IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Big Data and Cloud computing: Review and future trends

Meble Varghese[1*], Victor Jose[2]

[1] Research scholar, Department of Computer Science, Noorul Islam University, Nagarcoil, India
[2] Department of Computer Science, Noorul Islam University, Nagarcoil, India

[*]*Corresponding Author:meblevarghese@gmail.com, Tel.: +91-9447987870*

Available online at: www.ijcseonline.org

Accepted: 18/Dec/2018, Published: 31/Dec/2018

***Abstract*—** Cloud computing is a powerful technology used to perform massive scale computing. Cloud computing avoids high cost of hardware, software and other resources for computing in the local machine. Now a day's massive amount of big data is generated through cloud system. Big data is complex in nature so analyzing it requires large computational infrastructure like cloud systems. In this paper, the use of big data in cloud system is reviewed. The comparison of big data and various cloud platforms are also discussed. Different issues of big data and how cloud platforms are addressed the above said issues are considered here .Furthermore, research challenges related to storage and securities are investigated. Finally open challenges that require high degree of research inputs and efforts are also summarized**.**

***Keywords*—**Bigdata,CloudComputing,Cloudsecurity

I. INTRODUCTION

Due to the continuous increase of organizations, companies and introduction of new technologies like internet of things (IoT) and multimedia, large amount of data is being created every day. Moreover organizations are capturing additional data apart from traditional data like transactional data for processing. Unstructured data such as web data for customer level behavior ,text data such as electronic mail ,e-news, Feeds from Facebook, location and time data ,sensor data and smart grid, social network data are referred as big data .

Big data has evolved as globally accepted technology for academic institutions, government organizations and industry. However big data Concept is different from other technologies due to the volume of data, number of transactions and data sources .This require innovative technologies to handle them. A major challenge for researchers to find an appropriate platform for big data analysis. Here we examine the feasibly of cloud platform for big data analysis.

For performing large scale and complex computing as in enterprise applications, cloud systems has become popular and cloud technology has transformed as a major technology. The major highlights of cloud system includes virtualized resources, parallel computing, scalable resources, flexible data service integration and enhanced protection for data and services. Cloud computing can reduce the cost incurred for computerization, reduced cost for system maintenance, enhanced management and user access. As per the above said features, various applications have been developed under cloud platforms. As a result of this, the first big data developer's implemented clusters of Hadoop in a large scale and adaptable computing paradigms provided by vendors such as Cloudera, MapR and Google. Furthermore, Virtualization technology in cloud computing is the underlying technology for many big data platform attributes[1] like analyze,access,store and manage distributed computing elements.

The goal of this study is to give an insight to the status of big data in cloud computing and define the features and classifications of big data .The relationship of cloud computing and big data and cloud technology for big data analytics is also discussed in this study. Furthermore, the research challenges which focuses on storage and security is also discussed. Several research issues which require good amount of research inputs are also discussed.

A. Big data-Definition

Big data is a term used to mention large amount of data which is not easy to access, analyze, store and process using conventional data base systems. The characteristics of big data is complex and requires substantial efforts and processing to get proper insights of data. The name big data is comparatively new in the research community's .Even though considerable background studies have been done in this area. For instance [1] defined big data as huge-volume, unformatted, autonomous sources with decentralized and distributed control, and attempt to explore complex and gradually develop relationships among data. Meanwhile [2]

and [3] defined big data featured by three Vs: volume, variety and velocity [4] identify that big data is only not defined by three V's instead defined by one more V namely value.in fact, the four V definition is widely accepted because it underlines meaning and completeness of big data.

(1) Volume mentions to the large scale data created from various sources and continue to increase. The data collection helps to create hidden information and pattern analysis.

(2) Variety points to the different categories of data collected from sources like smartphones, sensors, business apps, social media, data storage, public web, archives. The data collected from these sources are mostly in unstructured format. For example data generated from sources like internet, sensors, smartphones are unstructured data types.

(3) Velocity mentions the rate of data transfer. Transactions with high refreshing rate can generates huge data continuously which results data streams generated at high speed and time to process these data streams are very short. There is a transition from batch processing to real time streaming.

(4) Value is the paramount parameter of big data. It is the Process of finding hidden values from big data groups with large data sets and fast generation.

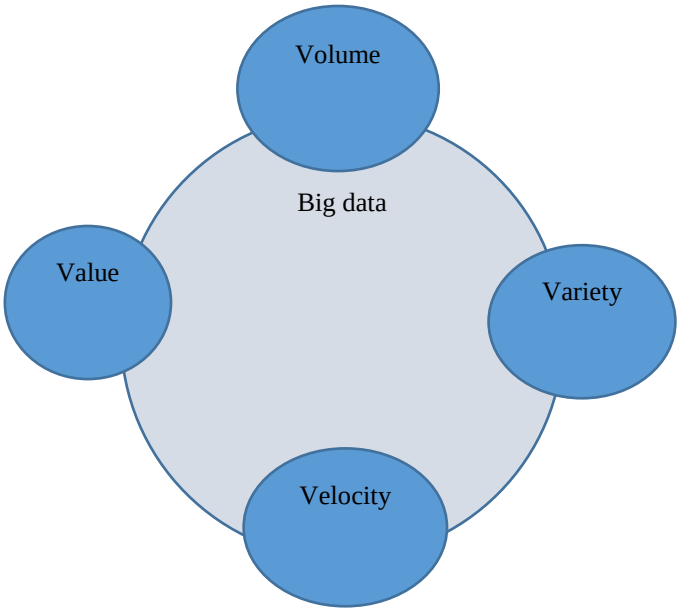

Fig.1 Big data Four V model

B. Cloud Computing

Cloud computing has emerged as one of the most popular technology in future business IT industry. Cloud computing offers reliable hardware, software and other resources over internet and remote data centres.Cloud computing has become the powerful architecture for large data storage, centralized computing, database and applications. The need to analyse, store and process large amount of data has lead companies to use cloud platforms. Large number of scientific application has been deployed in cloud platforms due to non-availability of computational power and other features in local server, minimum costs and the provision of parallel computing for people from different places.

Cloud computing is a framework for flexible, reliable and on request network access for many configured computing resources like server, network storage, application and services. Cloud computing has many favorable factors to facilitate the rapid growth of economies and technical challenges. Cloud computing provide a paradigm of facilitating the ownership to another company and permits companies to concentrate on the core business without concerned about the infrastructure for resources and computing. Moreover this offer an attractive environment for scientists and researchers for performing their experiments. Cloud service model consist of following service models.

- PaaS, points to different platforms of resources on a cloud for end users such as Googles App engine, APP42, Microsoft Azure, Amazon web, Acquia cloud services etc.
- SaaS refers to the end user applications that are run and managed by the service provider such as Google Docs, Gmail, online payroll etc.
- IaaS provide infra structure as a service and provide access to networking devices, work stations and data storage. It provides maximum level of flexibly and freedom for IT resources.Amazone EC2, Flexiscale, google compute engine are some popular examples for IaaS.

The wireless and mobile devices are widely accepted but the limitation of processing capabilities, storage capacities and battery lifetime makes cloud platform more acceptable for such network systems. This has shaped the process of emerging mobile cloud computing paradigm. Mobile cloud platforms allows users to outsource tasks to external sources.For example data can processed and stored outside mobile devices. Mobile cloud applications such as ICloud, Dropbox and Gmail have become widespread in industry now. The juniper research expects that cloud based applications will reach a tally of 9.5$billion approximately. However the peculiar nature of mobile devices and intrinsic

nature of wireless device have imposed some restrictions in computing and storage.

C. *Relation between Big data and Cloud computing*

Big data contains large volume of data and complex datasets so conventional data management tools and data cleaning and mining techniques is hard to apply in it.However cloud computing can provide the adequate tools for processing through the use of Hadoop, open source framework for distributed processing that manages processing of big data and storage for big data applications which runs in clustered systems. Big data, by its nature uses cloud computing based distributed storage technology instead of local memory storage attached to personal computer or electronic device. Cloud computing serves as service model for big data analytics apart from processing and computation of big data.

Jackson et al [3] has provided a survey on the programming models that supports big data analytics. It identifies MapReduce [4] cloud based model for distributed computing on large and complex data sets yet also makes it difficult to incorporate with languages like C++, java or python. Hadoop {Formatting Citation} is another cloud based model developed by Apache software written in java that supports distributed computing across clusters of commodity. Hadoop uses two components named as HDFS and MapReduce programming framework both work closely each other possible to co-deployed so that a single cluster is produced [5].Yibin Li et al [6]discussed the complexity of sensitive big data storage and proposed a secure distributed cloud storage for sensitive big data. The paper focuses on security issues of big data and proposed an intelligent cryptography approach for big data.

Yang.et.al [7] has reviewed in detailed manner of the challenges of big data possessed including analysis, visualization ,integration and architecture .It also discussed how cloud platforms such as massive parallel processing ,distributed databases and data mining grid are used to overcome the challenges.Fernandez.et.al [8] in his paper analyzed MapReduce and Hadoop cloud based platform for big data analytics. In the paper they have identified several libraries and software projects to address the programming model.Sookhak.et.al [9] proposed algebraic signature based efficient remote data auditing for big data in cloud system. They have proposed new data structure divide and conquer table (DCT) for large scale data storage.Venkata.et.al [10] had discussed the security issues of big data in cloud computing and proposed possible solutions in cloud computing security and Hadoop.Joonsang.et.al [11] discussed a trusted cloud computing based architecture for big data information management in smart grids.

Table1 Related studies that deals with big data in cloud systems

Reference	Title of paper	Objectives
7	Big data and cloud computing: innovation opportunities and challenges	To propose cloud platforms as the right platform for handling big data challenges
8	Big data with cloud computing: an insight on computing environment	To discuss features of Hadoop and MapReduce
9	Dynamic remote data auditing for securing big data storage in cloud computing	To discuss remote data auditing for dynamically updating data in cloud storage
10	Security Issues associated with big data in cloud computing	Discussed the security issues of big data in cloud computing and method for improving security of cloud computing
11	Optical storage arrays: a perspective for future big data storage	Discussed recent advancements in Nano photonics –enabled optical storage techniques.
13	Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates	Proposed fine-grained data updates which support authorized auditing and fine-grained update requests

II. ISSUES AND CHALLENGES

Although big data is accepted as an emerging study area by different organizations, it poses some issues and challenges.

A *Plumbing problem*

This problem is existed due to rate at which the data is created and stored every day. The digital universal is doubling every two year or approximately 41% every year [12] and is increasing remarkably faster than data storage, bandwidth and network connections. In 2020, based on IDC, the digital universe will comprise of 40,000 Exabyte's, and 70 % of that will be created or used by End users. Video on demand services occupy 30% of internet now a days. YouTube is creating more than72 hours of video per minute, which require approximately 17 Petabytes of new storage. Wireless devices and mobile devices also generate and consume tremendous data which is approximately 25 % of the internet traffic. Video will occupy for 86% of wireless traffic in 2020 based study conducted by Cisco. Sensor data such as control data, location based data, patient monitoring, traffic data are also generated by mobile devices. Internet of things based data is also generated by different house hold devices. The IDC report predicts 42% of data generated in 2020 will be accounted by machines [13].

B *Security*

Security of big data stored in cloud systems is evident in [14] and [15] and so many other articles. Harsh et.al[16] pointed out a state of the art security and privacy issues in big data

which is applied to health care industry.Vijey.et.al [17] pointed out that big data security is better by using quantum cryptography using groovers algorithm and authentication technique. They have predicted that in future light based quantum cryptography and pair Hand protocol will be the best. Existing work also done in the field of privacy preserving data mining. The basic idea here is to handle data cleaning and mining algorithms effectively without compromising the security of information [18).Lei.Xu[18] et.al view users in four classes ,for each users they have addressed privacy concerns and methods that can be adopted to protect sensitive data.

III. OPEN RESEARCH ISSUES

The amount of big data generated has exponentially increased in recent years. But the technologies and mechanisms for processing these data are limited. Only few tools are available for processing big data applications. State of the art technologies like MapReduce, Dryad, Pergel,MangoDB,Hbase have limited facility for storing and retrieving. Hadoop and MapReduce have limited query processing and lacks infrastructure for data processing and management. Some of the issues are discussed here.

A *Heterogeneous nature of Data*

The most important research issue of big data is heterogeneous nature of big data. The data collected from various places are not formatted. For instance, the data generated in social media, mobile cloud platforms, wireless devices are similar to pieces of text messages, videos and images. Cleaning and transforming such unformatted data before storing into warehousing is a challenging task. Understanding the meaning of unformatted data is important especially to understand meaningful information from it. Hadoop and MapReduce have limited support for distributed data processing of unformatted data. According to Erhard et.al [19] limited research work has done in the field of data cleaning .Data cleaning is required in data warehousing and query processing on unformatted data sources e.g.: in web based information systems.

B *Distributed Storage systems*

To store big data in cloud platforms various solutions have been proposed. However, capability of cloud storage to provide performance required for big data processing is a huge concern [20]. The data migration between servers and how data can be easily retrieved from cloud storage systems is also an issue to consider.

C *Data Analysis*

The selection of an appropriate model for data analysis is critical. Current algorithms for data analysis of big data is not efficient.Efficiant tools are required for processing data. The algorithms performance is not increasing linearly with increasing computational resources [21].The new problems in big data analysis is raised due to the unstructured data generated from different sources. The speed of stream data should analyzed and compared with historical data within a certain period of time.

D *Data Security*

In cloud computing several unsolved security issues are existing for big data storage and process. Security threats are increased by volume, variety and velocity of big data. Various threats exists for confidentiality, privacy, integrity and availability of data [21].Data security must be measured after data is outsourced to cloud service. The cloud systems must be analyzed at regular intervals to safeguard against threats. Policies for all user level access should be enforced. Encapsulating sensitive data and utilize a strong cryptography in cloud computing environment is vital and developing a secure algorithm for key exchange and management is essential here.Moreover, previously developed algorithms for hashing scheme for integrity of data are no longer used to large amounts of data.

IV. CONCLUSION AND FUTURE SCOPE

The size of the data generated from various sources are in exponential rise. The rise of big data is seen as an opportunities for companies as it is a way obtaining advantage over other competitors. The companies will get more revenue, more customers and optimized operation costs if they are able to extract information from data which exists in the business. However, big data analytics is still a challenging and time consuming task that requires high cost software, huge computational infrastructure and effort. Cloud systems alleviates these issues by providing the state of the art infrastructure, on demand computational model for institutions. Furthermore it allows resources to be available on demand and scale up and down whenever required. In this study, we have proposed the classification of big data, conceptual view of big data .The relation between cloud computing and big data is reviewed here.This work looked at the issues of big data as i) the plumbing issue and security issues at different levels. It was observed from the study that the data is continue to increase every year and the heterogeneous nature of big data causes difficult to analyze it.Cloud is used in industry and cloud technologies are required to address the rise of big data.it is evident from our study that data security is major concern in the future as well.

REFERENCES

[1] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, "Data mining with big data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, 2014.

[2] M. Chen, S. Mao, and Y. Liu, "Big Data : A Survey," no. January, pp. 171–209, 2014.

[3] J. C. Jackson, V. Vijayakumar, A. Quadir, and C. Bharathi, "Survey on Programming Models and Environments for Cluster , Cloud , and Grid Computing that defends Big Data," Procedia - Procedia Comput. Sci., vol. 50, pp. 517–523, 2015.

[4] G. Yang, "The Application of MapReduce in the Cloud Computing," 2011.

[5] V. Qxudgkd, L. Ndoedqgl, J. Frp, M. Ylw, and D. F. Lq,"A brief introduction on Big Data 5 V's characteristics and Hadoop Technology" vol. 48, no. Iccc, pp. 319–324, 2015.

[6] Y. Li, K. Gai, L. Qiu, M. Qiu, and H. Zhao, "Intelligent cryptography approach for secure distributed big data storage in cloud computing," Inf. Sci. (Ny)., vol. 0, pp. 1–13, 2016.

[7] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big Data and cloud computing : innovation opportunities and challenges," Int. J. Digit. Earth, vol. 0, no. 0, pp. 1–41, 2016**.**

[8] A. Fernández et al., "Big Data with Cloud Computing : an insight on the computing environment , MapReduce , and programming frameworks," vol. 4, no. October, 2014.

[9] M. Sookhak, A. Gani, M. K. Khan, and R. Buyya, "Dynamic remote data auditing for securing big data storage in cloud computing," Inf. Sci. (Ny)., 2015Lin C., Lee B., "*Exploration of Routing Protocols in Wireless Mesh Network*", In the Proceedings of the 2015 IEEE Symposium on Colossal Big Data Analysis and Networking Security, Canada, pp.**111-117**, **2015**.

[10] V. N. Inukollu, S. Arsi, and S. R. Ravuri, "SECURITY I SSUES A SSOCIATED WITH B IG DATA IN CLOUD COMPUTING," vol. 6, no. 3, pp. 45–56, 2014.

[11] M. Gu, X. Li, and Y. Cao, "Optical storage arrays : a perspective for future big data storage," no. March, 2014.

[12] C. Liu, J. Chen, S. Member, L. T. Yang, and X. Zhang, "Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates," vol. 25, no. 9, pp. 2234–2244, 2014.

[13] J. P. D. Comput, M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, "Big Data computing and clouds : Trends and future directions," J. Parallel Distrib. Comput., vol. 79–80, pp. 3–15, 2015.

[14] G. Lafuente, "The big data security," Netw. Secur., vol. 2015, no. 1, pp. 12–14, 2015.

[15] Ertaul, L. and Singhal, S. 2009. Security Challenges in Cloud Computing. California State University, East Bay.

[16] H. K. Patil and R. Seshadri, "Big data security and privacy issues in healthcare," pp. 775–778, 2014.

[17] V. Thayananthan and A. Albeshri, "Big data security issues based on quantum cryptography and privacy with authentication for mobile data center," Procedia - Procedia Comput. Sci., vol. 50, pp. 149–156, 2015.

[18] L. E. I. Xu, C. Jiang, and J. Wang, "Information Security in Big Data : Privacy and Data Mining," pp. 1149–1176, 2014.

[19] C. Ji, Y. Li, W. Qiu, U. Awada, and K. Li, "Big Data Processing in Cloud Computing Environments," pp. 17–23, 2012.

[20] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," pp. 1–20, 2014.

[21] L. E. I. Xu, C. Jiang, and J. Wang, "Information Security in Big Data : Privacy and Data Mining," pp. 1149–1176, 2014.

