Challenges of Cloud on Big Data

K. Monica

Assistant Professor, Godavari Institute of Engineering & Technology

ABSTRACT:

we are living in a world where people needs everything to be fast, they need high data rate networks, they like fast bike rides they need huge likes for their post in social media. For this faster generation the services need to be fast and need to accommodate huge data. For this purpose Big Data is being introduced where it refers huge amount of data. Traditionally , it is the storage that varies from MB to GB, but as the technology went on increasing and digitization is on its run data is being collected from multiple sources like mobile phone, social media, and other sources too. In order to maintain that huge data the Big data came into existence which can accommodate literally data in PB and ZB. Well as the data comes into play the service is also a key role to be played. In order to manage, maintain and also provide proper services the cloud comes into action as we can say it is a package of things which can provide services, platform and infrastructure too. So in this paper, mainly we'll be concentrating on the challenges of Cloud on Big Data.

Keywords:*Big Data, Cloud computing, resources and v5,etc;*

1) INTRODUCTION:

Information is the outcome of data collected without being structured and without undergoing KDD process. In order to maintain such data it is complicated by using traditional database. As the sources cannot useit prior processing. Information represents data after processing and analysis [4]. The technology has been developed and used in all aspects of life, increasing the demand for storing and processing more data. As a result, several systems have been developed including cloud computing that support big data. While big data is responsible for data storage and processing, the cloud provides a reliable, accessible, and scalable environment for big data systems to function [1]. Big data is defined as the quantity of digital data produced from different sources of technology for example, sensors, digitizers, scanners, numerical modeling, mobile phones, Internet, videos, e-mails and social networks. The data types include texts, geometries, images, videos, sounds and combinations of each. Such data can be directly or indirectly related to geospatial information [2].Cloud computing refers to on-demand computer resources and systems available across the network that can provide a number of integrated computing services without local resources to facilitate user access. These resources include data storage capacity, backup and self-synchronization [4]. Most IT Infrastructure computing consist of services that are provided and delivered through public centers and servers based on them. Here, clouds appear as individual access points for the computing needs of the consumer. It is generally expected for commercial offers to meet the QoS requirements of customers or consumers, and typically include service level agreements (SLAs) [5]. They are an online storage model where data are stored on multiple virtual servers, rather than being hosted on a specific server, and are usually provided by a third party. The hosting companies, which have advanced data centers, rent spaces that are stored in a cloud to their customers in line with their needs [6].

2) LITERATURE REVIEW:

[1] Neves, Pedro Caldeira, Bradley Schmerl, Jorge Bernardino, and Javier Cámara. "Big Data in Cloud Computing: features and issues."

The term big data arose under the explosive increase of global data as a technology that is able to store and process big and varied volumes of data, providing both enterprises and science with deep insights over its clients/experiments. Cloud computing provides a reliable, fault-tolerant, available and scalable environment to harbour big data distributed management systems. Within the context of this paper we present an overview of both technologies and cases of success when integrating big data and cloud frameworks. Although big data solves much of our current problems it still presents some gaps and issues that raise concern and need improvement. Security, privacy, scalability, data governance policies, data heterogeneity, disaster recoverymechanisms, and other challenges are yet to be addressed.

[2] Lopez, Xavier. "Big data and advanced spatial analytics." In Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications, p. 5. ACM, 2012.

Today's business and government organizations are challenged when trying to manage and analyze information from enterprise databases, streaming servers, social media and open source. This is compounded by the complexity of integrating diverse data types (relational, text, spatial, images, spreadsheets) and their representations (customers, products, suppliers, events, and locations) - all of which need to be understood and re-purposed in different contexts. Identifying meaningful patterns across these different information sources is non-trivial. Moreover, conventional IT tools, such as conventional data warehousing and business intelligence alone, are insufficient at handling the volumes, velocity and variety of content at hand. A new framework and associated tools are needed. Dr. Lopez outlines how data scientists and analysts are applying Spatial and Semantic Web concepts to make sense of this Big Data stream. He will describe new approaches oriented toward search, discovery, linking, and analyzing information on the Web, and throughout the enterprise. The role of Map Reduce is described, as is importance of engineered systems to simplify the creation and configuration of Big Data environments. The key take away is use of spatial and linked open data concepts to enhance content alignment, interoperability, discovery and analytics in the Big Data stream.

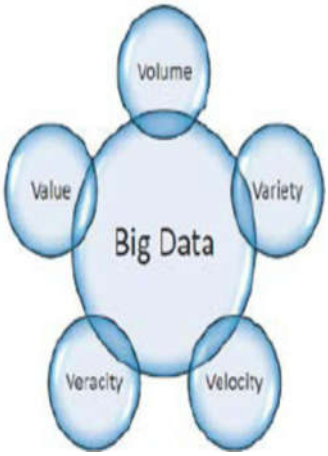
[3] Kshetri, Nir. "Cloud computing in developing economies." Computer 43, no. 10 (2010): 47-55.

The developing world's cloud computing sector has received considerable attention from global and local IT players, national governments, and international agencies. For example, IBM has established cloud computing centers in China, India, Vietnam, Brazil, and South Korea. Other global cloud players such as Microsoft, VMware, Salesforce, Dell, and Parallels are actively searching for opportunities in the developing world. Cloud-related venture capital and other investments are also flowing into developing economies. It is probably fair to say that in no other major technological innovations has the developing world received this level of attention.

3) BIG DATA

Big data comes and is composed through electronics operations from multiple sources. It requires proper processing power and high capabilities for analysis [5]. The importance of big data lies in the analytical use which can help generate an informed decision to provide better and faster services [6].The term big data is called on the huge amount of high-speed big data of different types; this data cannot be processed and stored in regular computers. The main characteristics of big data, called V's 5 As in Figure 1 , can be summed up in the fact that the issue is not only about the volume of data, other dimensions of big data, known as 'five Vs', are as follows:

1. ***Volume:***It represents the amount of data produced from multiple sources which show the huge data in numbers by zeta bytes. The volume is most evident dimension inwhat concerns to big data.
2. ***Variety:*** It represents data types, with, increasing the number of Internet users everywhere, smart phones and social networks users, the familiar form of data has changed from structured data in databases to unstructured data that includes a large number of formats such as images, audio and video clips, SMS, and GPS data [7].
3. ***Velocity:*** It represents the speed of data frequency from different sources, that is, the speed of data production such as Twitter and Facebook. The huge increase in data volume and their frequency dictates the need for a system that ensures super-speed data analysis.
4. ***Veracity:*** It represents the quality of the data, it shows the accuracy of the data and the confidence in the data content. The quality of the data captured can vary greatly, which affects the accuracy of analysis. Although there is wide agreement on the potential value of big data, the data is almost worthless if it is not accurate [8].
5. ***Value:*** It represents the value of big data, i.e. it shows the importance of data after analysis. This is due to the fact that the data on its own is almost worthless. The value lies in careful analysis of the exact data, the information and ideas it provides. The value is the final stage that comes after processing volume, velocity, variety, contrast, validity and visualization [9]



Characteristics of Big Data

3.1 The type and nature of the data:

Data in general is a set of values that are in the form of numbers, letters, symbols and other forms where they are concerned with a particular idea and subject .The data does not make sense without analysis, and is, therefore, compiled for use. It represents input, while information is output after processing, i.e. data is entered into the system first, then processed until it comes out in the form of useful information that has a clear meaning and against which decisions are made.

Big data comes from multiple sources including sensors and free texts such as social media, unstructured data, metadata and other geospatial data collected from web logs, GPS, medical devices, etc. [10]. The big data is gathered from different sources ,so it is in several forms, including:

1. ***Structured data:***
   It is the organized data in the form of tables or databases to be processed.
2. ***Unstructured data:***
   It represents the biggest proportion of data; it is the data that people generate daily as texts, images, videos, messages, log records, click-streams,etc.
3. ***Semi-structured data:***
   It is regarded a kind of structured data but not designed in tables or databases, for example XML documents or JSON [11].

4) CLOUD COMPUTING:

Cloud computing types are classified on the basis of two models: cloud computing service models and cloud computing deployment models.

- ***Software as a service (SAAS)****:* Cloud service providersprovide various software applications to users who can use them without installing them on their computer. The user is not responsible for anything other than adjusting the settings and customizing the service as appropriate to his needs. SAAS helps big-data clients to perform data.
- ***Platform as a service (PAAS)****:* Cloud service providersprovide platforms, tools and other services to users, where the cloud service provider manages everything else, including the operating system and middleware., with resources that enable you to deliver everything from simple cloud-based apps to sophisticated.
- ***Infrastructure as a service (IAAS)****:* Cloud serviceproviders provide infrastructure such as storage, computing capacity, etc. is a form of cloud computing that provides virtualized computing resources over the Internet , In an IaaS model, a third-party provider hosts hardware, software, servers, storage and other infrastructure components on behalf of its users [15][16].
- ***DaaS*** *:* It is the alternative cloud computing model, as itdiffers from traditional models like (SAAS, IAAS, PAAS) in providing data to users through the network, as data is considered the value of this model [17] in conjunction with cloud computing based on solving some of the challenges in managing a huge amount of data. For these reasons, DaaS is closely related to big data whose technologies must be utilized [18]. DaaS provides highly efficient methods of data distribution and processing. DaaS is closely related to SaaS

(storage as a service) and SaaS (software as a service) which can be combined with one of these models or both of them [19].

5) CHALLENGES:

In certain domains, such as social media and health information, as more data is accumulated about individuals, there is a fear that certain organizations will know too much about individuals. Developing algorithms that randomize personal data among a large data set enough to ensure privacy is a key research problem. Perhaps the biggest threat to personal security is the unregulated accumulation of data by numerous social media companies. This data represents a severe security concern, especially when many individuals so willingly surrender such information. Questions of accuracy, dissemination, expiration, and access abound. Clearly, some big data must be secured with respect to privacy and security laws and regulations. International Data Corporation suggested five levels of increasing security [12]: privacy, compliance-driven, custodial, confidential, and lockdown. Further research is required to clearly define these security levels and map them against both current law and current analytics. For example, in Face book, one can restrict pages to 'friends'. But, if Face book runs an analytic over its databases to extract all the friend's linkages in an expanding graph, at what security level should that analytic operate? e.g., how many of an individual's friends should be revealed by such an analytic at a given level if the individual (has the ability to and) has marked those friends at certain security levels? With the increase in the use of big data in business, many companies are wrestling with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset; therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and national security. Meeting the challenges presented by big data will be difficult. The variety of data being generated is also expanding, and organizations capability to capture and process this data is limited. Current technology, architecture management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data. In the distributed systems world, "Big Data" started to become a major issue in the late 1990"s due to the impact of the world-wide Web and a resulting need to index and query its rapidly mushrooming content. Database technology (including parallel databases) was considered for the task, but was found to be neither well-suited nor cost-effective [9] for those purposes.

Google's technical response to the challenges of Web-scale data management and analysis was simple, by database standards, but kicked off what has become the modern "Big Data" revolution in the systems world [13]. To handle the challenge of Web-scale storage, the Google File System (GFS) was created [14]. To handle the challenge of processing the data in such large files, Google pioneered its Map Reduce programming model and platform.

This model, characterized by some as "parallel programming for dummies", enabled Google developers to process large collections of data by writing two user-defined functions, map and reduce, that the Map Reduce framework applies to the instances(map) and sorted groups of instances that share a common key (reduce) similar to the sort of partitioned parallelism utilized in shared-nothing parallel query processing. Taking

Google's GFS and Map Reduce papers as rough technical specifications, open-source equivalents were developed, and the Apache Hadoop Map Reduce platform and its underlying file system (HDFS, the Hadoop Distributed File System) were born. Popular languages include Pig from Yahoo! , Jaql from IBM [14], and Hive from Facebook. Microsoft's technologies include a parallel runtime system called Dryad and two higher-level programming models, Dryad LINQ and the SQLlike SCOPE language, which utilizes Dryad under the covers. Interestingly, Microsoft has also recently announced that its future "Big Data" strategy includes support for Hadoop .

The challenges of security in cloud computing environments can be categorized into network level, user authentication level, data level, and generic issues. Individuals so willingly surrender such information. Questions of accuracy, dissemination, expiration, and access abound. Clearly, some big data must be secured with respect to privacy and security laws and regulations. International Data Corporation suggested five levels of increasing security: privacy, compliance-driven, custodial, confidential, and lockdown. Further research is required to clearly define these security levels and map them against both current law and current analytics. For example, in Face book, one can restrict pages to 'friends'. But, if Face book runs an analytic over its databases to extract all the friend's linkages in an expanding graph, at what security level should that analytic operate? e.g., how many of an individual's friends should be revealed by such an analytic at a given level if the individual (has the ability to and) has marked those friends at certain security levels? With the increase in the use of big data in business, many companies are wrestling with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset; therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and national security. Meeting the challenges presented by big data will be difficult. The variety of data being generated is also expanding, and organizations capability to capture and process this data is limited. Current technology, architecture management and analysis approaches are unable to cope with the flood of data, and organizations will need to change the way they think about, plan, govern, manage, process and report on data to realize the potential of big data. In the distributed systems world, "Big Data" started to become a major issue in the late 1990"s due to the impact of the world-wide Web and a resulting need to index and query its rapidly mushrooming content. Database technology (including parallel databases) was considered for the task, but was found to be neither well-suited nor cost-effective for those purposes.

6) CONCLUSION:

Big data is the "new" business and social science frontier. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to massage and process data. Moreover, it has become clear that "more data is not just more data", but that "more data is different". "Big data" is just the beginning of the problem. Technology evolution and placement guarantee that in a few years more data will be available in a year than has been collected since the dawn of man. If Facebook and Twitter are producing, collectively, around 50 gigabytes of data per day, and tripling every year, within a few years (perhaps 2-4) we are indeed facing the challenge of "big data becoming really big data". In this work, we have done in-depth reviews on recent efforts dedicated to big data and big data networking. We have reviewed the progresses in fundamental big data technologies, important aspects of big data networking, and

security in cloud computing such as new challenges and opportunities, resource management and performance optimizations are also introduced and discussed with independent viewpoints.

7) REFERENCES:

[1] Neves, Pedro Caldeira, Bradley Schmerl, Jorge Bernardino, and Javier Cámara. "Big Data in Cloud Computing: features and issues."

[2] Lopez, Xavier. "Big data and advanced spatial analytics." In Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications, p. 5. ACM, 2012.

[3] Kshetri, Nir. "Cloud computing in developing economies." Computer 43, no. 10 (2010): 47-55.

[4] Charmaz, K., and A. Bryant. "The SAGE Handbook of Grounded Theory: Paperback Edition." (2010).

[5] Boyd, D., & Crawford, K. (2011, September). Six provocations for big data. In A decade in internet time: Symposium on the dynamics of the internet and society (Vol. 21). Oxford: Oxford Internet Institute.

[6] SHAN, Y. C., Chao, L. V., ZHANG, Q. Y., & TIAN, X. Y. (2017). Research on Mechanism of Early Warning of Health Management Based on Cloud Computing and Big Data. In Proceedings of the 23rd International Conference on Industrial Engineering and Engineering Management 2016 (pp. 291-294). Atlantis Press, Paris.

[7] Parvin Ahmadi Doval Amiri and Mina Rahbari Gavgani, 2016. A Review on Relationship and Challenges of Cloud Computing And Big Data: Methods of Analysis and Data Transfer. *Asian Journalof Information Technology, 15: 2516-2525*

[8] Chen, Min, et al. Big data: related technologies,challenges and future prospects. Heidelberg: Springer, 2014.

[9] Demchenko, Yuri, et al. "Big security for big data: Addressing security challenges for the big data infrastructure."Workshop on Secure DataManagement. Springer, Cham, 2013".

[10] Liebowitz, J. (Ed.). (2014). *Bursting the big databubble: The case for intuition-based decision making*.CRC Press.

[11] Sremack, Joe. Big Data Forensics–Learning Hadoop Investigations. Packt Publishing Ltd, 2015.

[12] https://www.internetworldstats.com/stats.html

[13] SHAN, Y. C., Chao, L. V., ZHANG, Q. Y., & TIAN, X. Y. (2017). Research on Mechanism of Early Warning of Health Management Based on Cloud Computing and Big Data. In Proceedings of the 23rd International Conference on Industrial Engineering and Engineering Management 2016 (pp. 291-294). Atlantis Press, Paris.

[14] Parvin Ahmadi Doval Amiri and Mina Rahbari Gavgani, 2016. A Review on Relationship and Challenges of Cloud Computing And Big Data: Methods of Analysis and Data Transfer. *Asian Journalof Information Technology15:2516-252*

[15] Vacca, J. R. (Ed.). (2016). Cloud Computing Security: Foundations and Challenges. CRC Press.

[16] https://support.rackspace.com/how-to/understanding-the-cloud-computing-stack-saas-paas-iaas/

[17] Terzo, O., Ruiu, P., Bucci, E., & Xhafa, F. (2013, July). Data as a service (DaaS) for sharing and processing of large data collections in the cloud. In Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on (pp. 475-480). IEEE.

[18] Motahari-Nezhad, H. R., Stephenson, B., & Singhal, S. (2009). Outsourcing business to cloud computing services: Opportunities and challenges. *IEEE InternetComputing*, *10*(4), 1-17.

[19] Rajesh Saturi, Data as a Service (Daas) in Cloud Computing [Data-As-A-Service in the Age of Data] Data as a Service Daas in Cloud Computing, Global Journal of Computer Science and Technology Cloud & Distributed Volume 12 Issue 11 ,2012.