An integrated GIS platform architecture for spatiotemporal big data

Shaohua Wang, Yang Zhong, Erqi Wang

PII: S0167-739X(17)31928-3
DOI: https://doi.org/10.1016/j.future.2018.10.034
Reference: FUTURE 4539

To appear in: *Future Generation Computer Systems*

Received date : 1 September 2017
Revised date : 3 October 2018
Accepted date : 20 October 2018

An Integrated GIS Platform Architecture for Spatiotemporal Big Data

Shaohua Wang[a,b], Yang Zhong[c], Erqi Wang[d,e,f]

[a]Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing

[b]Department of Geography, University of California, Santa Barbara, CA

[c]Claremont Graduate University, Claremont, CA

[d]SuperMap Software Co., Ltd., Beijing 100015, China

[e]Beijing Engineering Technology Research Center of Geographical Information Core Software and Application, Beijing 100015, China

[f]National Administration of Surveying, Mapping and Geoinformation, Beijing

Abstract:

With the increase in smart devices, spatiotemporal data has grown exponentially. To deal with challenges caused by an increase data requires a scalable and efficient architecture that can store, query, analyze, and visualize spatiotemporal big data. This paper describes a Cloud-terminal integrated GIS platform architecture designed to meet the requirements of processing and analyzing spatiotemporal big data. Cloud-terminal Integration GIS is developed according to the architecture. Extensive experiments deployed on the internal organization cluster using real-time data sets showed that the SuperMap GIS spatiotemporal big data engine achieved excellent performance.

Keywords: Spatiotemporal Big Data; Distributed Computing Framework; Cloud-terminal Integration GIS; SuperMap GIS

1.Introduction

With the advancement of information technology, the demands of processing, analyzing and visualizing spatiotemporal big data have dramatically increased [1]. In this era of big data, the Geographic Information System (GIS) is facing new challenges. To overcome the difficulties caused by big data, GIS must evolve its technologies to cope with big data [2, 3].

Some of the challenges for GIS include analyzing and processing the spatiotemporal big data, clustering and distributing spatial big data, indexing and managing big data, and computing and visualizing the big data in the system while maintaining a high performance [4, 5]. Currently, popular big data platforms (such as Hadoop and Spark) do not have the capacity to perform of spatial analysis, spatial computation, or spatial data mining. To recognize breakthroughs and innovations for the large scale spatial data of distributed storage and management, distributed spatial computing, real-time big data processing, and visualization [6], it is necessary for GIS to integrate the general big data technology.

In the face of the increase in data volume and the growing number of data types, traditional relational database is prone to bottleneck problems, such as low storage efficiency, weak concurrent access ability, and difficulty in horizontal scaling. It is imperative to develop new spatial data storage technology [7, 8]. Container technology (such as Docker) facilitates rapid and large-scale deployment of GIS. Optimal synchronization and discovery mechanisms in load-balancing provide support for the dynamic scaling and disaster recovery of GIS services.

If a GIS system seeks to consume data to perform queries or generate maps, the output data from Spark must be converted and transferred into GIS platforms. The process is typically time and storage-consuming. Additionally, the traditional GIS system only executes the computing tasks in the job queue, it cannot process streaming data. The conventional GIS software and stand-alone processing architecture cannot be analyzed a large volume (For example, over 1 billion records) of spatiotemporal big data. Moreover, these integration processes require high specs of computer hardware and a rewriting of most of the algorithms for big data in GIS [9, 10].

In this paper, we focus on the design and implementation of an integrated GIS platform architecture for spatiotemporal big data. The paper is organized into four parts. Related work is illustrated in section 2. Section 3 introduces our integrated GIS platform architecture. Section 4 discusses its implementation. Case tests and results follow in section 5.

2.Related work

Hadoop extensions like Hadoop-GIS[11-13] and SpatialHadoop [13-15] support spatial query using a MapReduce framework. The issue with theses extensions, however, is that they save intermediate results to the disk reducing effiency.

One advantage that Spark's framework has over Hadoop is its speed. The memory-based parallel computing architecture performs better than the MapReduce model in Hadoop. With the use of RDD, distributed computing leads to better performance by two orders of magnitude. Additionally, Spark offers more support to big data computing, its enhanced stream processing, graph computing, and machine learning sub-systems are versatile. These are the reasons that Spark based framework was chosen in this study. A few solutions of processing spatial data such as GeoSpark [16, 17], SpatialSpark [18, 19], LocationSpark [20], and Simba [21] offer limited functions within Spark.

There are two ways for Spark to execute GIS functions. The first approach is to have GIS computing run outside of Spark, this allows for managing task orders and visualizing the output of analysis. The other approach is to have it run internally. Using this method, we can perform a various of tasks, including generating a spatial index, executing the spatial query, and performing spatial analysis and computing. Considering GIS core features and applications, we prefer the latter: running GIS directly inside the Spark framework so as to take full advantage of its potential.

Most of the big data frameworks such as Spark, HDFS, MongoDB, and ZooKeeper are based on Linux[22]. In Windows environments, these frameworks are being mostly used for research and study purposes. Hence, the best way to have GIS and big data framework work together seamlessly is to have a cross-platform GIS system. A cross-platform GIS system can directly support Linux from its core functions but also work in Windows environments. GIS functions have to be cross-platform[23].

In spatiotemporal data computing, the system is required to process a large volume of data and to manage dynamic changes. Spatial online analytical processing for real-time data based on SpatialHadoop is usability[24], it can be improved by Apache Storm[25] or Spark Streaming[26, 27]. Also, cloud computing demands high processing performance and the capability to manage dynamic changes. To take full advantage of the optimizing cloud computing, the ability to support the virtual machine or Docker's quick deployment is also critical in ensuring the high efficiency of this spatiotemporal analysis engine.

Our research develops an integrated GIS platform architecture which enables spatiotemporal big data storage, processing, visualization and analysis.

3. An Integrated GIS Platform Architecture for Spatiotemporal Big Data

A variety of big data platforms yielded low spatial data storage, spatial analysis and spatiotemporal visualization performance. We proposed an integrated GIS platform architecture for spatiotemporal big data (Fig. 1), which contains large-scale virtual storage, a distributed computing framework, cloud computing and integration, stream data processing, 3D and virtual reality, being rapidly applied across the multi-terminal, the open source community, and container and continuous delivery.

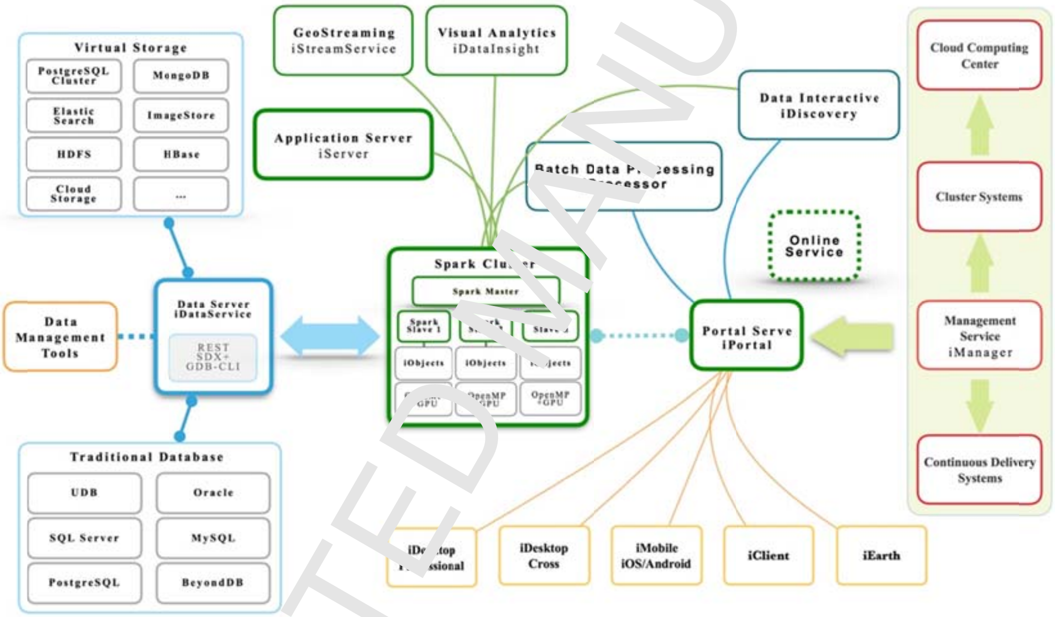

Fig. 1. Architecture of the GIS platform for spatiotemporal big data

3.1 Massive spatial virtual storage

In big data systems, a critical issue is data storage. As the data are being generated with high data type variety and low value density, the traditional file systems and databases can no longer maintain a high performance while continuing to satisfy big data storage requirements. In recent years, technologies and solutions in virtual storage have emerged, many of these have been widely used by internet platforms. For geospatial data, there is also a need to evolve traditional file systems and relational database storage solutions to distributed, virtual and software defined storage system so that storage scalability and processing capability can meet future challenges.

The virtual storage system can be classified into three categories: the distributed file system, the distributed relational database, and the NoSQL/NewSQL storage system. The distributed file system is mainly used to solve the issue of limited storage space and the high cost of a single machine system. Running concurrent I/O with multi-replication copies not only increases the computing bandwidth, but also enhances the system's load balance, error tolerance, and dynamic scalability. The system can be deployed in a cloud computing environment with the support of a large file size, memory cache, space sharing, and REST web services. One popular database of this type is Hadoop; Other similar systems include Ceph, and IPFS. The distributed relational database is mainly implemented by adding newly distributed clusters and distributed transaction processing features in traditional databases (Examples of implementation include PostgreSQL cluster, MySQL cluster, and CrateDB based on Docker technology). Because of the high compatibility with the original databases, these systems can better support SQL and transaction processing. Since the original management methods and software can still be applied, data migration and system scaling becomes easier. As most of these systems are open-source, the cost is relatively low; this is important especially when the system needs to be deployed in multi-nodes cluster environment. The NoSQL/NewSQL storage system focused on reducing the number of ACID transactions so that its data processing performance can be significantly improved. When managing various unstructured data, the system not only simplifies the development and maintenance processes, but also lowers the total cost of operation (TCO). This kind of solution has been widely used in many internet platforms as well, such as MongoDB, HBase, Cassandra, Redis, etc.

Today, many different virtual storage systems exist in various kinds of environment and are being used in diverse ways. How do we fully utilize the advantages of each system while enabling the sharing and transferring of resources between systems? How to provide a unified way of visiting, read and write data while having the ability to store data in diverse platforms so that the data become more valuable? To solve these issues, we designed and developed a virtual spatiotemporal integrated service system – DaaS (Data as a Service) based on a seamless integration of multi-source spatial data in SDX+ and the interface in GDB-CLI[28]. We implemented a unified REST service framework that can easily connect with multiple types of data storage systems and work with the existing connected database systems at the same time. This system supports distributed, multi-level spatial database storage services, and cloud/local data management in one portal. By using its unified data interface, the system can connect with Hadoop storage ecosystem, the MongoDB storage system, the PostgreSQL cluster, the MySQL cluster, and other existing databases (Fig. 2).

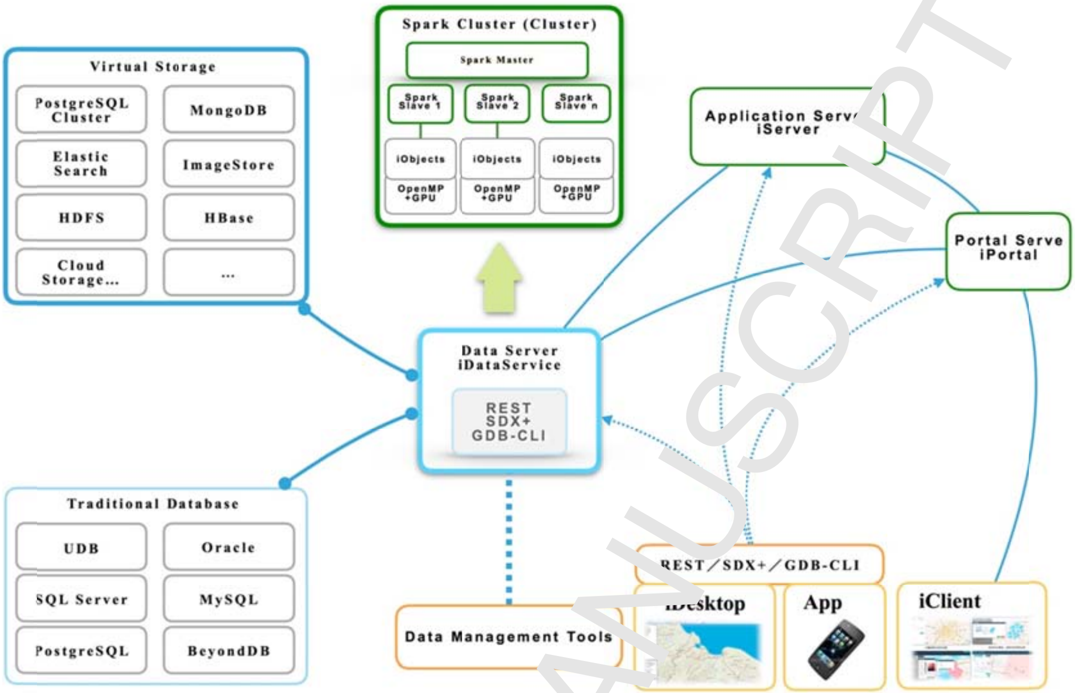

Fig. 2. From SDX+ to DaaS

With an increase of in the demand for storage space and an increase in maintenance cost, the value of data has been on the decline. If we can consume data within a reasonable amount of time, it may become a more precious asset. In contrast, if data is not being used properly, it can become a burden to a business. For example, without enough investment in data security, a company runs the risk of leaking sensitive data, which could be detrimental to the company. Simply owning data does not benefit a business. In fact, how productively data is used determines its value. The more that we consume data, the more value we can derive from it. Therefore, establishing a continuous data processing infrastructure to meet the needs of the application is vital. Additionally, maintaining and applying the data value is a critical aspect in developing a big data system.

3.2 A Distributed Computing Framework

When Moore's law reaches its end, it is difficult to pursue further processor speed by increasing the CPU's clock rate. Instead, multicore CPU becomes the new normal. By using multi thread and process technologies to manage and parallel process tasks or using graphic cards' CUDA and the OpenCL parallel computing mechanism, the system can break through the limitation of computing capability within a single CPU. In SuperMap 8C, the multi-thread support, the multi-process service, and the spatial analysis algorithm based on OpenMP, CUDA significantly improves the efficiency of spatial data processing and model analysis. It enables the object visualization capability to operate in real-time. With the merging of big data, computing power encounters its bottleneck. The multi core CPU and large-scale cluster system are needed to adapt the changes.

Designed for batch processing, MapReduce module in Hadoop is considered a pioneer in the new generation of distributed computing. However, it has a number of weaknesses. These weaknesses include a slow starting speed, complicated deployment, and an inability to perform regression computing. Built on a distributed memory computing model and on Flink which better supports stream computing the module has started to be replaced by Spark. The Hadoop/Spark open source ecosystem led by the Apache software foundation has become the standard in the big data field. Many business solutions have been built based on this framework. (These business solutions include the big data service clouds from Databricks, Amazon, IBM, and Oracle.)

With the advancement of GPS systems, satellite imagery, drone photography, and smart measurement devices, the requirements for spatial data storage and processing have been rapidly increasing. Thus, importing GIS functionality into the Spark framework to build an integrated distributed spatial and temporal data processing platform has taken on new importance. The latest SuperMap GIS platform provides full support to the Spark computing framework. It establishes a complete big data solution with three main components, a GIS core engine, a client SDK, and an application system. The GIS core engine can either be imported into the Spark environment as Scala or be implemented in a different frontend big data analytics software by supporting Python. By integrating iObjects for the Spark service into the iServer product series, a distributed spatial analysis model computing service can be exposed via REST. Its return results can be consumed and visualized easily at applications with iObjects, iDesktop, iDesktop Cross, iMobile, iClient, and with other 2D/3D linkage clients (Fig. 3).

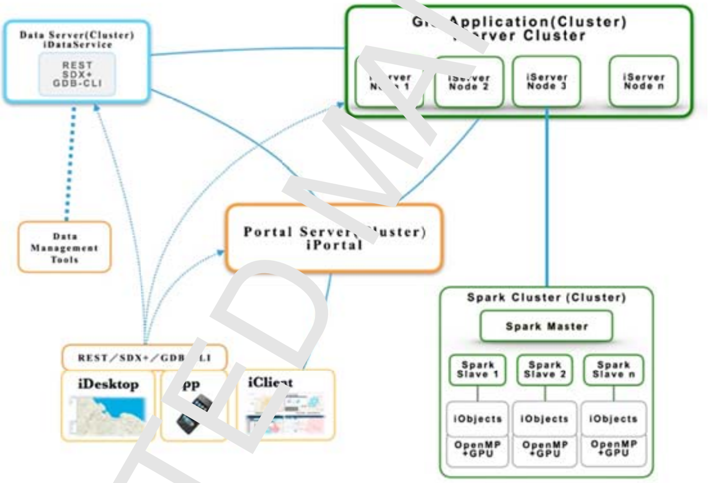

Fig. 3. The Architecture of massive GIS clusters

It's a huge advancement toward moving the GIS core functions from single core CPU to a distributed computing framework. With this move, the GIS system will be able to take full advantage of the capabilities of large-scale storage, distributed memory, cluster management, and its deployment brought on by modern computing hardware and data centers. This move will also solve issues in the traditional GIS software: such as a lack of storage and deficient computing power. It makes it possible to build a large-scale application system or to conduct spatial relationship research at a high accuracy level. We will likely be seeing numerous types of applications and breakthrough development in geography spatial models or algorithms. Not only will it bring GIScience and geography science to a new level, but it will also improve the efficiency in environmental management, disaster management, urban planning, etc.

3.3 Cloud Computing Integration

Cloud computing provides a set of models and methods for sharing computing resources. Allocating computing resources dynamically, not only enhances system utilization efficiency, but also makes gathering large scale computing power in a short amount of time possible. Amazon, Google, Microsoft, and IBM all provide cloud data center services on a large scale. In China, Ali Cloud, Baidu Cloud, and Tencent Cloud also offer diverse cloud computing services. In recent years, many startups have begun to provide services based on Docker technology, such as Qiniu and QingCloud. All of these cloud computing platforms allow users to manage the computing resources, lease resources based on demand and quickly establish a large scale cloud computing cluster. In the past, the traditional server leasing service was the main focus. Today, the distributed cloud computing cluster based on Hadoop/Spark has become the standard service of large data centers. With the rapid development of Docker container technology, the cloud computing service it is based on can further lower the cost of maintenance and provide more flexible and agile solutions to allocate and deploy resources. Services's migration between different data centers or between public cloud and private cloud centers, is also made significantly easier with Docker technology. To sum up, the cloud computing service has been moving from server leasing services based on the virtual machine to the distributed cluster services and micro services based on recent technologies such as Docker, Hadoop/Spark, etc.

In Docker, cloud services can be encapsulated by a business component as micro services and can be assembled based on demand during deployment. The Docker instance can be developed, tested, run, and deployed as needed within the public cloud, dedicated cloud, industry cloud, and private cloud in a streamlined way. This will greatly reduce the maintenance cost and development difficulty of a cloud computing service. A GIS cloud computing integration infrastructure must fully integrate with the Docker technology, and design, develop, and deploy the system based on the micro services concept model. SuperMap iServer, iExpress, iPortal and iManager have already been supporting Docker; The service based on its technical standards and micro services structure can be deployed to diverse cloud computing data centers. Other features like integration among different types of computing infrastructures along with functions for automatic management system are all included as well. Additionally, the enterprise user and the personal user can directly access these services via Dituhui or the online portal (Fig. 4).

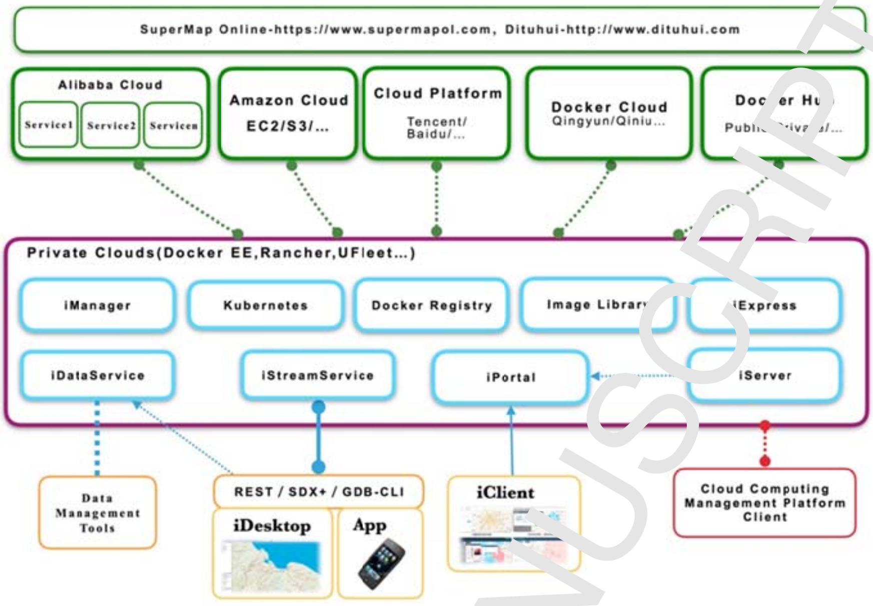

Fig. 4. The Micro Services Architecture based on Cloud Computing and Docker

By implementing the micro service infrastructure based on Docker, the GIS system can be deployed as a cloud computing module, and the multi-clouds integration and management can be unified. We can also fully integrate geospatial big data into the cloud computing infrastructure. All of these features have become the core capabilities of a modern data center and have even become essential system components in the smart city, environmental resources, and many other industries. It also serves the following core functions of geospatial data management, spatial pattern analysis, geospatial data visualization, API sharing, and other application services.

3.4 Streaming Data Processing

With the development of GIS technologies, the data sources of the GIS system have changed tremendously. In the past, data mainly came from traditional map digitization and measurement input via devices like the plane table, total station, etc. The common data format was the static vector map, which lacks update accuracy, and currency. The new measurement works extensively using photogrammetry method to collect the raw data. The main data sources include imagery, video, radar, and GPS data, which are generated from satellite, airplane, drone, and measurement vehicles. The latest devices such as panoramic camera, the street view camera, the observation satellite, and the LiDAR system are capable of retrieving omnidirectional images and spatial information. Some of these devices support streaming services so that the data can be dynamically transferred to users. Today, the traditional static data storage, static cartography, and scheduled data update methods have become less critical. It also leads to a tremendous change in traditional data storage, processing, analysis, and usage methods.

LiveGIS—a new feature of GIS—has the ability to generate, process, and consume live data via streaming[19]. Due to the changes in data types and the increase in processing data volume, the GIS system structure has been evolving to adapt this revolution. There are currently several preferred streaming practices using distributed computing as the system structure, using Spark Stream as the stream data framework, integrating message-oriented middleware such as Kafka, and combining the message receiving, processing, and high efficiency data storage with real-time spatial analysis as a spatial temporal integrated software platform which satisfies LiveGIS demands. There are already numerous successful solutions have been used in eCommerce, social media, logistics, and transportation industries. The latest SuperMap GIS platform has integrated this system solution with advanced GIS functions so that the streaming data can take advantage of GIS spatial analysis and the visualization feature. This platform greatly enriches the traditional GIS system's capability and usages.

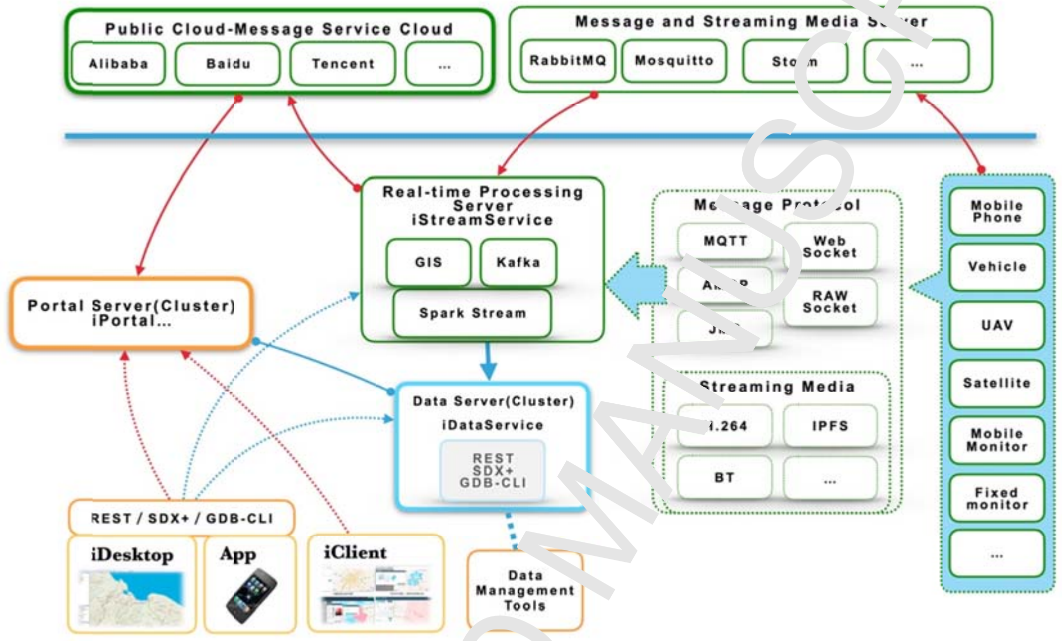

Fig. 5. The Process Diagram and Architecture for GeoStreaming data

The capability of backend processing along with the flexibility of mobile applications can provide a reliable platform for IoT and smart devices to process its spatial temporal data (Fig. 5). Not only is it scalable with the development of business scope, but it can also quickly migrate between environments. To sum up, it has become the core foundation of smart city's development and operation.

3.5 3D and Virtual Reality

In recent years, 3D related information technologies have been making great progress. With the advancement of graphic card processing capability, the supported software standards and techniques such as OpenGL, OpenCL, WebGL, etc. have rapidly evolved. The breakthrough of the VR/AR headset and glasses brings digital 3D application into a new era. Owing to the IT revolution, two critical improvements in GIS have been made: oblique photography integration and 2D-3D linkage capabilities. From retrieving full range 3D geospatial data to building models, the process of consuming data in the end terminal application has been streamlined. The SuperMap platforms' 3D GIS technique has been built in every product, providing comprehensive solution for importing data, publishing service, analyzing application, gaining web access, and improving mobile Apps. It is compatible with diverse server types, components, mobile platforms, web, desktop software, existing databases, cloud computing services, and other IT infrastructures.

By integrating BIM technology with GIS, we can further apply this GIS system to several micro management areas. For example, it can be used for building parts and managing component objects. We can also use it to develop the support system for smart buildings or IoT networks. By integrating VR/AR with GIS, the city planning and management can offer a richer user experience which could enhance the public service quality in land management, municipal administration, urban planning, etc. The mobile 3D GIS feature not only simplifies the data collection process, but also provides powerful on site management function. Furthermore, it creates a public IT platform that allows users to work on further spatial planning, application and optimization (Fig. 6).

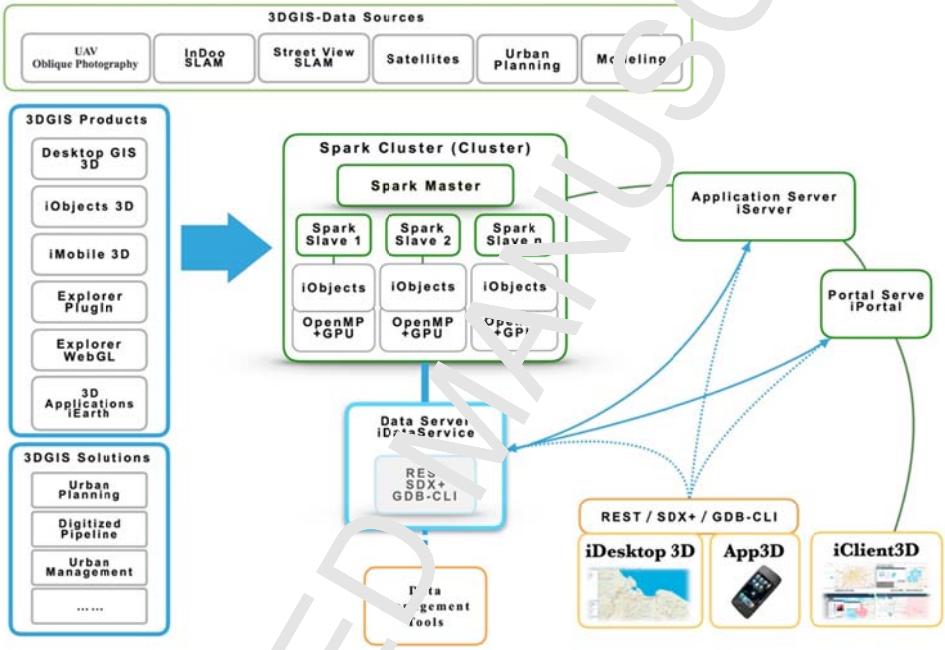

Fig. 6. The 3D GIS Technology and Solution

The current 3D GIS has become a core component in the spatial temporal big data system. However, the future 3D GIS will go beyond the current 3D GIS and simulate the real world. It will also support the actual instance model's bool operation. Additionally, importing the physics engine and the collision detection algorithm into GIS will make the simulation of the model and of the spatiotemporal environment more realistic. It will advance the business applications in planning, designing, pipeline, transportation, construction, etc. The future 3D GIS will also influence new advancements like high-accuracy navigation, the self-driving car, and airport management.

3.6 Quick Multi-Terminal Application

Software is like a magnifying glass for data value. The more that data is being used, the more value that it generates and the more software compatibility is required. Not only does the software need to have powerful data capability, it also has to have diverse application compatibility. Lastly, it should work in different environments and with all mobile devices. The client end can be classified in devices, operating systems, hardware infrastructures, and programming languages.

The more types of client side it supports, the more compatible it will be. It also means that more users are exposed to generating more value for the data.

The SuperMap GIS product family offers very rich client-end support. The iDesktop based on .NET and iDesktop built on Java can directly access cloud computing resources and a large volume of storage. It has the functions for professional GIS users to process data, generate maps, and analyze spatial patterns. The iClient provides WebGIS functions which are compatible with different browsers. Its functions including accessing server shared data, executing online analysis, and visualizing scenes can be used on multiple operating systems without plugin software installation. The iMobile not only provides SDK for iOS and Android development, it also supports YuanXin OS and other embedded operating systems. Since GIS functions are made to be easily accessible and portable, numerous applications have been developed on handheld platforms by SuperMap partners or other vehicle measurement devices to meet their own professional needs.

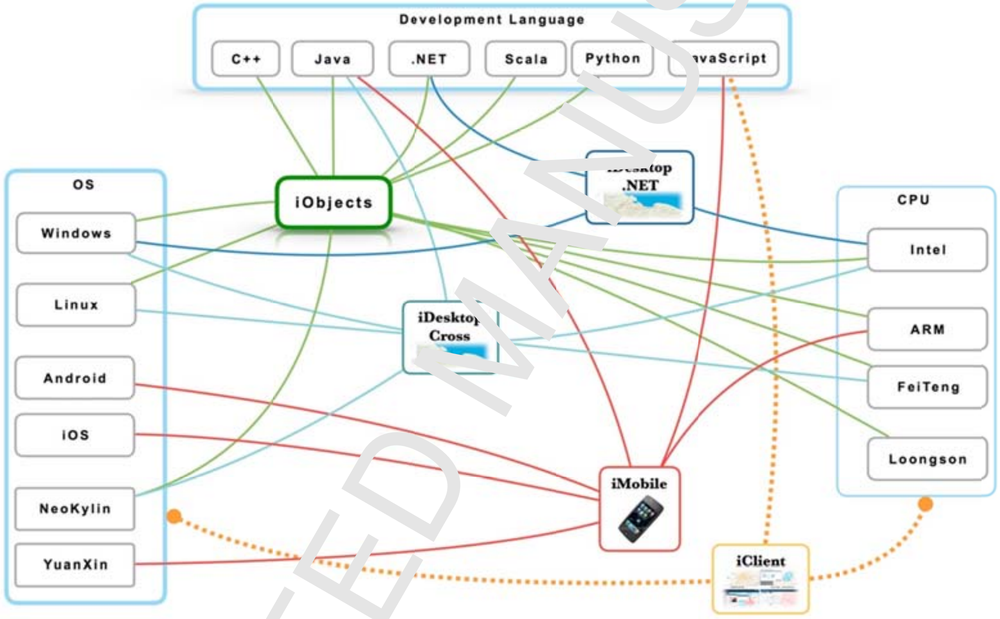

Fig. 7. The multi-client GIS products

SuperMap is the GIS platform that supports the largest number of terminals. It offers SDK on desktop, web, mobile, and supports accessing cloud services via API (Fig. 7). Users can develop all-purpose applications by using the given SDK and plugin framework from SuperMap. Many native Chinese CPU brands are supported, such as FeiTeng, and Loongson, along with the Chinese operating system, Kylin OS. In sum, the progress made in improving the compatibility of the GIS system will reveal system's extra potential from the big data and generate more data value.

3.7 Open source and the Open source community development

Open source and its communities are considered boosters for the modern software industry. Especially the emergence of github.com makes it possible to have developers around the world work on the same code repository and communicate with other developers, significantly improving software development productivity. In China, there are several platforms similar to Github, which offers source code management and sharing services (One such platform is

Oschina.net). The open source software and online source control model changes the traditional development of closed research and development. It makes the development process public and connects the developer with entire user community. Thus, in turn, enhances developers' productivity and the quality of their work. Large scale development cooperation becomes possible.

The open source software system represented by Apache Hadoop/Spark has become the essential foundation in the big data ecosystem. There are already many published and business solutions which were built based on this framework.

By integrating with the Apache big data ecosystem and working with the open source community, the GIS system can continue to evolve at its steady pace. The open source big data software provides unprecedented computing power via its distributed structure methodology. This leads to tremendous improvements in traditional GIS solutions. Adding geospatial data types to the open source big data solution also benefits the open source community users' usage of the spatial data. The latest version of the SuperMap GIS platform is built based on the Spark 2.x framework. It implements the big data analysis engine and the distributed computing feature. The big data processing, and analyzing efficiency greatly improves, and the maximum supported data size increases as well. Additionally, the open source iClient, iDesktop Cross, and iObjects' Python scripts can all process and publish big data (Fig. 8).

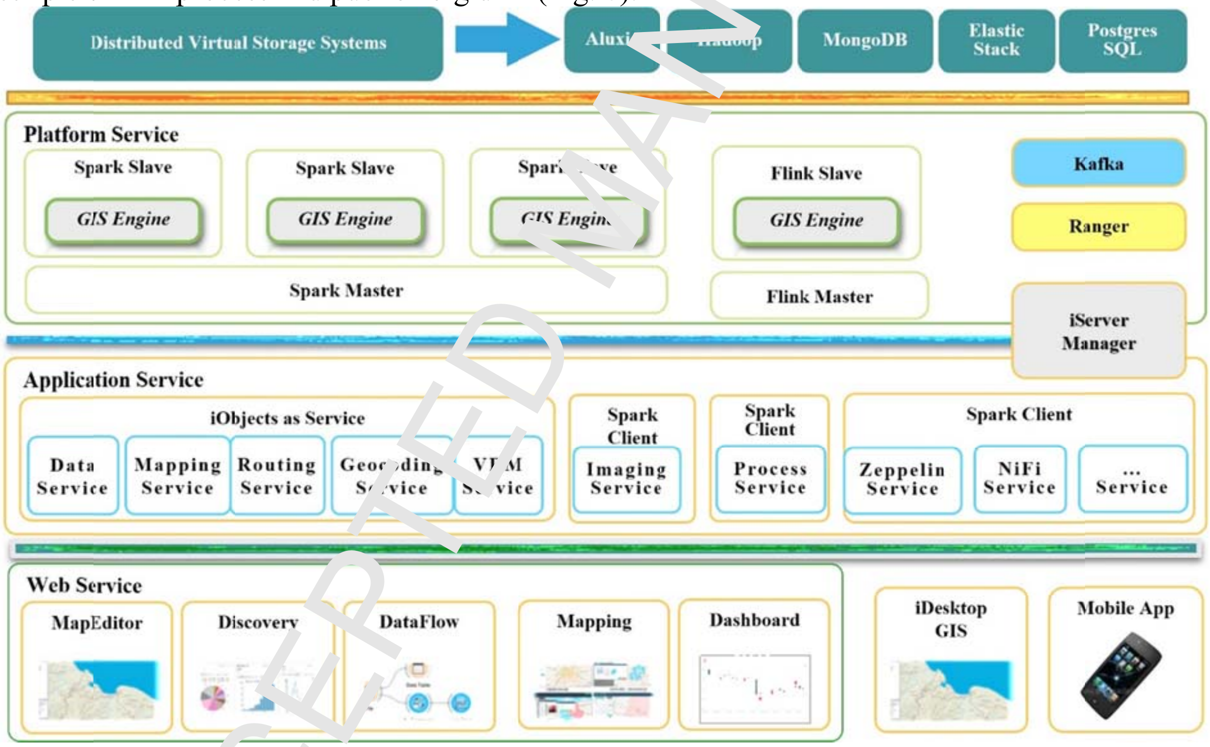

Fig. 8. The integration architecture of GIS and open source software

Open source and closed source software have their pros and cons. The advantages of open source software are strong developer involvement, fast update speed, and a high level of innovation. On the other hand, closed source software excels at testing, version control, persistence, and professional support. Today, it is rare to see a solution that is completely open source or closed source. Combining the advantage from the two to build the final solution will be the main realistic method.

3.8 Container Technology and Continuous Delivery

Boosting by internet technologies, the software development methodology has been evolved completely. With the emergence of Git/Gitlab/Github, distributed version control has replaced the traditional centralized software development methodology. Today, community development, public code review, auto testing, and continuous integration have become the standard development methodology. Compared with virtualization, the Docker container can be deployed at the bottom level of the system and run directly on top of the Linux kernel. Docker allows the user to compile software as a package and to isolate the running environment. By implementing Docker, a customizable micro services system framework can be easily established. Docker also shortens the system deployment time and simplifies the migration process between data centers. To meet the latest demands of the online platform, the continuous delivery concept and the DevOps method have been making great progress. The integration with the cluster management system such as Mesos, Kubernets, etc. has been developed as well. The automated processes on the container framework and the continuous delivery methodology greatly shorten the time of the software update and bugs-fixing and improve the software development's responsiveness. This quick iteration leads to a faster software innovation and to fewer system risks.

To implement the quick iteration, live testing, runtime verification, and controlled delivery features, we categorized the system into three deployment zones, the development zone, the validation zone, and the production zone. includes development tools, the use case library, and the testing system, in addition to testing data and the source code. The validation zone includes the validation data, the validation system, and the evaluation system. The production zone includes the production system, which contains the current running system and the latest updated system. This makes it possible for gray release, which is the act of migrating to the new version of the system via the AB testing method. To ensure the reliability of its GIS platform software and to solve the complications brought on by the multi-version issue, the SuperMap research team has developed the continuous delivery system, which covers the entire solution. It establishes the automated workflow of software development, integration, and testing. To move towards the online platform, the Dituhui and online service portal have gradually built the framework to support continuous delivery and DevOps. So far, SuperMap iServer, iExpress, iPortal, and iManager products have already endorsed Docker and the Micro service framework. These features have been integrated into some continuous delivery systems and have become some of the core components of the IT infrastructure (Fig. 9).

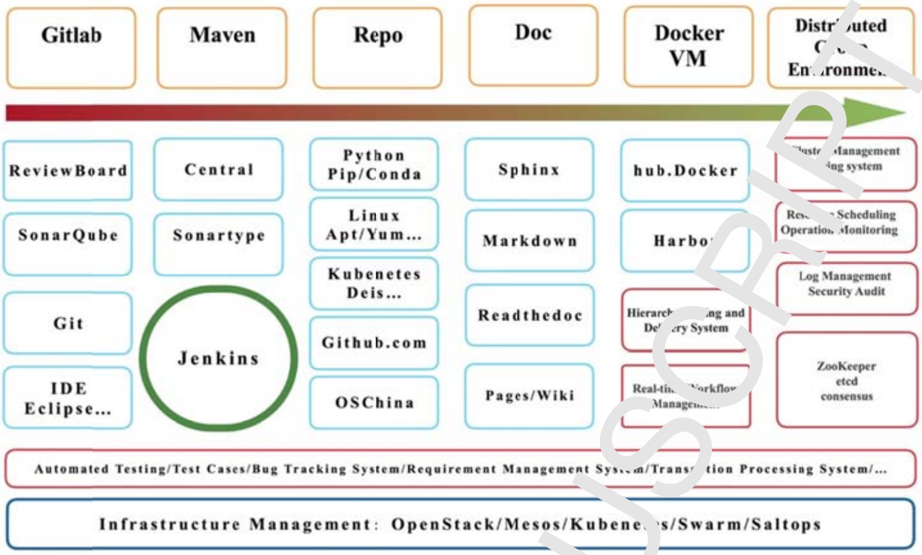
Gitlab
Maven
Repo
Doc
Docker VM
Distributed Cloud Environment
ReviewBoard
SonarQube
Git
IDE Eclipse...
Central
Sonartype
Jenkins
Python Pip/Conda
Linux Apt/Yum...
Kubenetes Deis...
Github.com
OSChina
Sphinx
Markdown
Readthedoc
Pages/Wiki
hub.Docker
Harbor
Hierarchical ... and Delivery System
Real-time Workflow Management
Cluster Management ... system
Resource Scheduling Operation Monitoring
Log Management Security Audit
ZooKeeper etcd consensus
Automated Testing/Test Cases/Bug Tracking System/Requirement Management System/Transaction Processing System/...
Infrastructure Management: OpenStack/Mesos/Kubenetes/Swarm/Saltops

Fig. 9. DevOps and Continuous Delivery flowchart

The spatial temporal big data system must have the capability to process the dynamic streaming data. Also, the software system itself should be able to evolve and update. The container and continuous delivery methods make software migration a smoother process, ensuring that the system and the data can be deployed as needed. This leads to better efficiency and improved system usability. The development, testing, validation, deployment, and production maintenance/management/update in traditional software development will all be integrated. The system will achieve a fast response time, runtime bug fixing, and update features that do not require downtime. In short, using the cloud computing framework, the virtualization technique, and the container technique to build a system with continuous delivery and DevOps workflow, will be the mainstream trend in future software development. It will become a necessary step to adapting to the challenges of big data.

4. Implementation of Cloud-terminal Integration GIS for Spatiotemporal Big Data

The SuperMap GIS platform is based on the Cloud-terminal integration technology, which consists of the intensive GIS cloud platforms, the diversified GIS terminals, and the integration of GIS cloud platforms and GIS terminals. The intensive GIS cloud platforms make full use of the cloud computing resources and provide high available GIS services. The diversified GIS terminals integrate technology of desktop GIS, web GIS, and mobile GIS to support construction of GIS applications across multi-terminal devices. The GIS cloud platforms and GIS terminals are integrated through technology of cloud-terminal interconnection and cooperation, realizing efficient interconnection and collaboration between clouds and terminals. In SuperMap GIS, the core GIS functions are based on the cross-platform universal GIS core (UGC) which was developed by standard C++. Because of the high compatibility of UGC with Spark and Scala, SuperMap GIS can be executed in Spark seamlessly.

The Cloud-terminal integration GIS platform for spatiotemporal big data that we have designed (see Fig. 10) not only builds the iObject for Java inside Spark, but also supports extended development. For this reason, SuperMap GIS can support using PostgreSQL, Elasticsearch, HDFS and MongoDB to store its data, using the open-sourced GIS cross-platform desktop (SuperMap iDesktop Cross) to effectively visualize the data, and using the GIS application server (SuperMap iServer) to publish its data as services. It also supports many high-performance cloud computing infrastructures such as Docker.

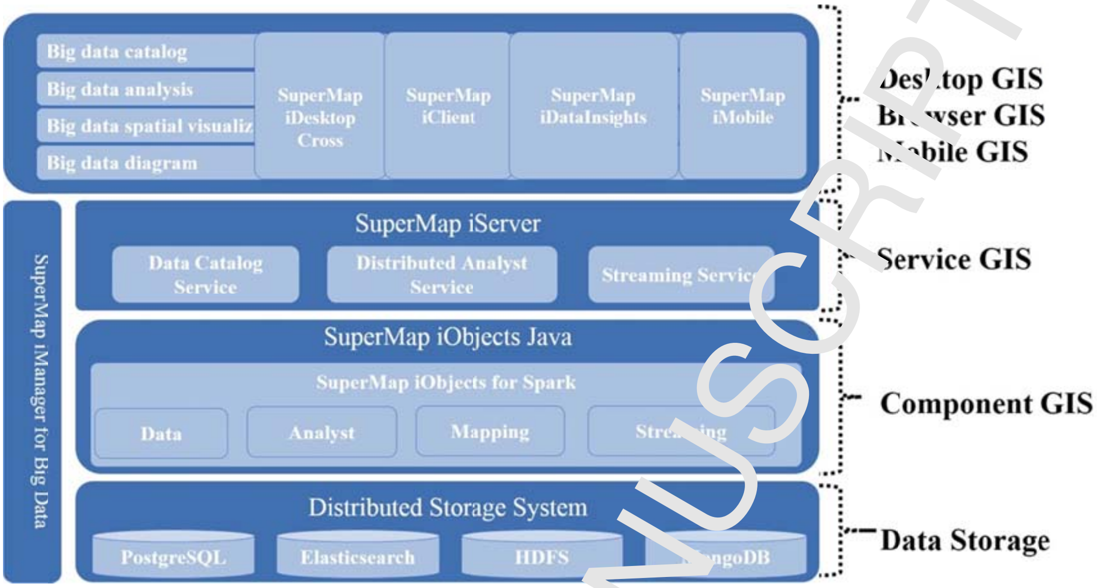

Fig. 10. The SuperMap GIS framework

Based on the SuperMap iObjects for Java and the latest Apache Spark framework, we implemented the spatiotemporal big data engine shown in Fig.11.

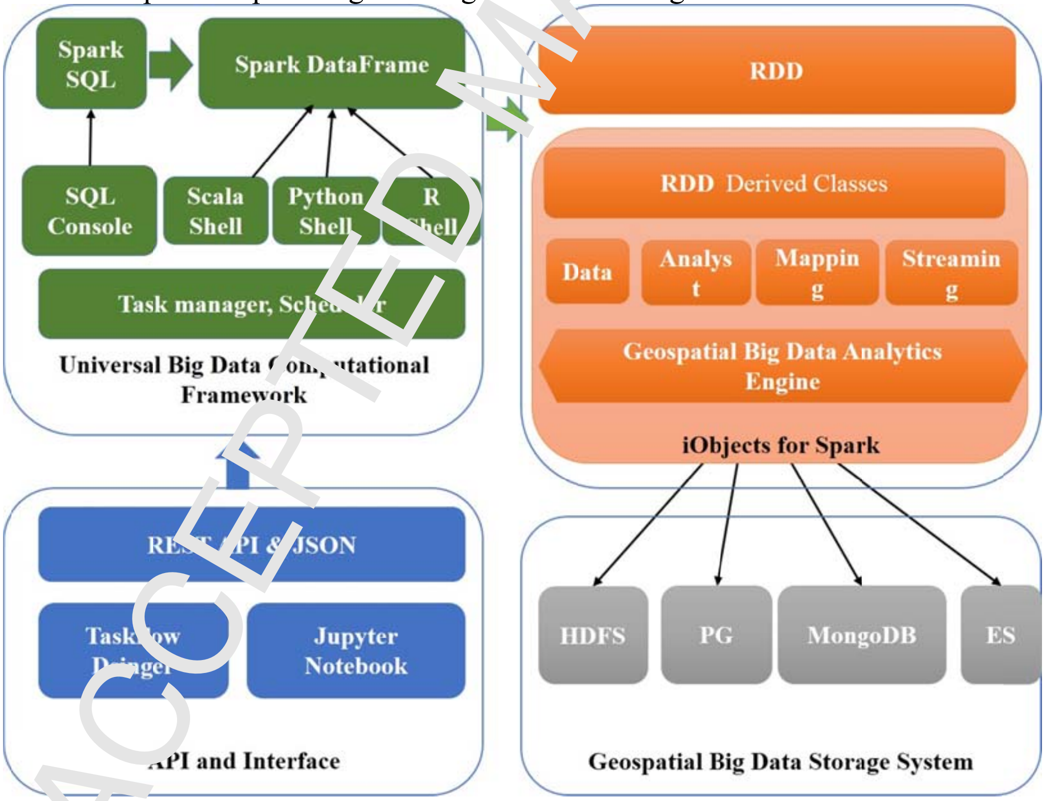

Fig. 11. Technology roadmap of the spatiotemporal big data engine

The engine uses HDFS, MongoDB, PostgreSQL and Elasticsearch to store spatiotemporal big data. It also supports the processing different data formats such as vector, image, and stream data. It is compatible with all of the functions in the existing GIS system, with an enhancement of its data services and spatial analysis functions. By using the distributed data storage system and Spark's cluster computing, GIS users can manage and analyze big data, which the traditional relation databases cannot accommodate.

Some customizations were made to the engine to extend Spark's Resilient Distributed Datasets (RDD). The engine can not only support traditional GIS analysis and basic computing (e.g., spatial query, nearby analysis, interpolation analysis, buffer zone analysis, and overlap analysis), but also many advanced GIS computing and data analyses like cluster analysis, density analysis, hot zone analysis, map matching, and traffic computing. By using Spark streaming, users are also able to analysis the spatial stream data. Additionally, with the built-in GIScript, an open-sourced data processing system, and the R data statistic tool, users can use Python and R to perform spatial data analysis and spatiotemporal data mining.

With this engine, iServer gains the distributed computing capability. The iDesktop users can utilize the web services offered by iServer to read and write the data from distributed cluster and to submit computing or analysis tasks back to the server.

Designed to run on the delegated multi-user infrastructure, the engine was deployed in the OpenStack cloud computing environment and in the Docker container. The Docker container makes it possible for the user to build and manage a distributed computing cluster quickly and effectively.

5. Experiments

To test the performance of the spatiotemporal big data processing and analysis in real-time using the architecture described in the previous section, we choose Apache Spark Streaming. For the storage of the real-time data, we used Elasticsearch. We propose a set of modules to build the program for processing the real-time data stream dynamically. Based on the combination of these modules and the Spark Streaming — to prove the increased usability of data processing and the substantially improved performance and throughput — we share an application example. We used over 1 billion records of global flight and around 0.15 billion records of worldwide shipping to carry out the experiments.

An Apache Spark cluster was constructed based on 5 machines. The operating system of single node is 64-bit Ubuntu Linux (16.04). And the CPU is Intel core i7-6700K. quad core processor. Other hardware configurations are quad core processors with a master frequency of 4 GHz, a hard disk of 1 TB, a memory of 16 GB and a hard disk of 16 GB. SuperMap iObjects Java and BDT Spark extension package are installed and configured on each node. Moreover, The Hadoop version is 2.7.3. and the Spark version is 2.1.0.

Fig.12 depicts the spatiotemporal big data stream processing in a flow chart. Real-time data, including location data (such as vessel or vehicle location), or other form of state data (such as air temperature, barometric pressure, PM2.5), can be received through Socket, HTTP and JMS. The parser allows CSV, JSON and GeoJSON data as input. Real-time big data computing components are used for receiving, filtering, mapping and processing. Output data can be exported to various locations and saved to Kafka/HDFS. Outputs can also be sent to JMS, Active MQ, and Socket. Results can be shown in many types of clients.

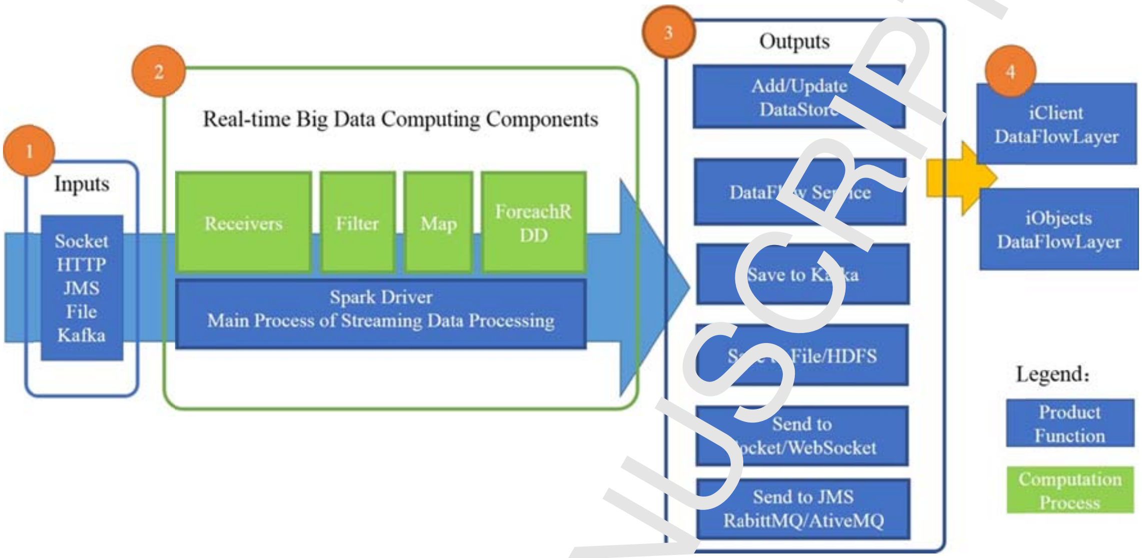

Fig. 12. Flow chart of spatiotemporal big data stream processing

Fig.13(a) depicts the spatial query results of global shipping based on Elasticsearch. Fig.13(b) displays where shipping routes were rebuilt

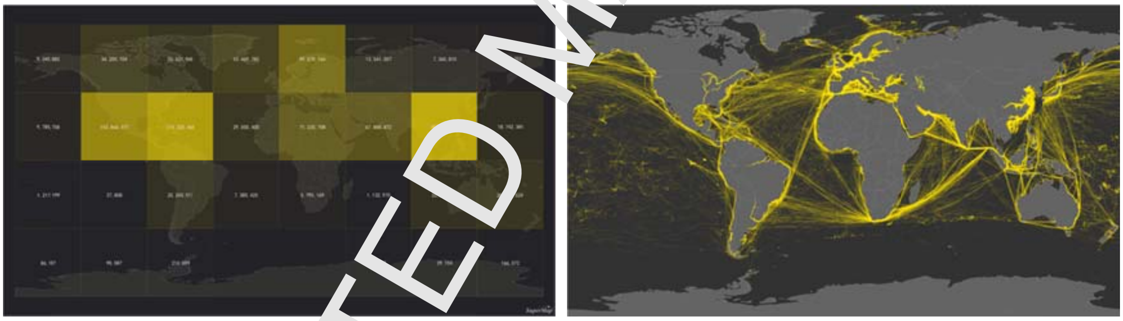

Fig 13. (a) Grid spatial query; (b) Rebuild shipping routes

Fig.14(a) shows real-time data of the monitor aircraft. A tracking map illustrates that 10 thousand long-distance air lines were tracked from 1 billion tracing points (Fig.14(b)). We use spatial association relation near, determine whether the time interval of the two planes is within 10 seconds. It takes two minutes of submitting the task, the analysis results are showed on the screen. But in traditional GIS, it takes more than thirty minutes to finish same task (Table 1).

Table 1: Performance result

Data	Operators	Traditional GIS (minutes)	Our Method (minutes)
0.15 billion points	Spatial query	3.23	0.45
0.15 billion points	Routes built	8.37	1.18

1 billion tracing points	Spatiotemporal query	31.12	2.01

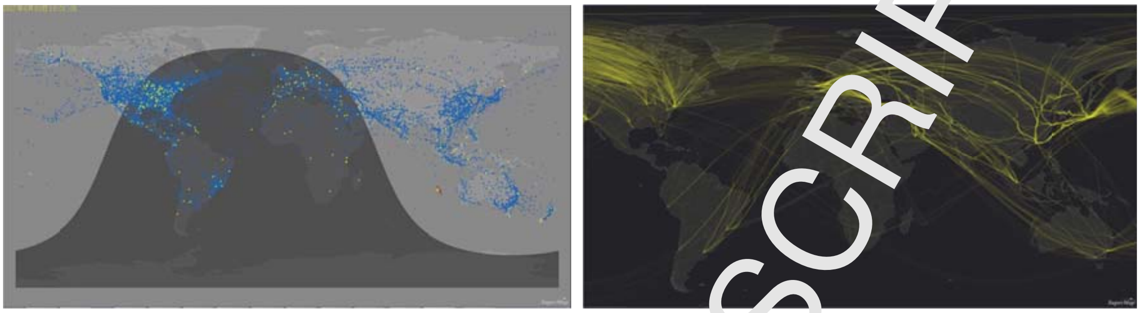

Fig 14. (a)Real-time data of monitor aircraft; (b) Rebuild tracking line from global flight

6. Conclusion and future work

In this paper, we propose an integrated GIS platform architecture for spatiotemporal big data, which contains large-scale virtual storage, a distributed computing framework, cloud computing and integration, stream data processing, 3D and virtual reality, being rapidly applied across the multi-terminal, the open source community, and the container and continuous delivery model. This makes GIS integrate into the mainstream of IT and provide geospatial information technology with unprecedented opportunities.

With the advancement in IT technology, key capabilities for processing spatiotemporal big data in GIS have improved tremendously. This progress has generated tremendous added data value for users. Large scale virtual storage technology allows users to maintain their data assets in long term, and maximize their data's value. The distributed computing architecture allows users to process a big volume of data instantly and to analyze a spatial model in a short amount of time. Cloud computing technologies enhance the resources management and maintenance in data centers. Its integration between the public cloud and the private cloud helps to better satisfy the requirements from professional users. The capability to manage stream data makes spatialized instant collaboration possible, and also makes a contribution to social network analysis, IoT systems, and smart city systems. It has solved the issues of storing and processing data. The 3D and virtual reality technologies create a 'digital image' of the real world. They provide digitalized solutions with high precision for spatial resources management, city operation, and city maintenance. It can improve management efficiency and service quality in these fields. The multi-type platforms make it possible to use spatiotemporal big data anytime and anywhere. They help to fully recognize the value of big data. The development of open source software and its community not only helps the technologies to evolve and related professions to grow, but also helps users to solve issues, such as limitations in system suppliers, high system risks, and high maintenance costs. The container (Docker, etc.) technology and continuous delivery model make the integrated DevOps process become possible. It shortens the cycles of software release and bug fixing, facilitates the problem solving of system complications caused by an increase in software scale, and reduces the cost of unit operation.

By integrating the latest big data technologies, software development methods, continuous delivery methodologies, 3D and virtual reality, and cloud computing with GIS, we have begun a new chapter in GIS. On one hand, it enhances the spatiotemporal analysis and visualization in various information systems, on the other hand, it provides a powerful geospatial data foundation to support diverse fields, such as resources, environment, energy, and city development. GIS has been transformed from the traditional applications for static cartography to the systems that could process and analyze dynamic, real-time spatiotemporal big data. It not only heralds greater potentials for sustainable development, geoformation science, and geography science, but also generates many unprecedented opportunities for related industries as well.

Acknowledgements

Our research was supported by the National Key R&D Plan (2016YFB0502004), the Independent Research Project of State Key Laboratory of Resources and Environmental Information Systems, Chinese Academy of Sciences (088RAC00YA), the Project of Beijing Excellent Talents (201500002685XG242), and the National Postdoctoral International Exchange Program (Grant No. 20150081).

References

[1] L. Qingquan, L. Deren, Big data GIS, Geomatics and Information Science of Wuhan University, 39 (2014) 641-644.
[2] W. Erqi, W. Shaohua, Technology Trends Prospects of the Future Development of GIS, Bulletin of Surveying and Mapping, 2015 (2015) 66-69.
[3] S. Li, S. Dragicevic, F.A. Castro, M. Sester, S. Winter, A. Coltekin, C. Pettit, B. Jiang, J. Haworth, A. Stein, Geospatial big data handling theory and methods: A review and research challenges, ISPRS Journal of Photogrammetry and Remote Sensing, 115 (2016) 119-133.
[4] C. Yang, Q. Huang, Z. Li, K. Liu, F. Hu, Big Data and cloud computing: innovation opportunities and challenges, International Journal of Digital Earth, 1 (2017) 13-53.
[5] C. Yang, M. Yu, F. Hu, Y. Jiang, Y. Li, Utilizing Cloud Computing to address big geospatial data challenges, Computers, Environment and Urban Systems, 61 (2017) 120-128.
[6] S. Wang, E. Zhong, E. Wang, Y. Zheng, W. Cai, S. Li, S. Gao, GISpark: A Geospatial Distributed Computing Platform for Spatiotemporal Big Data, in: AGU Fall Meeting Abstracts, 2016.
[7] Z. Li, C. Yang, B. Jin, M. Yu, K. Liu, M. Sun, M. Zhan, Enabling big geoscience data analytics with a cloud-based, MapReduce-enabled and service-oriented workflow framework, PloS one, 10 (2015) e0116781.
[8] Z. Li, C. Yang, K. Liu, F. Hu, B. Jin, Automatic Scaling Hadoop in the Cloud for Efficient Process of Big Geospatial Data, ISPRS International Journal of Geo-Information, 5 (2016) 173.
[9] D. Sui, Opportunities and impediments for open GIS, Transactions in GIS, 18 (2014) 1-24.
[10] S. Dianzhi, Y. Xinyue, G. Tian, Open GIS for big data: opportunities and impediments, Progress in Geography, 33 (2014) 723-737.
[11] A. Aji, X. Sun, H. Vo, Q. Liu, R. Lee, X. Zhang, J. Saltz, F. Wang, Demonstration of Hadoop-GIS: a spatial data warehousing system over MapReduce, in: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2013, pp. 528-531.
[12] F. Wang, R. Lee, Q. Liu, A. Aji, X. Zhang, J. Saltz, Hadoop-gis: A high performance query system for analytical medical imaging with mapreduce, Atlanta–USA: Technical report, Emory University, (2011) 1-13.

[13] H. Vo, A. Aji, F. Wang, Sato: A spatial data partitioning framework for scalable query processing, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2014, pp. 545-548.

[14] A. Eldawy, M.F. Mokbel, SpatialHadoop: A MapReduce framework for spatial data, in: Data Engineering (ICDE), 2015 IEEE 31st International Conference on, IEEE, 2015, pp. 1352-1363.

[15] S. Gao, L. Li, W. Li, K. Janowicz, Y. Zhang, Constructing gazetteers from volunteered big geo-data based on Hadoop, Computers, Environment and Urban Systems, 61 (2017) 172-186.

[16] J. Yu, J. Wu, M. Sarwat, Geospark: A cluster computing framework for processing large-scale spatial data, in: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2015, pp. 70.

[17] J. Yu, J. Wu, M. Sarwat, A demonstration of GeoSpark: A cluster computing framework for processing big spatial data, in: Data Engineering (ICDE), 2016 IEEE 32nd International Conference on, IEEE, 2016, pp. 1410-1413.

[18] J. Zhang, S. You, L. Gruenwald, Large-scale spatial data processing on GPUs and GPU-accelerated clusters, SIGSPATIAL Special, 6 (2015) 27-34.

[19] S. You, J. Zhang, L. Gruenwald, Large-scale spatial join query processing in cloud, in: Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference on, IEEE, 2015, pp. 34-41.

[20] M. Tang, Y. Yu, Q.M. Malluhi, M. Ouzzani, W.G. Aref, Locationspark: a distributed in-memory data management system for big spatial data, Proceedings of the VLDB Endowment, 9 (2016) 1565-1568.

[21] D. Xie, F. Li, B. Yao, G. Li, L. Zhou, M. Guo, Simba: Efficient in-memory spatial analytics, in: Proceedings of the 2016 International Conference on Management of Data, ACM, 2016, pp. 1071-1085.

[22] K. Zheng, Y. Fu, Research on vector spatial data storage schema based on Hadoop platform, International Journal of database Theory and Applications, 6 (2013) 85-94.

[23] C. Wenwen, W. Shaohua, Z. Ershun, H. Chenpu, X. Liu, Design and Implementation of a New Cross-platform Open Source GIS Desktop Software, Bulletin of Surveying and Mapping, 2017 (2017) 122-125.

[24] J.-G. Lee, M. Kang, Geospatial big data: challenges and opportunities, Big Data Research, 2 (2015) 74-81.

[25] F. Zhang, Y. Zheng, D. Xu, Z. Du, Y. Wang, R. Liu, X. Ye, Real-time spatial queries for moving objects using storm topology, ISPRS International Journal of Geo-Information, 5 (2016) 178.

[26] Z. Galić, Spatio-Temporal Data Streams and Big Data Paradigm, in: Spatio-Temporal Data Streams, Springer, 2016, pp. 47-69.

[27] Z. Galić, E. Mešković, D. Osmanović, Distributed processing of big mobility data as spatio-temporal data streams, GeoInformatica, 21 (2017) 263-291.

[28] L. Shaojun, Z. Ershun, W. Shaohua, Z. Xun, Z. Qin, X. Jiong, Research on Opening Geospatial database connectivity, in: Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on, IEEE, 2013, pp. 462-466.

[29] E. Zhong, Geocontrol and live geography: Some thoughts on the direction of GIS, Journal of Geo-Information Science, 15 (2013) 783-792.

Authors Bios.

Dr. Shaohua Wang

Shaohua Wang received a Diploma degree in mathematics from Beijing University of Chemical Technology, China, in 2006. He received Ph.D. degree in the field of Cartography and GIS from the University of Chinese Academy of Science in 2013. From 2013 to 2017, he is Postdoctoral Research Assistant at Institute of Geographic Sciences and Natural Resources Research, CAS. From 2016 to 2017, he is a visiting scholar in Geography Department of Geography, University of California, Santa Barbara. His major research interests include spatial cloud computing, spatiotemporal big data, spatial visualization and spatial optimization.

Mr. Yang Zhong

Yang Zhong is a M.S. student in Information Systems & Technology with a concentration in GIS Solutions Development at Claremont Graduate University, California. He earned his B.A. degree in Software Engineering from Beijing JiaoTong University, China. He is currently conducting research in geospatial data visualization and integrating social network data into GIS.

Prof. Erqi Wang

Erqi Wang received a Diploma degree and a Master degree both in Geography from Beijing Normal University, China. From 1998 to 2017, he is a Chief Architect at SuperMap Software Co., Ltd., Beijing. His primary research interests include spatial cloud computing, spatiotemporal big data, Open source.

Authors Photos.

Dr. Shaohua Wang

Mr. Yang Zhong

Prof. Erqi Wang

Highlights for Reviewers

Article entitled: An Integrated GIS Platform Architecture for Spatiotemporal Big Data

Highlights

In this paper, we propose an integrated GIS platform architecture designed to meet the requirements of processing and analyzing spatiotemporal big data.
Cloud-terminal Integration GIS for spatiotemporal big data is developed according the novel architecture.
The experiments about Stream processing for spatiotemporal big data showed SuperMap GIS spatiotemporal big data engine achieved excellent performance.