Proceedings of
National Conference on Innovative Trends in Computer Science Engineering (ITCSE-2015)
held at BRCMCET , Bahal on 4th April 2015

Role of Cloud Computing in Big Data

[1]Dr. Dinesh Kumar , [2]Dr. Rohit Bajaj
[1]Guru Kashi University , Punjab
[2]Chandigarh Engineering College , Punjab
[1]kdinesh.gku@gmail.com , [2]cecm.cse.rohitbajaj@gmail.com

Abstract- In recent years, advances in Web technology and the proliferation of sensors and mobile devices connected to the Internet have resulted in the generation of immense data sets available on the Web that need to be processed and stored. Cloud computing has emerged as a paradigm that promises to meet these requirements. Cloud computing is an extremely successful paradigm of service oriented computing, and has revolutionized the way computing infrastructure is abstracted and used. Cloud computing is associated with service provisioning and these services are based on processing and analysis of huge volume of data. Big data management in cloud computing environments is hereby proposed in recent years. Big data management in cloud computing environments needs information interoperations in a right way. At this point, explicit knowledge usage can play a critical role in cloud computing. It is very critical to deal with the worth information for effective problem solving and decision making. It is especially true when a variety of data types and users' requirements as well as large volumes of data are available. We are entering into a "big data" era. The breakthrough of big data technologies will not only resolve the aforementioned problems, but also promote the wide application of Cloud computing and the "Internet of Things" technologies. In this paper, we focus on discussing the development and pivotal technologies of big data, providing a comprehensive description of big data from several perspectives, including the development of big data, the current data-burst situation, the relationship between big data and Cloud computing, and big data technologies.

Keywords- Big Data; Cloud Computing; Mining; Speed

I. INTRODUCTION-THE BACKGROUND OF BIG DATA

Nowadays, information technology opens the door through which humans step into a smart society and leads to the development of modern services such as: Internet e-commerce, modern logistics, and e-finance. It also promotes the development of emerging industries, such as Telematics, Smart Grid, New Energy, Intelligent Transportation, Smart City, and High-End Equipment Manufacturing. Modern information technology is becoming the engine of the operation and development of all walks of life. But this engine is facing the huge challenge of big data [1]. Various types of business data are growing by exponential orders of magnitude. Problems such as data collection, storage, retrieval, analysis, and the application of data can no longer be solved by traditional information processing technologies. These issues have become great obstacles to the realization of a digital society, network society, and intelligent society. The New York Stock Exchange produces 1 terabyte (TB) of trading data every day; Twitter generates more than 7 TB of data every day; Facebook produces more than 10 TB of data every day; the Large Hadron Collider located at CERN produces about 15 PB of data every year. According to a study conducted by the well-known consulting firm International Data Corporation (IDC), the total global information volume of 2007 was about 165 exabytes (EB) of data. Even in 2009 when the global financial crisis happened, the global information volume reached 800 EB, which was an increase of 62% over the previous year. In the future, the data volume of the whole world will be doubled every 18 months [2]. The number will reach 35 (zettabytes) ZB in 2020, about 230 times the number in 2007, yet the written record of 5000 years of human history amounts to only 5 EB data. These statistics indicate the eras of TB, PB, and EB are all in the past; global data storage is formally entering the "Zetta era." Beginning in 2009, "big data" has become a buzzword of the Internet information technology industry. Most applications of big data in the beginning were in the Internet industry: the data on the Internet is increasing by 50% per year, doubling every 2 years. Most global Internet companies are aware of the advent of the "big Data" era and the great significance of data. In May 2011, McKinsey Global Institute published a report titled "Big data: The next frontier for innovation, competition, and productivity" [3], and since the report was released, "big data" has become a hot topic within the computer industry. According to the big data report released by Wikibon in 2011 [5], the big data market is on the eve of a growth spurt: the global market value of big data will reach $50 billion in the next five years. At the beginning of 2012, the total income of large data related software, hardware, and services were around $5 billion. As companies gradually realize that big data and its related analysis will form a new differentiation and competitive advantage and will improve operational efficiency, big data related technologies and services will see considerable development, and big data will gradually touch the ground and big data market will maintain a 58% compound annual growth rate over the next five years. Greg McDowell, an analyst with JMP Securities, said that the market of big data tools is expected to grow from $9 billion to $86 billion in 10 years. By 2020, investment in big data tools will account for 11% of overall corporate IT spending [12]. At present the industry does not have a unified definition of big data; big data has been defined in differing ways as follows by various parties

II. BIG DATA CHARACTERISTICS

Big data has four main characteristics: Volume, Velocity, Variety, and Value. (Referred to as "4V," referencing the huge amount of data volume, fast processing speed, various data types, and low-value density). Following are brief descriptions for each of these characteristics [11].

A. Volume

It refers to the large amount of data involved with big data. The scale of datasets keeps increasing from gigabytes (GB) to TB, then to the petabyte (PB) level; some even are

measured with exabytes (EB) and zettabytes (ZB). For instance, the video surveillance cameras of a medium-sized city in China can produce tens of TB data every day.

B. Variety

It indicates that the types of big data are complex. In the past, the data types that were generated or processed were simpler, and most of the data was structured. But now, with the emerging of new channels and technologies, such as social networking, the Internet of Things, mobile computing, and online advertising, much semi-structured or unstructured data is produced, in the form of text, XML, emails, blogs, and instant messages—as just a few examples—resulting in a surge of new data types. With the explosive growth of sensors, smart devices, and social collaborative technologies, the types of data are uncountable, including text, microblogs, sensor data, audio, video, click streams, log files, and so on.

C. Velocity

The velocity of data generation, processing, and analysis continues to accelerate. There are three reasons: the real-time nature of data creation, the demands from combining streaming data with business processes, and decision making processes. The velocity of data processing needs to be high, and processing capacity shifts from batch processing to stream processing. There is a "one-second rule" in the industry referring to a standard for the processing of big data, which shows the capability of big data processing and the essential difference between it and traditional data mining.

D. Value

Because of the enlarging scale, big data's value density per unit of data is constantly reducing; however, the overall value of the data is increasing. Big data is even compared to gold and oil, indicating big data contains unlimited commercial value. According to a prediction from IDC research reports, the big data technology and services market will rise from $3.2 billion in 2010 to $16.9 billion in 2015, will achieve an annual growth rate of 40%, and will be seven times the growth rate of the entire IT and communication industry [9]. By processing big data and discovering its potential commercial value, enormous commercial profits can be made. In specific applications, big data processing technologies can provide technical and platform support for pillar industries of the nation by analyzing, processing, and mining data for enterprises; extracting important information and knowledge; and then transforming it into useful models and applying them to the processes of research, production, operations, and sales. Meanwhile, many countries are strongly advocating the development of the "smart city" in the context of urbanization and information integration, focusing on improving people's livelihoods, enhancing the competitiveness of enterprises, and promoting the sustainable development of cities.

III. BIG DATA PROBLEMS

Big data is becoming an invisible "gold mine" for the potential value it contains. With the accumulation and growth of production, operations, management, monitoring, sales, customer services, and other types of data, as well as the increase of user numbers, analyzing the correlation patterns and trends from the large amount of data makes it possible to achieve efficient management, precision marketing. This can be a key to opening this "gold mine." However, traditional IT infrastructure and methods for data management and analysis cannot adapt to the rapid growth of big data [11]. We summarize the problems of big data into seven categories in Table 3.1.

TABLE 3.1 Big Data Problems

Classification of big data problems	*Description*
Speed	Import and export problems, Statistical analysis problems, Query and retrieval problems, Real-time response problems
Types and structures	Multisource problems, Heterogeneity problems, The original system's infrastructure problems
Volume and flexibility	Linear scaling problems, Dynamic scheduling problems
Cost	Cost difference between mainframe and PC servers, Cost control of the original system's adaptation
Value mining	Data analysis and mining, Actual benefit from data mining
Security and privacy	Structured and nonstructured, Data security, Privacy
Connectivity and data sharing	Data standards and interfaces, Protocols for sharing, Access control

IV. BIG DATA TECHNOLOGIES

Big data brings not only opportunities but also challenges. Traditional data processing has been unable to meet the massive real-time demand of big data; we need the new generation of information technology to deal with the outbreak of big data [10]. Table 4.1 classifies big data technologies into five categories.

A. Infrastructure support

Mainly includes infrastructure-level data center management, Cloud computing platforms, Cloud storage equipment and technology, network technology, and resource monitoring technology. Big data processing needs the support from Cloud data centers that have large-scale physical resources and Cloud computing platforms that have efficient scheduling and management functionalities.

B. Data acquisition

Data acquisition technology is a prerequisite for data processing; first we need the means of data acquisition for collecting the information and then we can apply top-layer data processing technologies to them. Besides the various types of sensors and other hardware and software equipment, data acquisition involves the ETL (extraction, transformation, loading) processing of data, which is actually pre processing.

C. Data storage

After collection and conversion, data needs to be stored and archived. Facing the large amounts of data, distributed file storage systems and distributed databases are generally used to distribute the data to multiple storage nodes, and are also needed to provide mechanisms such as backup, security, access interfaces, and protocols.

TABLE 4.1 GROUP OF BIG DATA TECHNOLOGIES

Classification of big data technologies	*Big data technologies and tools*
Infrastructure support	Cloud Computing Platform, Cloud Storage, Virtualization Technology, Network Technology, Resource Monitoring Technology
Data acquisition	Data Bus, ETL Tools
Data storage	Distributed File System, Relational Database, NoSQL Technology, Integration of Relational Databases and Non Relational Databases, In-Memory Database
Data computing	Data Queries, Statistics, and Analysis, Data Mining and Prediction, Graph Analysis, BI (Business Intelligence)
Display and interaction	Graphics and Reports, Visualization Tools, Augmented Reality Technology

D. Data computing

Data queries, statistics, analysis, forecasting, mining, graph analysis, business intelligence (BI), and other relevant technologies are collectively referred to as data computing technologies. Data computing technologies cover all aspects of data processing and utilize the core techniques of big data technology.

E. Display and interaction

Display of data and interaction with data are also essential in big data technologies, since data will eventually be utilized by people to provide decision making support for production, operation, and planning. Choosing an appropriate, vivid, and visual display can give a better understanding of the data, as well as its connotations and associated relationships, and can also help with the interpretation and effective use of the data, to fully exploit its value. For the means of display, in addition to traditional reporting forms and graphics, modern visualization tools and humancomputer interaction mechanisms—or even Augmented Reality (AR) technology, such as Google Glasses—can be used to create a seamless interface between data and reality [12].

V. BIG DATA TECHNOLOGIES AND CLOUD COMPUTING

Cloud computing has development greatly since 2007. Cloud computing core model is large-scale distributed computing, providing computing, storage, networking, and other resources to many users in service mode, and users can use them whenever they need them. Cloud computing offers enterprises and users high scalability, high availability, and high reliability. It can improve resource utilization efficiency and can reduce the cost of business information construction, investment, and maintenance. As the public Cloud services from Amazon, Google, and Microsoft become more sophisticated and better developed, more and more companies are migrating toward the Cloud computing platform. The outbreak of big data is a thorny problem encountered in social and informatization development. Because of the growth of data traffic and data volume, data formats are now multisource and heterogeneous, and they require real-time and accurate data processing. Big data can help us discover the potential value of large amounts of data. Traditional IT architecture is incapable of handling the big data problem, as there are many bottlenecks, such as: poor scalability; poor fault tolerance; low performance; difficulty in installation, deployment, and maintenance; and so on. Because of the rapid development of the Internet of Things, the Internet, and mobile communication network technology in recent years, the frequency and speed of data transmission has greatly accelerated [13]. This gives rise to the big data problem, and the derivative development and deep recycling use of data make the big data problem even more prominent.

Cloud computing and big data are complementary, forming a dialectical relationship. Cloud computing and the Internet of Things' widespread application is people's ultimate vision, and the rapid increase in big data is a thorny problem that is encountered during development. The former is a dream of humanity's pursuit of civilization; the latter is the bottleneck to be solved in social development. Cloud computing is a trend in technology development, while big data is an inevitable phenomenon of the rapid development of a modern information society.

To solve big data problems, we need modern means and Cloud computing technologies. The breakthrough of big data technologies can not only solve the practical problems, but can also make Cloud computing and the Internet of Things' technologies land on the ground and be promoted and applied in in-depth ways.

From the development of IT technologies, we can summarize a few patterns:

- The competition between Mainframe and personal PCs ended in the PC's triumph. The battle between Apple's iOS and the Android, and the open Android platform has taken over more than 2/3 of market share in only a couple of years. Nokia's Symbian operating system is on the brink of oblivion because it is not open. All of these situations indicate that modern IT technologies need to adopt the concept of openness and crowd sourcing to achieve rapid development.
- The collision of existing conventional technologies with Cloud computing technology is similar to the aforementioned situations; the advantage of Cloud computing technology is its utilization of the crowd sourcing theory and open-source architecture. Its construction is based on a distributed architecture of open platform and novel open-source technologies, which allow it to solve problems that the

existing centralized approach is difficult to solve or cannot solve. TaoBao, Tencent, and other large Internet companies once also relied on proprietary solutions provided by big companies such as Sun, Oracle, and EMC. Then they abandoned those platforms because of the cost and adopted open-source technologies. Their products have also, in turn, ultimately contributed to the open source community, reflecting the trend in information technology development.

- The traditional industry giants are shifting toward open-source architecture; this is a historic opportunity for others to compete. Traditional industry giants and large state enterprises—such as the National Grid, telecommunications, banking, and civil aviation—rely too heavily on sophisticated proprietary solutions provided by foreign companies for historical reasons, resulting in a pattern that lacks innovation and has been hijacked by foreign products. Analyzing from the perspective of the path and the plan to solve the big data problem, we must abandon the traditional IT architecture gradually, and must begin to utilize the new generation of information technology represented by Cloud technology.

Despite the fact that advanced Cloud computing technology originated mainly in the United States, because of open-source technology, the gap between Chinese technology and the advanced technology is not large. The urgent big data problem of applying Cloud computing technologies to large-scale industry is also China's historic opportunity to achieve breakthrough innovations, defeat monopolies, and catch up with international advanced technologies.

VI. SUMMARY

Big Data is the big frontier of today's information technology development. The Internet of Things, the Internet, and the rapid development of mobile communication networks have spawned big data problems and have created problems of speed, structure, volume, cost, value, security privacy, and interoperability. Traditional IT processing methods are impotent when faced with big data problems, because of their lack of scalability and efficiency. Big Data problems need to be solved by Cloud computing technology, while big data can also promote the practical use and implementation of Cloud computing technology. There is a complementary relationship between them. We focus on infrastructure support, data acquisition, data storage, data computing, data display, and interaction to describe several types of technology developed for big data, and then describe the challenges and opportunities of big data technology from a different angle from the scholars in the related fields. Big data technology is constantly growing with the surge of data volume and processing requirements, and it is affecting our daily habits and lifestyles.

REFERENCES

[1] Zikopoulos PC, Eaton C, DeRoos D, Deutsch T, Lapis G. Understanding big data. New York, NY: McGraw-Hill; 2012.
[2] Bell G, Hey T, Szalay A. Beyond the data deluge. Science 2009; 323(5919):1297.
[3] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: the next frontier for innovation, competition, and productivity. MacKinsey Global Institute; 2011.
[4] Gupta R, Gupta H, Mohania M. Cloud computing and big data analytics: what is new from databases perspective? Big data analytics. Berlin, Heidelberg: Springer; 2012. p. 4261.
[5] Li GJ, Cheng XQ. Research status and scientific thinking of big data. Bull Chin AcadSci 2012;27 (6):64757 (In Chinese).
[6] Khetrapal A, Ganesh V. HBase and hypertable for large scale distributed storage systems. Department of Computer Science, Purdue University; 2006.
[7] Labrinidis A, Jagadish HV. Challenges and opportunities with big data. Proc VLDB Endowment 2012;5(12):2032
[8] Foster I, Zhao Y, Raicu I, Shiyong L. Cloud computing and grid computing 360-degree compared. Grid computing environments workshop, IEEE, 2008. p. 110.
[9] Nurmi D, Wolski R, Grzegorczyk C, Obertelli G, Soman S, Youseff L, et al. The eucalyptus open-source cloud-computing system. Ninth IEEE/ACM international symposium on cluster computing and the grid, 2009. p. 12431.
[10] Keahey K, Freeman T. Contextualization: providing one-click virtual clusters. IEEE fourth international conference on eScience, 2008. p. 30108.
[11] http://scitechconnect.elsevier.com
[12] Ghemawat S, Gobioff H, Leung ST. The Google file system. Proceedings of the 19th ACM symposium on operating systems principles. New York, NY: ACM Press; 2003. p. 2943.
[13] Zheng QL, Fang M, Wang S, Wang XQ, Wu XW, Wang H. Scientific parallel computing based on MapReduce model. Micro Electron Comput 2009;26(8):13
[14] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM 2008; 51(1):107.