THE GEOSPATIAL DATA CLOUD: AN IMPLEMENTATION OF APPLYING CLOUD COMPUTING IN GEOSCIENCES*

Xuezhi Wang[1*]*, Jianghua Zhao*[1]*, Yuanchun Zhou*[1]*, Jianhui Li*[1]

[1]*Scientific Data Center, Computer Network Information Center, Chinese Academy of Sciences, Beijing, China*
**Email:* wxz@cnic.cn

ABSTRACT

The rapid growth in the volume of remote sensing data and its increasing computational requirements bring huge challenges for researchers as traditional systems cannot adequately satisfy the huge demand for service. Cloud computing has the advantage of high scalability and reliability, which can provide firm technical support. This paper proposes a highly scalable geospatial cloud platform named the Geospatial Data Cloud, which is constructed based on cloud computing. The architecture of the platform is first introduced, and then two subsystems, the cloud-based data management platform and the cloud-based data processing platform, are described.

Keywords: Remote sensing data, Cloud computing, Geospatial data cloud, Geosciences

1 INTRODUCTION

With the development of earth observation technology and GIS technology, massive geospatial data, especially remote sensing data, are accumulating and being used in various industries and research institutes. Geospatial data are more complex than other data types because they contain location information. Furthermore, because the volumes of geospatial data are huge, a large amount of space and time usage of remote sensing images is required for many applications, especially in the field of environmental science. For example, research on global climate change requires geospatial data covering more than individual countries or continents. It is difficult for individuals or researchers not only to obtain all the data they need but also to process the massive amount of geospatial data because of a lack of high performance computational infrastructure. How to effectively integrate the continuously growing geospatial data resources and provide the public with services such as spatial data downloading, storing, and analyzing has become an urgent problem.

The advent of cloud computing provides a potential solution for the problem described above. Since cloud computing was first put forward by IBM in 2007 (Sims, 2007), it has quickly become a hot spot in the IT industry. The development of cloud computing is based on the maturity of several technologies, including virtualization technology, distributed computing, gird computing, service-oriented architecture, web 2.0 technology, and so on. Cloud computing concentrates computation and storage resources in a core where high-performance machines are linked by high-bandwidth connections (Cui, Wu, & Zhang, 2010). Users access the carefully managed resources based on their practical needs. All of these are of high mobility and collaboration, which not only make Cloud computing an elastic and on-demand computer platform but also provide the theoretical basis and technology support for massive remote sensing storage and processing services.

Researchers at home and abroad have already begun to study how to apply cloud computing in massive geospatial data management and sharing services. Cary, Sun, Hristidis, and Rishe (2009) tried combining cloud computing with spatial data storage and processing and solved two key problems by using MapReduce framework, an R-Tree index for all kinds of spatial data and remote sensing image quality evaluations. Blower (2010) implemented a web map service on Google App Engine and described the difficulties in developing GIS systems on the public cloud. Huang, Yang, Nebert, Liu, and Wu (2010) demonstrated how to utilize cloud computing to support geoscience applications by describing the deployment and maintenance of the GEOSS Clearinghouse on the EC2 platform. Siládi, Huraj, Polčák, and Vesel (2012) applied cloud computing in GIS science by carrying out spatial data interpolating computing on the Amazon Elastic Compute Cloud. This was necessary because geospatial data processes are very time-consuming and data intensive, especially for large

----

* This paper was presented at the First Scientific Data Conference on Scientific Research, Big Data, and Data Science, organized by CODATA-China and held in Beijing on 24-25 February, 2014.

and complex datasets. Rezgui, Malik, and Yang (2013) also used spatial interpolation to assess the benefits of cloud computing for geoscience applications.

Wang, Wang, and Zhou (2009) proposed an interoperable spatial data object model and re-designed spatial indexing algorithms, such as Quad-Tree and R-Tree, to solve the drawbacks of spatial data storage in the common cloud computing platform. Liu, Guo, Jiang, Gong, et al. (2009) analyzed the necessity of adopting cloud computing in the remote sensing data processing service. And by studying the framework and key technologies of cloud computing, they implemented a prototype system named OpenRS-Cloud. Wang, Han, Tu, Dai, Zhou, and Song (2010) proposed that MapReduce was not suitable to express spatial computation due to unfit features and performance degradation and provided several solutions. Yang (2010) studied spatial data clustering service architecture and data storage model and network distribution methods and implemented a prototype system named TSS. Kang (2011) developed a prototype cloud based system to store and process high-resolution remote sensing images as well. Fang (2011) proposed a theoretical cloud computing framework to effectively process land resource services, based on mainstream theory and key technologies of cloud computing. Liu (2013) studied key technologies of cloud computing and implemented an electronic chart cloud service that achieved efficient management of global electronic chart data and provided personalized, flexible, and highly available service.

All these works have promoted improvement of the theoretical study of the combination of cloud computing with massive geospatial data service. However, some problems still exist. First, the systems developed are mostly unstable prototypes usually lacking detailed descriptions and with little content that is worth referencing. Second, the prototype systems implemented were tested with a small amount of data, not enough to validate the performance of a cloud-based system. Third, many studies used a public cloud platform such as Google, Amazon, Microsoft, etc. to develop their cloud based geospatial data platform. As the internal implementation mechanisms of the public cloud are unknown, it is difficult to further study the underlying combination of cloud computing with geosciences service. In addition, these public clouds are commercial, and users need to pay for the services.

The conditions of using cloud computing to provide the public with the ability to access and process massive geospatial data are demanding. Not only do the bandwidth and the scale of the organization need to be large but also a large economic investment and much time for development are required. The Scientific Data Center of the Computer Network Information Center, a research institute supported by the Chinese Academy of Sciences, has the ability and responsibility to undertake such a task. In 2007, the Scientific Data Center launched construction of the International Scientific Data Services Platform, which officially began providing services in 2008. The platform has been upgraded since then by applying cloud computing technology. It has been renamed the ‘Geospatial Data Cloud’ and is free for users to download massive geospatial data and use the data processing service. This paper illustrates the architecture of the Geospatial Data Cloud in detail and also describes key technologies used in implementing the cloud platform. The structure of the paper is as follows: Section 2 briefly introduces the multi-tiered architecture that the Geospatial Data Cloud development was based on. In Section 3, we introduce the cloud-based data collection platform. Section 4 introduces the cloud-based data processing platform. Conclusions and discussions on future work concerning the application of cloud computing in geospatial science are given in Section 5.

2 ARCHITECTURE

This section briefly describes the system architecture of the Geospatial Data Cloud. It is represented by a framework that includes the geospatial data cloud portal, geospatial application layer, the cloud computing environment, and basic infrastructure. This architecture is depicted in Figure 1.

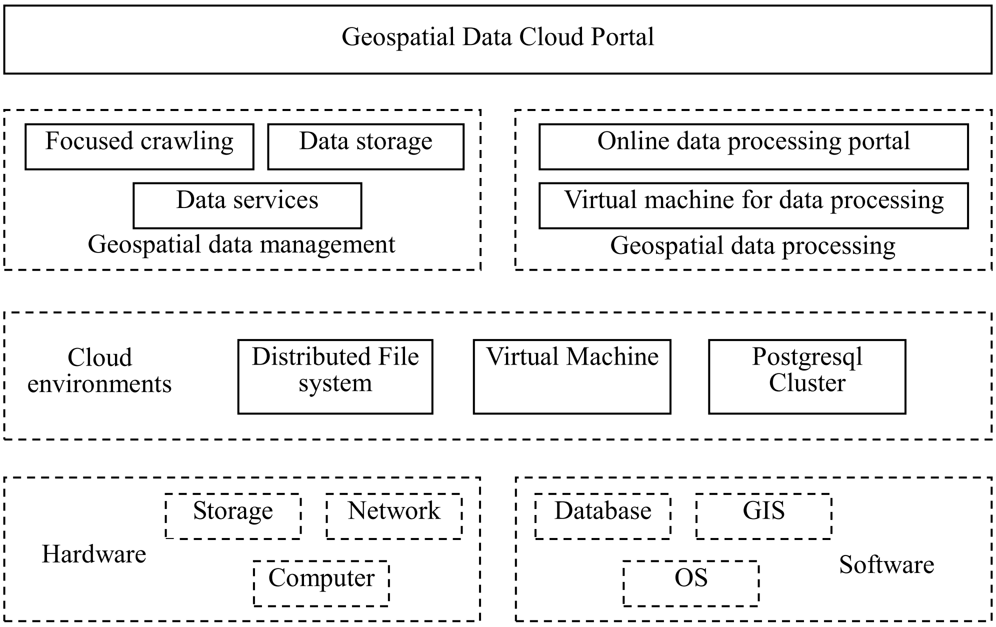


Figure 1. The system architecture of the Geospatial Data Cloud

2.1 Geospatial data cloud portal

On top of the architecture of the Geospatial Data Cloud Portal is a one-stop portal that provides a variety of services through a browser. Through this layer, any authorized user can use networked terminal equipment to log in to the cloud platform in accordance with a standard public application interface to enjoy cloud service from anywhere. All the users have to do is submit requests. The main functions of the portal are: data retrieval, batch data downloading, remote sensing image visualization, private space to store data, and so on.

2.2 Geospatial application layer

The geospatial application layer is the core component of the Geospatial Data Cloud because the cloud platform was developed specifically for geosciences. This layer contains two modules: the cloud-based data management platform and the cloud-based data processing platform. These not only optimize the cloud computing resources but also combine cloud computing with GIS technology.

2.3 Cloud environment layer

The cloud environment layer is the most important component in the Geospatial Data Cloud. In it, virtualization technology enables the cloud architecture to become scalable and elastic. In fact, cloud computing is actually a virtualized resource pool that provides storage and computing resources by way of the cloud host. It links up the basic hardware and software resources with upper applications.

In the Geospatial Data Cloud, a distributed file system is adopted when storing remote sensing data. A virtual server virtualizes the postgresql cluster and enables users to access all the underlying resources by way of the cloud host. Also, by scheduling tasks, it can locate all the nodes to support the execution of a given programming model.

2.4 Resource layer

The lowest layer of the stack is the basic infrastructure layer, which is fundamental to the platform. It consists of physical and basic resources including servers, storage, networks, databases, operating systems, and other software. As the Geospatial Data Cloud is the application of cloud computing in geosciences, common GIS software should be pre-installed.

This overview of the Geospatial Data Cloud provides a general idea of the use of cloud computing in geospatial sciences. The sections below mainly focus on two sub-platforms, the cloud-based data collection platform and the cloud-based data processing platform.

3 THE CLOUD-BASED DATA MANAGEMENT PLATFORM

The data management platform contains two important modules, the data collection module, which includes automatic metadata crawling and the data entities collection, and the data service module, which provides massive geospatial data storage and high effective data retrieval interfaces.

3.1 Data collection module

3.1.1 Focused crawling for geospatial metadata

Data is the basis of a cloud service platform. In the field of remote sensing, the sources of the data, for example, LANDSAT, MODIS, EO-1, DEM, NCAR, NOAA, LUCC, etc., are diverse. As data entities are often complex and heterogeneous, metadata are often used to describe data entities. In order to unify the expression of metadata and data entities, the authors of this paper studied data expression in geosciences, international common data representation, and data interoperability standards. Then we created a metadata information model that is locally compatible. On the basis of these works, a local metadata converter was developed to parse, transform, and extract information from metadata from diverse data resources.

The contents of the metadata include the resolution, source, format, and all items of the datasets. Users can locate and retrieve geospatial data entities through the description and retrieval mechanism of the metadata. As the remote sensing images are being continuously produced, the metadata need to be updated daily. In order to do so, the metadata crawl program has to be executed on time every day. The metadata updating process is shown in Figure 2.

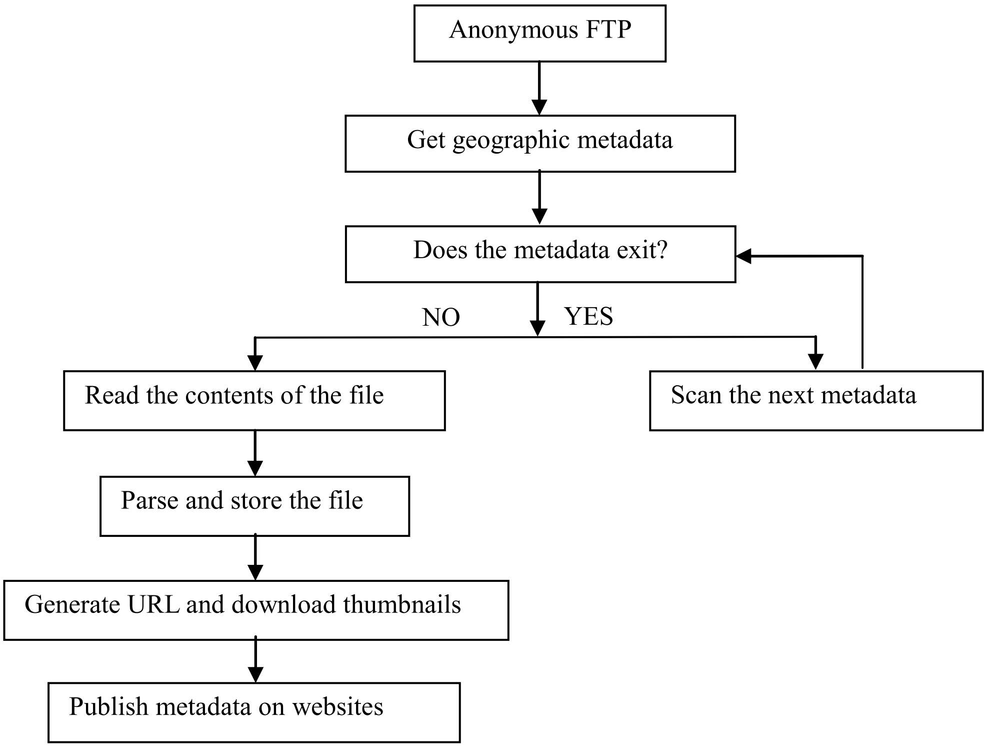


Figure 2. Metadata updating process

First, the process anonymously logs into the FTP websites and retrieves the metadata file. Then it determines whether the same metadata already exist. If the metadata exist, then it continues to scan the next file on the FTP website. If there are no such metadata, it reads the contents, parses, exports, and finally saves the new metadata in a database. At the same time, it generates URLs and downloads thumbnails. Finally, it publishes the metadata on websites for retrieval by users.

3.1.2 On-demand caching for data entities

In the Geospatial Data Cloud, data entities are crawled according to users' orders; this is called "on-demand" cache service for data entities. The process is depicted in Figure 3. First, users submit their orders, which then enter the order queue. Next, the platform caches data from the Internet and validates the data. If the data meet the requirements, they will be stored in the cloud. If the data are not qualified, then other data are cached from the Internet. Finally, an email is sent to notify users that they can download the data requested in their order from the cloud platform.

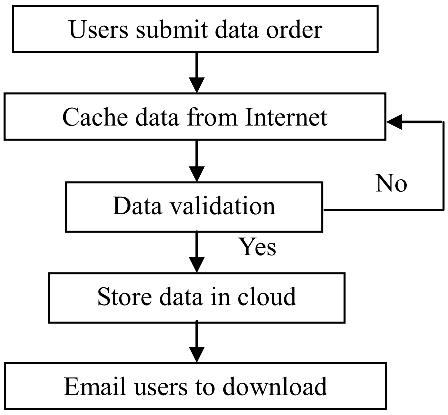


Figure 3. Data entities collection process

3.2 Data service module

After crawling the metadata from the Internet, the data service module reads header files and exports detailed metadata. After adding the corresponding product label (year, month) and generating three types of files, txt, browse, and scale, it stores the core metadata in PostgreSQL and PostGIS through clustering technology. Then, it publishes the metadata in the corresponding server and file directory. Finally, it establishes a hierarchical spatial index, which makes user retrieval very convenient.

As the open source implementation of GFS, the Moose File System is used to storing large quantities of data entities. The Moose File System is a network distributed system. It fragments the data entities and stores the data blocks in different storage servers..It is not only simple to be configured and installed but also has high reliability. When users download files, the file name is used as the key to call the cloud storage data access interface.

The data retrieval efficiency is an important indicator when evaluating a cloud service platform. The Geospatial Data Cloud provides three methods of data retrieval. The first one is implemented by using geocoding technology. When users input the name of the desired region, the platform will transform the natural language describing the location into geographic coordinates, and then by means of a spatial indexing system, the data can be delivered promptly to the user. The second method allows users to retrieve data by vector geometry, which uses the PostGIS spatial index. The third method provides users with the ability to search data by typing the name of administrative divisions or other attributes. All three methods together offer users great convenience in finding appropriate data.

4 CLOUD-BASED DATA PROCESSING PLATFORM

4.1 Online data-processing portal

In addition to using the massive data management service, users also need to process data in order to discover useful geospatial information for further research. Traditional remote sensing data processing methods usually require downloading the data first to personal computers and then using related software to process. This not only needs a lot of storage space but also requires relatively high computing performance.

The Geospatial Data Cloud provides a computing infrastructure that can support data processing over the Internet. Users do not have to download all the data that need to be processed. They just have to define the model they intend to use. All the processing will be executed on the cloud. The architecture of our cloud-based data processing system is depicted in Figure 4.

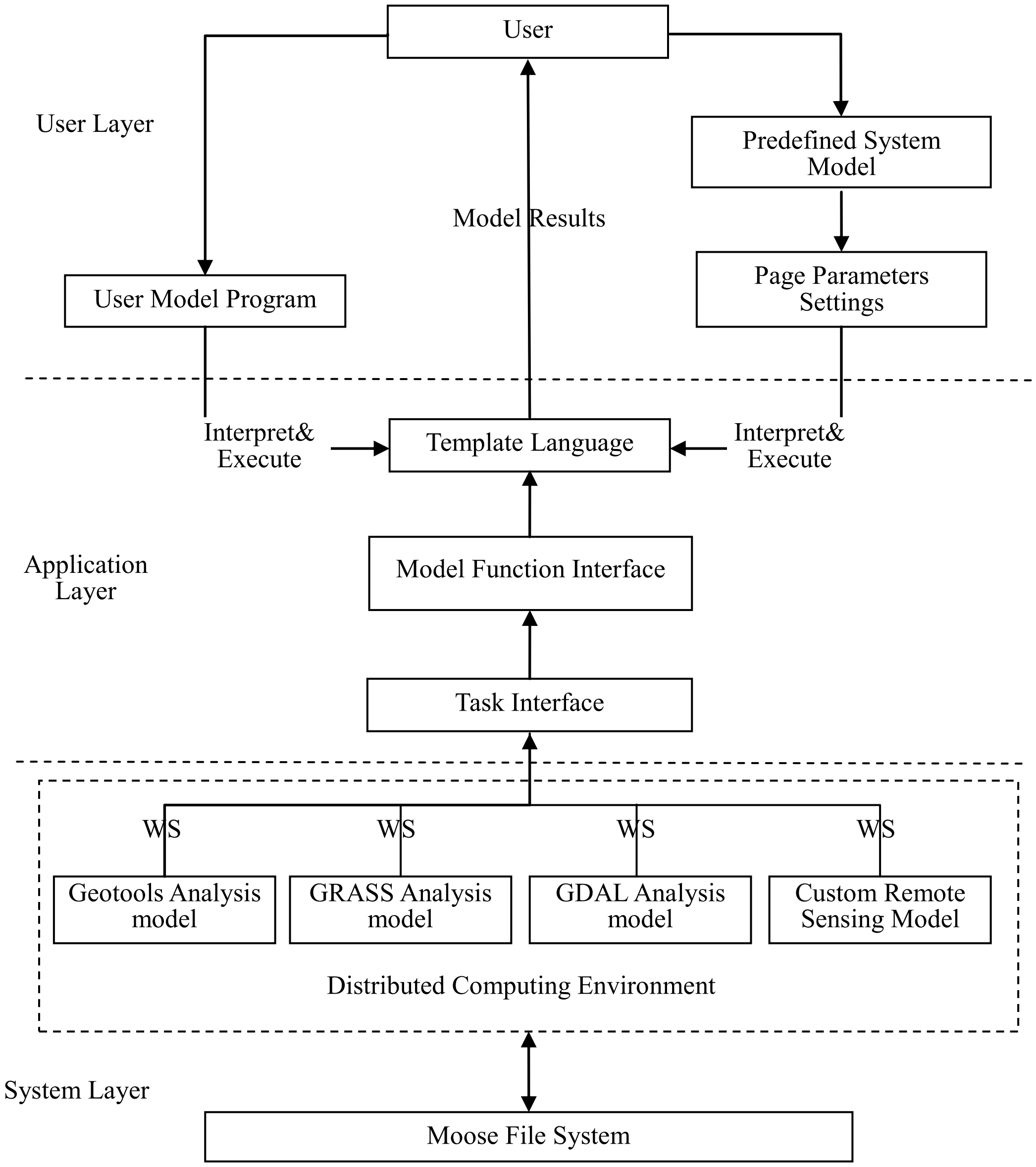


Figure 4. Cloud-based data processing system architecture

The architecture of the online data processing module consists of three modules, the system layer, application layer, and user layer. In the system layer, the Moose file system is adopted to build the storage environment that stores the system tasks, models, and related data. Analysis tools such as GDAL (Geospatial Data Abstraction Library), GRASS (Geographic Resources Analysis Support System), Geotools, and other custom remote sensing models are integrated. Python is used to package the analysis tools. Communication is implemented using WebServices technology.

Because the spatial analysis tools of remote sensing are usually organized according to a model calculation formula, taking into account the generality of the model, it is necessary to define the model description language. In the application layer, Python is used to describe the models. Each model has a corresponding Python program that is executed by web applications, implemented by jython. In the user layer, model parameters are set on the web page. This layer consists of three components, parameters description language, parameters entity template, and parameter template interpretation engine. Among these, XML is adopted as the parameters description language that describes related metadata information, the quantity and type of parameters, and the execution path of the models. The parameters entity template is described by jsp and js script, and the parameter template interpretation engine is developed by a Java framework called spring. The parameter template interpretation engine automatically generates web pages according to the parameters descriptions and corresponding parameter template entity template. After the information has been submitted, the model will be interpreted and executed.

4.2 Virtual machine for data-processing

The online data processing module is implemented by using virtualization technology. In the logical architecture of the virtual data processing machine depicted in Figure 5, the distributed file system and Linux cluster are used to build the storage environment. The SMB server cluster is deployed to enable Windows and Linux users to access the large-scale remote sensing data in read-only format. In the cloud host, commercial or open source GIS software that are most commonly used to process remote sensing images are pre-installed. All users have to do to use the data and computing resources is to log in on their own virtual machine.

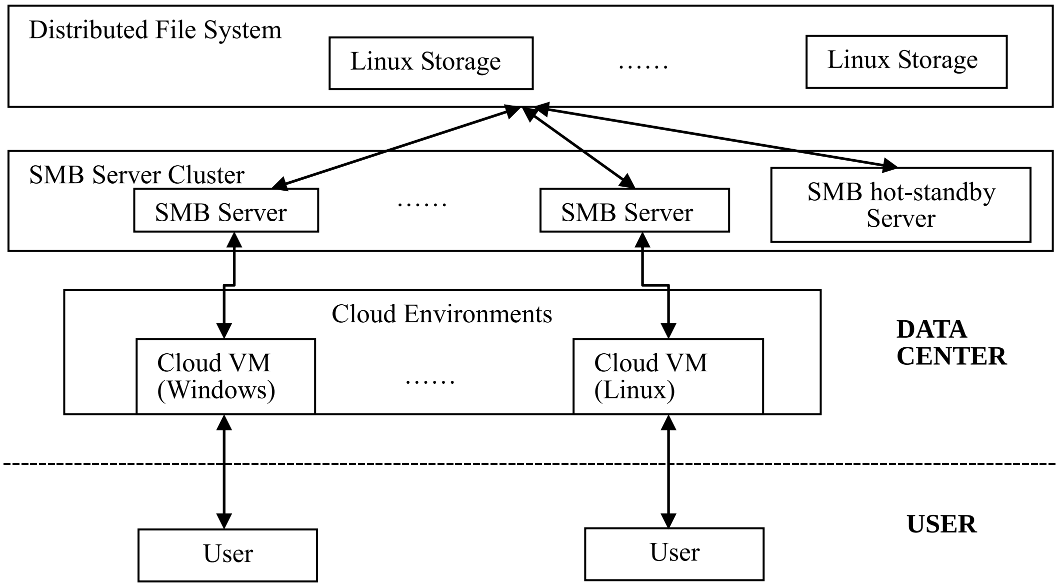


Figure 5. Logical architecture of the virtual data-processing machine

The process of users accessing the computing and service resources is shown in Figure 6. Users first submit data access requests. After being audited by the system administrator, they can generate their own cloud host, which enables them to use the storage and computing resources on the cloud.

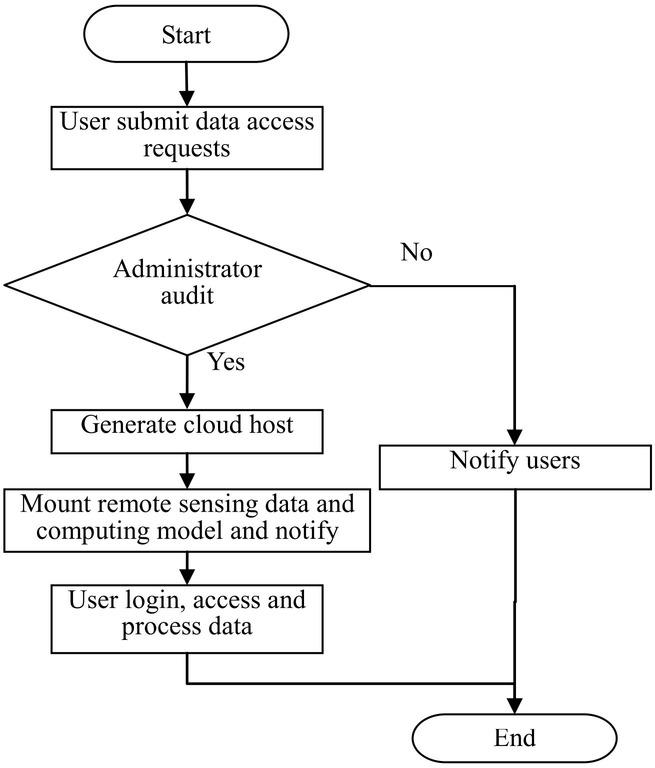
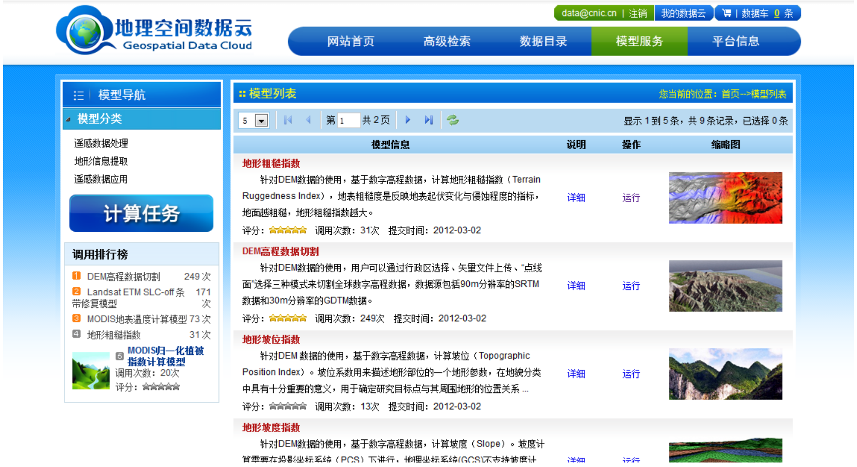

Start
User submit data access requests
Administrator audit
No
Yes
Generate cloud host
Notify users
Mount remote sensing data and computing model and notify
User login, access and process data
End

Figure 6. The process of accessing resource by users

5 CONCLUSION

The Geospatial Data Cloud provides users with large scale, remote sensing data resources, elastic computing and storage resources, and a basic infrastructure environment by using a cloud service. The modules that compose the platform are loosely coupled. As a result, it is easy to locate failure points and find bottlenecks during the system's performance, and thus maintenance and optimization of the system are facilitated. In addition, many other aspects of improvement have been made in the Geospatial Data Cloud. First, diverse and huge data sources are available. In fact, there are over 350TB data entities provided for free. As the data retrieval is fast, it is very convenient and time-saving for researchers to acquire data. Second, nine data processing models are provided. Thus, users do not have to purchase high performance computational equipment and install data processing software, such as ENVI, which are costly. Finally, all the services are free to all users. Nowadays, there are more than 70,000 users of the Geospatial Data Cloud. Figure 7 illustrates the interface of the Geospatial Data Cloud data processing model. Results of the model computing can be seen and downloaded at “user space”.



调用排行榜

1 DEM高程数据切割 249 次
2 Landsat ETM SLC-off条带修复模型 171 次
3 MODIS地表温度计算模型 73 次
4 地形粗糙指数 31 次
5 MODIS归一化植被指数计算模型
调用次数：20次
评分：

显示 1 到 5 条，共 9 条记录，已选择 0 条

模型信息	说明	操作	缩略图
地形粗糙指数 针对DEM数据的使用，基于数字高程数据，计算地形粗糙指数（Terrain Ruggedness Index），地表粗糙度是反映地表起伏变化与侵蚀程度的指标，地面越粗糙，地形粗糙指数越大。 评分：★★★★★ 调用次数：31次 提交时间：2012-03-02	详细	运行	
DEM高程数据切割 针对DEM数据的使用，用户可以通过行政区选择、矢量文件上传、"点线面"选择三种模式来切割全球数字高程数据，数据源包括90m分辨率的SRTM数据和30m分辨率的GDTM数据。 评分：★★★★★ 调用次数：249次 提交时间：2012-03-02	详细	运行	
地形坡位指数 针对DEM 数据的使用，基于数字高程数据，计算坡位（Topographic Position Index）。坡位系数用来描述地形部位的一个地形参数，在地貌分类中具有十分重要的意义，用于确定研究目标点与其周围地形的位置关系 ... 评分：★★★★★ 调用次数：13次 提交时间：2012-03-02	详细	运行	
地形坡度指数 针对DEM数据的使用，基于数字高程数据，计算坡度（Slope）。坡度计算需要在投影坐标系统（PCS）下进行，地理坐标系统(GCS)不支持坡度计	详细	运行	

Figure 7. The data processing model interface

The adoption of cloud computing in the field of remote sensing is definitely appealing. However, more work still needs to be done to improve this platform in the future. First, the process of customizing models and computing tasks needs to be more humane. Also, as a data cloud platform, the Geospatial Data Cloud should import more and more high-quality data to users.

6 REFERENCES

Blower, J. D. (2010) GIS in the cloud: implementing a Web Map Service on Google App Engine. In *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, ACM, p 34.

Cary, A., Sun, Z., Hristidis, V., & Rishe, N. (2009) Experiences on processing spatial data with mapreduce. In *Scientific and Statistical Database Management*, Springer: Berlin Heidelberg, pp 302-319.

Cui, D., Wu, Y., & Zhang, Q. (2010) Massive spatial data processing model based on cloud computing model. In *2010 Third International Joint Conference on Computational Science and Optimization (CSO) 2*, IEEE, pp 347-350.

Fang, L. (2011) *A Cloud-computing-based Study on the Key Technologies to Implement the Practical Platform for Efficient Processing of Land Resources Services*, PhD thesis, ZheJiang University, Hangzhou, Zhejiang, China.

Huang, Q., Yang, C., Nebert, D., Liu, K., & Wu, H. (2010) Cloud computing for geosciences: deployment of GEOSS clearinghouse on Amazon's EC2. In *Proceedings of the ACM SIGSPATIAL International Workshop on High Performance and Distributed Geographic Information Systems*, pp 35-38.

Kang, J.F. (2011) *Technologies of Storage and Efficient Management on Cloud Computing for High Resolution Remote Sensing Image*, PhD thesis, ZheJiang University, Hangzhou, Zhejiang, China.

Liu, C.Y. (2013) *Research and Practice on Key Technologies of Electronic Chart Cloud Service*, PLA Information Engineering University, Zhengzhou, Henan, China.

Liu, Y., Guo, W., Jiang, W. S., & Gong, J. Y. (2009) Research of remote sensing service based on cloud computing mode. *Application Research of Computers 26*(9), pp 3428-3431.

Rezgui, A., Malik, Z., & Yang, C. (2013) High-resolution spatial interpolation on cloud platforms. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pp 377-382.

Siládi, V., Huraj, L., Polčák, N., & Vesel, E. (2012) A parallel processing of spatial data interpolation on computing cloud. In *Proceedings of the Fifth Balkan Conference in Informatics,* ACM, pp 193-198.

Sims, K. (2007) IBM introduces ready-to-use cloud computing collaboration services get clients started with cloud computing. Retrieved September 10, 2014 from the World Wide Web: http://www-03.ibm.com/press/us/en/pressrelease/22613.wss

Wang, K., Han, J., Tu, B., Dai, J., Zhou, W., & Song, X. (2010) Accelerating spatial data processing with mapreduce. In *2010 IEEE 16th International Conference on Parallel and Distributed Systems (ICPADS),* pp 229-236.

Wang, Y., Wang, S., & Zhou, D. (2009) Retrieving and indexing spatial data in the cloud computing environment. In *Cloud Computing,* Springer: Berlin Heidelberg, pp 322-331.

Yang, J.Q. (2010) *Research on Spatial Data Clustering Services Architecture and Algorithms*, PhD thesis, PLA Information Engineering University, Zhengzhou, Henan, China.

(Article history: Received 10 August 2014, Accepted 9 October, Available online 6 October 2014)