Cloud Computing and Big Data for Oil and Gas Industry Application in China

Yang Zhifeng[1,2,3]*, Han Fei[4], Feng Xuehui[3], Yuan Qi[3], Cao Zhen[3], Zhang Yidan[5]

[1]Postdoctoral Workstation, Xinjiang Oilfield Company, Petrochina, Karamay, Xinjiang, China.
[2]State Key Laboratory of Petroleum Resource and Prospecting, China University of Petroleum, Beijing, China.
[3]Sugon Information Industry Co., Ltd., Beijing, China.
[4]Lenovo (Beijing) Co., Ltd., Beijing, China.
[5] Xinjiang Oilfield Company, petroChina, Karamay, Xinjiang, China.

* Corresponding author. Email: yangzhifeng_2005@126.com
Manuscript submitted February 12, 2019; accepted April 13, 2019.
doi: 10.17706/jcp.14.4.268-282

Abstract: The oil and gas industry is a complex data-driven industry with compute-intensive, data-intensive and business-intensive features. Cloud computing and big data have a broad application prospect in the oil and gas industry. This research aims to highlight the cloud computing and big data issues and challenges from the informatization in oil and gas industry. In this paper, the distributed cloud storage architecture and its applications for seismic data of oil and gas industry are focused on first. Then,cloud desktop for oil and gas industry applications are also introduced in terms of efficiency, security and usability. Finally, big data architecture and security issues of oil and gas industry are analyzed. Cloud computing and big data architectures have advantages in many aspects, such as system scalability, reliability, and serviceability. This paper also provides a brief description for the future development of Cloud computing and big data in oil and gas industry. Cloud computing and big data can provide convenient information sharing and high quality service for oil and gas industry.

Key words: Big data, cloud computing, cloud desktop, oil and gas industry, China.

1. Introduction

The evolution of Cloud computing over the past few years is potentially one of the major advances in the history of computer field. At present, a lot of research work on cloud computing and big data has been applied at the domestic and foreign field. Google was the first company to implement cloud computing services. IBM released the "blue Cloud platform schedule" in 2007. Microsoft developed Windows live online services. The governments of all countries also pay more attention about cloud computing. The U.S. government tax monitoring platform had been deployed on the Amazon Cloud computing platform and became the first government in the world to use commercial cloud computing services. The purpose of cloud computing is to reduce costs and help users focus their critical application and avoid the constraint of traditional IT technique. Cloud computing transfer the risks for over-provisioning or under-provisioning of the Cloud computing vendors, who mitigates that risk by statistical analysis of customer group. The cloud computing vendors provide quality and convenient services for IT client at a relatively low price [1].

With the rapid development of information technology in recently years, cloud computing and big data technology has also been introduced into the oil and gas industry. The upstream oil and gas industry

engages in exploration and production, midstream includes storage and transportation, downstream evolves refining and chemical industry as well as sales. Oil exploration and production has the characteristics of large data volume, high I / O read and write performance, high reliability and linear expansion of data storage nodes. The application of new generation cloud platform will help meet these challenges. Cloud computing integrates application software platform of big data. It provides service for different departments of oil and gas industry. Cloud computing provides a unified data storage and data access interface for the seismic exploration, drilling and other department. It provides information services and establishes an integrated business information platform. Mykola Gordii articulated the global oil and gas industry investment in public cloud would rise from USD 1billion in 2012 to more than USD 2billion in 2014[2]. Spending on Cloud computing and big data technologies in the next few years may climb as high as 42 billion/year [3].Based on the rapid development of Cloud computing and big data technologies in China, the new Cloud storage, Cloud desktop and big data architecture of oil and gas industry are presented in this study.

2. Cloud Computing

2.1. Cloud Computing Definition

Cloud computing is defined by the National institute of standard and Technology (NIST) as a model for enabling convenience, on-demand network access to share pool of configurable computing resources(such as servers, storage device, network, application, and services). Cloud computing architecture can be rapidly deployment and released with the characteristic of low management cost and interaction service function"[4]. Cloud computing, which integrates grid computing, distributed computing, network storage, virtualization, load balancing and other traditional computer technology, is an integrated product [5]. Distribution services of large number scattered computers is the main feature of Cloud Computing [6]. Cloud computing has an enterprise data center which enables to migrate resource to meet application requirements and provides access to all storage devices. Cloud computing is appearing as a model in support of "everything-as-a-service" architecture, hardware, software, network are available for users in the form of services [7].

Xu summarized the characteristics of Cloud computing as follows, low cost information technologies, real time dynamic resource deployment, flexibility, standardization, and ultra-large scale [8].The advantage of Cloud computing in oil and gas industry can be summarized as follows: first, it improves the resource utilization. Second, Cloud computing provides free download, installation and time limitless service which can reduce the cost of IT industry. Third, Cloud computing has unparalleled advantage in real time data backup, high reliability for oil and gas industry information management. Fourth, the virtual data center of cloud computing can reduce energy consumption and carbon emission which exactly meet the demand for green IT era.

2.2. System Architecture of Cloud Computing

According to NIST, cloud computing services usually includes three models: SAAS (software as a service) [9], PAAS (Platform as a service) and IAAS (infrastructure as a service) [10], [11]. At a conceptual level, there are essentially three types of Cloud architecture: public cloud, private cloud and hybrid cloud [12]-[14]. Public Clouds constructed by commercial providers which offer an internet-accessible interface for building and managing computing resource within their own physical domains [10], [15]. Private Clouds are flexible for managing and getting rid of these kinds of threats. The purpose of private clouds is "to sell computing resource by publicly accessible interfaces from the internet, but to give local user a flexible and agile private infrastructure to run service workloads within their managed domains". Hybrid Cloud is a

very useful bridge that customers can move some computing resource to public Cloud while still hanging on to legacy systems [10]. Feblowitz articulated that the public cloud is a deployment model where the Cloud is open to a largely unrestricted potential client [16]. The private cloud is designed for restricting access to a single enterprise. Hybrid cloud solution, which combines private cloud with public cloud resources, is a smart way to reap many benefits from the public cloud while ensure data security.

Xu Weiping suggested that the critical application of oil and gas industry should be deployed in the form of enterprise private Cloud [8]. Non-critical application, which can get flexibility and low cost management from SAAS vendor's service, should be deployed on public cloud. Enterprise hybrid cloud mod, which combines private Cloud with public Cloud, can gain a balance between safety and efficiency. The author confirmed that the cloud computing of information management system in oil and gas industry can not achieve overnight. First, we should build private cloud to meet specific business needs of working environment in the oil and gas industry. Second, we should continue to gather service and data expansion in platform. Third, with the continued data extension, establish a Chinese "oil industry cloud" to service for the industry as a whole and even other countries.

3. Big Data

3.1. Big Data Definition

Big data is information that is so large, complex and fast moving. It's difficult to handle using everyday data management tools [17]. Typically data sets are so large and complex that they require advanced data storage, management, analysis, and visualization technology [18]. David cameron indicated that big data integrated a variety of IT technologies, including regular analysis, large scale parallel databases, memory computing, fast retrieval, natural language processing and the statistical analysis of large datasets [19]. J.Johnston proposed that big data revealed hidden laws and unknown correlations according to the exhibited data set [20].

The first definition of big data was proposed by its three features, Volume, Velocity and Variety. Based on data quality, IBM has added the fourth V called Veracity. While, oracle has added the fourth V called Value, for emphasizing the value of big data [21]. A recent Mckinsey global institute report defines big data as "data-sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze". Big data technologies are essentially based on ApacheTM Hadoop project which is open-source software for reliability, scalability and distributed computing [22].

3.2. System Architecture of Big Data

The main application scenarios for big data include batch processing, interactive analysis and streaming processing. Batch processing has the characteristics of large data volume, high precision and low data value density. Data is clear and limited, and transferred in block mode during the batch processing. The data is stored first and then analyzed. MapReduce is a very important batch-processing module, the data processing time is relatively longer. Batch processing is mainly used in the Internet of things, cloud computing and the Internet. The representative system of batch processing is GFS, MapReduce and HDFS.

The system operator realizes one question and one answer mode by the way of man-machine interaction data processing. The data files in stored system are processed and modified at near real time, and the results are processed immediately. Interactive data processing is mainly used in information processing systems (OLTP, OLAP, Hive and Pig) and the Internet (search engine, email, real time communication tools, social networks, Microblog and blog). Typical databases for interactive analysis include NoSQL, HBase and MongoDB.

Stream processing mainly includes streaming data processing (log real-time acquisition) and interactive

data processing. Streaming data structures are diverse with timestamp or ordering attributes. Streaming data are new data or data streams, unknown or unlimited in advance, non-stored or non-predictable in memory. The data are processed quickly and the results are obtained instantly. The data stream can reach at the second-millisecond level. The typical application of streaming data processing is data collection (log, sensor and webpage).

4. Application of Cloud Computing and Big Data in the Oil and Gas Industry

Oil and gas industry face a lot of questions such as data integration, achievement sharing and work collaboration. Cloud computing provides not only massive data storage and sharing solutions, but also integrates various types of hardware and software resources, to meet the informatization requirement of oil and gas industry. At the same time, the data integration technology is used to manage all kinds of data and information in a unified and effective way, and break the restrictions on accessing different departments among oil and gas industry. Cloud storage combines various information islands to achieve a unified data information service and sharing mechanism.

The cloud computing of oil and gas industry is mainly including research cloud and desktop cloud. Research cloud improves work efficiency of exploration and production. Cloud computing can realize collaborative work and remote visualization in the oil company's headquarters, branch company, and even remote place. Petroleum companies will leverage cloud services to enhance existing super-computing capabilities and process massive seismic data generated by ultra-sensitive seismic sensors, and reduce imaging analysis time. The cloud computing in midstream and downstream oil and gas industry is mainly concentrated on the cloud-desktop application, management and online software service.

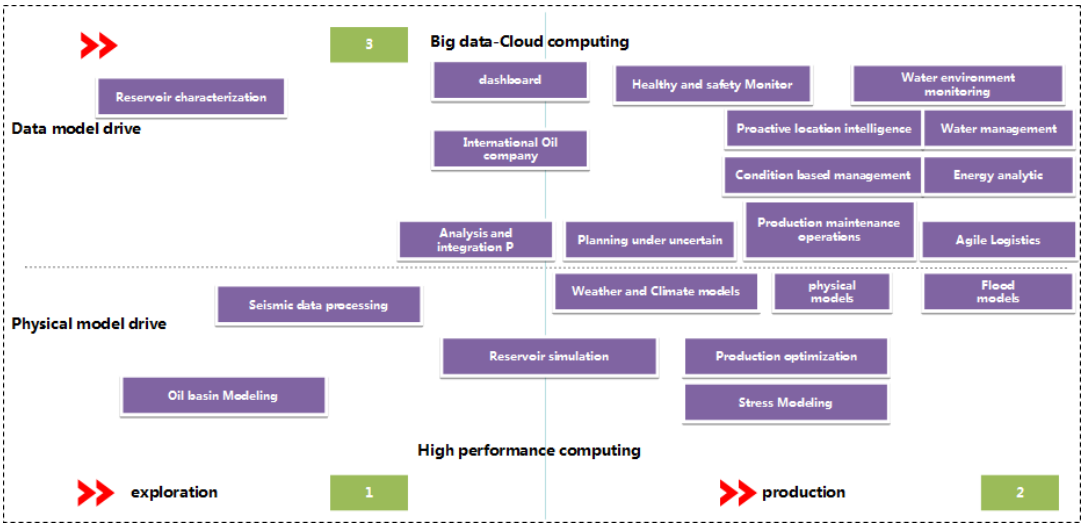


Fig. 1. Cloud computing and big data of oil and gas industry.

4.1. Cloud Storage Application of Oil and Gas Industry

With the broad application of 3D seismic exploration and high-precision 3D seismic acquisition techniques, the data volume of petroleum exploration is very huge. Seismic data is a kind of big data. Seismic data management requires an effective system, as well as high performance and large capacity storage device. The seismic data storage is a vital part of geophysical exploration. It has the characteristics of fast storage, large capacity, strong environment adaptability, high reliability and long storage cycle. In recent years, the cloud storage technology for big data has become more and more popular. Cloud storage is a system that provides data storage and business access. Based on the functions of cluster application,

grid technology, and distributed file system, cloud storage will integrate a large number of different storage devices. The large-scale distributed cloud storage system has massive storage space, and supports the flexible, high efficient, high-performance and reliable service for file sharing storage platform and high concurrent access. At present, the research on cloud storage is mainly focused on the security of data storage and cloud storage architecture. With the development of cloud storage technology, many scholars apply cloud storage technology on the critical application systems in different industries.

Naresh vurukonda was concerned about the security of data storage in cloud computing, and proposed an effective solution for data storage in the field of privacy and encryption under cloud environment [23]. Swapnali More proposed public audit architecture for privacy protection based on the third party audit platform. Cloud computing has the characteristics of privacy protection, public audit and data integrity [24]. Li concerned on the security issues of cloud data storage,and proposed a way to prevent cloud administrator to gain sensitive data from clients [25]. Wang studied the data replication techniques of distributed storage system and evaluated its performance [26]. Wu developed a cloud storage system named MingCloud with high availability and high performance. The cloud storage system has the architectural features of the Master/Slave [27]. Li used dual active Master nodes and multi service heartbeat detection algorithm to build a distributed and high availability framework for smart grid design [28]. Liao built distributed storage architecture for multimedia file segmentation and editing storage with video servers cluster. Base on the business characteristics of the relay satellite system [29], Chen design a cloud storage architecture. The cloud storage system meets the demand of massive data storage, and solves the bottleneck of large capacity data storage in the relay satellite system [30].

4.1.1. Comparison of distributed cloud storage technology

Table 1. Comparison of GFS, HDFS, Lustre and Parastor File System

GFS	HDFS	Lustre	Parastor
Master Node+ chunck Node	Name Node+ Data Node	single or multiple meta-data node,single or multiple object storage service,client	2~128 meta-data nodes,3~4096 data nodes
More Read、more write mode,any file write,the default size of block is 64MB	Write alone, more read mode, only attachment mode,the default size of block is128MB	Multi-client concurrent large files read and write scenarios, write performance is better than read performance	Large file and small file scenarios
Master receives heart beat from chunk server	Name node receives heart beat from data node	RAID1 or RAID5/6 of storage nodes provides reliability	Index node configuration RAID6,data node configuration copy and erase code
Google	Developed early by Yahoo, now an open source architecture	Open source storage based object	Independent research and development
C、C++	Java	C	

With the expansion and changes of big data in application requirements, the traditional GFS, HDFS and Lustre distributed file systems have obvious shortcomings in the requirements of massive file processing scenarios and data storage fault tolerance. Compared with the traditional distributed file system, this paper introduces Parastor, new and efficient distributed parallel cloud storage architecture. Parastor is an

asymmetric storage system for massive unstructured data processing. It can provide TB/s high speed bandwidth and EB level storage space, with ultra-strong scale out capacity. It is far beyond the traditional NAS, SAN and traditional distributed storage in terms of system capacity and linear aggregate bandwidth performance. It has some typical features, such as high reliability, scalability, flexibility and excellent storage performance (Table 1).

4.1.2. Cloud storage application in oil and gas exploration

Petroleum exploration has the characteristics of single source, compute-intensive and complicated process. To meet the storage requirement of the oil and gas industry, a distributed parallel cloud storage system is used to build storage architecture with high stability, high security and sustainability. It can meet the demand of massive seismic data storage in oil and gas industry, the high I/O reading and writing bandwidth and the high floating point computing performance. The cloud storage architecture has four important modules, including storage subsystem, computing subsystem, network subsystem and seismic data processing and interpretation subsystem. The storage subsystem includes the management node, meta-data node and data node. The POSIX protocol is configured for high performance computing application scenarios, with higher bandwidth and consistency in the data cache. Computation subsystem is a critical part of oil exploration and production, mainly engages in seismic data processing. Blade servers and fat node servers are usually selected, in order to meet user's demand for high floating point computing capacity, high scalability and high system memory bandwidth of the seismic data processing. The entire architecture selects 10GbE, FDR or EDR high-speed network, to ensure data transmission rate and meet customer requirement for high-speed network. The workstation shows the results of seismic data processing, and explains the strata and faults after the processing of seismic data. It provides information on drilling decision analysis for geology engineers.

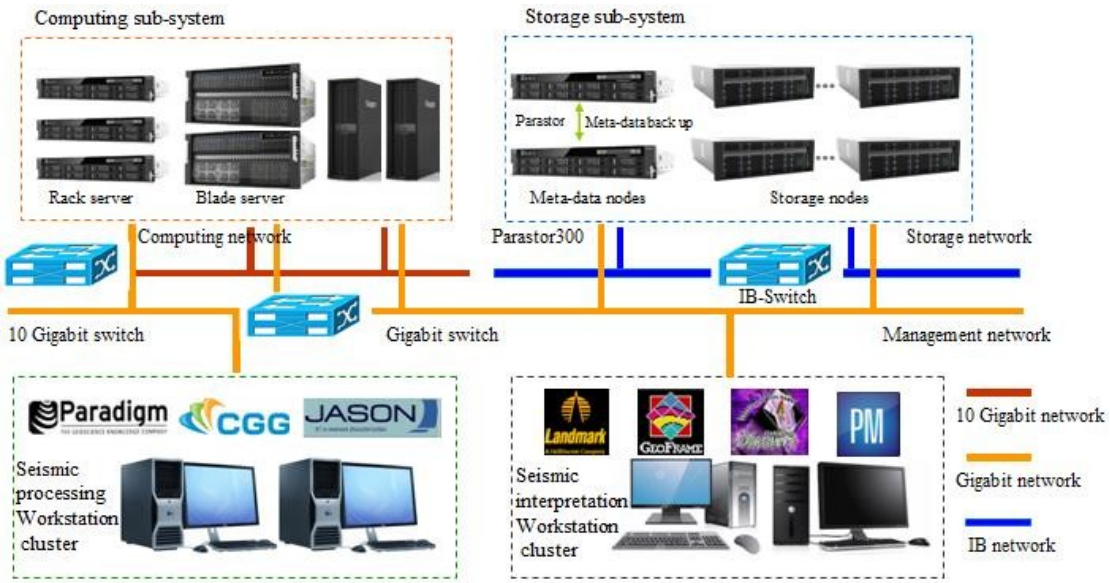


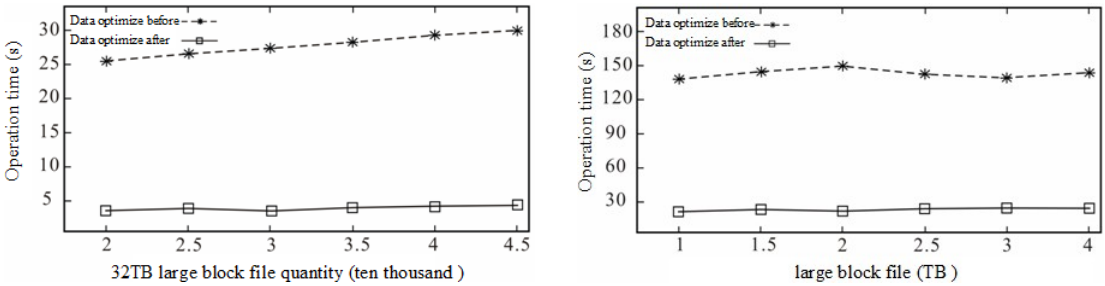
Fig. 2. Distributed storage architecture of seismic data storage.

4.1.3. Test result

BGP is the largest specialized engineering and technology service company for geophysical exploration in China. Its independent research integrated system of seismic data processing and interpretation is GeoEast. Seismic exploration business of BGP has five features, massive data, GB-level I / O throughput, complex business scenarios, high reliability and rapidly growth of data volume. According to the business characteristics of BGP, the parameters of Parastor distributed parallel cloud storage integrated system are

optimized.

On the basis of performance optimization, a set of distributed parallel cloud storage system architecture is deployed. The architecture includes two dual -active redundant meta-data nodes (responsible for meta-data access, storage system monitoring and management) and six data nodes (responsible for data access requirement, and equipped with 216TB bare capacity). The hard disks use an erasure code protection strategy of 8 + 2: 1, with up to 75% utilization of storage space. The test results of the write / read aggregate bandwidth and I / O throughput are up to 5GB / s. IOPS is 120000 times. IO response time is 9ms (Table 2). The result shows that the distributed parallel cloud storage system meets the high reliability requirements of the high speed access and storage system of oil field.



(a) Changes in Ls-l monitoring business data (b) Single file large block random read scene

Fig. 3. Business performance optimization of parastor.

Table 2. Performance Test of Parastor Distributed Parallel Cloud Storage System

Test performance	Aggregate bandwidth(write)	Aggregate bandwidth(read)	I/O through put	IOPS	IO response time
Test software	IO zone	IO zone	IO zone	vdbench	vdbench
Test result	5GB/s	5GB/s	5GB/s	120000	9ms

4.2. Cloud Desktop

4.2.1. Cloud desktop virtualization technology

Cloud computing provides cloud desktop services, cooperation office platform and remote mobile office for the IT managers. It realizes the centralized deployment and sharing application of the portal system in the oil field. Virtual desktop has a broad application prospect. The virtual desktop has the ability to access the desktop anytime and anywhere. The virtual desktop has the characteristics of reducing the management and maintenance cost of hardware and software. VDI, the virtual desktop infrastructure, is running in the operating system of the server. The user accessed the virtual desktop by computing protocol from thin client. The user's access the desktop is like accessing a traditional local installed desktop. Virtual desktop technology needs to provide good man-machine interaction ability. Most desktop applications need frequent interaction from users. With the broad application of computers and electronic entertainment devices, large 3D games and video playback are becoming more and more popular. The requirements of these applications are increasing, and the ability to display high performance and smooth graphics has become the inevitable requirement for virtual desktop system. The traditional VDI is weak in performance and image display. Its imaging does not use GPU, which are generated by virtual machine software. So there are some performance defects in traditional VDI. GPU mainly have floating point operation and parallel operation capacity. Its floating-point operation and parallel computing speed are more than 100 times faster than CPU. After using the GPU virtual technology, virtual machines running on the server can share the same or multiple GPU processors for graphics processing. This safe and efficient way of desktop access is being pursued by more and more users. Therefore, the virtual GPU replaces the

server for image processing, which not only improves application performance but also enables VDI to meet the needs of more users.

4.2.2. Characteristic of SCADA system

The SCADA system is the key business system in the middle stream oil and gas insustry. The SCADA system collects real time data and realizes local or remote control. SCADA system carries on the comprehensive, real-time monitoring for production running process. It provides the necessary reference data for production, dispatching and management. The software of the business system includes 3 parts, the computer operating system, the SCADA system, and the application software. The data transmission throughput is large and the real-time performance is strong. According to rough statistics, the total number of parallel monitoring data is close to millions, and the time precision is usually millisecond. The hardware of the SCADA system includes 3 parts, the host computer system, the lower computer system, the automation instrument and the executive mechanism. The host computer system can operate and control the instructions from other devices, such as the SCADA server and PLC. The lower computer system is responsible for sending data to the upper device and executing operation instructions from sensors and actuators. Data transmission is asymmetrical. The data collected from the lower computer system to host computer system is massive.

4.2.3. Cloud desktop architecture of midstream in oil and gas industry

GPU servers, which are deployed with virtual desktop clusters, are used as the underlying hardware devices. Hot standby node is configured to ensure reliability, continuity and stability for user business system. The GPU servers are used to realize the graphical virtualization of the SCADA system. Use desktop virtualization software to pool the underlying physical resources. Each virtual desktop is allocated 50GB to deploy the operating system, 50GB for data storage and 300GB for system management. The virtual desktop and the customer's SCADA monitoring system are deployed in virtual machine. Virtual desktop terminals are deployed in thin clients cluster. Thin client terminal clusters access the core switching network of SCADA. Independent display devices are used to monitor the oil pipeline system and the crude oil pipeline system.

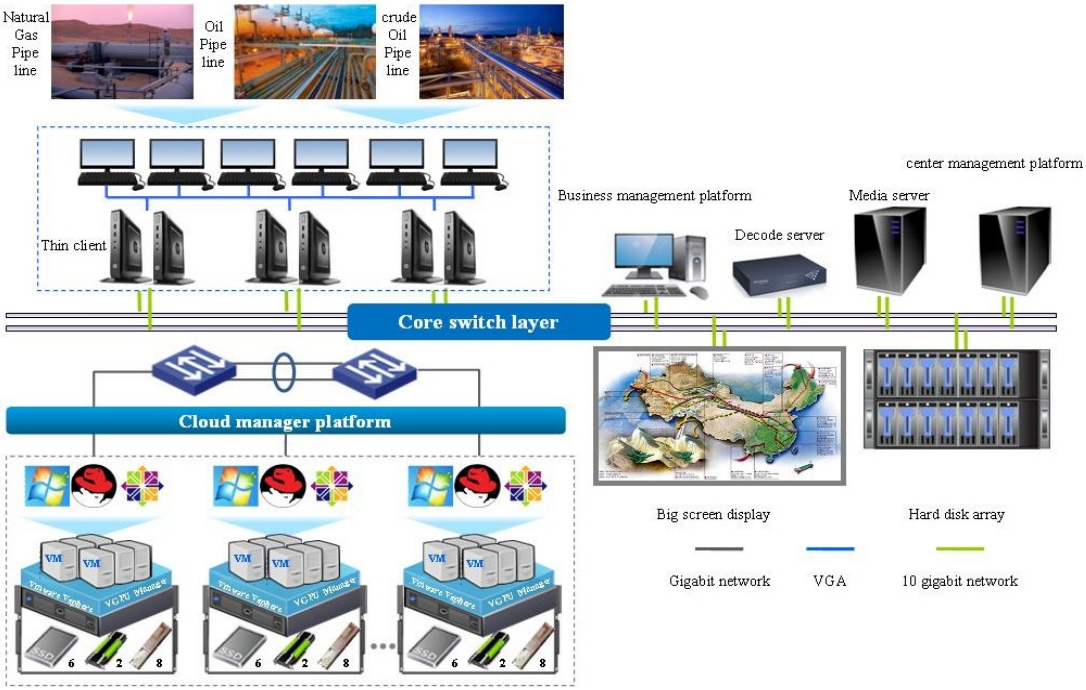
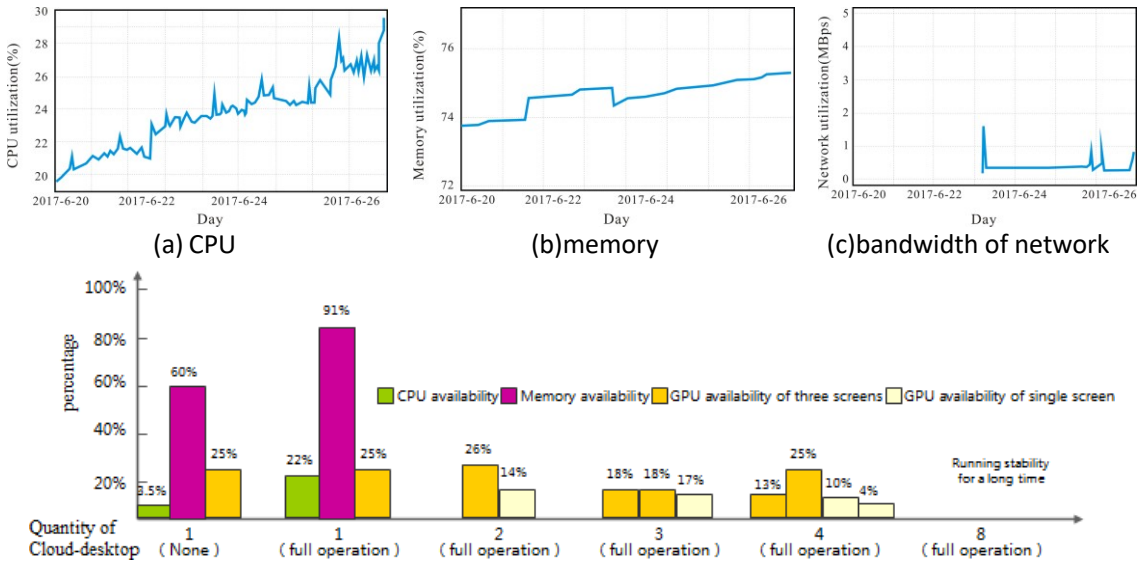


Fig. 4. Cloud desktop architecture of midstream in oil and gas industry.

4.2.4. Test result

A single virtual desktop access case, a virtual machine is allocated 4 virtual cores, 8GB memory. CPU utilization rate is 3.5% ~ 22%, memory utilization rate is 60% ~ 91%, GPU utilization rate is 25%, and network bandwidth is 16KB ~ 870KB. The number of virtual desktops gradually increased to 8, the system runs steadily for a long time. The results of the experiment show that each user needs at least 20-30IOPS. Each Cloud-desktop user needs at least 2GB graphic memory. After using the virtual desktop scheme, the PCO-IP protocol is selected to achieve a more fluent effect and optimize the bandwidth of the virtual desktop. The cloud architecture effectively reduces the operation and maintenance costs for IT managers in data center and speeds up the system deployment time. The cloud desktop simplifies the critical IT management process, reduces the risk of the security department and effectively prevents data leakage.



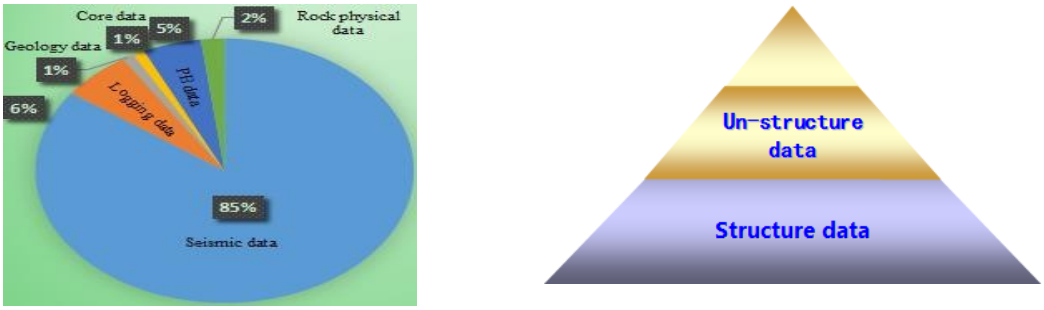
(a) CPU (b)memory (c)bandwidth of network

(d)Virtual desktop running status

Fig. 5. Test result of cloud desktop in oil and gas industry .

4.3. Big Data Application in Oil and Gas Industry

Petroleum industry is complex data-driven business with data volumes growing exponentially [16]. Data volumes can predict potential oil and gas resources, guide hydraulic fracturing of oil field, and enhance oil recovery. While seismic data sets are notoriously large and cumbersome, many other departments of the oil and gas industry also generate massive data [31]. The capacity of modern seismic data center can easily reach as much as 20 PB [32]. Petroleum industry data are usually unstructured or semi-structured data such as email, documents, spreadsheets, images, voice, video and multimedia.



(a) data scale (b) data type of oil and gas industry

Fig. 6. Data characteristic of big data in oil and gas industry.

4.3.1. Upstream oil and gas industry

Seismic data scale of oil exploration and production have already reached above PB level. The upstream oil and gas industry is not strange for big data. Thus,big data analysis techniques is first applied in the upstream oil and gas industry . Then, big data gradually played a critical role in the pipeline transportation, refining and chemical and sales department. The big data research in the upstream oil and gas industry are mainly on seismic data processing, intelligent decision-making support system, drilling based on multiple conditions recognition, production data processing and predictive analytics. Currently, in the process of analysis and forecasting, oil and gas companies couldn't establish relational data model with uncertain variable data sets. Thus, they begin to explore regularities and find solutions by big data analysis techniques. Petroleum companies, service companies and drilling contractors use thousands of sensors which are installed on the ground and underground to provide continuous real-time data acquisition, equipment monitoring and job schedule [33]. Big data comes from sensors, space and GPS coordinates, meteorological services, seismic surveys and various measurement equipment are used for managing surveying, processing and imaging, exploration planning, reservoir modeling, and other business system.

Abdelkader Baaziz articulated that big data can benefit oil and gas industry, includes automatic recognition of seismic trace, reservoir modeling, enhanced oil and gas recovery and predictive analysis [34]. The oil companies use big data techniques to accurately identify abnormal conditions and avoid employee injury and equipment damage. Geological engineers, who want to gain maximize asset utilization and improve production efficiency, collected real-time data from well site, wellhead and underground sensors to optimize drilling techniques, monitoring operation and provide precise geological guidance. Data analysis of related fields can increase efficiency of drilling and completion. In petroleum exploration and production industry, the oil companies use big data techniques to mine the critical factors that reduce oil production, such as geological data, production data, operation data and power consumption data. In the field of petroleum engineering, massive data of historical production well are analyzed by big data technique. Oil recovery identification model is established. Thus the automatic warning of the production well is achieved, the problem is instantly discovered and the production efficiency of oil well is further improved.

4.3.2. Midstream oil and gas industry

In the midstream oil and gas industry, big data mainly focuses on the oil and gas transportation business. Big data technology is mainly applied on data acquisition and monitoring of SCADA system. It collects, manages, analyzes and presents real-time and accurate information from oil and gas pipelines. Thus, big data provides monitoring service and realizes intelligent dispatching in every aspects of oil and gas transportation. Big data technology will make full use of the results of information technology and digitization, strengthen the supervision service and control quality. Big data captured from SCADA systems, drill heads, flow sensors, or other condition sensors is transmitted in near real time. The data is streaming processing rather than batch processing. Data types of SCADA system are structured, unstructured, or semi-structured. Structured data can be time series or transaction data that is conductive to a structured relational database.

4.3.3. Downstream oil and gas industry

In the downstream oil and gas industry, the application of big data mainly focuses on gas station management, sales activity analysis, customer behavior and preferences analysis and marketing strategy. Big-data provides more personalized, accurate marketing services, meanwhile avoids the commercial risk. The oil companies can use big data analysis technology for describing customer portraits in the downstream oil and gas industry.

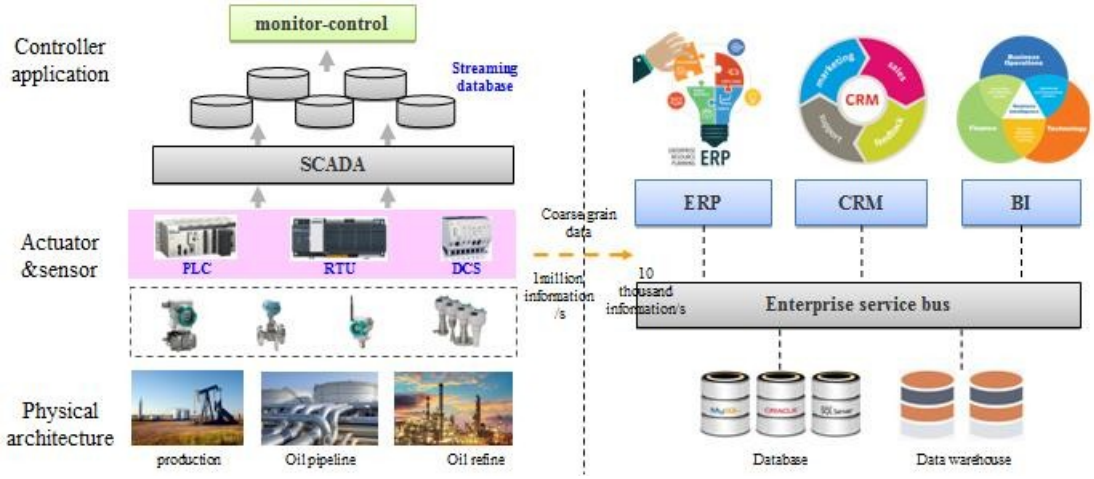


Fig. 7. Application field of big data in oil and gas industry.

4.3.4. Big data architecture of oil and gas industry

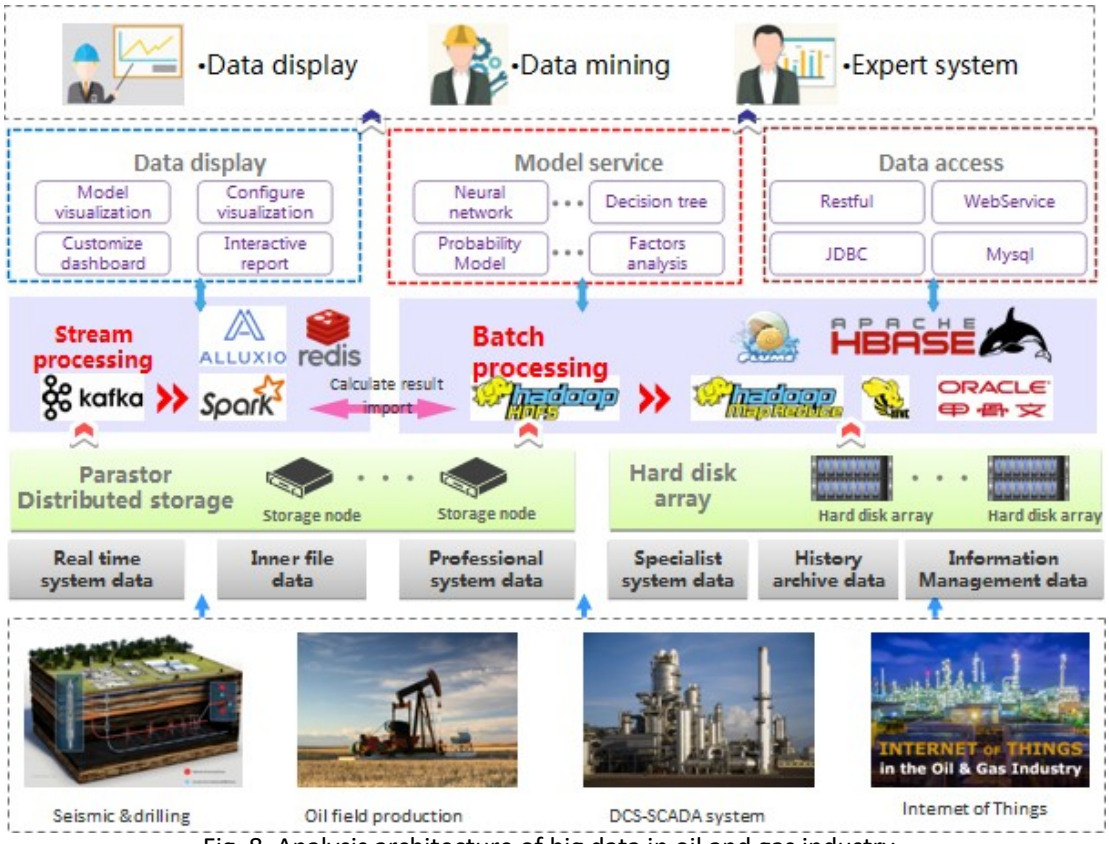


Fig. 8. Analysis architecture of big data in oil and gas industry.

Petroleum companies use tape libraries to archive and disk arrays to store business data and enterprise office data. Big data storage architecture of oil and gas industry uses traditional disk array and distributed cloud storage solution. The largest seismic data volume is stored in the cloud storage architecture, and other structure data are stored in the disk array. First,data cleaning and data standardize are processed, followed by big data analysis. Spark streaming of Hadoop ecosystem is used to access real-time data of the upstream oil and gas industry such as actuators and sensors of SCADA system. The results of memory computing can be imported into Hadoop for relational data analysis and batch processing. The big data

architecture will establish a data model to provide exploration and production decision-making for geologists and oil engineers. The marketing data of oil enterprises are stored in distributed storage, and the user's consumption behavior is described by correlation analysis from Hadoop. After the analysis, the result of big data is displayed on the big screen. The value of data mining will provide analysis and decision for the exploration and production.

5. Cloud Computing and Big Data: State of the Art and Challenge

5.1. Big Data in Cloud Computing

Big data has the following characteristics: large capacity, fast speed, and safe storage. The storage requirement of the big data can not be separated from the cloud computing. The high speed big data can only be processed in the waiting time by cloud computing. At the same time, cloud computing is a feasible way to analyze and understand big data. Big data value can be discovered through data mining. Its potential value can be found from low value density data, and the implementation of big data mining technology is inseparable from cloud computing. In a word, cloud computing is the core support technology on big data processing. Cloud computing provides processing capabilities for big data. Cloud computing based big data analysis is a service model in which elements of the big data analysis process are provided through public and private Cloud.

5.2. Current Challenge

However, the Cloud computing and big data in oil and gas industry are still in its infancy stages with some important technical and institutional challenge. Robert K.perrons analyzed the application of cloud computing in the upstream oil and gas industry [13]. The authors pointed out that data security, the dependence of traditional large seismic data sets, and significant investments in traditional information technology framework are some of the challenges for the migration applications to public cloud in the upstream oil and gas industry. Feblowitz studied Cloud migration in the upstream oil and gas industry [16]. The authors proposed that two issues deserve consideration, when upstream oil and gas business systems were migrated to the public cloud. First, the safety of sensitive data such as seismic and logging data should be considered. Second, the computation resource consumption of critical business system, such as 3D rendering and graphics acceleration, is very huge. Yuan clarified the issue of data security in the upstream oil and gas industry should be first considered [35]. The big data of oil and gas industry faces two very important challenges. One is the data storage, and the other is data analysis and processing. Petroleum industry data includes seismic data, drilling data, logging data, as well as production data of oil and gas wells. The data source layer is responsible for the collection of these heterogeneous data. The format and form of the data require the professional archiving and quality management. The data source layer provides data resources and application for different departments and different projects. It also provides the original data source for the upper-layer cloud database.

5.2.1. Unclear requirement

The technique of the big data mining and analysis in oil & gas industry are still stay in the experimental stage. Only a handful companies have adopted big data in the oil field [36]. This is mainly due to lack of understanding of big data technology. There are usually two views about big data. On the one hand, IT administrators don‘t fully understand big data techniques and working methods are still dependent on traditional technology. On the other hand, with the arrival of big data era, all the previous business can use big data and intelligent methods together. Jill Feblowitz articulated that most of the data are still unused and cannot be effective sharing[16].Although there has been greater attention in recent years to establish standards and data models, such as PPDM, SEG-Y, WITSML, and ProDML, etc.

5.2.2. Data sharing and security

Under the traditional information industry development mode, we need to purchase a lot of hardware and licenses for the corresponding software. We need more professional information technology maintainers. Due to the limitation of China's traditional management mode and application system development model, data and application systems are usually closely- coupled modes. The degree of data sharing is not high, and the information island question is acute. Marcos D.Assuncao discussed current research achievement of data management [37]. There are still many open challenges in data management. How to handle an increasing volume of data? Especially when the data are unstructured, how should we quickly mine data value? How to efficiently recognize and store important information from unstructured data?

Data security is a vital part for petroleum companies. The protection of data is the strategic resource to protect the survival and development of the enterprise. Data quality standards are inconsistent. The authenticity and availability of data are not guaranteed. It causes great difficulties during data acquisition.

6. Conclusion

Our case study has revealed that the development process of China's Cloud computing and big data in oil and gas industry. Opportunities and challenges of Cloud computing and big data for oil and gas industry applications are described. Parastor, distributed parallel storage architecture, can provide high quality cloud storage service. It could meet the demand of massive data storage for the growing petroleum exploration system, and solve the bottleneck of massive data storage. The cloud computing of oil and gas industry is mainly including research cloud and desktop cloud. Desktop cloud applications mainly refer to the office desktop and monitor desktop of oil and gas storage and transportation. The big data research in the upstream oil and gas industry are mainly on seismic data processing, intelligent decision-making support system, drilling based multiple conditions recognition and production data processing. In the midstream oil and gas industry, big data mainly focuses on oil and gas storage and transportation. In the downstream oil and gas industry, the application of big data mainly focuses on the gas station management, sales activity analysis, and customer behavior and preference. Cloud computing and big data technology can provide convenient information sharing and high quality service for oil and gas industry.

Abbreviations

NIST：National institute of standard and Technology; SAAS: software as a service; PAAS: Platform as a service; IAAS: infrastructure as a service; GFS: google file system; HDFS: hadoop file system; OLTP: on-line transaction process; OLAP: on-line analytical process; NAS: network attached storage; SAN: storage area network; FDR: fourteen data rate; EDR: enhance data rate; BGP: bureau of geophysical prospecting INC.,china national petroleum corporation; VDI: virtual desktop infrastructure; GPU: graphic processing unit; SCADA: supervisory control and data acquisition; PPDM: profession petroleum data management; SEG: society of exploration geophysicists; WITSML: well information transmission standard mark language

Acknowledgment

The authors thanks Zhang yulong for his initial contribution.

References

[1] Armbrust, M. (2009). Above the clouds, A berkeley view of cloud computing. *Science, 53(4),* 50-58.

[2] Gordij., & Mykola. (2013). Use of cloud computing in oil and gas industry. *Geomatics and Environmental Engineering, 7(2),* 35-41.

[3] Buyya, R., Pandey, S., & Vecchiola, C. (2009). Cloudbus toolkit for market-oriented cloud computing. *Computer Science, 5931,* 24-44.

[4] Garfinkel, S. L. (2011). Cloud computing defined business impact report series. *Technology Review Magazine.*

[5] Vaquero, L. M., Rodero-Merino, L., & Lindner, M. (2008). A break in the clouds, towards a cloud definition. *Acm Sigcomm Computer Communication Review, 39(1),* 50-55.

[6] Zheng, L., Chen, S., & Hu, Y. (2011). Applications of cloud computing in the smart grid. *Proceedings of International Conference on Artificial Intelligence, Management Science and Electronic Commerce* (pp. 203-206).

[7] Lenk, A., Klems, M., & Nimis, J. (2009). What's inside the Cloud? An architectural map of the Cloud landscape. *IEEE Computer Society,* 23-31.

[8] Xu, W. P., & Zhao, H. (2013). Research of cloud computing information management mode for oil enterprise. *Applied Mechanics & Materials,* 336-338.

[9] Janssen, M., & Joha, A. (2011). Challenges for adopting cloud-based software as a service (SAAS) in the public sector. *European Conference on Information Systems.*

[10] Sotomayor, B., Montero, R. S., & Liorente, I. M. (2009). Virtual infrastructure management in private and hybrid clouds. *IEEE Internet Computing, 13(5),* 14-22.

[11] Owens, D. (2010). Securing elasticity in the cloud. *ACM, 53(53),* 46-51.

[12] Abokhodair, N., Taylor, H., & Hasegawa, J. (2016). *Heading for the Clouds, Implications for Cloud Computing Adopters,* 1-9.

[13] Perrons, R. K., & Hems, A. (2013). Cloud computing in the upstream oil & gas industry, a proposed way forward. *Energy Policy, 56,* 732-737.

[14] Geczy, P., Izumi, N., & Hasida, K. (2012). Cloudsourcing, managing cloud adoption. *Global Journal of Business Research, 6(2)*, 57-70.

[15] Hofmann, P., & Dan, W. (2010). Cloud computing, the limits of public clouds for business applications. *IEEE Internet Computing, 14(6),* 90-93.

[16] Feblowitz, J. (2011). Oil and gas, into the cloud? *Journal of Petroleum Technology, 63(5),* 32-33.

[17] Hems, A., Soofi, A., & Perez, E. (2013). Drilling for new business value, how innovative oil and gas companies are using big data to outmaneuver the competition. *A Microsoft White Paper.*

[18] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics ,from big data to big impact. *Society for Information Management and the Management Information System Research Center, 36(4),* 1165-1188.

[19] Cameron, D. (2014). Big data in exploration and production, silicon snake-oil, magic bullet, or useful tool. *Society of Petroleum Engineers.*

[20] Johnston, J., & Guichard, A. (2015). New findings in drilling and wells using big data analytics. *Proceedings of Offshore Technology Conference.*

[21] Baaziz, A., & Quoniam, L. (2013). How to use big data technologies to optimize operations in Upstream petroleum industry. *International Journal of Innovation, 1(1).*

[22] What Is Apache Hadoop? Apache™ Hadoop® Website, last update. Retrieved from http,//hadoop.apache.org/#What Is+Apache+Hadoop%3F

[23] Vurukonda, N., & Rao, B. T. (2016). A study on data storage security issues in cloud computing. *Procedia Computer Science, (92),* 128-135.

[24] More, S., & Chaudhari, S. (2016). Third party public auditing scheme for cloud storage. *Procedia Computer Science, 79,* 69-76.

[25] Li, Y., Gai, K., & Qiu, L. [2016]. Intelligent cryptography approach for secure distributed big data storage

in cloud computing. *Information Sciences, 387(C),* 103-115.

[26] Wang, Y., & Li, S. (2006). Research and performance evaluation of data replication technology in distributed storage systems. *Computers & Mathematics with Application, 51(11),* 1625-1632.

[27] Wu, J. Y., Fu, J. Q., & Ping, L. D. (2011). Study on the P2P cloud storage system. *Acta Electronica Sinica, 39(5),* 1100-1107.

[28] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information systems, 47, 98-115.

[29] Donghui, L. I., Shen, L., & Fang, F. High availability framework of distribution cloud storage. *Computer & Digital Engineering, 42(1),* 76-80.

[30] Liao, X., & Jin, H. (2005). A new distributed storage scheme for cluster video server. *Journal of Systems Architecture, (51),* 79-94.

[31] Tadapaneni, N. R. (2018). Cloud Computing: Opportunities And Challenges. International Journal of Technical Research and Applications. SSRN Electronic Journal.10.2139/ssrn.3563342

[32] Chen, H., Zhai, Z., & Gao, S. (2016). Design of ground cloud storage structure for the tracking and relay satellite system. *Journal of Spacecraft TT &C Technology, 35(5),* 392-399.

[33] Perrons, R. K. (2010). Perdidoties together shell digital oil field technologies. *World Oil , 231(5),* 43-49.

[34] Beckwith, R. (2011). Managing big data, cloud computing and co-location centers. *Journal of Petroleum Technology, 63(10),* 42-45.

[35] Brule, M. (2013). Tapping the power of big data for the oil and gas industry. *IBM Software White Paper for Petroleum Industry.*

[36] Baaziz, A., & Quoniam, L. (2013). How to use big data technologies to optimize operations in upstream petroleum industry. *Social Science Electronic Publishing, 1(1),* 19-25.

[37] Yuan, H., Mahdavi, M., & Paul, D. (2011). Security, digital oil field or digital nightmare? *Journal of Petroleum Technology, 63(8),* 16-18.

[38] Nicholson, R. (2012). Big data in the oil & gas industry. *IDC Energy Insights.*

[39] Assuncao, M. D., Calheiros, R. N., & Bianchi, S. (2015). Big data computing and clouds, trends and future directions. *Journal of Parallel & Distributed Computing, 79-80,* 3-15.

Yang Zhifeng was born in China in 1987. In 2012, he had graduated from China University of Petroleum (Hua Dong), obtained the bachelor and master degree in geology. In 2016, he had graduated from China University of petroleum (Bei Jing), obtained the Ph.D degree in reservoir exploration and production. At present, he is working in the postdoctoral workstation of Xinjiang Oilfield Company, PetroChina. Main research directions: cloud computing, big data management, data aggregation and information development of petroleum industry.

Han fei was born in China in 1985. In 2017, she had graduated from China University of Petroleum (Bei jing), obtained the Ph.D degree in computer science. At present, she is working in Lenovo (Beijing) Co., Ltd., Beijing. Main research directions: high performance computing, big data and artifical Intelligence.