PROCEEDINGS OF SCIENCE

Research based on Big Data and Cloud Computing

Xiaoru Chen
South China Institute of Software Engineering of Guang Zhou University
Guangzhou, 510990, China
E-mail: 479170369@qq.com

Lijun Chen
South China Institute of Software Engineering of Guang Zhou University
Guangzhou, 510990, China
E-mail: 372158286@qq.com

The term big data arose along with the explosive increase of global data as a technology that is able to store and process big and varied volumes of data, providing both enterprise and science with deep insights over its clients/experiments. Cloud computing offers a reliable, fault-tolerant, available and scalable environment to harbor big data-distributed management systems.With regard to cloud computing and big data-related areas, researches have been conducted on how to use cloud computing to process digital information.. This paper focuses on the application of cloud computing and big data to build the relevant channels, of the adaption to new requirements, the relevant laws to solve the data processing and technical scheme of space data and geographic science, its competency to deal with the challenges of the current situation, and innovative research in the future. Big data processing speeds up the diversity of data transformation, data value, and evaluation of the significance of digital research and the value added by application of big data and cloud.

ISCC2017
16-17 December 2017
Guangzhou, China

1. Introduction

In recent years, there has been an increasing demand to store and process more and more data, in domains such as finance, science, and government. Systems that support big data and host them using cloud computing, have been developed and used successfully. Whereas, big data is responsible for storing and processing data, cloud provides a reliable, fault-tolerant, available and scalable environment to ensure proper performance of big data systems.

Both business and scientific areas view big data, and in particular big data analytics, as a way to correlate data, find patterns and predict new trends. Therefore, there is a huge interest in leveraging these technologies, as they can provide businesses with a competitive advantage, and science with ways to aggregate and summarize data from experiments such as those performed at the Large Hadron Collider(LHC). To be able to fulfil the current requirements, big data systems must be available, fault tolerant, scalable and elastic.

In the progress of global integration of digital current, stress has been put on comprehensive investigation on the big data related challenges , technical challenges of cloud computing and big data, the current technical status of the two, how to solve the problem during data integration, the current and the future research. The corresponding situation Figure 1 shows.

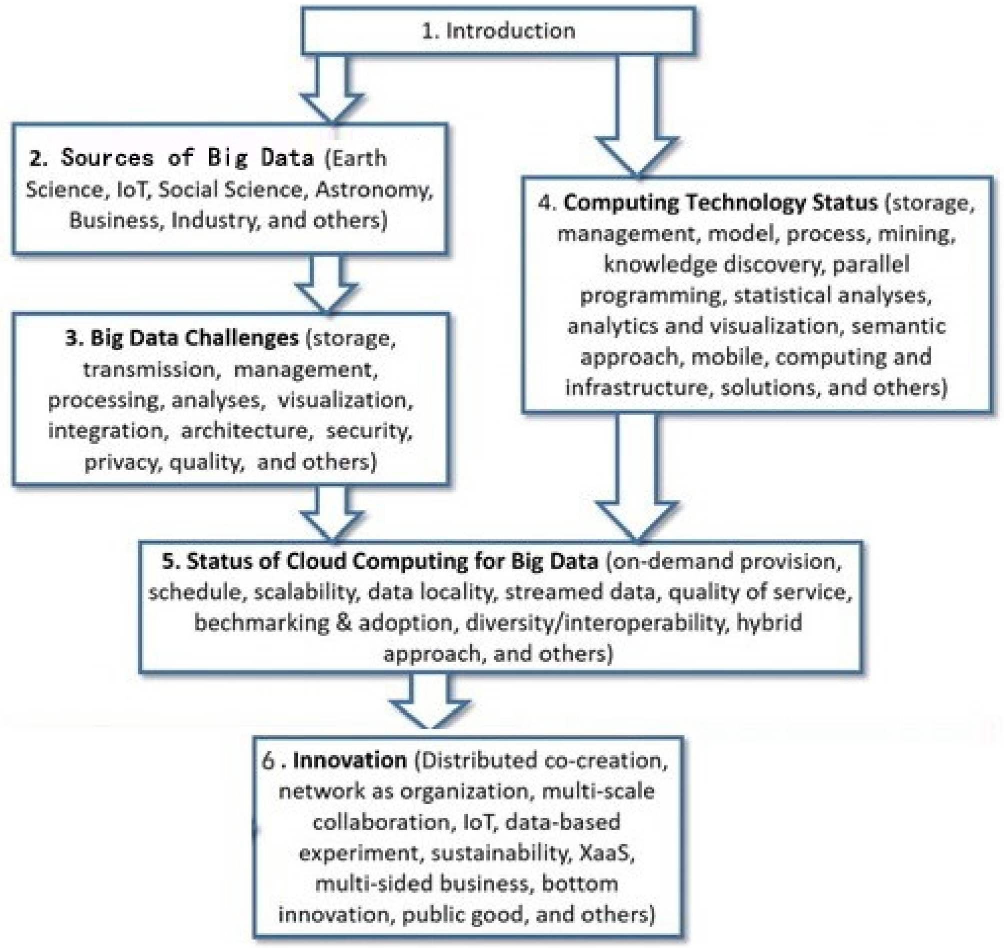
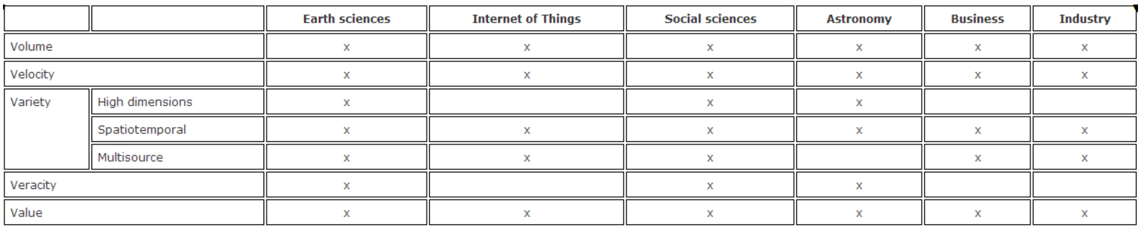

1. Introduction
2. Sources of Big Data (Earth Science, IoT, Social Science, Astronomy, Business, Industry, and others)
3. Big Data Challenges (storage, transmission, management, processing, analyses, visualization, integration, architecture, security, privacy, quality, and others)
4. Computing Technology Status (storage, management, model, process, mining, knowledge discovery, parallel programming, statistical analyses, analytics and visualization, semantic approach, mobile, computing and infrastructure, solutions, and others)
5. Status of Cloud Computing for Big Data (on-demand provision, schedule, scalability, data locality, streamed data, quality of service, bechmarking & adoption, diversity/interoperability, hybrid approach, and others)
6 . Innovation (Distributed co-creation, network as organization, multi-scale collaboration, IoT, data-based experiment, sustainability, XaaS, multi-sided business, bottom innovation, public good, and others)

Figure 1: Tackling Big Data challenges with cloud computing for innovation.

2. Sources of Big Data

Data exists everywhere, no matter it is factory machinery data, transaction data or people related, valuable information can be extracted through the analysis of data processing, . At the same time, some social websites have hundreds of millions of twitter messages each day, regardless of text, pictures or video. The amount of data in one day now exceeds the amount of data in a whole last year. The amount of data is shown in Table 1.



		Earth sciences	Internet of Things	Social sciences	Astronomy	Business	Industry
Volume		x	x	x	x	x	x
Velocity		x	x	x	x	x	x
Variety	High dimensions	x		x	x		
	Spatiotemporal	x	x	x	x	x	x
	Multisource	x	x	x		x	x
Veracity		x		x	x		
Value		x	x	x	x	x	x

Table 1: Big Data in Different Domains as Denoted with an 'x'.

The machine specific data-feedback information, equipment operation at the same time, also includes information related to sensor data, log file, the laboratory test data. The commodity trading process produces huge data information, including the products itself, product packaging, product price, deposit and final payment information, to name a few. Every second, every day in a huge amount of transaction information, big data either in traditional or structured storage at present advocate non-structured storage; the information contains the value of RMB2. There are also structured data with unstructured data, even more structured data storage. The classic database could or the new data model, can take different data formats and datatypes.

Interactive social network or network application can set up their own log information in the form of pure text type, as well as pictures or video clips, with the purpose of enhancing network performance regarding information specific operation experience and customer relationship.

For every business of various size, they need to know the benefits by big data, how to deal with the increasing amount of big data and information content in order to enhance the research work related to big data, and promote big data to serve themselves and the enterprises.

2.1 Instrument Information

The information of the machine instrument is conveyed in real time. It is conveyed by means of receiving and firing the relevant sensors of the machine, with some of the sensors processed automatically or manually.

2.2 The Interaction of Large Data

The big social data generated between human, human and machine, machine and machine, machine and related function produces a large amount of data, which can adopt quantitative methods for extraction, or qualitative observation or observation space in advance. In addition, it does not matter which channel is used for the corresponding analysis nor the geographical position since the key is the data itself.

2.3 Commercial Information Data

The current business information data can be stored in a relational database, the original table can also be archived in the unstructured database, huge data information can be a complete record of business information exchanges of a unit, for different data content and data information is stored in both tables, text, pictures and short video, etc.. Different document formats and different document forms can be parallel processing or serial processing. Get the information related to the content.

2.4 Network File

Network storage is fulfilled in different document forms. Either static or dynamic page is stored Rebecca, and the storage can be in the form of audio, video, documents. More interests are paid in the network file, the corresponding data mining, the search technology of different data analysis and the pattern recognition. The information content value is driven from numerous data in advance.

2.5 Information Delivery

Broadcast information through audio mode can perform video communication, by means of powerful computing capability and broadband network. For either digital or analog information, data processing and concurrent processes related to scheduling, the final form of solution would deliver smooth performance information.

3. Big Data Technology Challenges

In the utilization of big data, many practical applications fail to make full use of the big data technology.. For instance, about 50% of the projects cannot be completed in time. Then what are the technical challenges or problems of big data at present?

(1) Difficulty in Mastering Hadoop

In spite of the advantages of Hadoop software, the specific use and management is relatively complicated, which leads to the necessity of learning a lot of knowledge and data of the operating system in order to master the technology. However, it gets better when dealing with a large number of internal data resources, whether relying on technology or not. While a sharing platform of big data is built, it becomes a very important content.

(2) Extensibility

For large data, the key is to be able to handle the demand on zoom. Many organizations do not have to consider the speed of the development and evolution of large data project. To certain extent, the suspension of the project brings in additional resources and reduces the time of data analysis. Big data workload tends burst, so it is difficult to predict the resources allocation. The degree of challenges for data solutions vary. Cloud solution turns to be easier and faster than the internalsolution.

nBusinesses start realizing the talent shortage. Not only the data scientists have limited access to data, but also successful achievement of a big data project requires a complex development team, data scientists and analysts who have abundant knowledge to identify valuable insights. Many big data suppliers are trying to provide most of the management with their own educational resources to overcome the challenges of big data.

(3)Convenience in operation

The design of big data means the need for a very clear business objectives and collecting the corresponding data source and channels. In addition, through the key mode of determining the business value of specific data including useless data, tedious data without any meaning.

Meanwhile, it requires a lot of manpower and resources for maintenance, such as user input error messages, duplicate information, or even incorrect input. To optimize the maintenance cost and information content data need. Maintenance through the optimization algorithm.

(4)Security of large data

The insurance of the safety of very large, massive data is very essential for the registration and authentication of information from the user. The legitimacy of registration leads to verification and certification on the user, to set different access pages according to different user permissions, accounting for various stages of the information cos. Training, maintenance and extension all needs the corresponding information in the cloud, to formulate the corresponding service agreement, charging mode and other information.

4. Cloud Computing

The structure of cloud computing system is composed of following parts: user interface, user cloud service directory, system management, service deployment tools, monitoring and measurement of the cloud configuration and system management, service registration, user identification, billing request regarding user permission, load balancing, resource management in fault detection, fault recovery and monitoring statistics, calculating the resource pool in the bottom level of resource virtualization, resource pool, and finally expanding the village at the bottom of the system, completing the corresponding infrastructure deployment. The three-layer architecture of the cloud is shown in Figure 2.

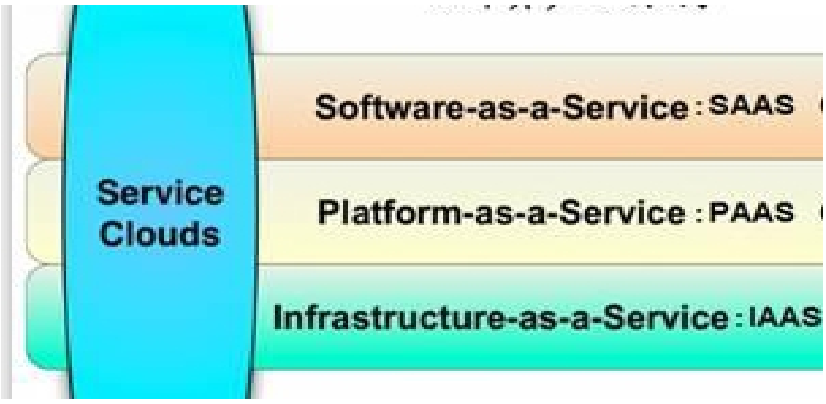

Software-as-a-Service：SAAS
Service Clouds
Platform-as-a-Service：PAAS
Infrastructure-as-a-Service：IAAS

Figure 2: The three-layer architecture of the cloud

5.Cloud Computing for Large Data Application

the challenges on the security of cloud computing environments fall under several levels: the network level, which includes dealing with network protocols and network security such as distributed nodes, distributed data, and communications between the nodes; authentication level where the user handles encryption/decryption techniques, authentication methods such as contract administrative rights, authentication of applications and nodes, and logging entry; the data level concerning data integrity and availability as well as data protection and data distribution. Cloud computing follows the policy of resource sharing, while the privacy of data is very important due to the challenges like integrity, access authorization, and availability (of backup / replication). Data integrity ensures that data is not corrupted or tampered with during communication. Authorized access prevents data from infiltration attacks while backups and replicas allow efficient access to data even in case of technical error in the cloud.

Big data is now encountering several challenges on data sets, processing and management. When dealing with huge amounts of data, there are the challenges such as volume, variety, velocity and verification, which are also known as 5V of big data [74]. In addition, in the field of computer networks, the cost of communications is a major concern compared with the cost of processing the same data, with the challenge on minimizing the communication cost while meeting the storage requirements and getting additional data from the general cloud [75]. The factors and challenges that affect the processing of big data in a timely manner are the bandwidth and latency. The challenges can be summarized based on the relationship between big data and cloud computing.

(1)Data storage technology: cloud computing is fully related to the data storage space to increase the application of big data into the cloud server, and storage space requirements can be met by the improving cloud storage technology, then effectively cope with the traditional server errors of hard disk reading and writing, hard disk damage, low competency of error correction. . The high quantity of data is in need of storage improvement..

(2) Variety of data: big data naturally increases and varies, which is the result of the growth of almost unlimited data sources. The growth leads to the heterogeneous nature of big data. Data from multiple sources of different types and representations are highly interrelated. They have incompatible shapes and are inconsistent. A user can store data in structured, semi-structured, or unstructured formats. Structured data format is suitable for today's database systems, while semi-structured data formats are only suitable. Unstructured data is inappropriate because it contains a complex format that is difficult to represent in rows and columns.

(3) Data transfer: The data goes through several stages-data collection, input, processing, and output. Big data transfer is a challenge, so data compression techniques have to be lessened to lower the volume, where data volume is a hindrance for speed transfer. It also affects the cost, while cloud computing provides distributed storage resources and data transfer on high-speed lines, which is to reduce costs through virtual resources and resource use at user's request.

(4) Privacy and data ownership: The cloud is an open environment and the user's role in monitoring is limited. Privacy and security are quite a challenge for big data. Big data and cloud computing come together in practice. According to (IDC) estimates , around 40% of global data can be accessed via cloud computing by 2020. Cloud computing provides strong storage, calculation and distribution capability to support big data processing. As such, there is a strong demand to investigate the privacy of information and challenges on security in both cloud computing and big data.

5.1 What is Big Data's Relationship to the Cloud?

The correspondence between big data and cloud computing environment reflects the mutual relationship. This is done through the cloud computing features to handle big data, the resources provided by cloud computing, the resource service provided for many users where the various physical and virtual resources are automatically set and reset upon request. Cloud computing has access from anywhere to data resources that are spread all over the world by using a (public) cloud to allow those sources faster access for storage. The nature of big data is generated by technologies and locations worldwide, so the cloud resource service provides and helps in the collection and storage of large amounts of data resulting from the use of technologies.

The cloud computing structure can expand the solid equipment to accommodate small or big data volumes. The cloud can be expanded to handle big amounts of data by dividing the data into parts, which is automatically done in IAAS. Expanding the environment is the requirement per big data. Cloud computing has the advantage of helping to reduce costs by paying for the value of the resources used, which helps to develop big data. Flexibility is also regarded as a requirement for big data. When more storage for data is needed, or a large number of virtual machines is expected to be handled in a single time, the cloud platform can dynamically expand to meet proper storage needs. . For error tolerance, the cloud helps to deal with big data in the extraction and storage process. Error tolerance helps SLAs, as well as QOS levels. Service level agreements specify different rules to regulate availability of cloud service.

Big companies such as Yahoo, Google, and Facebook offer web-based services, and the amount of data they routinely collect through online user interactions has overwhelmingly traditional IT capabilities. Therefore, the development of infrastructure components has to be conducted. Apache Hadoop has been introduced as a benchmark for managing big amounts of unstructured data. Apache Hadoop is an open platform- distributed software for storing and processing data. By using Hadoop, large amounts (pet bytes) of servers can be stored, while effectively scaling performance in terms of cost. MapReduce is based on the distribution of a data set among multiple servers, and partial results are then reassembled.

Big data is characterized by its diversity. ETL technology, which deals with data diversity, represents several functions such as extraction, conversion, and loading. These functions are integrated into one tool to pull data from one database and place it in another one. It helps to convert databases from one to another.

Big data relies on the integrity to be effective. If big data is stored at the local level, it will take a huge amount of work to manually merge all data to manage. This can be realized by the cloud, enabling one site to store and manage all commercial data. In this way, one source of the truth can be generated, without exhausting time and resources to manually merge the data.

Cloud computing offers features and benefits to big data through ease of use, access to resources, low cost in resource utilization on supply and demand, and reduces the involvement of solid equipment used to handle big data. Both big data and the cloud aim to increase the value of a company while reducing investment costs. The cloud reduces the cost of managing local software, while big data reduces investment costs by encouraging more prudent business decisions. It seems only natural that these two concepts together provide greater value to companies.

Any system in technology must pass through several main stages. The computer system follows the input, processing and output model. Input is done through devices and then processed through the CPU. Thus, the results of the information are delivered. In the relationship between the data and cloud computing, data is stored in external and remote storage units. On the other hand, , the data in the computer system is stored internally or locally. Therefore, the relationship between the data and cloud computing represents the input, processing and output model as in Figure 2. The big data is accessed through devices such as the mouse, cellular devices and other smart devices. Processing is carried out through the tools and techniques used by the cloud computing in providing service, and the outputs are the results, which represents the value of data after processing.

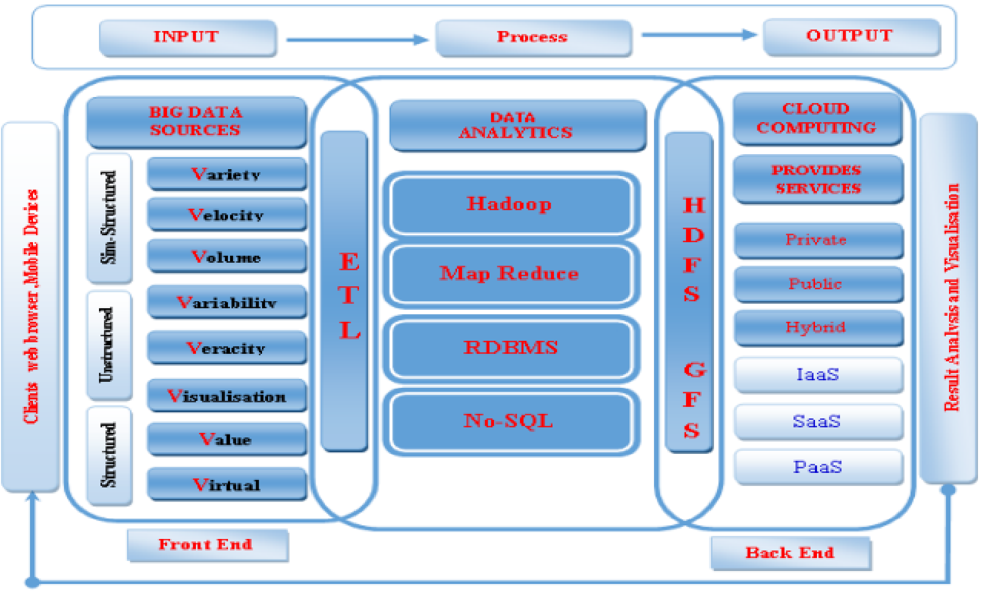


Figure 3:A Model Showing the Relationship between Big Data and Cloud Computing

The input and output model defines input, output and processing tasks required to convert input to output. Input represents the flow of data and raw materials. The processing step includes all tasks required to transform inputs. The output is data flowing from the transformation process.

5.2 Common Factor between Cloud Computing and Big Data

The Internet of Things (IoT) represents the new concept of the internet , which enables communication among several parties. These parties include smart devices, mobile devices, sensors and other, which are considered as an effective communication among all elements of architecture so that it can rapidly deploy applications and process, as well as to analyze data to make decisions as quickly as possible. The architecture represents several systems: objects, gates, network infrastructure, cloud infrastructure. Internet objects can benefit from the scalability and performance of cloud computing infrastructure. In fact, Internet applications produce large amounts of data and consist of multiple computer components upon request.

The Internet of Things (IoT) is going to generate a massive amount of data and this in turn puts a huge strain on Internet Infrastructure (IT). As a result, this forces companies to find solutions to minimize the pressure and solve their problems of transferring large amounts of data. However, cloud computing has played a major role in IT, by migrating its data operations to the cloud. Many cloud providers allow data to be transmitted either over traditional internet connection or via a dedicated link. The real purpose of cloud computing and IoT increases efficiency in daily tasks and has a complementary relationship. The Internet of things generates huge amounts of data, and cloud computing provides a pathway for these data to navigate. By storing data in the cloud, most companies find it possible to access large amounts of big data through the cloud. Moreover, internet of things is all parts of a continuum. It is difficult to consider internet things without thinking about the cloud and the cloud without analyzing the big data. The data generated is stored in the cloud computing, which is the only technology suitable for filtering, analysis, storage and access to IoT and other data in effective ways, as these data constitute large quantities must be analyzed. As a conclusion, the object is a common factor between the erased cloud and big data.

5.3 Common Points between Big Data and the Cloud

The cloud computing environment consists of several user terminals and service provider. The big data comes from both sides-the user collects the data and the big data is produced in the process of dealing with the technology tools. The role of the service provider is to save, store and process the big data upon the user's request, so cloud computing represents the big data infrastructure. The service provider must ensure that the users have on-demand resources or access to their data, systems and applications on a regular basis, with availability throughout the service without any interruption.

(1) Data, whether small or big, require storage, processing and security, but the volume and capacity of data requirements differ in accordance with the volume of the data. Therefore, the cloud computing must provide storage, processing and security for big data in its environment. The cloud environment is scalable and uses sophisticated high-end data management techniques and security policies as the service provider protects and manages data.

(2) Cloud computing provides security, independent on data volume but availability of security and protection for small and big data. The service provider guarantees complete confidentiality of user data of all kinds and only allows access to authorized users. Therefore, identity management and access control must be provided for information and service resources, according to user needs. Users can connect to the network in these resources through a simple software interface that simplifies and ignores many internal details and processes.

(3) Cloud computing saves the cost of storing and processing data to the user through the availability of geographically dispersed servers and the availability of virtual server technology. The service provider must ensure that the devices and equipment are sufficiently available and restricted by an integrated and documented entry system for reference when needed. Cloud computing offers high-level applications and software, regardless of the efficiency of the devices being used, as it depends on the strength of the network servers but not on the personal resources of specific device. The user can benefit from the cloud service, regardless of the efficiency.

(4) Cloud computing is considered as a distributed system over a geographical distance. One of the examples is the general cloud, where resources are distributed widely. This makes it easier for the user to speed up access to the data. Thus, cloud computing is based on solving the problem of geographical divergence between devices and resources.

(5) Cloud computing is characterized by its continuity, i.e. the ability to withstand failure by providing resources even in the absence of defect in the components. The nature of the cloud is that it is geographically distributed, so there is a high probability of errors. These events increase the demand for failure tolerance techniques to achieve reliability.

All these points represent the relationship between big data and cloud computing, as it shows the important requirements for the continuous increase in the growth of big data and provides the appropriate environment to deal with big data.

6. Conclusion

there has been the conclusion that the relationship between them is complementary. Big data and cloud computing constitute an integrated model in the world of distributed network technology. The development of big data and their requirements is a factor that motivates service providers in the cloud for continuous development, as the relationship between them is built on the product, the storage and processing as a common factor. Big data represents the product and the cloud represents the container. The big data is concerned with the capacities of cloud computing. On the other hand, cloud computing is interested in the type and source of big data. Depending on the relationship between them, cloud computing represents an environment of flexibly distributed resources that uses advanced techniques in the processing and management of data and yet reduces the cost. All these characteristics show that cloud computing has an integrated relationship with big data. Big data and cloud computing are moving towards rapid progress to keep pace with progress in technology requirements and users.

References

[1]*Correlated network data publication via differential privacy*[J] . Rui Chen,Benjamin C. M. Fung,Philip S. Yu,Bipin C. Desai. The VLDB Journal . 2014 (4)

[2]*Competition of wireless providers for atomic users*[J] . Vojislav Gaji?,Jianwei Huang,Bixio Rimoldi. IEEE/ACM Transactions on Networking (TON) . 2014 (2)

[3]*Game theory meets network security and privacy*[J] . Mohammad Hossein Manshaei,Quanyan Zhu,Tansu Alpcan,Tamer Bac?ar,Jean-Pierre Hubaux. ACM Computing Surveys (CSUR) . 2013 (3)

[4]*Privacy vulnerability of published anonymous mobility trace*s[J] . Chris Y. T. Ma,David K. Y. Yau,Nung Kwan Yip,Nageswara S. V. Rao. IEEE/ACM Transactions on Networking (TON) . 2013 (3)

[5]*User k -anonymity for privacy preserving data mining of query logs*[J] . Information Processing and Management . 2011 (3)

[6]*How bad are selfish investments in network security?*[J] . Libin Jiang,Venkat Anantharam,Jean Walrand. IEEE/ACM Transactions on Networking (TON) . 2011 (2)

[7]*Output privacy in data mining*[J] . Ting Wang,Ling Liu. ACM Transactions on Database Systems (TODS) . 2011 (1)

[8]*Privacy-aware location data publishing*[J] . Haibo Hu,Jianliang Xu,Sai Tung On,Jing Du,Joseph Kee-Yin Ng. ACM Transactions on Database Systems (TODS) . 2010 (3)

[9]*Privacy-preserving data publishing*[J] . Benjamin C. M. Fung,Ke Wang,Rui Chen,Philip S. Yu. ACM Computing Surveys (CSUR) . 2010 (4)

[10]*Privacy-preserving data mining: A feature set partitioning approach*[J] . Nissim Matatov,Lior Rokach,Oded Maimon. Information Sciences . 2010 (14)

[11]*TEE: A virtual DRTM based execution environment for secure cloud-end computing*[J] . Weiqi Dai,Hai Jin,Deqing Zou,Shouhuai Xu,Weide Zheng,Lei Shi,Laurence Tianruo Yang. Future Generation Computer Systems . 2014

[12]*Trust mechanisms for cloud computing*[J] . Jingwei Huang,David M Nicol. Journal of Cloud Computing . 2013 (1)

[13]*Cloud services certification*[J] . Ali Sunyaev,Stephan Schneider. Communications of the ACM . 2013 (2)

[14]*Effective Ways of Secure, Private and Trusted Cloud Computing*[J] . Kumar, Pardeep,Sehgal, Vivek Kumar,Chauhan, Durg Singh,Gupta, P K,Diwakar, Manoj. International Journal of Computer Science Issues (IJCSI) . 2011 (3)

[14]D*esign and verification of a lightweight reliable virtual machine monitor for a many-core architecture*[J] . Yuehua Dai,Yi Shi,Yong Qi,Jianbao Ren,Peijian Wang. Frontiers of Computer Science . 2013 (1)

[15]*Establishing Trust in Cloud Computing*[J] . Khan, Khaled M,Malluhi, Qutaibah. IT Professional Magazine . 2010 (5)