26th ACM SIGSPATIAL Int. Conf. on Advances in Geo. Info. Sys. (Washington) pp 604-607[CrossRef]
[15] Davies A J and Guinotte J M 2019 Global Habitat Suitability for Framework-Forming Cold-Water Corals - Environmental Data[CrossRef]

Acknowledgments

The authors would like to express appreciations to Large-Scale Scientific Information Systems research group at Jacobs University Bremen for their valuable contributions on related standards and reference implementations. This research was supported by the National Key R&D Program of China (NO. 2019YFE0127100), and
Natural Science Foundation of Fujian Province of China (No.2019J0006). The RF model training in this paper has been done on the supercomputing system in the Supercomputing Center of Fujian of China.